

# 1

# ECONOMETRICS

---

## 1.1 INTRODUCTION

This book will present an introductory survey of econometrics. We will discuss the fundamental ideas that define the methodology and examine a large number of specific models, tools and methods that econometricians use in analyzing data. This chapter will introduce the central ideas that are the paradigm of econometrics. Section 1.2 defines the field and notes the role that theory plays in motivating econometric practice. Section 1.3 discusses the types of applications that are the focus of econometric analyses. The process of econometric modeling is presented in Section 1.4 with a classic application, Keynes's consumption function. A broad outline of the book is presented in Section 1.5. Section 1.6 notes some specific aspects of the presentation, including the use of numerical examples and the mathematical notation that will be used throughout the book.

## 1.2 THE PARADIGM OF ECONOMETRICS

In the first issue of *Econometrica*, Ragnar Frisch (1933) said of the Econometric Society that

its main object shall be to promote studies that aim at a unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems and that are penetrated by constructive and rigorous thinking similar to that which has come to dominate the natural sciences. But there are several aspects of the quantitative approach to economics, and no single one of these aspects taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a definitely quantitative character. Nor should econometrics be taken as synonymous [*sic*] with the application of mathematics to economics. Experience has shown that each of these three viewpoints, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the *unification* of all three that is powerful. And it is this unification that constitutes econometrics.

The Society responded to an unprecedented accumulation of statistical information. They saw a need to establish a body of principles that could organize what would otherwise become a bewildering mass of data. Neither the pillars nor the objectives of econometrics have changed in the years since this editorial appeared. Econometrics concerns itself with the application of mathematical statistics and the tools of statistical

## 2 PART I ♦ The Linear Regression Model

inference to the empirical measurement of relationships postulated by an underlying theory.

The crucial role that econometrics plays in economics has grown over time. The Nobel Prize in Economic Sciences has recognized this contribution with numerous awards to econometricians, including the first which was given to (the same) Ragnar Frisch in 1969, Lawrence Klein in 1980, Trygve Haavelmo in 1989, James Heckman and Daniel McFadden in 2000, and Robert Engle and Clive Granger in 2003. The 2000 prize was noteworthy in that it celebrated the work of two scientists whose research was devoted to the marriage of behavioral theory and econometric modeling.

### **Example 1.1 Behavioral Models and the Nobel Laureates**

The pioneering work by both James Heckman and Dan McFadden rests firmly on a theoretical foundation of utility maximization.

For Heckman's, we begin with the standard theory of household utility maximization over consumption and leisure. The textbook model of utility maximization produces a demand for leisure time that translates into a supply function of labor. When home production (work in the home as opposed to the outside, formal labor market) is considered in the calculus, then desired "hours" of (formal) labor can be negative. An important conditioning variable is the "reservation" wage—the wage rate that will induce formal labor market participation. On the demand side of the labor market, we have firms that offer market wages that respond to such attributes as age, education, and experience. What can we learn about labor supply behavior based on observed market wages, these attributes and observed hours in the formal market? Less than it might seem, intuitively because our observed data omit half the market—the data on formal labor market activity are not randomly drawn from the whole population.

Heckman's observations about this implicit truncation of the distribution of hours or wages revolutionized the analysis of labor markets. Parallel interpretations have since guided analyses in every area of the social sciences. The analysis of policy interventions such as education initiatives, job training and employment policies, health insurance programs, market creation, financial regulation and a host of others is heavily influenced by Heckman's pioneering idea that when participation is part of the behavior being studied, the analyst must be cognizant of the impact of common influences in both the presence of the intervention and the outcome. We will visit the literature on sample selection and treatment/program evaluation in Chapter 18.

Textbook presentations of the theories of demand for goods that produce utility, since they deal in continuous variables, are conspicuously silent on the kinds of discrete choices that consumers make every day—what brand of product to choose, whether to buy a large commodity such as a car or a refrigerator, how to travel to work, whether to rent or buy a home, where to live, what candidate to vote for, and so on. Nonetheless, a model of "random utility" defined over the alternatives available to the consumer provides a theoretically sound platform for studying such choices. Important variables include, as always, income and relative prices. What can we learn about underlying preference structures from the discrete choices that consumers make? What must be assumed about these preferences to allow this kind of inference? What kinds of statistical models will allow us to draw inferences about preferences? McFadden's work on how commuters choose to travel to work, and on the underlying theory appropriate to this kind of modeling, has guided empirical research in discrete consumer choice for several decades. We will examine McFadden's models of discrete choice in Chapter 17.

The connection between underlying behavioral models and the modern practice of econometrics is increasingly strong. A useful distinction is made between *microeconometrics* and *macroeconometrics*. The former is characterized by its analysis of cross section and panel data and by its focus on individual consumers, firms, and micro-level decision makers. Practitioners rely heavily on the theoretical tools of microeconomics including utility maximization, profit maximization, and market equilibrium. The analyses

are directed at subtle, difficult questions that often require intricate formulations. A few applications are as follows:

- What are the likely effects on labor supply behavior of proposed negative income taxes? [Ashenfelter and Heckman (1974).]
- Does attending an elite college bring an expected payoff in expected lifetime income sufficient to justify the higher tuition? [Kreuger and Dale (1999) and Kreuger (2000).]
- Does a voluntary training program produce tangible benefits? Can these benefits be accurately measured? [Angrist (2001).]
- Do smaller class sizes bring real benefits in student performance? [Hanushek (1999), Hoxby (2000), Angrist and Lavy (1999).]
- Does the presence of health insurance induce individuals to make heavier use of the health care system—is moral hazard a measurable problem? [Riphahn et al. (2003).]

Macroeconomics is involved in the analysis of time-series data, usually of broad aggregates such as price levels, the money supply, exchange rates, output, investment, economic growth and so on. The boundaries are not sharp. For example, an application that we will examine in this text concerns spending patterns of municipalities, which rests somewhere between the two fields. The very large field of financial econometrics is concerned with long time-series data and occasionally vast panel data sets, but with a sharply focused orientation toward models of individual behavior. The analysis of market returns and exchange rate behavior is neither exclusively macro- nor microeconometric. (We will not be spending any time in this book on financial econometrics. For those with an interest in this field, I would recommend the celebrated work by Campbell, Lo, and Mackinlay (1997) or, for a more time-series-oriented approach, Tsay (2005).) Macroeconomic model builders rely on the interactions between economic agents and policy makers. For examples:

- Does a monetary policy regime that is strongly oriented toward controlling inflation impose a real cost in terms of lost output on the U.S. economy? [Cecchetti and Rich (2001).]
- Did 2001's largest federal tax cut in U.S. history contribute to or dampen the concurrent recession? Or was it irrelevant?

Each of these analyses would depart from a formal model of the process underlying the observed data.

### 1.3 THE PRACTICE OF ECONOMETRICS

We can make another useful distinction between *theoretical econometrics* and *applied econometrics*. Theorists develop new techniques for estimation and hypothesis testing and analyze the consequences of applying particular methods when the assumptions that justify those methods are not met. Applied econometricians are the users of these techniques and the analysts of data (“real world” and simulated). The distinction is far from sharp; practitioners routinely develop new analytical tools for the purposes of the

#### 4 PART I ♦ The Linear Regression Model

study that they are involved in. This book contains a large amount of econometric theory, but it is directed toward applied econometrics. I have attempted to survey techniques, admittedly some quite elaborate and intricate, that have seen wide use “in the field.”

Applied econometric methods will be used for estimation of important quantities, analysis of economic outcomes such as policy changes, markets or individual behavior, testing theories, and for forecasting. The last of these is an art and science in itself that is the subject of a vast library of sources. Although we will briefly discuss some aspects of forecasting, our interest in this text will be on estimation and analysis of models. The presentation, where there is a distinction to be made, will contain a blend of microeconometric and macroeconometric techniques and applications. It is also necessary to distinguish between *time-series analysis* (which is not our focus) and methods that primarily use time-series data. The former is, like forecasting, a growth industry served by its own literature in many fields. While we will employ some of the techniques of time-series analysis, we will spend relatively little time developing first principles.

### 1.4 ECONOMETRIC MODELING

Econometric analysis usually begins with a statement of a theoretical proposition. Consider, for example, a classic application by one of Frisch’s contemporaries:

**Example 1.2 Keynes’s Consumption Function**

From Keynes’s (1936) *General Theory of Employment, Interest and Money*:

We shall therefore define what we shall call the propensity to consume as the functional relationship  $f$  between  $X$ , a given level of income, and  $C$ , the expenditure on consumption out of the level of income, so that  $C = f(X)$ .

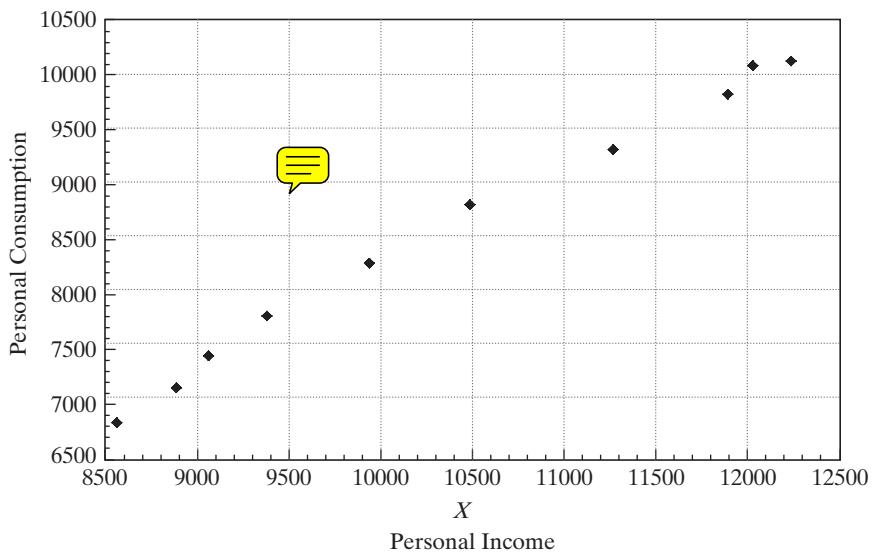
The amount that the community spends on consumption depends (i) partly on the amount of its income, (ii) partly on other objective attendant circumstances, and (iii) partly on the subjective needs and the psychological propensities and habits of the individuals composing it. The fundamental psychological law upon which we are entitled to depend with great confidence, both a priori from our knowledge of human nature and from the detailed facts of experience, is that men are disposed, as a rule and on the average, to increase their consumption as their income increases, but not by as much as the increase in their income. That is,  $\dots dC/dX$  is positive and less than unity.

But, apart from short period changes in the level of income, it is also obvious that a higher absolute level of income will tend as a rule to widen the gap between income and consumption. . . . These reasons will lead, as a rule, to a greater proportion of income being saved as real income increases.

The theory asserts a relationship between consumption and income,  $C = f(X)$ , and claims in the second paragraph that the marginal propensity to consume (MPC),  $dC/dX$ , is between zero and one.<sup>1</sup> The final paragraph asserts that the average propensity to consume (APC),  $C/X$ , falls as income rises, or  $d(C/X)/dX = (MPC - APC)/X < 0$ . It follows that  $MPC < APC$ . The most common formulation of the consumption function is a linear relationship,  $C = \alpha + X\beta$ , that satisfies Keynes’s “laws” if  $\beta$  lies between zero and one and if  $\alpha$  is greater than zero.

These theoretical propositions provide the basis for an econometric study. Given an appropriate data set, we could investigate whether the theory appears to be consistent with

<sup>1</sup>Modern economists are rarely this confident about their theories. More contemporary applications generally begin from first principles and behavioral axioms, rather than simple observation.



**FIGURE 1.1** Aggregate U.S. Consumption and Income Data, 2000–2009.

the observed “facts.” For example, we could see whether the linear specification appears to be a satisfactory description of the relationship between consumption and income, and, if so, whether  $\alpha$  is positive and  $\beta$  is between zero and one. Some issues that might be studied are (1) whether this relationship is stable through time or whether the parameters of the relationship change from one generation to the next (a change in the average propensity to save,  $1 - APC$ , might represent a fundamental change in the behavior of consumers in the economy); (2) whether there are systematic differences in the relationship across different countries, and, if so, what explains these differences; and (3) whether there are other factors that would improve the ability of the model to explain the relationship between consumption and income. For example, Figure 1.1 presents aggregate consumption and personal income in constant dollars for the U.S. for the 10 years of 2000–2009. (See Appendix Table F1.1.) Apparently, at least superficially, the data (the facts) are consistent with the theory. The relationship appears to be linear, albeit only approximately, the intercept of a line that lies close to most of the points is positive and the slope is less than one, although not by much. (However, if the line is fit by linear least squares regression, the intercept is negative, not positive.)

Economic theories such as Keynes’s are typically sharp and unambiguous. Models of demand, production, labor supply, individual choice, educational attainment, income and wages, investment, market equilibrium and aggregate consumption all specify precise, *deterministic* relationships. Dependent and independent variables are identified, a functional form is specified, and in most cases, at least a qualitative statement is made about the directions of effects that occur when independent variables in the model change. The model is only a simplification of reality. It will include the salient features of the relationship of interest but will leave unaccounted for influences that might well be present but are regarded as unimportant.

Correlations among economic variables are easily observable through descriptive statistics and techniques such as linear regression methods. The ultimate goal of the econometric model builder is often to uncover the deeper causal connections through

## 6 PART I ♦ The Linear Regression Model

elaborate structural, behavioral models. Note, for example, Keynes's use of the behavior of a "representative consumer" to motivate the behavior of macroeconomic variables such as income and consumption. Heckman's model of labor supply noted in Example 1.1 is framed in a model of individual behavior. Berry, Levinsohn and Pakes's (1995) detailed model of equilibrium pricing in the automobile market is another.

No model could hope to encompass the myriad essentially random aspects of economic life. It is thus also necessary to incorporate stochastic elements. As a consequence, observations on a variable will display variation attributable not only to differences in variables that are explicitly accounted for in the model, but also to the randomness of human behavior and the interaction of countless minor influences that are not. It is understood that the introduction of a random "disturbance" into a deterministic model is not intended merely to paper over its inadequacies. It is essential to examine the results of the study, in a sort of postmortem, to ensure that the allegedly random, unexplained factor is truly unexplainable. If it is not, the model is, in fact, inadequate. [In the example given earlier, the estimated constant term in the linear least squares regression is negative. Is the theory wrong, or is the finding due to random fluctuation in the data? Another possibility is that the theory is broadly correct, but the world changed between 1936 when Keynes devised his theory and 2000–2009 when the data (outcomes) were generated. Or, perhaps linear least squares is not the appropriate technique to use for this model, and that is responsible for the inconvenient result (the negative intercept).] The stochastic element endows the model with its statistical properties. Observations on the variable(s) under study are thus taken to be the outcomes of a random processes. With a sufficiently detailed stochastic structure and adequate data, the analysis will become a matter of deducing the properties of a probability distribution. The tools and methods of mathematical statistics will provide the operating principles.

A model (or theory) can never truly be confirmed unless it is made so broad as to include every possibility. But it may be subjected to ever more rigorous scrutiny and, in the face of contradictory evidence, refuted. A deterministic theory will be invalidated by a single contradictory observation. The introduction of stochastic elements into the model changes it from an exact statement to a probabilistic description about expected outcomes and carries with it an important implication. Only a preponderance of contradictory evidence can convincingly invalidate the probabilistic model, and what constitutes a "preponderance of evidence" is a matter of interpretation. Thus, the probabilistic model is less precise but at the same time, more robust.<sup>2</sup>

The techniques used in econometrics have been employed in a widening variety of fields, including political methodology, sociology [see, e.g., Long (1997) and DeMaris (2004)], health economics, medical research (how do we handle attrition from medical treatment studies?) environmental economics, economic geography, transportation engineering, and numerous others. Practitioners in these fields and many more are all heavy users of the techniques described in this text.

The process of econometric analysis departs from the specification of a theoretical relationship. We initially proceed on the optimistic assumption that we can obtain

---

<sup>2</sup>See Keuzenkamp and Magnus (1995) for a lengthy symposium on testing in econometrics.

precise measurements on all the variables in a correctly specified model. If the ideal conditions are met at every step, the subsequent analysis will be routine. Unfortunately, they rarely are. Some of the difficulties one can expect to encounter are the following:

- The data may be badly measured or may correspond only vaguely to the variables in the model. “The interest rate” is one example.
- Some of the variables may be inherently unmeasurable. “Expectations” is a case in point.
- The theory may make only a rough guess as to the correct form of the model, if it makes any at all, and we may be forced to choose from an embarrassingly long menu of possibilities.
- The assumed stochastic properties of the random terms in the model may be demonstrably violated, which may call into question the methods of estimation and inference procedures we have used.
- Some relevant variables may be missing from the model.
- The conditions under which data are collected leads to a sample of observations that is systematically unrepresentative of the population we wish to study.

The ensuing steps of the analysis consist of coping with these problems and attempting to extract whatever information is likely to be present in such obviously imperfect data. The methodology is that of mathematical statistics and economic theory. The product is an econometric model.

## 1.5 PLAN OF THE BOOK

Econometrics is a large and growing field. It is a challenge to chart a course through that field for the beginner. Our objective in this survey is to develop in detail a set of tools, then use those tools in applications. The following set of applications is large and will include many that readers will use in practice. But, it is not exhaustive. We will attempt to present our results in sufficient generality that the tools we develop here can be extended to other kinds of situations and applications not described here.

One possible approach is to organize (and orient) the areas of study by the type of data being analyzed—cross section, panel, discrete data, then time series being the obvious organization. Alternatively, we could distinguish at the outset between micro- and macro econometrics.<sup>3</sup> Ultimately, all of these will require a common set of tools, including, for example, the multiple regression model, the use of moment conditions for estimation, instrumental variables (IV) and maximum likelihood estimation. With that in mind, the organization of this book is as follows: The first half of the text develops

<sup>3</sup>An excellent reference on the former that is at a more advanced level than this book is Cameron and Trivedi (2005). As of this writing, there does not appear to be available a counterpart, large-scale pedagogical survey of macroeconomics that includes both econometric theory and applications. The numerous more focused studies include books such as Bårdesen, G., Eitrheim, Ø., Jansen, E. and Nymoen, R., *The Econometrics of Macroeconomic Modelling*, Oxford University Press, 2005 and survey papers such as Wallis, K., “Macroeconometric Models,” published in *Macroeconomic Policy: Iceland in an Era of Global Integration* (M. Gudmundsson, T.T. Herbertsson, and G. Zoëga, eds), pp.399–414. Reykjavik: University of Iceland Press, 2000 also at [http://www.ecomod.net/conferences/ecomod2001/papers.web/Wallis\\_Iceland.pdf](http://www.ecomod.net/conferences/ecomod2001/papers.web/Wallis_Iceland.pdf)

## 8 PART I ♦ The Linear Regression Model

fundamental results that are common to all the applications. The concept of multiple regression and the linear regression model in particular constitutes the underlying platform of most modeling, even if the linear model itself is not ultimately used as the empirical specification. This part of the text concludes with developments of IV estimation and the general topic of panel data modeling. The latter pulls together many features of modern econometrics, such as, again, IV estimation, modeling heterogeneity, and a rich variety of extensions of the linear model. The second half of the text presents a variety of topics. Part III is an overview of estimation methods. Finally, Parts IV and V present results from microeconomics and macroeconomics, respectively. The broad outline is as follows:

### I. Regression Modeling

Chapters 2 through 6 present the multiple linear regression model. We will discuss specification, estimation, and statistical inference. This part develops the ideas of estimation, robust analysis, functional form and principles of model specification.

### II. Generalized Regression, Instrumental Variables, and Panel Data

Chapter 7 extends the regression model to nonlinear functional forms. The method of instrumental variables is presented in Chapter 8. Chapters 9 and 10 introduce the generalized regression model and systems of regression models. This section ends with Chapter 11 on panel data methods.

### III. Estimation Methods

Chapters 12 through 16 present general results on different methods of estimation including GMM, maximum likelihood, and simulation based methods. Various estimation frameworks, including non- and semiparametric and Bayesian estimation are presented in Chapters 12 and 16.

### IV. Microeconomic Methods

Chapters 17 through 19 are about microeconomics, discrete choice modeling and limited dependent variables, and the analysis of data on events—how many occur in a given setting and when they occur. Chapters 17 to 19 are devoted to methods more suited to cross sections and panel data sets.

### V. Macroeconometric Methods

Chapters 20 to 23 focus on time-series modeling and macroeconomics.

### VI. Background Materials

Appendices A through E present background material on tools used in econometrics including matrix algebra, probability and distribution theory, estimation, and asymptotic distribution theory. Appendix E presents results on computation. Appendices A through E are chapter-length surveys of the tools used in econometrics. Because it is assumed that the reader has some previous training in each of these topics, these summaries are included primarily for those who desire a refresher or a convenient reference. We do not anticipate that these appendices can substitute for a course in any of these subjects. The intent of these chapters is to provide a reasonably concise summary of the results, nearly all of which are explicitly used elsewhere in the book. The data sets used in the numerical examples are described in Appendix F. The actual data sets and other supplementary materials can be downloaded from the author's web site for the text: <http://pages.stern.nyu.edu/~wgreen/Text/>. Useful tables related to commonly used probability distributions are given in Appendix G.

## 1.6 PRELIMINARIES

Before beginning, we note some specific aspects of the presentation in the text.

### 1.6.1 NUMERICAL EXAMPLES

There are many numerical examples given throughout the discussion. Most of these are either self-contained exercises or extracts from published studies. In general, their purpose is to provide a limited application to illustrate a method or model. The reader can, if they wish, replicate them with the data sets provided. This will generally not entail attempting to replicate the full published study. Rather, we use the data sets to provide applications that relate to the published study in a limited, manageable fashion that also focuses on a particular technique, model or tool. Thus, Riphahn, Wambach, and Million (2003) provide a very useful, manageable (though relatively large) laboratory data set that the reader can use to explore some issues in health econometrics. The exercises also suggest more extensive analyses, again in some cases based on published studies.

### 1.6.2 SOFTWARE AND REPLICATION

As noted in the preface, there are now many powerful computer programs that can be used for the computations described in this book. In most cases, the examples presented can be replicated with any modern package, whether the user is employing a high level integrated program such as *NLOGIT*, *Stata* or *SAS*, or writing their own programs in languages such as *R*, *MatLab* or *Gauss*. The notable exception will be exercises based on simulation. Since, essentially, every package uses a different random number generator, it will generally not be possible to replicate exactly the examples in this text that use simulation (unless you are using the same computer program we are). Nonetheless, the differences that do emerge in such cases should be attributable to, essentially, minor random variation. You will be able to replicate the essential results and overall features in these applications with any of the software mentioned. We will return to this general issue of replicability at a few points in the text, including in Section 15.2 where we discuss methods of generating random samples for simulation based estimators.

### 1.6.3 NOTATIONAL CONVENTIONS

We will use vector and matrix notation and manipulations throughout the text. The following conventions will be used: A scalar variable will be denoted with an italic lowercase letter, such as  $y$  or  $x_{nK}$ . A column vector of scalar values will be denoted

by a boldface, lowercase letter, such as  $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$  and, likewise for,  $\mathbf{x}$ , and  $\mathbf{b}$ . The

dimensions of a column vector are always denoted as those of a matrix with one column, such as  $K \times 1$  or  $n \times 1$  and so on. A matrix will always be denoted by a boldface

## 10 PART I ♦ The Linear Regression Model

uppercase letter, such as the  $n \times K$  matrix,  $\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nK} \end{bmatrix}$ . Specific elements

in a matrix are always subscripted so that the first subscript gives the row and the second gives the column. Transposition of a vector or a matrix is denoted with a prime. A row vector is obtained by transposing a column vector. Thus,  $\boldsymbol{\beta}' = [\beta_1, \beta_2, \dots, \beta_K]$ . The product of a row and a column vector will always be denoted in a form such as  $\boldsymbol{\beta}'\mathbf{x} = \beta_1x_1 + \beta_2x_2 + \cdots + \beta_Kx_K$ . The elements in a matrix,  $\mathbf{X}$ , form a set of vectors. In terms of its columns,  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K]$ —each column is an  $n \times 1$  vector. The one possible, unfortunately unavoidable source of ambiguity is the notation necessary to denote a row of a matrix such as  $\mathbf{X}$ . The elements of the  $i$ th row of  $\mathbf{X}$  are the row vector,  $\mathbf{x}'_i = [x_{i1}, x_{i2}, \dots, x_{iK}]$ . When the matrix, such as  $\mathbf{X}$ , refers to a data matrix, we will prefer to use the “ $i$ ” subscript to denote observations, or the rows of the matrix and “ $k$ ” to denote the variables, or columns. As we note unfortunately, this would seem to imply that  $\mathbf{x}_i$ , the transpose of  $\mathbf{x}'_i$ , would be the  $i$ th column of  $\mathbf{X}$ , which will conflict with our notation. However, with no simple alternative notation available, we will maintain this convention with the understanding that  $\mathbf{x}'_i$  always refers to the row vector that is the  $i$ th row of an  $\mathbf{X}$  matrix. A discussion of the matrix algebra results used in the book is given in Appendix A. A particularly important set of arithmetic results about summation and the elements of the matrix product matrix,  $\mathbf{X}'\mathbf{X}$  appears in Section A.2.7.

## 2

# THE LINEAR REGRESSION MODEL

---

## 2.1 INTRODUCTION

Econometrics is concerned with *model building*. An intriguing point to begin the inquiry is to consider the question, “What is the model?” The statement of a “model” typically begins with an observation or a proposition that one variable “is caused by” another, or “varies with another,” or some qualitative statement about a relationship between a variable and one or more **covariates** that are expected to be related to the interesting one in question. The model might make a broad statement about behavior, such as the suggestion that individuals’ usage of the health care system depends on, for example, perceived health status, demographics such as income, age, and education, and the amount and type of insurance they have. It might come in the form of a verbal proposition, or even a picture such as a flowchart or **path diagram** that suggests directions of influence. The econometric model rarely springs forth in full bloom as a set of equations. Rather, it begins with an *idea* of some kind of relationship. The natural next step for the econometrician is to translate that idea into a set of equations, with a notion that some feature of that set of equations will answer interesting questions about the variable of interest. To continue our example, a more definite statement of the relationship between insurance and health care demanded might be able to answer, *how* does health care system utilization depend on insurance coverage? Specifically, is the relationship “positive”—all else equal, is an insured consumer more likely to “demand more health care,” or is it “negative”? And, ultimately, one might be interested in a more precise statement, “how much more (or less)?” This and the next several chapters will build up the set of tools that model builders use to pursue questions such as these using data and econometric methods.

From a purely statistical point of view, the researcher might have in mind a variable,  $y$ , broadly “demand for health care,  $H$ ,” and a vector of covariates,  $\mathbf{x}$  (income,  $I$ , insurance,  $T$ ), and a joint probability distribution of the three,  $p(H, I, T)$ . Stated in this form, the “relationship” is not posed in a particularly interesting fashion—what is the statistical process that produces health care demand, income, and insurance coverage. However, it is true that  $p(H, I, T) = p(H|I, T)p(I, T)$ , which decomposes the probability model for the joint process into two outcomes, the joint distribution of insurance coverage and income in the population and the distribution of “demand for health care” for a specific income and insurance coverage. From this perspective, the conditional distribution,  $p(H|I, T)$  holds some particular interest, while  $p(I, T)$ , the distribution of income and insurance coverage in the population is perhaps of secondary, or no interest. (On the other hand, from the same perspective, the conditional “demand” for insurance coverage, given income,  $p(T|I)$ , might also be interesting.) Continuing this line of

## 12 PART I ♦ The Linear Regression Model

thinking, the model builder is often interested not in joint variation of all the variables in the model, but in **conditional variation** of one of the variables related to the other.

The idea of the conditional distribution provides a useful starting point for thinking about a relationship between a variable of interest, a “ $y$ ,” and a set of variables, “ $x$ ,” that we think might bear some relationship to it. There is a question to be considered now that returns us to the issue of “what is the model?” What feature of the conditional distribution is of interest? The model builder, thinking in terms of features of the conditional distribution, often gravitates to the expected value, focusing attention on  $E[y|x]$ , that is, the **regression function**, which brings us to the subject of this chapter. For the preceding example, above, this might be natural if  $y$  were “doctor visits” as in an example examined at several points in the chapters to follow. If we were studying incomes,  $I$ , however, which often have a highly skewed distribution, then the mean might not be particularly interesting. Rather, the **conditional median**, for given ages,  $M[I|x]$ , might be a more interesting statistic. On the other hand, still considering the distribution of incomes (and still conditioning on age), other quantiles, such as the 20<sup>th</sup> percentile, or a poverty line defined as, say, the 5<sup>th</sup> percentile, might be more interesting yet. Finally, consider a study in finance, in which the variable of interest is asset returns. In at least some contexts, means are not interesting at all—it is variances, and conditional variances in particular, that are most interesting.

The point is that we begin the discussion of the regression model with an understanding of what we mean by “the model.” For the present, we will focus on the conditional mean which is usually the feature of interest. Once we establish how to analyze the regression function, we will use it as a useful departure point for studying other features, such as quantiles and variances. The **linear regression model** is the single most useful tool in the econometrician’s kit. Although to an increasing degree in contemporary research it is often only the departure point for the full analysis, it remains the device used to begin almost all empirical research. And, it is the lens through which relationships among variables are usually viewed. This chapter will develop the linear regression model. Here, we will detail the fundamental assumptions of the model. The next several chapters will discuss more elaborate specifications and complications that arise in the application of techniques that are based on the simple models presented here.

## 2.2 THE LINEAR REGRESSION MODEL

The **multiple linear regression model** is used to study the relationship between a **dependent variable** and one or more **independent variables**. The generic form of the linear regression model is

$$\begin{aligned} y &= f(x_1, x_2, \dots, x_K) + \varepsilon \\ &= x_1\beta_1 + x_2\beta_2 + \dots + x_K\beta_K + \varepsilon \end{aligned} \quad (2-1)$$

where  $y$  is the dependent or **explained** variable and  $x_1, \dots, x_K$  are the independent or **explanatory** variables. One’s theory will specify  $f(x_1, x_2, \dots, x_K)$ . This function is commonly called the **population regression equation** of  $y$  on  $x_1, \dots, x_K$ . In this setting,  $y$  is the **regressand** and  $x_k, k=1, \dots, K$  are the **regressors** or **covariates**. The underlying theory will specify the dependent and independent variables in the model. It is not always obvious which is appropriately defined as each of these—for example,

## CHAPTER 2 ♦ The Linear Regression Model 13

a demand equation,  $quantity = \beta_1 + price \times \beta_2 + income \times \beta_3 + \varepsilon$ , and an inverse demand equation,  $price = \gamma_1 + quantity \times \gamma_2 + income \times \gamma_3 + u$  are equally valid representations of a market. For modeling purposes, it will often prove useful to think in terms of “autonomous variation.” One can conceive of movement of the independent variables outside the relationships defined by the model while movement of the dependent variable is considered in response to some independent or exogenous stimulus.<sup>1</sup>

The term  $\varepsilon$  is a random **disturbance**, so named because it “disturbs” an otherwise stable relationship. The disturbance arises for several reasons, primarily because we cannot hope to capture every influence on an economic variable in a model, no matter how elaborate. The net effect, which can be positive or negative, of these omitted factors is captured in the disturbance. There are many other contributors to the disturbance in an empirical model. Probably the most significant is errors of measurement. It is easy to theorize about the relationships among precisely defined variables; it is quite another to obtain accurate measures of these variables. For example, the difficulty of obtaining reasonable measures of ~~per~~<sup>gross</sup> national product, interest rates, capital stocks, or, worse yet, flows of services from capital stocks is a recurrent theme in the empirical literature. At the extreme, there may be no observable counterpart to the theoretical variable. The literature on the permanent income model of consumption [e.g., Friedman (1957)] provides an interesting example.

We assume that each observation in a sample  $(y_i, x_{i1}, x_{i2}, \dots, x_{iK}), i = 1, \dots, n$ , is generated by an underlying process described by

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \dots + x_{iK}\beta_K + \varepsilon_i.$$

The observed value of  $y_i$  is the sum of two parts, a deterministic part and the random part,  $\varepsilon_i$ . Our objective is to estimate the unknown parameters of the model, use the data to study the validity of the theoretical propositions, and perhaps use the model to predict the variable  $y$ . How we proceed from here depends crucially on what we assume about the stochastic process that has led to our observations of the data in hand.

### Example 2.1 Keynes's Consumption Function

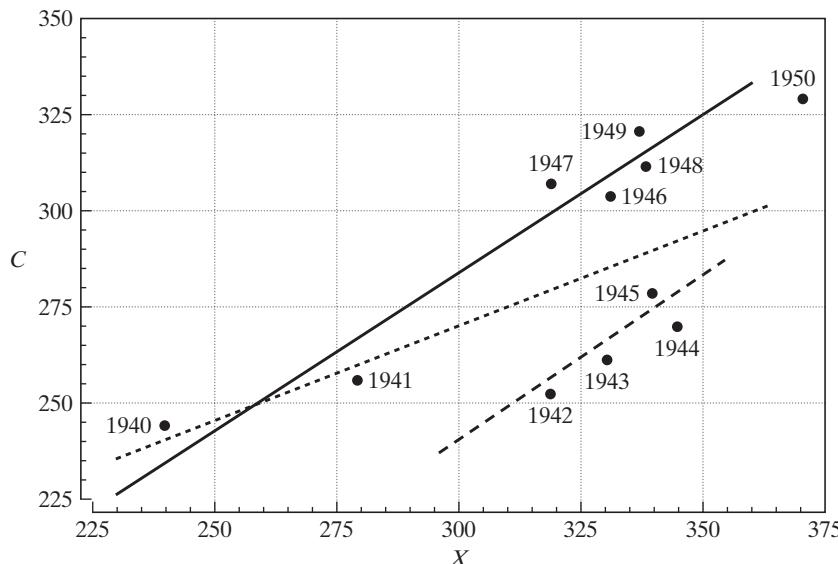
Example 1.2 discussed a model of consumption proposed by Keynes and his *General Theory* (1936). The theory that consumption,  $C$ , and income,  $X$ , are related certainly seems consistent with the observed “facts” in Figures 1.1 and 2.1. (These data are in Data Table F2.1.) Of course, the linear function is only approximate. Even ignoring the anomalous wartime years, consumption and income cannot be connected by any simple **deterministic relationship**. The linear model,  $C = \alpha + \beta X$ , is intended only to represent the salient features of this part of the economy. It is hopeless to attempt to capture every influence in the relationship. The next step is to incorporate the inherent randomness in its real-world counterpart. Thus, we write  $C = f(X, \varepsilon)$ , where  $\varepsilon$  is a stochastic element. It is important not to view  $\varepsilon$  as a catchall for the inadequacies of the model. The model including  $\varepsilon$  appears adequate for the data not including the war years, but for 1942–1945, something systematic clearly seems to be missing. Consumption in these years could not rise to rates historically consistent with these levels of income because of wartime rationing. A model meant to describe consumption in this period would have to accommodate this influence.

It remains to establish how the stochastic element will be incorporated in the equation. The most frequent approach is to assume that it is *additive*. Thus, we recast the equation

---

<sup>1</sup>By this definition, it would seem that in our demand relationship, only income would be an independent variable while both price and quantity would be dependent. That makes sense—in a market, price and quantity are determined at the same time, and do change only when something outside the market changes.

## 14 PART I ♦ The Linear Regression Model



**FIGURE 2.1** Consumption Data, 1940–1950.

in stochastic terms:  $C = \alpha + \beta X + \varepsilon$ . This equation is an empirical counterpart to Keynes's theoretical model. But, what of those anomalous years of rationing? If we were to ignore our intuition and attempt to "fit" a line to all these data—the next chapter will discuss at length how we should do that—we might arrive at the dotted line in the figure as our best guess. This line, however, is obviously being distorted by the rationing. A more appropriate specification for these data that accommodates both the stochastic nature of the data and the special circumstances of the years 1942–1945 might be one that shifts straight down in the war years,  $C = \alpha + \beta X + d_{\text{waryears}}\delta_w + \varepsilon$ , where the new variable,  $d_{\text{waryears}}$  equals one in 1942–1945 and zero in other years and  $\delta_w < 0$ .

One of the most useful aspects of the multiple regression model is its ability to identify the independent effects of a set of variables on a dependent variable. Example 2.2 describes a common application.

### Example 2.2 Earnings and Education

A number of recent studies have analyzed the relationship between earnings and education. We would expect, on average, higher levels of education to be associated with higher incomes. The simple regression model

$$\text{earnings} = \beta_1 + \beta_2 \text{education} + \varepsilon,$$

however, neglects the fact that most people have higher incomes when they are older than when they are young, regardless of their education. Thus,  $\beta_2$  will overstate the marginal impact of education. If age and education are positively correlated, then the regression model will associate all the observed increases in income with increases in education. A better specification would account for the effect of age, as in

$$\text{earnings} = \beta_1 + \beta_2 \text{education} + \beta_3 \text{age} + \varepsilon.$$

It is often observed that income tends to rise less rapidly in the later earning years than in the early ones. To accommodate this possibility, we might extend the model to

$$\text{earnings} = \beta_1 + \beta_2 \text{education} + \beta_3 \text{age} + \beta_4 \text{age}^2 + \varepsilon.$$

We would expect  $\beta_3$  to be positive and  $\beta_4$  to be negative.

## CHAPTER 2 ♦ The Linear Regression Model 15

The crucial feature of this model is that it allows us to carry out a conceptual experiment that might not be observed in the actual data. In the example, we might like to (and could) compare the earnings of two individuals of the same age with different amounts of “education” even if the data set does not actually contain two such individuals. How education should be measured in this setting is a difficult problem. The study of the earnings of twins by Ashenfelter and Krueger (1994), which uses precisely this specification of the earnings equation, presents an interesting approach [Studies of twins and siblings have provided an interesting thread of research on the education and income relationship. Two other studies are Ashenfelter and Zimmerman (1997) and Bonjour, Cherkas, Haskel, Hawkes, and Spector (2003).]. We will examine this study in some detail in Section 8.5.3.

The experiment embodied in the earnings model thus far suggested is a comparison of two otherwise identical individuals who have different years of education. Under this interpretation, the “impact” of education would be  $\partial E[\text{Earnings}|\text{Age}, \text{Education}] / \partial \text{Education} = \beta_2$ . But, one might suggest that the experiment the analyst really has in mind is the truly unobservable impact of the additional year of education on a particular individual. To carry out the experiment, it would be necessary to observe the individual twice, once under circumstances that actually occur,  $\text{Education}_i$ , and a second time under the hypothetical (**counterfactual**) circumstance,  $\text{Education}_i + 1$ . If we consider  $\text{Education}$  in this example as a **treatment**, then the real objective of the experiment is to measure the **impact of the treatment on the treated**. The ability to infer this result from nonexperimental data that essentially compares “otherwise similar individuals will be examined in Chapter 18.”

A large literature has been devoted to another intriguing question on this subject. Education is not truly “independent” in this setting. Highly motivated individuals will choose to pursue more education (for example, by going to college or graduate school) than others. By the same token, highly motivated individuals may do things that, on average, lead them to have higher incomes. If so, does a positive  $\beta_2$  that suggests an association between income and education really measure the effect of education on income, or does it reflect the result of some underlying effect on both variables? We will revisit the issue in Chapter 18.<sup>2</sup>

## 2.3 ASSUMPTIONS OF THE LINEAR REGRESSION MODEL

The linear regression model consists of a set of assumptions about how a data set will be produced by an underlying “data generating process.” The theory will specify a deterministic relationship between the dependent variable and the independent variables. The assumptions that describe the form of the model and relationships among its parts and imply appropriate estimation and inference procedures are listed in Table 2.1.

### 2.3.1 LINEARITY OF THE REGRESSION MODEL

Let the column vector  $\mathbf{x}_k$  be the  $n$  observations on variable  $x_k$ ,  $k = 1, \dots, K$ , and assemble these data in an  $n \times K$  data matrix  $\mathbf{X}$ . In most contexts, the first column of  $\mathbf{X}$  is assumed to be a column of 1s so that  $\beta_1$  is the constant term in the model. Let  $\mathbf{y}$  be the  $n$  observations,  $y_1, \dots, y_n$ , and let  $\boldsymbol{\varepsilon}$  be the column vector containing the  $n$  disturbances.

<sup>2</sup>This model lays yet another trap for the practitioner. In a cross section, the higher incomes of the older individuals in the sample might tell an entirely different, perhaps macroeconomic story (a “cohort effect”) from the lower incomes of younger individuals as time and their incomes evolve. It is not necessarily possible to deduce the characteristics of incomes of younger people in the sample if they were older by comparing the older individuals in the sample to the younger ones. A parallel problem arises in the analysis of treatment effects that we will examine in Chapter 18.

## 16 PART I ♦ The Linear Regression Model

**TABLE 2.1** Assumptions of the Linear Regression Model

- A1. Linearity:**  $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$ . The model specifies a linear relationship between  $y$  and  $x_1, \dots, x_K$ .
- A2. Full rank:** There is no exact linear relationship among any of the independent variables in the model. This assumption will be necessary for estimation of the parameters of the model.
- A3. Exogeneity of the independent variables:**  $E[\varepsilon_i | x_{j1}, x_{j2}, \dots, x_{jK}] = 0$ . This states that the expected value of the disturbance at observation  $i$  in the sample is not a function of the independent variables observed at any observation, including this one. This means that the independent variables will not carry useful information for prediction of  $\varepsilon_i$ .
- A4. Homoscedasticity and nonautocorrelation:** Each disturbance,  $\varepsilon_i$  has the same finite variance,  $\sigma^2$ , and is uncorrelated with every other disturbance,  $\varepsilon_j$ . This assumption limits the generality of the model, and we will want to examine how to relax it in the chapters to follow.
- A5. Data generation:** The data in  $(x_{j1}, x_{j2}, \dots, x_{jK})$  may be any mixture of constants and random variables. The crucial elements for present purposes are the strict mean independence assumption A3 and the implicit variance independence assumption in A4. Analysis will be done conditionally on the observed  $\mathbf{X}$ , so whether the elements in  $\mathbf{X}$  are fixed constants or random draws from a stochastic process will not influence the results. In later, more advanced treatments, we will want to be more specific about the possible relationship between  $\varepsilon_i$  and  $\mathbf{x}_j$ .
- A6. Normal distribution:** The disturbances are normally distributed. Once again, this is a convenience that we will dispense with after some analysis of its implications.

The model in (2-1) as it applies to all  $n$  observations can now be written

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_K\beta_K + \boldsymbol{\varepsilon}, \quad (2-2)$$

or in the form of Assumption 1,

$$\text{ASSUMPTION: } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \quad (2-3)$$

### A NOTATIONAL CONVENTION

Henceforth, to avoid a possibly confusing and cumbersome notation, we will use a boldface  $\mathbf{x}$  to denote a column or a row of  $\mathbf{X}$ . Which of these applies will be clear from the context. In (2-2),  $\mathbf{x}_k$  is the  $k$ th column of  $\mathbf{X}$ . Subscripts  $j$  and  $k$  will be used to denote columns (variables). It will often be convenient to refer to a single observation in (2-3), which we would write

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i. \quad (2-4)$$

Subscripts  $i$  and  $t$  will generally be used to denote rows (observations) of  $\mathbf{X}$ . In (2-4),  $\mathbf{x}_i$  is a column vector that is the transpose of the  $i$ th  $1 \times K$  row of  $\mathbf{X}$ .

Our primary interest is in estimation and inference about the parameter vector  $\boldsymbol{\beta}$ . Note that the simple regression model in Example 2.1 is a special case in which  $\mathbf{X}$  has only two columns, the first of which is a column of 1s. The assumption of linearity of the regression model includes the additive disturbance. For the regression to be linear in the sense described here, it must be of the form in (2-1) either in the original variables or after some suitable transformation. For example, the model

$$y = Ax^\beta e^\varepsilon$$

## CHAPTER 2 ♦ The Linear Regression Model 17

is linear (after taking logs on both sides of the equation), whereas

$$y = Ax^\beta + \varepsilon$$

is not. The observed dependent variable is thus the sum of two components, a deterministic element  $\alpha + \beta x$  and a random variable  $\varepsilon$ . It is worth emphasizing that neither of the two parts is directly observed because  $\alpha$  and  $\beta$  are unknown.

The linearity assumption is not so narrow as it might first appear. In the regression context, *linearity* refers to the manner in which the parameters and the disturbance enter the equation, not necessarily to the relationship among the variables. For example, the equations  $y = \alpha + \beta x + \varepsilon$ ,  $y = \alpha + \beta \cos(x) + \varepsilon$ ,  $y = \alpha + \beta/x + \varepsilon$ , and  $y = \alpha + \beta \ln x + \varepsilon$  are all linear in some function of  $x$  by the definition we have used here. In the examples, only  $x$  has been transformed, but  $y$  could have been as well, as in  $y = Ax^\beta e^\varepsilon$ , which is a linear relationship in the logs of  $x$  and  $y$ ;  $\ln y = \alpha + \beta \ln x + \varepsilon$ . The variety of functions is unlimited. This aspect of the model is used in a number of commonly used functional forms. For example, the **loglinear model** is

$$\ln y = \beta_1 + \beta_2 \ln x_2 + \beta_3 \ln x_3 + \cdots + \beta_K \ln x_K + \varepsilon.$$

This equation is also known as the **constant elasticity** form as in this equation, the elasticity of  $y$  with respect to changes in  $x$  is  $\partial \ln y / \partial \ln x_k = \beta_k$ , which does not vary with  $x_k$ . The loglinear form is often used in models of demand and production. Different values of  $\beta$  produce widely varying functions.

### Example 2.3 The U.S. Gasoline Market

Data on the U.S. gasoline market for the years 1953–2004 are given in Table F2.2 in Appendix F. We will use these data to obtain, among other things, estimates of the income, own price, and cross-price elasticities of demand in this market. These data also present an interesting question on the issue of holding “all other things constant,” that was suggested in Example 2.2. In particular, consider a somewhat abbreviated model of per capita gasoline consumption:

$$\ln(G/pop) = \beta_1 + \beta_2 \ln(Income/pop) + \beta_3 \ln price_G + \beta_4 \ln P_{newcars} + \beta_5 \ln P_{usedcars} + \varepsilon.$$

This model will provide estimates of the income and price elasticities of demand for gasoline and an estimate of the elasticity of demand with respect to the prices of new and used cars. What should we expect for the sign of  $\beta_4$ ? Cars and gasoline are complementary goods, so if the prices of new cars rise, *ceteris paribus*, gasoline consumption should fall. Or should it? If the prices of new cars rise, then consumers will buy fewer of them; they will keep their used cars longer and buy fewer new cars. If older cars use more gasoline than newer ones, then the rise in the prices of new cars would lead to higher gasoline consumption than otherwise, not lower. We can use the multiple regression model and the gasoline data to attempt to answer the question.

A **semilog** model is often used to model growth rates:

$$\ln y_t = \mathbf{x}'_t \boldsymbol{\beta} + \delta t + \varepsilon_t.$$

In this model, the autononE(at least not explained by the model itself) proportional, per period growth rate is  $d/\ln y/dt = \delta$ . Other variations of the general form

$$f(y_t) = g(\mathbf{x}'_t \boldsymbol{\beta} + \varepsilon_t)$$

will allow a tremendous variety of functional forms, all of which fit into our definition of a linear model.

## 18 PART I ♦ The Linear Regression Model

The linear regression model is sometimes interpreted as an approximation to some unknown, underlying function. (See Section A.8.1 for discussion.) By this interpretation, however, the linear model, even with quadratic terms, is fairly limited in that such an approximation is likely to be useful only over a small range of variation of the independent variables. The translog model discussed in Example 2.4, in contrast, has proved far more effective as an approximating function.

### Example 2.4 The Translog Model

Modern studies of demand and production are usually done with a **flexible functional form**. Flexible functional forms are used in econometrics because they allow analysts to model complex features of the production function, such as elasticities of substitution, which are functions of the second derivatives of production, cost, or utility functions. The linear model restricts these to equal zero, whereas the loglinear model (e.g., the Cobb-Douglas model) restricts the interesting elasticities to the uninteresting values of  $-1$  or  $+1$ . The most popular flexible functional form is the **translog model**, which is often interpreted as a second-order approximation to an unknown functional form. [See Berndt and Christensen (1973).] One way to derive it is as follows. We first write  $y = g(x_1, \dots, x_K)$ . Then,  $\ln y = \ln g(\dots) = f(\dots)$ . Since by a trivial transformation  $x_k = \exp(\ln x_k)$ , we interpret the function as a function of the logarithms of the  $x$ 's. Thus,  $\ln y = f(\ln x_1, \dots, \ln x_K)$ .

Now, expand this function in a second-order Taylor series around the point  $\mathbf{x} = [1, 1, \dots, 1]'$  so that at the expansion point, the log of each variable is a convenient zero. Then

$$\begin{aligned}\ln y &= f(\mathbf{0}) + \sum_{k=1}^K [\partial f(\cdot)/\partial \ln x_k]_{|\ln x=0} \ln x_k \\ &\quad + \frac{1}{2} \sum_{k=1}^K \sum_{l=1}^K [\partial^2 f(\cdot)/\partial \ln x_k \partial \ln x_l]_{|\ln x=0} \ln x_k \ln x_l + \varepsilon.\end{aligned}$$

The disturbance in this model is assumed to embody the familiar factors and the error of approximation to the unknown function. Since the function and its derivatives evaluated at the fixed value  $\mathbf{0}$  are constants, we interpret them as the coefficients and write

$$\ln y = \beta_0 + \sum_{k=1}^K \beta_k \ln x_k + \frac{1}{2} \sum_{k=1}^K \sum_{l=1}^K \gamma_{kl} \ln x_k \ln x_l + \varepsilon.$$

This model is linear by our definition but can, in fact, mimic an impressive amount of curvature when it is used to approximate another function. An interesting feature of this formulation is that the loglinear model is a special case,  $\gamma_{kl} = 0$ . Also, there is an interesting test of the underlying theory possible because if the underlying function were assumed to be continuous and twice continuously differentiable, then by Young's theorem it must be true that  $\gamma_{kl} = \gamma_{lk}$ . We will see in Chapter 10 how this feature is studied in practice.

Despite its great flexibility, the linear model will not accommodate all the situations we will encounter in practice. In Example 14.10 and Chapter 19, we will examine the regression model for doctor visits that was suggested in the introduction to this chapter. An appropriate model that describes the number of visits has conditional mean function  $E[y|x] = \exp(x_i \beta)$ . It is tempting to linearize this directly by taking logs, since  $\ln E[y|x] = x_i \beta$ . But,  $\ln E[y|x]$  is not equal to  $E[\ln y|x]$ . In that setting,  $y_i$  can equal zero (and does for most of the sample), so  $x_i \beta$  (which can be negative) is not an appropriate model for  $\ln y_i$  (which does not exist) nor for  $y_i$ , which cannot be negative. The methods we consider in this chapter are not appropriate for estimating the parameters of such a model. Relatively straightforward techniques have been developed for nonlinear models such as this, however. We shall treat them in detail in Chapter 11.

### 2.3.2 FULL RANK

Assumption 2 is that there are no exact linear relationships among the variables.

 ASSUMPTION:  $\mathbf{X}$  is an  $n \times K$  matrix with rank  $K$ . (2-5)

Hence,  $\mathbf{X}$  has full column rank; the columns of  $\mathbf{X}$  are linearly independent and there are at least  $K$  observations. [See (A-42) and the surrounding text.] This assumption is known as an **identification condition**. To see the need for this assumption, consider an example.

#### Example 2.5 Short Rank

Suppose that a cross-section model specifies that consumption,  $C$ , relates to income as follows:

$$C = \beta_1 + \beta_2 \text{ nonlabor income} + \beta_3 \text{ salary} + \beta_4 \text{ total income} + \varepsilon,$$

where *total income* is exactly equal to *salary* plus *nonlabor income*. Clearly, there is an exact linear dependency in the model. Now let

$$\begin{aligned}\beta'_2 &= \beta_2 + a, \\ \beta'_3 &= \beta_3 + a,\end{aligned}$$

and

$$\beta'_4 = \beta_4 - a,$$

where  $a$  is any number. Then the exact same value appears on the right-hand side of  $C$  if we substitute  $\beta'_2$ ,  $\beta'_3$ , and  $\beta'_4$  for  $\beta_2$ ,  $\beta_3$ , and  $\beta_4$ . Obviously, there is no way to estimate the parameters of this model.

If there are fewer than  $K$  observations, then  $\mathbf{X}$  cannot have **full rank**. Hence, we make the (redundant) assumption that  $n$  is at least as large as  $K$ .

In a two-variable linear model with a constant term, the full rank assumption means that there must be variation in the regressor  $x$ . If there is no variation in  $x$ , then all our observations will lie on a vertical line. This situation does not invalidate the other assumptions of the model; presumably, it is a flaw in the data set. The possibility that this suggests is that we *could* have drawn a sample in which there was variation in  $x$ , but in this instance, we did not. Thus, the model still applies, but we cannot learn about it from the data set in hand.

#### Example 2.6 An Inestimable Model

In Example 3.4, we will consider a model for the sale price of Monet paintings. Theorists and observers have different models for how prices of paintings at auction are determined. One (naïve) student of the subject suggests the model

$$\begin{aligned}\ln \text{Price} &= \beta_1 + \beta_2 \ln \text{Size} + \beta_3 \ln \text{Aspect Ratio} + \beta_4 \ln \text{Height} + \varepsilon \\ &= \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon,\end{aligned}$$

where  $\text{Size} = \text{Width} \times \text{Height}$  and  $\text{Aspect Ratio} = \text{Width}/\text{Height}$ . By simple arithmetic, we can see that this model shares the problem found with the consumption model in Example 2.5—in this case,  $x_2 - x_4 = x_3 + x_4$ . So, this model is, like the previous one, not estimable—it is not identified. It is useful to think of the problem from a different perspective here (so to speak). In the linear model, it must be possible for the variables to vary linearly independently. But, in this instance, while it is possible for any pair of the three covariates to vary independently, the three together cannot. The “model,” that is, the theory, is an entirely reasonable model

## 20 PART I ♦ The Linear Regression Model

as it stands. Art buyers might very well consider all three of these features in their valuation of a Monet painting. However, it is not possible to learn about that from the observed data, at least not with this linear regression model.

### 2.3.3 REGRESSION

The disturbance is assumed to have conditional expected value zero at every observation, which we write as

$$E[\varepsilon_i | \mathbf{X}] = 0. \quad (2-6)$$

For the full set of observations, we write Assumption 3 as

$$\text{ASSUMPTION: } E[\boldsymbol{\varepsilon} | \mathbf{X}] = \begin{bmatrix} E[\varepsilon_1 | \mathbf{X}] \\ E[\varepsilon_2 | \mathbf{X}] \\ \vdots \\ E[\varepsilon_n | \mathbf{X}] \end{bmatrix} = \mathbf{0}. \quad (2-7)$$

There is a subtle point in this discussion that the observant reader might have noted. In (2-7), the left-hand side states, in principle, that the mean of each  $\varepsilon_i$  *conditioned on all observations  $\mathbf{x}_i$*  is zero. This conditional mean assumption states, in words, that no observations on  $\mathbf{x}$  convey information about the expected value of the disturbance. It is conceivable—for example, in a time-series setting—that although  $\mathbf{x}_i$  might provide no information about  $E[\varepsilon_i | \cdot]$ ,  $\mathbf{x}_j$  at some other observation, such as in the next time period, might. Our assumption at this point is that there is no information about  $E[\varepsilon_i | \cdot]$  contained in any observation  $\mathbf{x}_j$ . Later, when we extend the model, we will study the implications of dropping this assumption. [See Wooldridge (1995).] We will also assume that the disturbances convey no information about each other. That is,  $E[\varepsilon_i | \varepsilon_1, \dots, \varepsilon_{i-1}, \varepsilon_{i+1}, \dots, \varepsilon_n] = 0$ . In sum, at this point, we have assumed that the disturbances are purely random draws from some population.

The zero conditional mean implies that the unconditional mean is also zero, since

$$E[\varepsilon_i] = E_{\mathbf{x}}[E[\varepsilon_i | \mathbf{X}]] = E_{\mathbf{x}}[0] = 0.$$

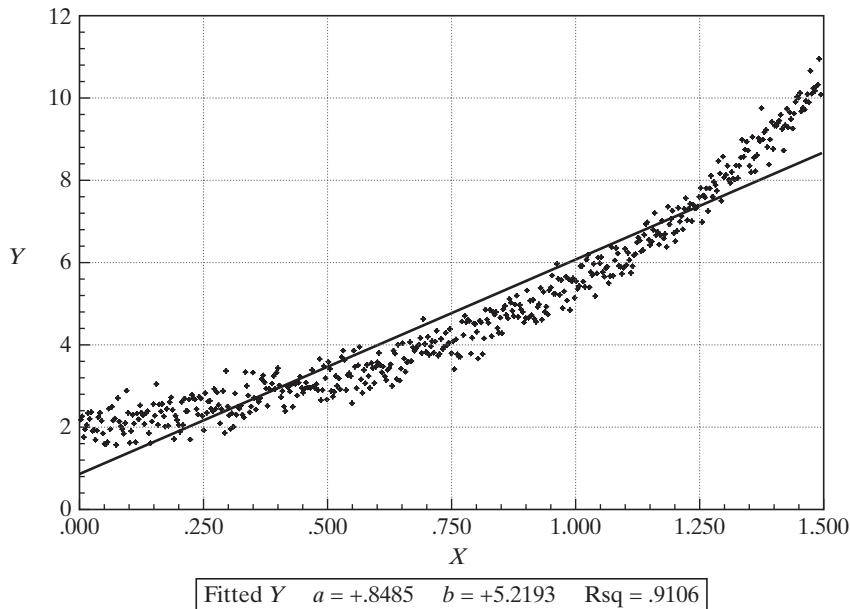
Since, for each  $\varepsilon_i$ ,  $\text{Cov}[E[\varepsilon_i | \mathbf{X}], \mathbf{X}] = \text{Cov}[\varepsilon_i, \mathbf{X}]$ , Assumption 3 implies that  $\text{Cov}[\varepsilon_i, \mathbf{X}] = 0$  for all  $i$ . The converse is not true;  $E[\varepsilon_i] = 0$  does not imply that  $E[\varepsilon_i | \mathbf{x}_i] = 0$ . Example 2.7 illustrates the difference.

#### **Example 2.7 Nonzero Conditional Mean of the Disturbances**

Figure 2.2 illustrates the important difference between  $E[\varepsilon_i] = 0$  and  $E[\varepsilon_i | x_i] = 0$ . The overall mean of the disturbances in the sample is zero, but the mean for specific ranges of  $x$  is distinctly nonzero. A pattern such as this in observed data would serve as a useful indicator that the assumption of the linear regression should be questioned. In this particular case, the true conditional mean function (which the researcher would not know in advance) is actually  $E[y|x] = 1 + \exp(1.5x)$ . The sample data are suggesting that the linear model is not appropriate for these data. This possibility is pursued in an application in Example 6.6.

In most cases, the zero overall mean assumption is not restrictive. Consider a two-variable model and suppose that the mean of  $\varepsilon$  is  $\mu \neq 0$ . Then  $\alpha + \beta x + \varepsilon$  is the same as  $(\alpha + \mu) + \beta x + (\varepsilon - \mu)$ . Letting  $\alpha' = \alpha + \mu$  and  $\varepsilon' = \varepsilon - \mu$  produces the original model. For an application, see the discussion of frontier production functions in Chapter 18. But, if the original model does not contain a constant term, then assuming  $E[\varepsilon_i] = 0$  could be

## CHAPTER 2 ♦ The Linear Regression Model 21



**FIGURE 2.2** Disturbances with Nonzero Conditional Mean and Zero Unconditional Mean.

substantive. This suggests that there is a potential problem in models without constant terms. As a general rule, regression models should not be specified without constant terms unless this is specifically dictated by the underlying theory.<sup>3</sup> Arguably, if we have reason to specify that the mean of the disturbance is something other than zero, we should build it into the systematic part of the regression, leaving in the disturbance only the unknown part of  $\varepsilon$ . Assumption 3 also implies that

$$E[\mathbf{y} | \mathbf{X}] = \mathbf{X}\boldsymbol{\beta}. \quad (2-8)$$

Assumptions 1 and 3 comprise the *linear regression model*. The **regression** of  $\mathbf{y}$  on  $\mathbf{X}$  is the conditional mean,  $E[\mathbf{y} | \mathbf{X}]$ , so that without Assumption 3,  $\mathbf{X}\boldsymbol{\beta}$  is *not* the conditional mean function.

The remaining assumptions will more completely specify the characteristics of the disturbances in the model and state the conditions under which the sample observations on  $\mathbf{x}$  are obtained.

#### 2.3.4 SPHERICAL DISTURBANCES

The fourth assumption concerns the variances and covariances of the disturbances:

$$\text{Var}[\varepsilon_i | \mathbf{X}] = \sigma^2, \quad \text{for all } i = 1, \dots, n,$$

<sup>3</sup>Models that describe first differences of variables might well be specified without constants. Consider  $y_t - y_{t-1}$ . If there is a constant term  $\alpha$  on the right-hand side of the equation, then  $y_t$  is a function of  $\alpha t$ , which is an explosive regressor. Models with linear time trends merit special treatment in the time-series literature. We will return to this issue in Chapter 21.

## 22 PART I ♦ The Linear Regression Model

and

$$\text{Cov}[\varepsilon_i, \varepsilon_j | \mathbf{X}] = 0, \quad \text{for all } i \neq j.$$

Constant variance is labeled **homoscedasticity**. Consider a model that describes the profits of firms in an industry as a function of, say, size. Even accounting for size, measured in dollar terms, the profits of large firms will exhibit greater variation than those of smaller firms. The homoscedasticity assumption would be inappropriate here. Survey data on household expenditure patterns often display marked **heteroscedasticity**, even after accounting for income and household size.

Uncorrelatedness across observations is labeled generically **nonautocorrelation**. In Figure 2.1, there is some suggestion that the disturbances might not be truly independent across observations. Although the number of observations is limited, it does appear that, on average, each disturbance tends to be followed by one with the same sign. This “inertia” is precisely what is meant by **autocorrelation**, and it is assumed away at this point. Methods of handling autocorrelation in economic data occupy a large proportion of the literature and will be treated at length in Chapter 20. Note that nonautocorrelation does not imply that observations  $y_i$  and  $y_j$  are uncorrelated. The assumption is that *deviations* of observations from their expected values are uncorrelated.

The two assumptions imply that

$$\begin{aligned} E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}] &= \begin{bmatrix} E[\varepsilon_1\varepsilon_1 | \mathbf{X}] & E[\varepsilon_1\varepsilon_2 | \mathbf{X}] & \cdots & E[\varepsilon_1\varepsilon_n | \mathbf{X}] \\ E[\varepsilon_2\varepsilon_1 | \mathbf{X}] & E[\varepsilon_2\varepsilon_2 | \mathbf{X}] & \cdots & E[\varepsilon_2\varepsilon_n | \mathbf{X}] \\ \vdots & \vdots & \vdots & \vdots \\ E[\varepsilon_n\varepsilon_1 | \mathbf{X}] & E[\varepsilon_n\varepsilon_2 | \mathbf{X}] & \cdots & E[\varepsilon_n\varepsilon_n | \mathbf{X}] \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ & \vdots & & \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}, \end{aligned}$$

which we summarize in Assumption 4:

ASSUMPTION:  $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}] = \sigma^2 \mathbf{I}$ .

(2-9)

By using the variance decomposition formula in (B-69), we find

$$\text{Var}[\boldsymbol{\varepsilon}] = E[\text{Var}[\boldsymbol{\varepsilon} | \mathbf{X}]] + \text{Var}[E[\boldsymbol{\varepsilon} | \mathbf{X}]] = \sigma^2 \mathbf{I}.$$

Once again, we should emphasize that this assumption describes the information about the variances and covariances among the disturbances that is provided by the independent variables. For the present, we assume that there is none. We will also drop this assumption later when we enrich the regression model. We are also assuming that the disturbances themselves provide no information about the variances and covariances. Although a minor issue at this point, it will become crucial in our treatment of time-series applications. Models such as  $\text{Var}[\varepsilon_t | \varepsilon_{t-1}] = \sigma^2 + \alpha\varepsilon_{t-1}^2$ , a “GARCH” model (see Chapter 20), do not violate our conditional variance assumption, but do assume that  $\text{Var}[\varepsilon_t | \varepsilon_{t-1}] \neq \text{Var}[\varepsilon_t]$ .

## CHAPTER 2 ♦ The Linear Regression Model 23

Disturbances that meet the assumptions of homoscedasticity and nonautocorrelation are sometimes called **spherical disturbances**.<sup>4</sup>

### 2.3.5 DATA GENERATING PROCESS FOR THE REGRESSORS

It is common to assume that  $\mathbf{x}_i$  is nonstochastic, as it would be in an experimental situation. Here the analyst chooses the values of the regressors and then observes  $y_i$ . This process might apply, for example, in an agricultural experiment in which  $y_i$  is yield and  $\mathbf{x}_i$  is fertilizer concentration and water applied. The assumption of **nonstochastic regressors** at this point would be a mathematical convenience. With it, we could use the results of elementary statistics to obtain our results by treating the vector  $\mathbf{x}_i$  simply as a known constant in the probability distribution of  $y_i$ . With this simplification, Assumptions A3 and A4 would be made unconditional and the counterparts would now simply state that the probability distribution of  $\varepsilon_i$  involves none of the constants in  $\mathbf{X}$ .

Social scientists are almost never able to analyze experimental data, and relatively few of their models are built around nonrandom regressors. Clearly, for example, in any model of the macroeconomy, it would be difficult to defend such an asymmetric treatment of aggregate data. Realistically, we have to allow the data on  $\mathbf{x}_i$  to be random the same as  $y_i$ , so an alternative formulation is to assume that  $\mathbf{x}_i$  is a random vector and our formal assumption concerns the nature of the random process that produces  $\mathbf{x}_i$ . If  $\mathbf{x}_i$  is taken to be a random vector, then Assumptions 1 through 4 become a statement about the joint distribution of  $y_i$  and  $\mathbf{x}_i$ . The precise nature of the regressor and how we view the sampling process will be a major determinant of our derivation of the statistical properties of our estimators and test statistics. In the end, the crucial assumption is Assumption 3, the uncorrelatedness of  $\mathbf{X}$  and  $\boldsymbol{\varepsilon}$ . Now, we do note that this alternative is not completely satisfactory either, since  $\mathbf{X}$  may well contain nonstochastic elements, including a constant, a time trend, and dummy variables that mark specific episodes in time. This makes for an ambiguous conclusion, but there is a straightforward and economically useful way out of it. We will assume that  $\mathbf{X}$  can be a mixture of constants and random variables, and the mean and variance of  $\varepsilon_i$  are both independent of all elements of  $\mathbf{X}$ .

ASSUMPTION:  $\mathbf{X}$  may be fixed or random.

(2-10)

### 2.3.6 NORMALITY

It is convenient to assume that the disturbances are **normally distributed**, with zero mean and constant variance. That is, we add normality of the distribution to Assumptions 3 and 4.

ASSUMPTION:  $\boldsymbol{\varepsilon} | \mathbf{X} \sim N[\mathbf{0}, \sigma^2 \mathbf{I}]$ .

(2-11)

<sup>4</sup>The term will describe the multivariate normal distribution; see (B-95). If  $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$  in the multivariate normal density, then the equation  $f(\mathbf{x}) = c$  is the formula for a “ball” centered at  $\mu$  with radius  $\sigma$  in  $n$ -dimensional space. The name *spherical* is used whether or not the normal distribution is assumed; sometimes the “spherical normal” distribution is assumed explicitly.

## 24 PART I ♦ The Linear Regression Model

In view of our description of the source of  $\epsilon$ , the conditions of the central limit theorem will generally apply, at least approximately, and the normality assumption will be reasonable in most settings. A useful implication of Assumption 6 is that it implies that observations on  $\epsilon_i$  are statistically independent as well as uncorrelated. [See the third point in Section B.9, (B-97) and (B-99).] **Normality** is often viewed as an unnecessary and possibly inappropriate addition to the regression model. Except in those cases in which some alternative distribution is explicitly assumed, as in the stochastic frontier model discussed in Chapter 18, the normality assumption is probably quite reasonable.

Normality is not necessary to obtain many of the results we use in multiple regression analysis, although it will enable us to obtain several exact statistical results. It does prove useful in constructing confidence intervals and test statistics, as shown in Section 4.5 and Chapter 5. Later, it will be possible to relax this assumption and retain most of the statistical results we obtain here. (See Sections 4.4 and 5.6.)

### 2.3.7 INDEPENDENCE

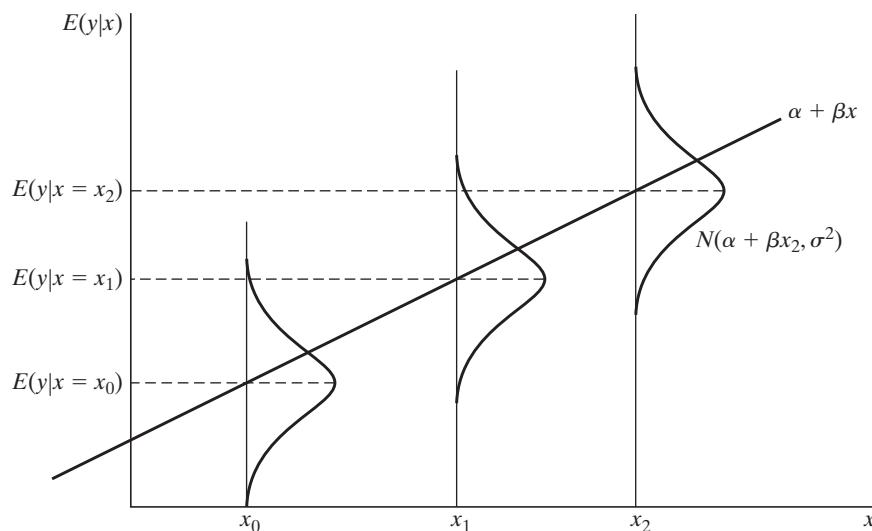
The term “independent” has been used several ways in this chapter.

In Section 2.2, the right-hand-side variables in the model are denoted the independent variables. Here, the notion of independence refers to the sources of variation. In the context of the model, the variation in the independent variables arises from sources that are outside of the process being described. Thus, in our health services vs. income example in the introduction, we have suggested a theory for how variation in demand for services is associated with variation in income. But, we have not suggested an explanation of the sample variation in incomes; income is assumed to vary for reasons that are outside the scope of the model.

The assumption in (2-6),  $E[\epsilon_i | \mathbf{X}] = 0$ , is **mean independence**. Its implication is that variation in the disturbances in our data is not explained by variation in the independent variables. We have also assumed in Section 2.3.4 that the disturbances are uncorrelated with each other (Assumption A4 in Table 2.1). This implies that  $E[\epsilon_i | \epsilon_j] = 0$  when  $i \neq j$ —the disturbances are also mean independent of each other. Conditional normality of the disturbances assumed in Section 2.3.6 (Assumption A6) implies that they are **statistically independent** of each other, which is a stronger result than mean independence.

Finally, Section 2.3.2 discusses the **linear independence** of the columns of the data matrix,  $\mathbf{X}$ . The notion of independence here is an algebraic one relating to the column rank of  $\mathbf{X}$ . In this instance, the underlying interpretation is that it must be possible for the variables in the model to vary linearly independently of each other. Thus, in Example 2.6, we find that it is not possible for the logs of surface area, aspect ratio, and height of a painting all to vary independently of one another. The modeling implication is that if the variables cannot vary independently of each other, then it is not possible to analyze them in a linear regression model that assumes the variables can each vary while holding the others constant. There is an ambiguity in this discussion of independence of the variables. We have both age and age squared in a model in Example 2.2. These cannot vary independently, but there is no obstacle to formulating a regression model containing both age and age squared. The resolution is that age and age squared, though not *functionally* independent, are *linearly* independent. That is the crucial assumption in the linear regression model.

## CHAPTER 2 ♦ The Linear Regression Model 25

**FIGURE 2.3** The Classical Regression Model.**2.4 SUMMARY AND CONCLUSIONS**

This chapter has framed the linear regression model, the basic platform for model building in econometrics. The assumptions of the classical regression model are summarized in Figure 2.3, which shows the two-variable case.

**Key Terms and Concepts**

- Autocorrelation
- Conditional median
- Conditional variation
- Constant elasticity
- Counter factual
- Covariate
- Dependent variable
- Deterministic relationship
- Disturbance
- Exogeneity
- Explained variable
- Explanatory variable
- Flexible functional form
- Full rank
- Heteroscedasticity
- Homoscedasticity
- Identification condition
- Impact of treatment on the treated
- Independent variable
- Linear independence
- Linear regression model
- Loglinear model
- Mean independence
- Multiple linear regression model
- Nonautocorrelation
- Nonstochastic regressors
- Normality
- Normally distributed
- Path diagram
- Population regression equation
- Regressand
- Regression
- Regressor
- Second-order effects
- Semilog
- Spherical disturbances
- Translog model



# 3

## LEAST SQUARES

---

### 3.1 INTRODUCTION

Chapter 2 defined the linear regression model as a set of characteristics of the population that underlies an observed sample of data. There are a number of different approaches to estimation of the parameters of the model. For a variety of practical and theoretical reasons that we will explore as we progress through the next several chapters, the method of least squares has long been the most popular. Moreover, in most cases in which some other estimation method is found to be preferable, least squares remains the benchmark approach, and often, the preferred method ultimately amounts to a modification of least squares. In this chapter, we begin the analysis of this important set of results by presenting a useful set of algebraic tools.

### 3.2 LEAST SQUARES REGRESSION

The unknown parameters of the stochastic relation  $y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i$  are the objects of estimation. It is necessary to distinguish between population quantities, such as  $\boldsymbol{\beta}$  and  $\varepsilon_i$ , and sample estimates of them, denoted  $\mathbf{b}$  and  $e_i$ . The **population regression** is  $E[y_i | \mathbf{x}_i] = \mathbf{x}'_i \boldsymbol{\beta}$ , whereas our estimate of  $E[y_i | \mathbf{x}_i]$  is denoted

$$\hat{y}_i = \mathbf{x}'_i \mathbf{b}.$$

The **disturbance** associated with the  $i$ th data point is

$$\varepsilon_i = y_i - \mathbf{x}'_i \boldsymbol{\beta}.$$

For any value of  $\mathbf{b}$ , we shall estimate  $\varepsilon_i$  with the **residual**,

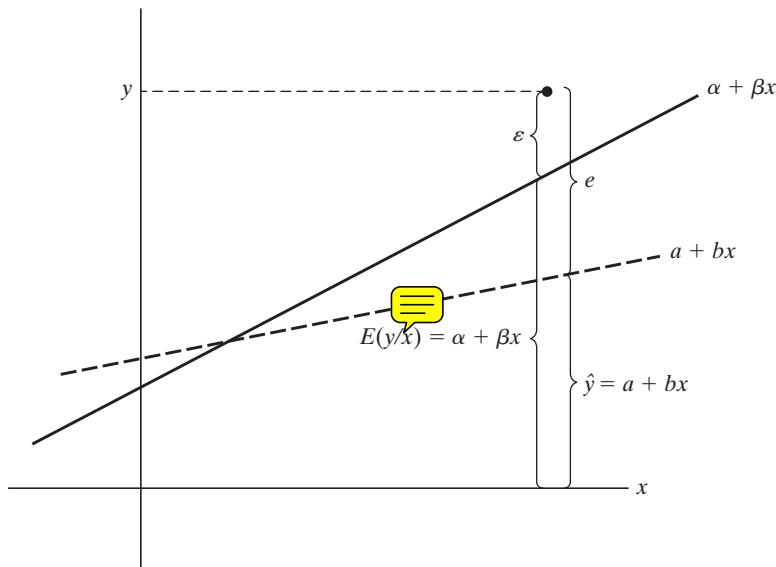
$$e_i = y_i - \mathbf{x}'_i \mathbf{b}.$$

From the definitions,

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i = \mathbf{x}'_i \mathbf{b} + e_i.$$

These equations are summarized for the two variable regression in Figure 3.1.

The **population quantity**  $\boldsymbol{\beta}$  is a vector of unknown parameters of the probability distribution of  $y_i$  whose values we hope to estimate with our sample data,  $(y_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ . This is a problem of statistical inference. It is instructive, however, to begin by considering the purely algebraic problem of choosing a vector  $\mathbf{b}$  so that the fitted line  $\mathbf{x}'_i \mathbf{b}$  is close to the data points. The measure of closeness constitutes a **fitting criterion**.



**FIGURE 3.1** Population and Sample Regression.

Although numerous candidates have been suggested, the one used most frequently is **least squares**.<sup>1</sup>

### 3.2.1 THE LEAST SQUARES COEFFICIENT VECTOR

The least squares coefficient vector minimizes the sum of squared residuals:

$$\sum_{i=1}^n e_{i0}^2 = \sum_{i=1}^n (y_i - \mathbf{x}'_i \mathbf{b}_0)^2, \quad (3-1)$$

where  $\mathbf{b}_0$  denotes the choice for the coefficient vector. In matrix terms, minimizing the sum of squares in (3-1) requires us to choose  $\mathbf{b}_0$  to

$$\text{Minimize}_{\mathbf{b}_0} S(\mathbf{b}_0) = \mathbf{e}'_0 \mathbf{e}_0 = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'(\mathbf{y} - \mathbf{X}\mathbf{b}_0). \quad (3-2)$$

Expanding this gives

$$\mathbf{e}'_0 \mathbf{e}_0 = \mathbf{y}'\mathbf{y} - \mathbf{b}'_0 \mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b}_0 + \mathbf{b}'_0 \mathbf{X}'\mathbf{X}\mathbf{b}_0 \quad (3-3)$$

or

$$S(\mathbf{b}_0) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b}_0 + \mathbf{b}'_0 \mathbf{X}'\mathbf{X}\mathbf{b}_0.$$

The necessary condition for a minimum is

$$\frac{\partial S(\mathbf{b}_0)}{\partial \mathbf{b}_0} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}_0 = \mathbf{0}.^2 \quad (3-4)$$

<sup>1</sup>We have yet to establish that the practical approach of fitting the line as closely as possible to the data by least squares leads to estimates with good statistical properties. This makes intuitive sense and is, indeed, the case. We shall return to the statistical issues in Chapter 4.

<sup>2</sup>See Appendix A.8 for discussion of calculus results involving matrices and vectors.

## 28 PART I ♦ The Linear Regression Model

Let  $\mathbf{b}$  be the solution. Then, after manipulating (3-4), we find that  $\mathbf{b}$  satisfies the **least squares normal equations**,

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}. \quad (3-5)$$

If the inverse of  $\mathbf{X}'\mathbf{X}$  exists, which follows from the full column rank assumption (Assumption A2 in Section 2.3), then the solution is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (3-6)$$

For this solution to minimize the sum of squares,

$$\frac{\partial^2 S(\mathbf{b}_0)}{\partial \mathbf{b}_0 \partial \mathbf{b}'_0} = 2\mathbf{X}'\mathbf{X}$$

must be a positive definite matrix. Let  $q = \mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c}$  for some arbitrary nonzero vector  $\mathbf{c}$ . Then

$$q = \mathbf{v}'\mathbf{v} = \sum_{i=1}^n v_i^2, \quad \text{where } \mathbf{v} = \mathbf{X}\mathbf{c}.$$

Unless every element of  $\mathbf{v}$  is zero,  $q$  is positive. But if  $\mathbf{v}$  could be zero, then  $\mathbf{v}$  would be a linear combination of the columns of  $\mathbf{X}$  that equals  $\mathbf{0}$ , which contradicts the assumption that  $\mathbf{X}$  has full column rank. Since  $\mathbf{c}$  is arbitrary,  $q$  is positive for every nonzero  $\mathbf{c}$ , which establishes that  $2\mathbf{X}'\mathbf{X}$  is positive definite. Therefore, if  $\mathbf{X}$  has full column rank, then the least squares solution  $\mathbf{b}$  is unique and minimizes the sum of squared residuals.

### 3.2.2 APPLICATION: AN INVESTMENT EQUATION

To illustrate the computations in a multiple regression, we consider an example based on the macroeconomic data in Appendix Table F3.1. To estimate an investment equation, we first convert the investment and GNP series in Table F3.1 to real terms by dividing them by the CPI and then scale the two series so that they are measured in trillions of dollars. The other variables in the regression are a time trend ( $1, 2, \dots$ ), an interest rate, and the rate of inflation computed as the percentage change in the CPI. These produce the data matrices listed in Table 3.1. Consider first a regression of real investment on a constant, the time trend, and real GNP, which correspond to  $x_1, x_2$ , and  $x_3$ . (For reasons to be discussed in Chapter 23, this is probably not a well-specified equation for these macroeconomic variables. It will suffice for a simple numerical example, however.) Inserting the specific variables of the example into (3-5), we have

$$\begin{aligned} b_1 n + b_2 \sum_i T_i + b_3 \sum_i G_i &= \sum_i Y_i, \\ b_1 \sum_i T_i + b_2 \sum_i T_i^2 + b_3 \sum_i T_i G_i &= \sum_i T_i Y_i, \\ b_1 \sum_i G_i + b_2 \sum_i T_i G_i + b_3 \sum_i G_i^2 &= \sum_i G_i Y_i. \end{aligned}$$

A solution can be obtained by first dividing the first equation by  $n$  and rearranging it to obtain

$$\begin{aligned} b_1 &= \bar{Y} - b_2 \bar{T} - b_3 \bar{G} \\ &= 0.20333 - b_2 \times 8 - b_3 \times 1.2873. \end{aligned} \quad (3-7)$$

**TABLE 3.1** Data Matrices

<i>Real Investment (Y)</i>	<i>Constant (I)</i>	<i>Trend (T)</i>	<i>Real GNP (G)</i>	<i>Interest Rate (R)</i>	<i>Inflation Rate (P)</i>
0.161	1	1	1.058	5.16	4.40
0.172	1	2	1.088	5.87	5.15
0.158	1	3	1.086	5.95	5.37
0.173	1	4	1.122	4.88	4.99
0.195	1	5	1.186	4.50	4.16
0.217	1	6	1.254	6.44	5.75
0.199	1	7	1.246	7.83	8.82
$y = 0.163$	$X = 1$	8	1.232	6.25	9.31
0.195	1	9	1.298	5.50	5.21
0.231	1	10	1.370	5.46	5.83
0.257	1	11	1.439	7.46	7.40
0.259	1	12	1.479	10.28	8.64
0.225	1	13	1.474	11.77	9.31
0.241	1	14	1.503	13.42	9.44
0.204	1	15	1.475	11.02	5.99

*Note:* Subsequent results are based on these values. Slightly different results are obtained if the raw data in Table F3.1 are input to the computer program and transformed internally.

Insert this solution in the second and third equations, and rearrange terms again to yield a set of two equations:

$$\begin{aligned} b_2 \Sigma_i (T_i - \bar{T})^2 + b_3 \Sigma_i (T_i - \bar{T})(G_i - \bar{G}) &= \Sigma_i (T_i - \bar{T})(Y_i - \bar{Y}), \\ b_2 \Sigma_i (T_i - \bar{T})(G_i - \bar{G}) + b_3 \Sigma_i (G_i - \bar{G})^2 &= \Sigma_i (G_i - \bar{G})(Y_i - \bar{Y}). \end{aligned} \quad (3-8)$$

This result shows the nature of the solution for the slopes, which can be computed from the sums of squares and cross products of the deviations of the variables. Letting lowercase letters indicate variables measured as deviations from the sample means, we find that the least squares solutions for  $b_2$  and  $b_3$  are

$$\begin{aligned} b_2 &= \frac{\Sigma_i t_i y_i \Sigma_i g_i^2 - \Sigma_i g_i y_i \Sigma_i t_i g_i}{\Sigma_i t_i^2 \Sigma_i g_i^2 - (\Sigma_i g_i t_i)^2} = \frac{1.6040(0.359609) - 0.066196(9.82)}{280(0.359609) - (9.82)^2} = -0.0171984, \\ b_3 &= \frac{\Sigma_i g_i y_i \Sigma_i t_i^2 - \Sigma_i t_i y_i \Sigma_i t_i g_i}{\Sigma_i t_i^2 \Sigma_i g_i^2 - (\Sigma_i g_i t_i)^2} = \frac{0.066196(280) - 1.6040(9.82)}{280(0.359609) - (9.82)^2} = 0.653723. \end{aligned}$$

With these solutions in hand,  $b_1$  can now be computed using (3-7);  $b_1 = -0.500639$ .

Suppose that we just regressed investment on the constant and GNP, omitting the time trend. At least some of the correlation we observe in the data will be explainable because both investment and real GNP have an obvious time trend. Consider how this shows up in the regression computation. Denoting by " $b_{yx}$ " the slope in the simple, **bivariate regression** of variable  $y$  on a constant and the variable  $x$ , we find that the slope in this reduced regression would be

$$b_{yg} = \frac{\Sigma_i g_i y_i}{\Sigma_i g_i^2} = 0.184078. \quad (3-9)$$

### 30 PART I ♦ The Linear Regression Model

Now divide both the numerator and denominator in the expression for  $b_3$  by  $\Sigma_i t_i^2 \Sigma_i g_i^2$ . By manipulating it a bit and using the definition of the sample correlation between  $G$  and  $T$ ,  $r_{gt}^2 = (\Sigma_i g_i t_i)^2 / (\Sigma_i g_i^2 \Sigma_i t_i^2)$ , and defining  $b_{yt}$  and  $b_{tg}$  likewise, we obtain

$$b_{yg-t} = \frac{b_{yg}}{1 - r_{gt}^2} - \frac{b_{yt} b_{tg}}{1 - r_{gt}^2} = 0.653723. \quad (3-10)$$

(The notation “ $b_{yg-t}$ ” used on the left-hand side is interpreted to mean the slope in the regression of  $y$  on  $g$  “in the presence of  $t$ .”) The slope in the **multiple regression** differs from that in the simple regression by including a correction that accounts for the influence of the additional variable  $t$  on both  $Y$  and  $G$ . For a striking example of this effect, in the simple regression of real investment on a time trend,  $b_{yt} = 1.604/280 = 0.0057286$ , a positive number that reflects the upward trend apparent in the data. But, in the multiple regression, after we account for the influence of GNP on real investment, the slope on the time trend is  $-0.0171984$ , indicating instead a downward trend. The general result for a three-variable regression in which  $x_1$  is a constant term is

$$b_{y2.3} = \frac{b_{y2} - b_{y3} b_{32}}{1 - r_{23}^2}. \quad (3-11)$$

It is clear from this expression that the magnitudes of  $b_{y2.3}$  and  $b_{y2}$  can be quite different. They need not even have the same sign.

In practice, you will never actually compute a multiple regression by hand or with a calculator. For a regression with more than three variables, the tools of matrix algebra are indispensable (as is a computer). Consider, for example, an enlarged model of investment that includes—in addition to the constant, time trend, and GNP—an interest rate and the rate of inflation. Least squares requires the simultaneous solution of five normal equations. Letting  $\mathbf{X}$  and  $\mathbf{y}$  denote the full data matrices shown previously, the normal equations in (3-5) are

$$\begin{bmatrix} 15.000 & 120.00 & 19.310 & 111.79 & 99.770 \\ 120.000 & 1240.0 & 164.30 & 1035.9 & 875.60 \\ 19.310 & 164.30 & 25.218 & 148.98 & 131.22 \\ 111.79 & 1035.9 & 148.98 & 953.86 & 799.02 \\ 99.770 & 875.60 & 131.22 & 799.02 & 716.67 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} 3.0500 \\ 26.004 \\ 3.9926 \\ 23.521 \\ 20.732 \end{bmatrix}.$$

The solution is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (-0.50907, -0.01658, 0.67038, -0.002326, -0.00009401)'.$$

#### 3.2.3 ALGEBRAIC ASPECTS OF THE LEAST SQUARES SOLUTION

The normal equations are

$$\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{X}'\mathbf{y} = -\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = -\mathbf{X}'\mathbf{e} = \mathbf{0}. \quad (3-12)$$

Hence, for every column  $\mathbf{x}_k$  of  $\mathbf{X}$ ,  $\mathbf{x}_k' \mathbf{e} = 0$ . If the first column of  $\mathbf{X}$  is a column of 1s, which we denote  $\mathbf{i}$ , then there are three implications.

## CHAPTER 3 ♦ Least Squares 31

1. *The least squares residuals sum to zero.* This implication follows from  $\mathbf{x}_i' \mathbf{e} = \mathbf{i}' \mathbf{e} = \sum_i e_i = 0$ .
2. *The regression hyperplane passes through the point of means of the data.* The first normal equation implies that  $\bar{y} = \bar{\mathbf{x}}' \mathbf{b}$ .
3. *The mean of the fitted values from the regression equals the mean of the actual values.* This implication follows from point 1 because the fitted values are just  $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$ .

It is important to note that none of these results need hold if the regression does not contain a constant term.

### 3.2.4 PROJECTION

The vector of least squares residuals is

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}. \quad (3-13)$$

Inserting the result in (3-6) for  $\mathbf{b}$  gives

$$\mathbf{e} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = \mathbf{M}\mathbf{y}. \quad (3-14)$$

The  $n \times n$  matrix  $\mathbf{M}$  defined in (3-14) is fundamental in regression analysis. You can easily show that  $\mathbf{M}$  is both symmetric ( $\mathbf{M} = \mathbf{M}'$ ) and idempotent ( $\mathbf{M} = \mathbf{M}^2$ ). In view of (3-13), we can interpret  $\mathbf{M}$  as a matrix that produces the vector of least squares residuals in the regression of  $\mathbf{y}$  on  $\mathbf{X}$  when it premultiplies any vector  $\mathbf{y}$ . (It will be convenient later on to refer to this matrix as a “**residual maker**.”) It follows that

$$\mathbf{M}\mathbf{X} = \mathbf{0}. \quad (3-15)$$

One way to interpret this result is that if  $\mathbf{X}$  is regressed on  $\mathbf{X}$ , a perfect fit will result and the residuals will be zero.

Finally, (3-13) implies that  $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ , which is the sample analog to (2-3). (See Figure 3.1 as well.) The least squares results partition  $\mathbf{y}$  into two parts, the fitted values  $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$  and the residuals  $\mathbf{e}$ . [See Section A.3.7, especially (A-54).] Since  $\mathbf{M}\mathbf{X} = \mathbf{0}$ , these two parts are orthogonal. Now, given (3-13),

$$\hat{\mathbf{y}} = \mathbf{y} - \mathbf{e} = (\mathbf{I} - \mathbf{M})\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}\mathbf{y}. \quad (3-16)$$

The matrix  $\mathbf{P}$ , which is also symmetric and idempotent, is a **projection matrix**. It is the matrix formed from  $\mathbf{X}$  such that when a vector  $\mathbf{y}$  is premultiplied by  $\mathbf{P}$ , the result is the fitted values in the least squares regression of  $\mathbf{y}$  on  $\mathbf{X}$ . This is also the **projection** of the vector  $\mathbf{y}$  into the column space of  $\mathbf{X}$ . (See Sections A3.5 and A3.7.) By multiplying it out, you will find that, like  $\mathbf{M}$ ,  $\mathbf{P}$  is symmetric and idempotent. Given the earlier results, it also follows that  $\mathbf{M}$  and  $\mathbf{P}$  are orthogonal;

$$\mathbf{PM} = \mathbf{MP} = \mathbf{0}.$$

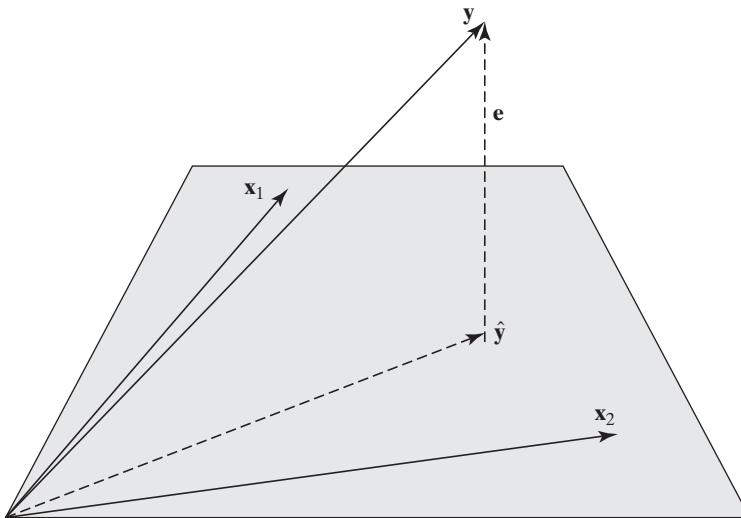
Finally, as might be expected from (3-15)

$$\mathbf{PX} = \mathbf{X}.$$

As a consequence of (3-14) and (3-16), we can see that least squares partitions the vector  $\mathbf{y}$  into two orthogonal parts,

$$\mathbf{y} = \mathbf{Py} + \mathbf{My} = \text{projection} + \text{residual}.$$

### 32 PART I ♦ The Linear Regression Model



**FIGURE 3.2** Projection of  $\mathbf{y}$  into the Column Space of  $\mathbf{X}$ .

The result is illustrated in Figure 3.2 for the two variable case. The gray shaded plane is the column space of  $\mathbf{X}$ . The projection and residual are the orthogonal dotted rays. We can also see the Pythagorean theorem at work in the sums of squares,

$$\begin{aligned}\mathbf{y}'\mathbf{y} &= \mathbf{y}'\mathbf{P}'\mathbf{P}\mathbf{y} + \mathbf{y}'\mathbf{M}'\mathbf{M}\mathbf{y} \\ &= \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e}\end{aligned}$$

In manipulating equations involving least squares results, the following equivalent expressions for the sum of squared residuals are often useful:

$$\begin{aligned}\mathbf{e}'\mathbf{e} &= \mathbf{y}'\mathbf{M}'\mathbf{M}\mathbf{y} = \mathbf{y}'\mathbf{M}\mathbf{y} = \mathbf{y}'\mathbf{e} = \mathbf{e}'\mathbf{y}, \\ \mathbf{e}'\mathbf{e} &= \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b}.\end{aligned}$$

### 3.3 PARTITIONED REGRESSION AND PARTIAL REGRESSION

It is common to specify a multiple regression model when, in fact, interest centers on only one or a subset of the full set of variables. Consider the earnings equation discussed in Example 2.2. Although we are primarily interested in the association of earnings and education, age is, of necessity, included in the model. The question we consider here is what computations are involved in obtaining, in isolation, the coefficients of a subset of the variables in a multiple regression (for example, the coefficient of education in the aforementioned regression).

Suppose that the regression involves two sets of variables,  $\mathbf{X}_1$  and  $\mathbf{X}_2$ . Thus,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

What is the algebraic solution for  $\mathbf{b}_2$ ? The **normal equations** are

$$\begin{aligned} (1) \quad & \left[ \begin{matrix} \mathbf{X}'_1 \mathbf{X}_1 & \mathbf{X}'_1 \mathbf{X}_2 \\ \mathbf{X}'_2 \mathbf{X}_1 & \mathbf{X}'_2 \mathbf{X}_2 \end{matrix} \right] \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'_1 \mathbf{y} \\ \mathbf{X}'_2 \mathbf{y} \end{bmatrix}. \\ (2) \quad & \end{aligned} \quad (3-17)$$

A solution can be obtained by using the partitioned inverse matrix of (A-74). Alternatively, (1) and (2) in (3-17) can be manipulated directly to solve for  $\mathbf{b}_2$ . We first solve (1) for  $\mathbf{b}_1$ :

$$\mathbf{b}_1 = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{y} - (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{X}_2 \mathbf{b}_2 = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 (\mathbf{y} - \mathbf{X}_2 \mathbf{b}_2). \quad (3-18)$$

This solution states that  $\mathbf{b}_1$  is the set of coefficients in the regression of  $\mathbf{y}$  on  $\mathbf{X}_1$ , minus a correction vector. We digress briefly to examine an important result embedded in (3-18). Suppose that  $\mathbf{X}'_1 \mathbf{X}_2 = \mathbf{0}$ . Then,  $\mathbf{b}_1 = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{y}$ , which is simply the coefficient vector in the regression of  $\mathbf{y}$  on  $\mathbf{X}_1$ . The general result is given in the following theorem.

### **THEOREM 3.1 Orthogonal Partitioned Regression**

*In the multiple linear least squares regression of  $\mathbf{y}$  on two sets of variables  $\mathbf{X}_1$  and  $\mathbf{X}_2$ , if the two sets of variables are orthogonal, then the separate coefficient vectors can be obtained by separate regressions of  $\mathbf{y}$  on  $\mathbf{X}_1$  alone and  $\mathbf{y}$  on  $\mathbf{X}_2$  alone.*

**Proof:** The assumption of the theorem is that  $\mathbf{X}'_1 \mathbf{X}_2 = \mathbf{0}$  in the normal equations in (3-17). Inserting this assumption into (3-18) produces the immediate solution for  $\mathbf{b}_1 = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{y}$  and likewise for  $\mathbf{b}_2$ .

If the two sets of variables  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are not orthogonal, then the solution for  $\mathbf{b}_1$  and  $\mathbf{b}_2$  found by (3-17) and (3-18) is more involved than just the simple regressions in Theorem 3.1. The more general solution is given by the following theorem, which appeared in the first volume of *Econometrica*:<sup>3</sup>

### **THEOREM 3.2 Frisch–Waugh (1933)–Lovell (1963) Theorem**

*In the linear least squares regression of vector  $\mathbf{y}$  on two sets of variables,  $\mathbf{X}_1$  and  $\mathbf{X}_2$ , the subvector  $\mathbf{b}_2$  is the set of coefficients obtained when the residuals from a regression of  $\mathbf{y}$  on  $\mathbf{X}_1$  alone are regressed on the set of residuals obtained when each column of  $\mathbf{X}_2$  is regressed on  $\mathbf{X}_1$ .*

<sup>3</sup>The theorem, such as it was, appeared in the introduction to the paper, “The partial trend regression method can never, indeed, achieve anything which the individual trend method cannot, because the two methods lead by definition to identically the same results.” Thus, Frisch and Waugh were concerned with the (lack of) difference between a regression of a variable  $\mathbf{y}$  on a time trend variable,  $t$ , and another variable,  $\mathbf{x}$ , compared to the regression of a detrended  $\mathbf{y}$  on a detrended  $\mathbf{x}$ , where detrending meant computing the residuals of the respective variable on a constant and the time trend,  $t$ . A concise statement of the theorem, and its matrix formulation were added later by Lovell (1963).

### 34 PART I ♦ The Linear Regression Model

To prove Theorem 3.2, begin from equation (2) in (3-17), which is

$$\mathbf{X}_2' \mathbf{X}_1 \mathbf{b}_1 + \mathbf{X}_2' \mathbf{X}_2 \mathbf{b}_2 = \mathbf{X}_2' \mathbf{y}.$$

Now, insert the result for  $\mathbf{b}_1$  that appears in (3-18) into this result. This produces

$$\mathbf{X}_2' \mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{y} - \mathbf{X}_2' \mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{X}_2 \mathbf{b}_2 + \mathbf{X}_2' \mathbf{X}_2 \mathbf{b}_2 = \mathbf{X}_2' \mathbf{y}.$$

After collecting terms, the solution is

$$\begin{aligned} \mathbf{b}_2 &= [\mathbf{X}_2' (\mathbf{I} - \mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1') \mathbf{X}_2]^{-1} [\mathbf{X}_2' (\mathbf{I} - \mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1') \mathbf{y}] \\ &= (\mathbf{X}_2' \mathbf{M}_1 \mathbf{X}_2)^{-1} (\mathbf{X}_2' \mathbf{M}_1 \mathbf{y}). \end{aligned} \quad (3-19)$$

The matrix appearing in the parentheses inside each set of square brackets is the “residual maker” defined in (3-14), in this case defined for a regression on the columns of  $\mathbf{X}_1$ . Thus,  $\mathbf{M}_1 \mathbf{X}_2$  is a matrix of residuals; each column of  $\mathbf{M}_1 \mathbf{X}_2$  is a vector of residuals in the regression of the corresponding column of  $\mathbf{X}_2$  on the variables in  $\mathbf{X}_1$ . By exploiting the fact that  $\mathbf{M}_1$ , like  $\mathbf{M}$ , is symmetric and idempotent, we can rewrite (3-19) as

$$\mathbf{b}_2 = (\mathbf{X}_2^* \mathbf{X}_2^*)^{-1} \mathbf{X}_2^* \mathbf{y}^*, \quad (3-20)$$

where

$$\mathbf{X}_2^* = \mathbf{M}_1 \mathbf{X}_2 \quad \text{and} \quad \mathbf{y}^* = \mathbf{M}_1 \mathbf{y}.$$

This result is fundamental in regression analysis.

This process is commonly called **partialing out** or **netting out** the effect of  $\mathbf{X}_1$ . For this reason, the coefficients in a multiple regression are often called the **partial regression coefficients**. The application of this theorem to the computation of a single coefficient as suggested at the beginning of this section is detailed in the following:

Consider the regression of  $\mathbf{y}$  on a set of variables  $\mathbf{X}$  and an additional variable  $\mathbf{z}$ . Denote the coefficients  $\mathbf{b}$  and  $c$ .

Sequence OK here?

#### COROLLARY 3.3.1 Individual Regression Coefficients

The coefficient on  $\mathbf{z}$  in a multiple regression of  $\mathbf{y}$  on  $\mathbf{W} = [\mathbf{X}, \mathbf{z}]$  is computed as  $c = (\mathbf{z}' \mathbf{M} \mathbf{z})^{-1} (\mathbf{z}' \mathbf{M} \mathbf{y}) = (\mathbf{z}^* \mathbf{z}^*)^{-1} \mathbf{z}^* \mathbf{y}^*$  where  $\mathbf{z}^*$  and  $\mathbf{y}^*$  are the residual vectors from least squares regressions of  $\mathbf{z}$  and  $\mathbf{y}$  on  $\mathbf{X}$ ;  $\mathbf{z}^* = \mathbf{M} \mathbf{z}$  and  $\mathbf{y}^* = \mathbf{M} \mathbf{y}$  where  $\mathbf{M}$  is defined in (3-14).

**Proof:** This is an application of Theorem 3.2 in which  $\mathbf{X}_1$  is  $\mathbf{X}$  and  $\mathbf{X}_2$  is  $\mathbf{z}$ .

In terms of Example 2.2, we could obtain the coefficient on education in the multiple regression by first regressing earnings and education on age (or age and age squared) and then using the residuals from these regressions in a simple regression. In a classic application of this latter observation, Frisch and Waugh (1933) (who are credited with the result) noted that in a time-series setting, the same results were obtained whether a regression was fitted with a time-trend variable or the data were first “detrended” by netting out the effect of time, as noted earlier, and using just the detrended data in a simple regression.<sup>4</sup>

<sup>4</sup>Recall our earlier investment example.

## CHAPTER 3 ♦ Least Squares 35

As an application of these results, consider the case in which  $\mathbf{X}_1$  is  $\mathbf{i}$ , a constant term that is a column of 1s in the first column of  $\mathbf{X}$ . The solution for  $\mathbf{b}_2$  in this case will then be the slopes in a regression that contains a constant term. Using Theorem 3.2 the vector of residuals for any variable in  $\mathbf{X}_2$  in this case will be

$$\begin{aligned}
 \mathbf{x}_* &= \mathbf{x} - \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{x} \\
 &= \mathbf{x} - \mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}'\mathbf{x} \\
 &= \mathbf{x} - \mathbf{i}(1/n)\mathbf{i}'\mathbf{x} \\
 &= \mathbf{x} - \mathbf{i}\bar{\mathbf{x}} \\
 &= \mathbf{M}^0\mathbf{x}.
 \end{aligned} \tag{3-21}$$

(See Section A.5.4 where we have developed this result purely algebraically.) For this case, then, the residuals are deviations from the sample mean. Therefore, each column of  $\mathbf{M}_1\mathbf{X}_2$  is the original variable, now in the form of deviations from the mean. This general result is summarized in the following corollary.


**COROLLARY 3.3.2** Regression with a Constant Term

*The slopes in a multiple regression that contains a constant term are obtained by transforming the data to deviations from their means and then regressing the variable  $y$  in deviation form on the explanatory variables, also in deviation form.*

[We used this result in (3-8).] Having obtained the coefficients on  $\mathbf{X}_2$ , how can we recover the coefficients on  $\mathbf{X}_1$  (the constant term)? One way is to repeat the exercise while reversing the roles of  $\mathbf{X}_1$  and  $\mathbf{X}_2$ . But there is an easier way. We have already solved for  $\mathbf{b}_2$ . Therefore, we can use (3-18) in a solution for  $\mathbf{b}_1$ . If  $\mathbf{X}_1$  is just a column of 1s, then the first of these produces the familiar result

$$b_1 = \bar{y} - \bar{x}_2 b_2 - \cdots - \bar{x}_K b_K$$

[which is used in (3-7)].

Theorem 3.2 and Corollaries 3.2.1 and 3.2.2 produce a useful interpretation of the partitioned regression when the model contains a constant term. According to Theorem 3.1, if the columns of  $\mathbf{X}$  are orthogonal, that is,  $\mathbf{x}'_k\mathbf{x}_m = 0$  for columns  $k$  and  $m$ , then the separate regression coefficients in the regression of  $\mathbf{y}$  on  $\mathbf{X}$  when  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K]$  are simply  $\mathbf{x}'_k\mathbf{y}/\mathbf{x}'_k\mathbf{x}_k$ . When the regression contains a constant term, we can compute the multiple regression coefficients by regression of  $\mathbf{y}$  in mean deviation form on the columns of  $\mathbf{X}$ , also in deviations from their means. In this instance, the “orthogonality” of the columns means that the sample covariances (and correlations) of the variables are zero. The result is another theorem:

### 36 PART I ♦ The Linear Regression Model

#### **THEOREM 3.3 Orthogonal Regression**

If the multiple regression of  $\mathbf{y}$  on  $\mathbf{X}$  contains a constant term and the variables in the regression are uncorrelated, then the multiple regression slopes are the same as the slopes in the individual simple regressions of  $\mathbf{y}$  on a constant and each variable in turn.

**Proof:** The result follows from Theorems 3.1 and 3.2.

### 3.4 PARTIAL REGRESSION AND PARTIAL CORRELATION COEFFICIENTS

The use of multiple regression involves a conceptual experiment that we might not be able to carry out in practice, the *ceteris paribus* analysis familiar in economics. To pursue Example 2.2, a regression equation relating earnings to age and education enables us to do the conceptual experiment of comparing the earnings of two individuals of the same age with different education levels, *even if the sample contains no such pair of individuals*. It is this characteristic of the regression that is implied by the term partial regression coefficients. The way we obtain this result, as we have seen, is first to regress income and education on age and then to compute the residuals from this regression. By construction, age will not have any power in explaining variation in these residuals. Therefore, any correlation between income and education after this “purging” is independent of (or after removing the effect of) age.

The same principle can be applied to the correlation between two variables. To continue our example, to what extent can we assert that this correlation reflects a direct relationship rather than that both income and education tend, on average, to rise as individuals become older? To find out, we would use a **partial correlation coefficient**, which is computed along the same lines as the partial regression coefficient. In the context of our example, the partial correlation coefficient between income and education, controlling for the effect of age, is obtained as follows:

1.  $y_*$  = the residuals in a regression of income on a constant and age.
2.  $z_*$  = the residuals in a regression of education on a constant and age.
3. The partial correlation  $r_{yz}^*$  is the simple correlation between  $y_*$  and  $z_*$ .

This calculation might seem to require a formidable amount of computation. Using Corollary 3.2.1, the two residual vectors in points 1 and 2 are  $\mathbf{y}_* = \mathbf{My}$  and  $\mathbf{z}_* = \mathbf{Mz}$  where  $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  is the residual maker defined in (3-14). We will assume that there is a constant term in  $\mathbf{X}$  so that the vectors of residuals  $\mathbf{y}_*$  and  $\mathbf{z}_*$  have zero sample means. Then, the square of the partial correlation coefficient is

$$r_{yz}^{*2} = \frac{(\mathbf{z}'_* \mathbf{y}_*)^2}{(\mathbf{z}'_* \mathbf{z}_*)(\mathbf{y}'_* \mathbf{y}_*)}.$$

There is a convenient shortcut. Once the multiple regression is computed, the  $t$  ratio in (5-13) for testing the hypothesis that the coefficient equals zero (e.g., the last column of

Table 4.1) can be used to compute

$$r_{yz}^{*2} = \frac{t_z^2}{t_z^2 + \text{degrees of freedom}}, \quad (3-22)$$

where the **degrees of freedom** is equal to  $n - (K + 1)$ . The proof of this less than perfectly intuitive result will be useful to illustrate some results on partitioned regression. We will rely on two useful theorems from least squares algebra. The first isolates a particular diagonal element of the inverse of a moment matrix such as  $(\mathbf{X}'\mathbf{X})^{-1}$ .

### THEOREM 3.4 Diagonal Elements of the Inverse of a Moment Matrix

Let  $\mathbf{W}$  denote the partitioned matrix  $[\mathbf{X}, \mathbf{z}]$ —that is, the  $K$  columns of  $\mathbf{X}$  plus an additional column labeled  $\mathbf{z}$ . The last diagonal element of  $(\mathbf{W}'\mathbf{W})^{-1}$  is  $(\mathbf{z}'\mathbf{M}\mathbf{z})^{-1} = (\mathbf{z}'_*\mathbf{z}_*)^{-1}$  where  $\mathbf{z}_* = \mathbf{M}\mathbf{z}$  and  $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ .

**Proof:** This is an application of the partitioned inverse formula in (A-74) where  $\mathbf{A}_{11} = \mathbf{X}'\mathbf{X}$ ,  $\mathbf{A}_{12} = \mathbf{X}'\mathbf{z}$ ,  $\mathbf{A}_{21} = \mathbf{z}'\mathbf{X}$ , and  $\mathbf{A}_{22} = \mathbf{z}'\mathbf{z}$ . Note that this theorem generalizes the development in Section A.2.8, where  $\mathbf{X}$  contains only a constant term,  $\mathbf{i}$ .

We can use Theorem 3.4 to establish the result in (3-22). Let  $c$  and  $\mathbf{u}$  denote the coefficient on  $\mathbf{z}$  and the vector of residuals in the multiple regression of  $\mathbf{y}$  on  $\mathbf{W} = [\mathbf{X}, \mathbf{z}]$ , respectively. Then, by definition, the squared  $t$  ratio in (3-22) is

$$t_z^2 = \frac{c^2}{\left[ \frac{\mathbf{u}'\mathbf{u}}{n - (K + 1)} \right] (\mathbf{W}'\mathbf{W})_{K+1,K+1}^{-1}}$$

where  $(\mathbf{W}'\mathbf{W})_{K+1,K+1}^{-1}$  is the  $(K + 1)$  (last) diagonal element of  $(\mathbf{W}'\mathbf{W})^{-1}$ . (The bracketed term appears in (4-17). We are using only the algebraic result at this point.) The theorem states that this element of the matrix equals  $(\mathbf{z}'_*\mathbf{z}_*)^{-1}$ . From Corollary 3.2.1, we also have that  $c^2 = [(\mathbf{z}'_*\mathbf{y}_*)/(\mathbf{z}'_*\mathbf{z}_*)]^2$ . For convenience, let  $DF = n - (K + 1)$ . Then,

$$t_z^2 = \frac{(\mathbf{z}'_*\mathbf{y}_*)^2 / (\mathbf{z}'_*\mathbf{z}_*)^2}{(\mathbf{u}'\mathbf{u}/DF)/(\mathbf{z}'_*\mathbf{z}_*)} = \frac{(\mathbf{z}'_*\mathbf{y}_*)^2 DF}{(\mathbf{u}'\mathbf{u})(\mathbf{z}'_*\mathbf{z}_*)}.$$

It follows that the result in (3-22) is equivalent to

$$\frac{t_z^2}{t_z^2 + DF} = \frac{\frac{(\mathbf{z}'_*\mathbf{y}_*)^2 DF}{(\mathbf{u}'\mathbf{u})(\mathbf{z}'_*\mathbf{z}_*)}}{\frac{(\mathbf{z}'_*\mathbf{y}_*)^2 DF}{(\mathbf{u}'\mathbf{u})(\mathbf{z}'_*\mathbf{z}_*)} + DF} = \frac{\frac{(\mathbf{z}'_*\mathbf{y}_*)^2}{(\mathbf{u}'\mathbf{u})(\mathbf{z}'_*\mathbf{z}_*)}}{\frac{(\mathbf{z}'_*\mathbf{y}_*)^2}{(\mathbf{u}'\mathbf{u})(\mathbf{z}'_*\mathbf{z}_*)} + 1} = \frac{(\mathbf{z}'_*\mathbf{y}_*)^2}{(\mathbf{z}'_*\mathbf{y}_*)^2 + (\mathbf{u}'\mathbf{u})(\mathbf{z}'_*\mathbf{z}_*)}.$$

Divide numerator and denominator by  $(\mathbf{z}'_*\mathbf{z}_*)$   $(\mathbf{y}'_*\mathbf{y}_*)$  to obtain

$$\frac{t_z^2}{t_z^2 + DF} = \frac{(\mathbf{z}'_*\mathbf{y}_*)^2 / (\mathbf{z}'_*\mathbf{z}_*)(\mathbf{y}'_*\mathbf{y}_*)}{(\mathbf{z}'_*\mathbf{y}_*)^2 / (\mathbf{z}'_*\mathbf{z}_*)(\mathbf{y}'_*\mathbf{y}_*) + (\mathbf{u}'\mathbf{u})(\mathbf{z}'_*\mathbf{z}_*)(\mathbf{y}'_*\mathbf{y}_*)} = \frac{r_{yz}^{*2}}{r_{yz}^{*2} + (\mathbf{u}'\mathbf{u})/(\mathbf{y}'_*\mathbf{y}_*)}. \quad (3-23)$$

### 38 PART I ♦ The Linear Regression Model

We will now use a second theorem to manipulate  $\mathbf{u}'\mathbf{u}$  and complete the derivation. The result we need is given in Theorem 3.5.

#### THEOREM 3.5 Change in the Sum of Squares When a Variable is Added to a Regression

If  $\mathbf{e}'\mathbf{e}$  is the sum of squared residuals when  $\mathbf{y}$  is regressed on  $\mathbf{X}$  and  $\mathbf{u}'\mathbf{u}$  is the sum of squared residuals when  $\mathbf{y}$  is regressed on  $\mathbf{X}$  and  $\mathbf{z}$ , then

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} - c^2(\mathbf{z}'_*\mathbf{z}_*) \leq \mathbf{e}'\mathbf{e}, \quad (3-24)$$

where  $c$  is the coefficient on  $\mathbf{z}$  in the long regression of  $\mathbf{y}$  on  $[\mathbf{X}, \mathbf{z}]$  and  $\mathbf{z}_* = \mathbf{M}\mathbf{z}$  is the vector of residuals when  $\mathbf{z}$  is regressed on  $\mathbf{X}$ .

**Proof:** In the long regression of  $\mathbf{y}$  on  $\mathbf{X}$  and  $\mathbf{z}$ , the vector of residuals is  $\mathbf{u} = \mathbf{y} - \mathbf{X}\mathbf{d} - \mathbf{zc}$ . Note that unless  $\mathbf{X}'\mathbf{z} = \mathbf{0}$ ,  $\mathbf{d}$  will not equal  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ . (See Sec. 4.3.2.) Moreover, unless  $c = 0$ ,  $\mathbf{u}$  will not equal  $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$ . From Corollary 3.3.1,  $c = (\mathbf{z}'_*\mathbf{z}_*)^{-1}(\mathbf{z}'_*\mathbf{y}_*)$ . From (3-18), we also have that the coefficients on  $\mathbf{X}$  in this long regression are

$$\mathbf{d} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{zc}) = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{zc}.$$

Inserting this expression for  $\mathbf{d}$  in that for  $\mathbf{u}$  gives

$$\mathbf{u} = \mathbf{y} - \mathbf{X}\mathbf{b} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{zc} - \mathbf{zc} = \mathbf{e} - \mathbf{M}\mathbf{zc} = \mathbf{e} - \mathbf{z}_*c.$$

Then,

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} + c^2(\mathbf{z}'_*\mathbf{z}_*) - 2c(\mathbf{z}'_*\mathbf{e})$$

But,  $\mathbf{e} = \mathbf{My} = \mathbf{y}_*$  and  $\mathbf{z}'_*\mathbf{e} = \mathbf{z}'_*\mathbf{y}_* = c(\mathbf{z}'_*\mathbf{z}_*)$ . Inserting this result in  $\mathbf{u}'\mathbf{u}$  immediately above gives the result in the theorem.

Returning to the derivation, then,  $\mathbf{e}'\mathbf{e} = \mathbf{y}'_*\mathbf{y}_*$  and  $c^2(\mathbf{z}'_*\mathbf{z}_*) = (\mathbf{z}'_*\mathbf{y}_*)^2/(\mathbf{z}'_*\mathbf{z}_*)$ . Therefore,

$$\frac{\mathbf{u}'\mathbf{u}}{\mathbf{y}'_*\mathbf{y}_*} = \frac{\mathbf{y}'_*\mathbf{y}_* - (\mathbf{z}'_*\mathbf{y}_*)^2/\mathbf{z}'_*\mathbf{z}_*}{\mathbf{y}'_*\mathbf{y}_*} = 1 - r_{yz}^{*2}$$

Inserting this in the denominator of (3.2.3) produces the result we sought.

#### Example 3.1 Partial Correlations

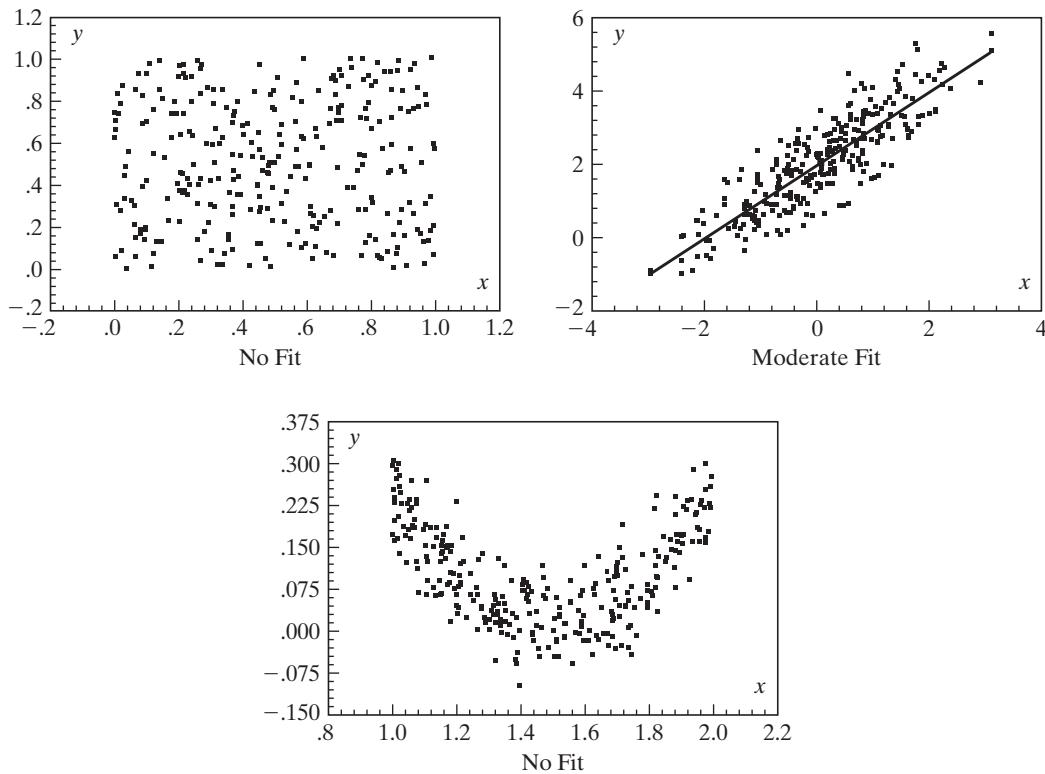
For the data in the app in Section 3.2.2, the simple correlations between investment and the regressors  $r_{ik}$  and the partial correlations  $r_{ik}^*$  between investment and the four regressors (given the other variables) are listed in Table 3.2. As is clear from the table, there is no necessary relation between the simple and partial correlation coefficients. One thing worth noting is the signs of the coefficients. The signs of the partial correlation coefficients are the same as the signs of the respective regression coefficients, three of which are negative. All the simple correlation coefficients are positive because of the latent “effect” of time.

**TABLE 3.2** Correlations of Investment with Other Variables

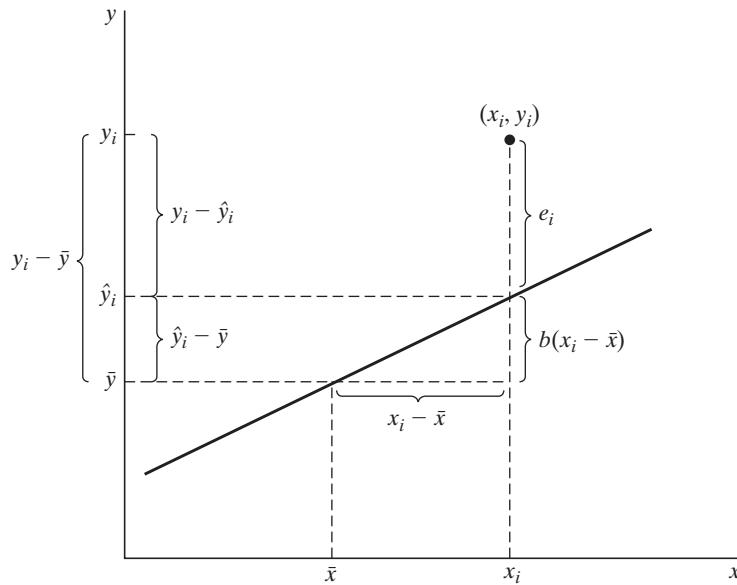
	<i>Simple Correlation</i>	<i>Partial Correlation</i>
Time	0.7496	-0.9360
GNP	0.8632	0.9680
Interest	0.5871	-0.5167
Inflation	0.4777	-0.0221

### 3.5 GOODNESS OF FIT AND THE ANALYSIS OF VARIANCE

The original fitting criterion, the sum of squared residuals, suggests a measure of the fit of the regression line to the data. However, as can easily be verified, the sum of squared residuals can be scaled arbitrarily just by multiplying all the values of  $y$  by the desired scale factor. Since the fitted values of the regression are based on the values of  $x$ , we might ask instead whether *variation* in  $x$  is a good predictor of *variation* in  $y$ . Figure 3.3 shows three possible cases for a simple linear regression model. The measure of fit described here embodies both the fitting criterion and the covariation of  $y$  and  $x$ .

**FIGURE 3.3** Sample Data.

## 40 PART I ♦ The Linear Regression Model



**FIGURE 3.4** Decomposition of  $y_i$ .

Variation of the dependent variable is defined in terms of deviations from its mean,  $(y_i - \bar{y})$ . The **total variation** in  $y$  is the sum of squared deviations:

$$\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2.$$

In terms of the regression equation, we may write the full set of observations as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} = \hat{\mathbf{y}} + \mathbf{e}.$$

For an individual observation, we have

$$y_i = \hat{y}_i + e_i = \mathbf{x}'_i \mathbf{b} + e_i.$$

If the regression contains a constant term, then the residuals will sum to zero and the mean of the predicted values of  $y_i$  will equal the mean of the actual values. Subtracting  $\bar{y}$  from both sides and using this result and result 2 in Section 3.2.3 gives

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + e_i = (\mathbf{x}'_i - \bar{\mathbf{x}}') \mathbf{b} + e_i.$$

Figure 3.4 illustrates the computation for the two-variable regression. Intuitively, the regression would appear to fit well if the deviations of  $y$  from its mean are more largely accounted for by deviations of  $x$  from its mean than by the residuals. Since both terms in this decomposition sum to zero, to quantify this fit, we use the sums of squares instead. For the full set of observations, we have

$$\mathbf{M}^0 \mathbf{y} = \mathbf{M}^0 \mathbf{X} \mathbf{b} + \mathbf{M}^0 \mathbf{e},$$

where  $\mathbf{M}^0$  is the  $n \times n$  idempotent matrix that transforms observations into deviations from sample means. (See (3-21) and Section A.2.8.) The column of  $\mathbf{M}^0 \mathbf{X}$  corresponding to the constant term is zero, and, since the residuals already have mean zero,  $\mathbf{M}^0 \mathbf{e} = \mathbf{e}$ .

## CHAPTER 3 ♦ Least Squares 41

Then, since  $\mathbf{e}'\mathbf{M}^0\mathbf{X} = \mathbf{e}'\mathbf{X} = \mathbf{0}$ , the total sum of squares is

$$\mathbf{y}'\mathbf{M}^0\mathbf{y} = \mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\mathbf{b} + \mathbf{e}'\mathbf{e}.$$

Write this as total sum of squares = regression sum of squares + error sum of squares, or

$$\text{SST} = \text{SSR} + \text{SSE}. \quad (3-25)$$

(Note that this is the same partitioning that appears at the end of Section 3.2.4.)

We can now obtain a measure of how well the regression line fits the data by using the

$$\text{coefficient of determination: } \frac{\text{SSR}}{\text{SST}} = \frac{\mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\mathbf{b}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}}. \quad (3-26)$$

The coefficient of determination is denoted  $R^2$ . As we have shown, it must be between 0 and 1, and it measures the proportion of the total variation in  $y$  that is accounted for by variation in the regressors. It equals zero if the regression is a horizontal line, that is, if all the elements of  $\mathbf{b}$  except the constant term are zero. In this case, the predicted values of  $y$  are always  $\bar{y}$ , so deviations of  $\mathbf{x}$  from its mean do not translate into different predictions for  $y$ . As such,  $\mathbf{x}$  has no explanatory power. The other extreme,  $R^2 = 1$ , occurs if the values of  $\mathbf{x}$  and  $y$  all lie in the same hyperplane (on a straight line for a two variable regression) so that the residuals are all zero. If all the values of  $y_i$  lie on a vertical line, then  $R^2$  has no meaning and cannot be computed.

Regression analysis is often used for forecasting. In this case, we are interested in how well the regression model predicts movements in the dependent variable. With this in mind, an equivalent way to compute  $R^2$  is also useful. First

$$\mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\mathbf{b} = \hat{\mathbf{y}}'\mathbf{M}^0\hat{\mathbf{y}},$$

but  $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$ ,  $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$ ,  $\mathbf{M}^0\mathbf{e} = \mathbf{e}$ , and  $\mathbf{X}'\mathbf{e} = \mathbf{0}$ , so  $\hat{\mathbf{y}}'\mathbf{M}^0\hat{\mathbf{y}} = \hat{\mathbf{y}}'\mathbf{M}^0\mathbf{y}$ . Multiply  $R^2 = \hat{\mathbf{y}}'\mathbf{M}^0\hat{\mathbf{y}}/\mathbf{y}'\mathbf{M}^0\mathbf{y} = \hat{\mathbf{y}}'\mathbf{M}^0\mathbf{y}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$  by  $1 = \hat{\mathbf{y}}'\mathbf{M}^0\mathbf{y}/\hat{\mathbf{y}}'\mathbf{M}^0\hat{\mathbf{y}}$  to obtain

$$R^2 = \frac{[\sum_i(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})]^2}{[\sum_i(y_i - \bar{y})^2][\sum_i(\hat{y}_i - \bar{\hat{y}})^2]}, \quad (3-27)$$

which is the squared correlation between the observed values of  $y$  and the predictions produced by the estimated regression equation.

### Example 3.2 Fit of a Consumption Function

The data plotted in Figure 2.1 are listed in Appendix Table F2.1. For these data, where  $y$  is  $C$  and  $x$  is  $X$ , we have  $\bar{y} = 273.2727$ ,  $\bar{x} = 323.2727$ ,  $S_{yy} = 12,618.182$ ,  $S_{xx} = 12,300.182$ ,  $S_{xy} = 8,423.182$  so  $\text{SST} = 12,618.182$ ,  $b = 8,423.182/12,300.182 = 0.6848014$ ,  $\text{SSR} = b^2 S_{xx} = 5,768.2068$ , and  $\text{SSE} = \text{SST} - \text{SSR} = 6,849.975$ . Then  $R^2 = b^2 S_{xx}/\text{SST} = 0.457135$ . As can be seen in Figure 2.1, this is a moderate fit, although it is not particularly good for aggregate time-series data. On the other hand, it is clear that not accounting for the anomalous wartime data has degraded the fit of the model. This value is the  $R^2$  for the model indicated by the dotted line in the figure. By simply omitting the years 1942–1945 from the sample and doing these computations with the remaining seven observations—the heavy solid line—we obtain an  $R^2$  of 0.93697. Alternatively, by creating a variable  $WAR$  which equals 1 in the years 1942–1945 and zero otherwise and including this in the model, which produces the model shown by the two solid lines, the  $R^2$  rises to 0.94639.

We can summarize the calculation of  $R^2$  in an **analysis of variance** table, which might appear as shown in Table 3.3.

## 42 PART I ♦ The Linear Regression Model

**TABLE 3.3** Analysis of Variance

	<i>Source</i>	<i>Degrees of Freedom</i>	<i>Mean Square</i>
Regression	$\mathbf{b}'\mathbf{X}'\mathbf{y} - n\bar{y}^2$	$K - 1$ (assuming a constant term)	
Residual	$\mathbf{e}'\mathbf{e}$	$n - K$	$s^2$
Total	$\mathbf{y}'\mathbf{y} - n\bar{y}^2$	$n - 1$	$S_{yy}/(n - 1) = s_y^2$
Coefficient of determination		$R^2 = 1 - \mathbf{e}'\mathbf{e}/(\mathbf{y}'\mathbf{y} - n\bar{y}^2)$	

**TABLE 3.4** Analysis of Variance for the Investment Equation

	<i>Source</i>	<i>Degrees of Freedom</i>	<i>Mean Square</i>
Regression	0.0159025	4	0.003976
Residual	0.0004508	10	0.00004508
Total	0.016353	14	0.0011681
$R^2 = 0.0159025/0.016353 = 0.97245$			

### Example 3.3 Analysis of Variance for an Investment Equation

The analysis of variance table for the investment equation of Section 3.2.2 is given in Table 3.4.

#### 3.5.1 THE ADJUSTED R-SQUARED AND A MEASURE OF FIT

There are some problems with the use of  $R^2$  in analyzing goodness of fit. The first concerns the number of degrees of freedom used up in estimating the parameters. [See (3-22) and Table 3.3.]  $R^2$  will never decrease when another variable is added to a regression equation. Equation (3-23) provides a convenient means for us to establish this result. Once again, we are comparing a regression of  $\mathbf{y}$  on  $\mathbf{X}$  with sum of squared residuals  $\mathbf{e}'\mathbf{e}$  to a regression of  $\mathbf{y}$  on  $\mathbf{X}$  and an additional variable  $\mathbf{z}$ , which produces sum of squared residuals  $\mathbf{u}'\mathbf{u}$ . Recall the vectors of residuals  $\mathbf{z}_* = \mathbf{M}\mathbf{z}$  and  $\mathbf{y}_* = \mathbf{My} = \mathbf{e}$ , which implies that  $\mathbf{e}'\mathbf{e} = (\mathbf{y}'\mathbf{y}_*)$ . Let  $c$  be the coefficient on  $\mathbf{z}$  in the longer regression. Then  $c = (\mathbf{z}'_*\mathbf{z}_*)^{-1}(\mathbf{z}'_*\mathbf{y}_*)$ , and inserting this in (3-24) produces

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} - \frac{(\mathbf{z}'_*\mathbf{y}_*)^2}{(\mathbf{z}'_*\mathbf{z}_*)} = \mathbf{e}'\mathbf{e}(1 - r_{yz}^{*2}), \quad (3-28)$$

where  $r_{yz}^*$  is the partial correlation between  $\mathbf{y}$  and  $\mathbf{z}$ , controlling for  $\mathbf{X}$ . Now divide through both sides of the equality by  $\mathbf{y}'\mathbf{M}^0\mathbf{y}$ . From (3-26),  $\mathbf{u}'\mathbf{u}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$  is  $(1 - R_{Xz}^2)$  for the regression on  $\mathbf{X}$  and  $\mathbf{z}$  and  $\mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$  is  $(1 - R_X^2)$ . Rearranging the result produces the following:

#### THEOREM 3.6 Change in $R^2$ When a Variable Is Added to a Regression

Let  $R_{Xz}^2$  be the coefficient of determination in the regression of  $\mathbf{y}$  on  $\mathbf{X}$  and an additional variable  $\mathbf{z}$ , let  $R_X^2$  be the same for the regression of  $\mathbf{y}$  on  $\mathbf{X}$  alone, and let  $r_{yz}^*$  be the partial correlation between  $\mathbf{y}$  and  $\mathbf{z}$ , controlling for  $\mathbf{X}$ . Then

$$R_{Xz}^2 = R_X^2 + (1 - R_X^2)r_{yz}^{*2}. \quad (3-29)$$

Thus, the  $R^2$  in the longer regression cannot be smaller. It is tempting to exploit this result by just adding variables to the model;  $R^2$  will continue to rise to its limit of 1.<sup>5</sup> The **adjusted  $R^2$**  (for degrees of freedom), which incorporates a penalty for these results is computed as follows<sup>6</sup>:

$$\bar{R}^2 = 1 - \frac{\mathbf{e}'\mathbf{e}/(n - K)}{\mathbf{y}'\mathbf{M}^0\mathbf{y}/(n - 1)}. \quad (3-30)$$

For computational purposes, the connection between  $R^2$  and  $\bar{R}^2$  is

$$\bar{R}^2 = 1 - \frac{n - 1}{n - K}(1 - R^2).$$

The adjusted  $R^2$  may decline when a variable is added to the set of independent variables. Indeed,  $\bar{R}^2$  may even be negative. To consider an admittedly extreme case, suppose that  $\mathbf{x}$  and  $\mathbf{y}$  have a sample correlation of zero. Then the adjusted  $R^2$  will equal  $-1/(n - 2)$ . [Thus, the name “adjusted  $R$ -squared” is a bit misleading—as can be seen in (3-30),  $\bar{R}^2$  is not actually computed as the square of any quantity.] Whether  $\bar{R}^2$  rises or falls depends on whether the contribution of the new variable to the fit of the regression more than offsets the correction for the loss of an additional degree of freedom. The general result (the proof of which is left as an exercise) is as follows.

**THEOREM 3.7 Change in  $\bar{R}^2$  When a Variable Is Added to a Regression**

*In a multiple regression,  $\bar{R}^2$  will fall (rise) when the variable  $x$  is deleted from the regression if the square of the t ratio associated with this variable is greater (less) than 1.*

We have shown that  $R^2$  will never fall when a variable is added to the regression. We now consider this result more generally. The change in the residual sum of squares when a set of variables  $\mathbf{X}_2$  is added to the regression is

$$\mathbf{e}'_{1,2}\mathbf{e}_{1,2} = \mathbf{e}'_1\mathbf{e}_1 - \mathbf{b}'_2\mathbf{X}'_2\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2,$$

where we use subscript 1 to indicate the regression based on  $\mathbf{X}_1$  alone and 1,2 to indicate the use of both  $\mathbf{X}_1$  and  $\mathbf{X}_2$ . The coefficient vector  $\mathbf{b}_2$  is the coefficients on  $\mathbf{X}_2$  in the multiple regression of  $\mathbf{y}$  on  $\mathbf{X}_1$  and  $\mathbf{X}_2$ . [See (3-19) and (3-20) for definitions of  $\mathbf{b}_2$  and  $\mathbf{M}_1$ .] Therefore,

$$R^2_{1,2} = 1 - \frac{\mathbf{e}'_1\mathbf{e}_1 - \mathbf{b}'_2\mathbf{X}'_2\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} = R^2_1 + \frac{\mathbf{b}'_2\mathbf{X}'_2\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2}{\mathbf{y}'\mathbf{M}^0\mathbf{y}},$$

<sup>5</sup>This result comes at a cost, however. The parameter estimates become progressively less precise as we do so. We will pursue this result in Chapter 4.

<sup>6</sup>This measure is sometimes advocated on the basis of the unbiasedness of the two quantities in the fraction. Since the ratio is not an unbiased estimator of any population quantity, it is difficult to justify the adjustment on this basis.

#### 44 PART I ♦ The Linear Regression Model

which is greater than  $R_1^2$  unless  $\mathbf{b}_2$  equals zero. ( $\mathbf{M}_1 \mathbf{X}_2$  could not be zero unless  $\mathbf{X}_2$  was a linear function of  $\mathbf{X}_1$ , in which case the regression on  $\mathbf{X}_1$  and  $\mathbf{X}_2$  could not be computed.) This equation can be manipulated a bit further to obtain

$$R_{1,2}^2 = R_1^2 + \frac{\mathbf{y}' \mathbf{M}_1 \mathbf{y}}{\mathbf{y}' \mathbf{M}^0 \mathbf{y}} \frac{\mathbf{b}_2' \mathbf{X}_2' \mathbf{M}_1 \mathbf{X}_2 \mathbf{b}_2}{\mathbf{y}' \mathbf{M}_1 \mathbf{y}}.$$

But  $\mathbf{y}' \mathbf{M}_1 \mathbf{y} = \mathbf{e}'_1 \mathbf{e}_1$ , so the first term in the product is  $1 - R_1^2$ . The second is the **multiple correlation** in the regression of  $\mathbf{M}_1 \mathbf{y}$  on  $\mathbf{M}_1 \mathbf{X}_2$ , or the partial correlation (after the effect of  $\mathbf{X}_1$  is removed) in the regression of  $\mathbf{y}$  on  $\mathbf{X}_2$ . Collecting terms, we have

$$R_{1,2}^2 = R_1^2 + (1 - R_1^2) r_{y2-1}^2.$$

[This is the multivariate counterpart to (3-29).]

Therefore, it is possible to push  $R^2$  as high as desired just by adding regressors. This possibility motivates the use of the adjusted  $R^2$  in (3-30), instead of  $R^2$  as a method of choosing among alternative models. Since  $\bar{R}^2$  incorporates a penalty for reducing the degrees of freedom while still revealing an improvement in fit, one possibility is to choose the specification that maximizes  $\bar{R}^2$ . It has been suggested that the adjusted  $R^2$  does not penalize the loss of degrees of freedom heavily enough.<sup>7</sup> Some alternatives that have been proposed for comparing models (which we index by  $j$ ) are

$$\tilde{R}_j^2 = 1 - \frac{n + K_j}{n - K_j} (1 - R_j^2),$$

which minimizes Amemiya's (1985) **prediction criterion**,

$$PC_j = \frac{\mathbf{e}'_j \mathbf{e}_j}{n - K_j} \left( 1 + \frac{K_j}{n} \right) = s_j^2 \left( 1 + \frac{K_j}{n} \right)$$

and the Akaike and Bayesian information criteria which are given in (5-43) and (5-44).<sup>8</sup>

##### 3.5.2 R-SQUARED AND THE CONSTANT TERM IN THE MODEL

A second difficulty with  $R^2$  concerns the constant term in the model. The proof that  $0 \leq R^2 \leq 1$  requires  $\mathbf{X}$  to contain a column of 1s. If not, then (1)  $\mathbf{M}^0 \mathbf{e} \neq \mathbf{e}$  and (2)  $\mathbf{e}' \mathbf{M}^0 \mathbf{X} \neq \mathbf{0}$  and the term  $2\mathbf{e}' \mathbf{M}^0 \mathbf{X} \mathbf{b}$  in  $\mathbf{y}' \mathbf{M}^0 \mathbf{y} = (\mathbf{M}^0 \mathbf{X} \mathbf{b} + \mathbf{M}^0 \mathbf{e})' (\mathbf{M}^0 \mathbf{X} \mathbf{b} + \mathbf{M}^0 \mathbf{e})$  in the preceding expansion will not drop out. Consequently, when we compute

$$R^2 = 1 - \frac{\mathbf{e}' \mathbf{e}}{\mathbf{y}' \mathbf{M}^0 \mathbf{y}},$$

the result is unpredictable. It will never be higher and can be far lower than the same figure computed for the regression with a constant term included. It can even be negative.

<sup>7</sup>See, for example, Amemiya (1985, pp. 50–51).

<sup>8</sup>Most authors and computer programs report the logs of these prediction criteria.

Computer packages differ in their computation of  $R^2$ . An alternative computation,

$$R^2 = \frac{\mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{y}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}},$$

is equally problematic. Again, this calculation will differ from the one obtained with the constant term included; this time,  $R^2$  may be larger than 1. Some computer packages bypass these difficulties by reporting a third “ $R^2$ ,” the squared sample correlation between the actual values of  $y$  and the fitted values from the regression. This approach could be deceptive. If the regression contains a constant term, then, as we have seen, all three computations give the same answer. Even if not, this last one will still produce a value between zero and one. But, it is not a proportion of variation explained. On the other hand, for the purpose of comparing models, this squared correlation might well be a useful descriptive device. It is important for users of computer packages to be aware of how the reported  $R^2$  is computed. Indeed, some packages will give a warning in the results when a regression is fit without a constant or by some technique other than linear least squares.

### 3.5.3 COMPARING MODELS

The value of  $R^2$  we obtained for the consumption function in Example 3.2 seems high in an absolute sense. Is it? Unfortunately, there is no absolute basis for comparison. In fact, in using aggregate time-series data, coefficients of determination this high are routine. In terms of the values one normally encounters in cross sections, an  $R^2$  of 0.5 is relatively high. Coefficients of determination in cross sections of individual data as high as 0.2 are sometimes noteworthy. The point of this discussion is that whether a regression line provides a good fit to a body of data depends on the setting.

Little can be said about the relative quality of fits of regression lines in different contexts or in different data sets even if they are supposedly generated by the same data generating mechanism. One must be careful, however, even in a single context, to be sure to use the same basis for comparison for competing models. Usually, this concern is about how the dependent variable is computed. For example, a perennial question concerns whether a linear or loglinear model fits the data better. Unfortunately, the question cannot be answered with a direct comparison. An  $R^2$  for the linear regression model is different from an  $R^2$  for the loglinear model. Variation in  $y$  is different from variation in  $\ln y$ . The latter  $R^2$  will typically be larger, but this does not imply that the loglinear model is a better fit in some absolute sense.

It is worth emphasizing that  $R^2$  is a measure of *linear* association between  $x$  and  $y$ . For example, the third panel of Figure 3.3 shows data that might arise from the model

$$y_i = \alpha + \beta(x_i - \gamma)^2 + \varepsilon_i.$$

(The constant  $\gamma$  allows  $x$  to be distributed about some value other than zero.) The relationship between  $y$  and  $x$  in this model is nonlinear, and a linear regression would find no fit.

A final word of caution is in order. The interpretation of  $R^2$  as a proportion of variation explained is dependent on the use of least squares to compute the fitted

## 46 PART I ♦ The Linear Regression Model

values. It is always correct to write

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + e_i$$

regardless of how  $\hat{y}_i$  is computed. Thus, one might use  $\hat{y}_i = \exp(\widehat{\ln y_i})$  from a loglinear model in computing the sum of squares on the two sides, however, the cross-product term vanishes only if least squares is used to compute the fitted values and if the model contains a constant term. Thus, the cross-product term has been ignored in computing  $R^2$  for the loglinear model. Only in the case of least squares applied to a linear equation with a constant term can  $R^2$  be interpreted as the proportion of variation in  $y$  explained by variation in  $\mathbf{x}$ . An analogous computation can be done without computing deviations from means if the regression does not contain a constant term. Other purely algebraic artifacts will crop up in regressions without a constant, however. For example, the value of  $R^2$  will change when the same constant is added to each observation on  $y$ , but it is obvious that nothing fundamental has changed in the regression relationship. One should be wary (even skeptical) in the calculation and interpretation of fit measures for regressions without constant terms.

## 3.6 LINEARLY TRANSFORMED REGRESSION

As a final application of the tools developed in this chapter, we examine a purely algebraic result that is very useful for understanding the computation of linear regression models. In the regression of  $\mathbf{y}$  on  $\mathbf{X}$ , suppose the columns of  $\mathbf{X}$  are linearly transformed. Common applications would include changes in the units of measurement, say by changing units of currency, hours to minutes, or distances in miles to kilometers. Example 3.4 suggests a slightly more involved case

### **Example 3.4 Art Appreciation**

Theory 1 of the determination of the auction prices of Monet paintings holds that the price is determined by the dimensions (width,  $W$  and height,  $H$ ) of the painting,

$$\begin{aligned}\ln P &= \beta_1(1) + \beta_2 \ln W + \beta_3 \ln H + \varepsilon \\ &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.\end{aligned}$$

Theory 2 claims, instead, that art buyers are interested specifically in surface area and aspect ratio,

$$\begin{aligned}\ln P &= \gamma_1(1) + \gamma_2 \ln(WH) + \gamma_3 \ln(W/H) + \varepsilon \\ &= \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \varepsilon.\end{aligned}$$

It is evident that  $z_1 = x_1$ ,  $z_2 = x_2 + x_3$  and  $z_3 = x_2 - x_3$ . In matrix terms,  $\mathbf{Z} = \mathbf{XP}$  where

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & -1 \end{bmatrix}.$$

The effect of a transformation on the linear regression of  $\mathbf{y}$  on  $\mathbf{X}$  compared to that of  $\mathbf{y}$  on  $\mathbf{Z}$  is given by Theorem 3.8.

**THEOREM 3.8 Transformed Variables**

In the linear regression of  $\mathbf{y}$  on  $\mathbf{Z} = \mathbf{XP}$  where  $\mathbf{P}$  is a nonsingular matrix that transforms the columns of  $\mathbf{X}$ , the coefficients will equal  $\mathbf{P}^{-1}\mathbf{b}$  where  $\mathbf{b}$  is the vector of coefficients in the linear regression of  $\mathbf{y}$  on  $\mathbf{X}$ , and the  $R^2$  will be identical.

**Proof:** The coefficients are

$$\begin{aligned}\mathbf{d} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = [(\mathbf{XP})'(\mathbf{XP})]^{-1}(\mathbf{XP})'\mathbf{y} = (\mathbf{P}'\mathbf{X}'\mathbf{X}\mathbf{P})^{-1}\mathbf{P}'\mathbf{X}'\mathbf{y} \\ &= \mathbf{P}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{P}'\mathbf{y} = \mathbf{P}^{-1}\mathbf{b}.\end{aligned}$$

The vector of residuals is  $\mathbf{u} = \mathbf{y} - \mathbf{Z}(\mathbf{P}^{-1}\mathbf{b}) = \mathbf{y} - \mathbf{XPP}^{-1}\mathbf{b} = \mathbf{y} - \mathbf{Xb} = \mathbf{e}$ . Since the residuals are identical, the numerator of  $1 - R^2$  is the same, and the denominator is unchanged. This establishes the result.

This is a useful practical, algebraic result. For example, it simplifies the analysis in the first application suggested, changing the units of measurement. If an independent variable is scaled by a constant,  $p$ , the regression coefficient will be scaled by  $1/p$ . There is no need to recompute the regression.

### 3.7 SUMMARY AND CONCLUSIONS

This chapter has described the purely algebraic exercise of fitting a line (hyperplane) to a set of points using the method of least squares. We considered the primary problem first, using a data set of  $n$  observations on  $K$  variables. We then examined several aspects of the solution, including the nature of the projection and residual maker matrices and several useful algebraic results relating to the computation of the residuals and their sum of squares. We also examined the difference between gross or simple regression and correlation and multiple regression by defining “partial regression coefficients” and “partial correlation coefficients.” The Frisch–Waugh–Lovell theorem (3.2) is a fundamentally useful tool in regression analysis which enables us to obtain in closed form the expression for a subvector of a vector of regression coefficients. We examined several aspects of the partitioned regression, including how the fit of the regression model changes when variables are added to it or removed from it. Finally, we took a closer look at the conventional measure of how well the fitted regression line predicts or “fits” the data.

#### Key Terms and Concepts

- Adjusted  $R^2$
- Analysis of variance
- Bivariate regression
- Coefficient of determination
- Degrees of Freedom
- Disturbance
- Fitting criterion
- Frisch–Waugh theorem
- Goodness of fit
- Least squares
- Least squares normal equations
- Moment matrix
- Multiple correlation
- Multiple regression
- Netting out
- Normal equations
- Orthogonal regression
- Partial correlation coefficient
- Partial regression coefficient



These KT are appearing either in heads or in titles, so we have left it as is as per design.  
Please suggest .



## 48 PART I ♦ The Linear Regression Model

- Partialing out
- Population regression
- Residual maker
- Partitioned regression
- Projection
- Total variation
- Prediction criterion
- Projection matrix
- Population quantity
- Residual

### Exercises

1. **The two variable regression.** For the regression model  $y = \alpha + \beta x + \varepsilon$ ,
  - a. Show that the least squares normal equations imply  $\sum_i e_i = 0$  and  $\sum_i x_i e_i = 0$ .
  - b. Show that the solution for the constant term is  $a = \bar{y} - b\bar{x}$ .
  - c. Show that the solution for  $b$  is  $b = [\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})]/[\sum_{i=1}^n (x_i - \bar{x})^2]$ .
  - d. Prove that these two values uniquely minimize the sum of squares by showing that the diagonal elements of the second derivatives matrix of the sum of squares with respect to the parameters are both positive and that the determinant is  $4n[(\sum_{i=1}^n x_i^2) - n\bar{x}^2] = 4n[\sum_{i=1}^n (x_i - \bar{x})^2]$ , which is positive unless all values of  $x$  are the same.
2. **Change in the sum of squares.** Suppose that  $\mathbf{b}$  is the least squares coefficient vector in the regression of  $\mathbf{y}$  on  $\mathbf{X}$  and that  $\mathbf{c}$  is any other  $K \times 1$  vector. Prove that the difference in the two sums of squared residuals is

$$(\mathbf{y} - \mathbf{X}\mathbf{c})'(\mathbf{y} - \mathbf{X}\mathbf{c}) - (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) = (\mathbf{c} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{c} - \mathbf{b}).$$

Prove that this difference is positive.



3. **Linear transformations of the data.** Consider the least squares regression of  $\mathbf{y}$  on  $K$  variables (with a constant)  $\mathbf{X}$ . Consider an alternative set of regressors  $\mathbf{Z} = \mathbf{XP}$ , where  $\mathbf{P}$  is a nonsingular matrix. Thus, each column of  $\mathbf{Z}$  is a mixture of some of the columns of  $\mathbf{X}$ . Prove that the residual vectors in the regressions of  $\mathbf{y}$  on  $\mathbf{X}$  and  $\mathbf{y}$  on  $\mathbf{Z}$  are identical. What relevance does this have to the question of changing the fit of a regression by changing the units of measurement of the independent variables?
4. **Partial Frisch and Waugh.** In the least squares regression of  $\mathbf{y}$  on a constant and  $\mathbf{X}$ , to compute the regression coefficients on  $\mathbf{X}$ , we can first transform  $\mathbf{y}$  to deviations from the mean  $\bar{y}$  and, likewise, transform each column of  $\mathbf{X}$  to deviations from the respective column mean; second, regress the transformed  $\mathbf{y}$  on the transformed  $\mathbf{X}$  without a constant. Do we get the same result if we only transform  $\mathbf{y}$ ? What if we only transform  $\mathbf{X}$ ?
5. **Residual makers.** What is the result of the matrix product  $\mathbf{M}_1 \mathbf{M}$  where  $\mathbf{M}_1$  is defined in (3-19) and  $\mathbf{M}$  is defined in (3-14)?
6. **Adding an observation.** A data set consists of  $n$  observations on  $\mathbf{X}_n$  and  $\mathbf{y}_n$ . The least squares estimator based on these  $n$  observations is  $\mathbf{b}_n = (\mathbf{X}'_n \mathbf{X}_n)^{-1} \mathbf{X}'_n \mathbf{y}_n$ . Another observation,  $\mathbf{x}_s$  and  $y_s$ , becomes available. Prove that the least squares estimator computed using this additional observation is

$$\mathbf{b}_{n,s} = \mathbf{b}_n + \frac{1}{1 + \mathbf{x}'_s (\mathbf{X}'_n \mathbf{X}_n)^{-1} \mathbf{x}_s} (\mathbf{X}'_n \mathbf{X}_n)^{-1} \mathbf{x}_s (y_s - \mathbf{x}'_s \mathbf{b}_n).$$

Note that the last term is  $e_s$ , the residual from the prediction of  $y_s$  using the coefficients based on  $\mathbf{X}_n$  and  $\mathbf{b}_n$ . Conclude that the new data change the results of least squares only if the new observation on  $y$  cannot be perfectly predicted using the information already in hand.

7. **Deleting an observation.** A common strategy for handling a case in which an observation is missing data for one or more variables is to fill those missing variables with 0s and add a variable to the model that takes the value 1 for that one observation and 0 for all other observations. Show that this “strategy” is equivalent to discarding the observation as regards the computation of  $\mathbf{b}$  but it does have an effect on  $R^2$ . Consider the special case in which  $\mathbf{X}$  contains only a constant and one variable. Show that replacing missing values of  $x$  with the mean of the complete observations has the same effect as adding the new variable.
8. **Demand system estimation.** Let  $Y$  denote total expenditure on consumer durables, nondurables, and services and  $E_d$ ,  $E_n$ , and  $E_s$  are the expenditures on the three categories. As defined,  $Y = E_d + E_n + E_s$ . Now, consider the expenditure system

$$E_d = \alpha_d + \beta_d Y + \gamma_{dd} P_d + \gamma_{dn} P_n + \gamma_{ds} P_s + \varepsilon_d,$$

$$E_n = \alpha_n + \beta_n Y + \gamma_{nd} P_d + \gamma_{nn} P_n + \gamma_{ns} P_s + \varepsilon_n,$$

$$E_s = \alpha_s + \beta_s Y + \gamma_{sd} P_d + \gamma_{sn} P_n + \gamma_{ss} P_s + \varepsilon_s.$$

Prove that if all equations are estimated by ordinary least squares, then the sum of the expenditure coefficients will be 1 and the four other column sums in the preceding model will be zero.

9. **Change in adjusted  $R^2$ .** Prove that the adjusted  $R^2$  in (3-30) rises (falls) when variable  $\mathbf{x}_k$  is deleted from the regression if the square of the  $t$  ratio on  $\mathbf{x}_k$  in the multiple regression is less (greater) than 1.
10. **Regression without a constant.** Suppose that you estimate a multiple regression first with, then without, a constant. Whether the  $R^2$  is higher in the second case than the first will depend in part on how it is computed. Using the (relatively) standard method  $R^2 = 1 - (\mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{M}^0\mathbf{y})$ , which regression will have a higher  $R^2$ ?
11. Three variables,  $N$ ,  $D$ , and  $Y$ , all have zero means and unit variances. A fourth variable is  $C = N + D$ . In the regression of  $C$  on  $Y$ , the slope is 0.8. In the regression of  $C$  on  $N$ , the slope is 0.5. In the regression of  $D$  on  $Y$ , the slope is 0.4. What is the sum of squared residuals in the regression of  $C$  on  $D$ ? There are 21 observations and all moments are computed using  $1/(n - 1)$  as the divisor.
12. Using the matrices of sums of squares and cross products immediately preceding Section 3.2.3, compute the coefficients in the multiple regression of real investment on a constant, real GNP and the interest rate. Compute  $R^2$ .
13. In the December 1969, *American Economic Review* (pp. 886–896), Nathaniel Leff reports the following least squares regression results for a cross section study of the effect of age composition on savings in 74 countries in 1964:

$$\ln S/Y = 7.3439 + 0.1596 \ln Y/N + 0.0254 \ln G - 1.3520 \ln D_1 - 0.3990 \ln D_2$$


$$\ln S/N = 2.7851 + 1.1486 \ln Y/N + 0.0265 \ln G - 1.3438 \ln D_1 - 0.3966 \ln D_2$$

where  $S/Y$  = domestic savings ratio,  $S/N$  = per capita savings,  $Y/N$  = per capita income,  $D_1$  = percentage of the population under 15,  $D_2$  = percentage of the population over 64, and  $G$  = growth rate of per capita income. Are these results correct? Explain. [See Goldberger (1973) and Leff (1973) for discussion.]

## 50 PART I ♦ The Linear Regression Model

### **Application**

The data listed in Table 3.5 are extracted from Koop and Tobias's (2004) study of the relationship between wages and education, ability, and family characteristics. (See Appendix Table F3.2.) Their data set is a panel of 2,178 individuals with a total of 17,919 observations. Shown in the table are the first year and the time-invariant variables for the first 15 individuals in the sample. The variables are defined in the article.

Let  $\mathbf{X}_1$  equal a constant, education, experience, and ability (the individual's own characteristics). Let  $\mathbf{X}_2$  contain the mother's education, the father's education, and the number of siblings (the household characteristics). Let  $y$  be the wage.

- a. Compute the least squares regression coefficients in the regression of  $y$  on  $\mathbf{X}_1$ . Report the coefficients.
- b. Compute the least squares regression coefficients in the regression of  $y$  on  $\mathbf{X}_1$  and  $\mathbf{X}_2$ . Report the coefficients.
- c. Regress each of the three variables in  $\mathbf{X}_2$  on all the variables in  $\mathbf{X}_1$ . These new variables are  $\mathbf{X}_2^*$ . What are the sample means of these three variables? Explain the finding.
- d. Using (3-26), compute the  $R^2$  for the regression of  $y$  on  $\mathbf{X}_1$  and  $\mathbf{X}_2$ . Repeat the computation for the case in which the constant term is omitted from  $\mathbf{X}_1$ . What happens to  $R^2$ ?
- e. Compute the adjusted  $R^2$  for the full regression including the constant term. Interpret your result.
- f. Referring to the result in part c, regress  $y$  on  $\mathbf{X}_1$  and  $\mathbf{X}_2^*$ . How do your results compare to the results of the regression of  $y$  on  $\mathbf{X}_1$  and  $\mathbf{X}_2$ ? The comparison you are making is between the least squares coefficients when  $y$  is regressed on  $\mathbf{X}_1$  and  $\mathbf{M}_1\mathbf{X}_2$  and when  $y$  is regressed on  $\mathbf{X}_1$  and  $\mathbf{X}_2$ . Derive the result theoretically. (Your numerical results should match the theory, of course.)

**TABLE 3.5** Subsample from Koop and Tobias Data

Person	Education	Wage	Experience	Ability	Mother's education	Father's education	Siblings
1	13	1.82	1	1.00	12	12	1
2	15	2.14	4	1.50	12	12	1
3	10	1.56	1	-0.36	12	12	1
4	12	1.85	1	0.26	12	10	4
5	15	2.41	2	0.30	12	12	1
6	15	1.83	2	0.44	12	16	2
7	15	1.78	3	0.91	12	12	1
8	13	2.12	4	0.51	12	15	2
9	13	1.95	2	0.86	12	12	2
10	11	2.19	5	0.26	12	12	2
11	12	2.44	1	1.82	16	17	2
12	13	2.41	4	-1.30	13	12	5
13	12	2.07	3	-0.63	12	12	4
14	12	2.20	6	-0.36	10	12	2
15	12	2.12	3	0.28	10	12	3

## 4

# THE LEAST SQUARES ESTIMATOR

---

## 4.1 INTRODUCTION

Chapter 3 treated fitting the linear regression to the data by least squares as a purely algebraic exercise. In this chapter, we will examine in detail least squares as an **estimator** of the model parameters of the linear regression model (defined in Table 4.1). We begin in Section 4.2 by returning to the question raised but not answered in Footnote 1, Chapter 3, that is, why should we use least squares? We will then analyze the estimator in detail. There are other candidates for estimating  $\beta$ . For example, we might use the coefficients that minimize the sum of absolute values of the residuals. The question of which estimator to choose is based on the **statistical properties** of the candidates, such as unbiasedness, consistency, efficiency, and their sampling distributions. Section 4.3 considers **finite-sample properties** such as unbiasedness. The finite-sample properties of the least squares estimator are independent of the sample size. The linear model is one of relatively few settings in which definite statements can be made about the exact finite-sample properties of any estimator. In most cases, the only known properties are those that apply to large samples. Here, we can only approximate finite-sample behavior by using what we know about large-sample properties. Thus, in Section 4.4, we will examine the large-sample, or **asymptotic properties** of the least squares estimator of the regression model.<sup>1</sup>

Discussions of the properties of an estimator are largely concerned with **point estimation**—that is, in how to use the sample information as effectively as possible to produce the best single estimate of the model parameters. **Interval estimation**, considered in Section 4.5, is concerned with computing estimates that make explicit the uncertainty inherent in using randomly sampled data to estimate population quantities. We will consider some applications of interval estimation of parameters and some functions of parameters in Section 4.5. One of the most familiar applications of interval estimation is in using the model to predict the dependent variable and to provide a plausible range of uncertainty for that prediction. Section 4.6 considers prediction and forecasting using the estimated regression model.

The analysis assumes that the data in hand correspond to the assumptions of the model. In Section 4.7, we consider several practical problems that arise in analyzing nonexperimental data. Assumption A2, full rank of  $\mathbf{X}$ , is taken as a given. As we noted in Section 2.3.2, when this assumption is not met, the model is not estimable, regardless of the sample size. **Multicollinearity**, the near failure of this assumption in real-world

---

<sup>1</sup>This discussion will use our results on asymptotic distributions. It may be helpful to review Appendix D before proceeding to this material.

## 52 PART I ♦ The Linear Regression Model

**TABLE 4.1** Assumptions of the Classical Linear Regression Model

- A1. Linearity:**  $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K$
- A2. Full rank:** The  $n \times K$  sample data matrix,  $\mathbf{X}$ , has full column rank.
- A3. Exogeneity of the independent variables:**  $E[\varepsilon_i | x_{j1}, x_{j2}, \dots, x_{jK}] = 0, i, j = 1, \dots, n$ .  
There is no correlation between the disturbances and the independent variables.
- A4. Homoscedasticity and nonautocorrelation:** Each disturbance,  $\varepsilon_i$ , has the same variance,  $\sigma^2$ , and is uncorrelated with every other disturbance,  $\varepsilon_j$  conditioned on  $x$ .
- A5. Stochastic or nonstochastic data:**  $(x_{i1}, x_{i2}, \dots, x_{iK}) i = 1, \dots, n$ .
- A6. Normal distribution:** The disturbances are normally distributed.

data, is examined in Sections 4.7.1 to 4.7.3. Missing data have the potential to derail the entire analysis. The benign case in which missing values are simply manageable random gaps in the data set is considered in Section 4.7.4. The more complicated case of nonrandomly missing data is discussed in Chapter 18. Finally, the problem of badly measured data is examined in Section 4.7.5.

## 4.2 MOTIVATING LEAST SQUARES

Ease of computation is one reason that least squares is so popular. However, there are several other justifications for this technique. First, least squares is a natural approach to estimation, which makes explicit use of the structure of the model as laid out in the assumptions. Second, even if the true model is not a linear regression, the regression line fit by least squares is an optimal linear predictor for the dependent variable. Thus, it enjoys a sort of robustness that other estimators do not. Finally, under the very specific assumptions of the classical model, by one reasonable criterion, least squares will be the most efficient use of the data. We will consider each of these in turn.

### 4.2.1 THE POPULATION ORTHOGONALITY CONDITIONS

Let  $\mathbf{x}$  denote the vector of independent variables in the population regression model and for the moment, based on assumption A5, the data may be stochastic or nonstochastic. Assumption A3 states that the disturbances in the population are stochastically orthogonal to the independent variables in the model; that is,  $E[\varepsilon | \mathbf{x}] = 0$ . It follows that  $\text{Cov}[\mathbf{x}, \varepsilon] = \mathbf{0}$ . Since (by the law of iterated expectations—Theorem B.1)  $E_{\mathbf{x}}\{E[\varepsilon | \mathbf{x}]\} = E[\varepsilon] = 0$ , we may write this as

$$E_{\mathbf{x}}E_{\varepsilon}[\mathbf{x}\varepsilon] = E_{\mathbf{x}}E_y[\mathbf{x}(y - \mathbf{x}'\boldsymbol{\beta})] = \mathbf{0}$$

or

$$E_{\mathbf{x}}E_y[\mathbf{x}\mathbf{y}] = E_{\mathbf{x}}[\mathbf{x}\mathbf{x}']\boldsymbol{\beta}. \quad (4-1)$$

(The right-hand side is not a function of  $y$  so the expectation is taken only over  $\mathbf{x}$ .) Now, recall the least squares normal equations,  $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}$ . Divide this by  $n$  and write it as a summation to obtain

$$\left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i y_i \right) = \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right) \mathbf{b}. \quad (4-2)$$

## CHAPTER 4 ♦ The Least Squares Estimator 53

Equation (4-1) is a population relationship. Equation (4-2) is a sample analog. Assuming the conditions underlying the laws of large numbers presented in Appendix D are met, the sums on the left-hand and right-hand sides of (4-2) are estimators of their counterparts in (4-1). Thus, by using least squares, we are mimicking in the sample the relationship in the population. We'll return to this approach to estimation in Chapters 12 and 13 under the subject of GMM estimation.

### 4.2.2 MINIMUM MEAN SQUARED ERROR PREDICTOR

As an alternative approach, consider the problem of finding an **optimal linear predictor** for  $y$ . Once again, ignore Assumption A6 and, in addition, drop Assumption A1 that the conditional mean function,  $E[y | \mathbf{x}]$  is linear. For the criterion, we will use the mean squared error rule, so we seek the minimum mean squared error linear predictor of  $y$ , which we'll denote  $\mathbf{x}'\boldsymbol{\gamma}$ . The expected squared error of this predictor is

$$\text{MSE} = E_y E_{\mathbf{x}} [y - \mathbf{x}'\boldsymbol{\gamma}]^2.$$

This can be written as

$$\text{MSE} = E_{y,\mathbf{x}} \{y - E[y | \mathbf{x}]\}^2 + E_{y,\mathbf{x}} \{E[y | \mathbf{x}] - \mathbf{x}'\boldsymbol{\gamma}\}^2.$$

We seek the  $\boldsymbol{\gamma}$  that minimizes this expectation. The first term is not a function of  $\boldsymbol{\gamma}$ , so only the second term needs to be minimized. Note that this term is not a function of  $y$ , so the outer expectation is actually superfluous. But, we will need it shortly, so we will carry it for the present. The necessary condition is

$$\begin{aligned} \frac{\partial E_y E_{\mathbf{x}} \{[E(y | \mathbf{x}) - \mathbf{x}'\boldsymbol{\gamma}]^2\}}{\partial \boldsymbol{\gamma}} &= E_y E_{\mathbf{x}} \left\{ \frac{\partial [E(y | \mathbf{x}) - \mathbf{x}'\boldsymbol{\gamma}]^2}{\partial \boldsymbol{\gamma}} \right\} \\ &= -2E_y E_{\mathbf{x}} \{\mathbf{x}[E(y | \mathbf{x}) - \mathbf{x}'\boldsymbol{\gamma}]\} = \mathbf{0}. \end{aligned}$$

Note that we have interchanged the operations of expectation and differentiation in the middle step, since the range of integration is not a function of  $\boldsymbol{\gamma}$ . Finally, we have the equivalent condition

$$E_y E_{\mathbf{x}} [\mathbf{x} E(y | \mathbf{x})] = E_y E_{\mathbf{x}} [\mathbf{x} \mathbf{x}'] \boldsymbol{\gamma}.$$

The left-hand side of this result is  $E_{\mathbf{x}} E_y [\mathbf{x} E(y | \mathbf{x})] = \text{Cov}[\mathbf{x}, E(y | \mathbf{x})] + E[\mathbf{x}] E_{\mathbf{x}} [E(y | \mathbf{x})] = \text{Cov}[\mathbf{x}, y] + E[\mathbf{x}] E[y] = E_{\mathbf{x}} E_y [\mathbf{x} y]$ . (We have used Theorem B.2.) Therefore, the necessary condition for finding the minimum MSE predictor is

$$E_{\mathbf{x}} E_y [\mathbf{x} y] = E_{\mathbf{x}} E_y [\mathbf{x} \mathbf{x}'] \boldsymbol{\gamma}. \quad (4-3)$$

This is the same as (4-1), which takes us to the least squares condition once again. Assuming that these expectations exist, they would be estimated by the sums in (4-2), which means that regardless of the form of the conditional mean, least squares is an estimator of the coefficients of the minimum expected mean squared error linear predictor. We have yet to establish the conditions necessary for the if part of the theorem, but this is an opportune time to make it explicit:

## 54 PART I ♦ The Linear Regression Model

### THEOREM 4.1 Minimum Mean Squared Error Predictor

If the data generating mechanism generating  $(x_i, y_i)_{i=1,\dots,n}$  is such that the law of large numbers applies to the estimators in (4-2) of the matrices in (4-1), then the minimum expected squared error linear predictor of  $y_i$  is estimated by the least squares regression line.

#### 4.2.3 MINIMUM VARIANCE LINEAR UNBIASED ESTIMATION

Finally, consider the problem of finding a **linear unbiased estimator**. If we seek the one that has smallest variance, we will be led once again to least squares. This proposition will be proved in Section 4.3.5.

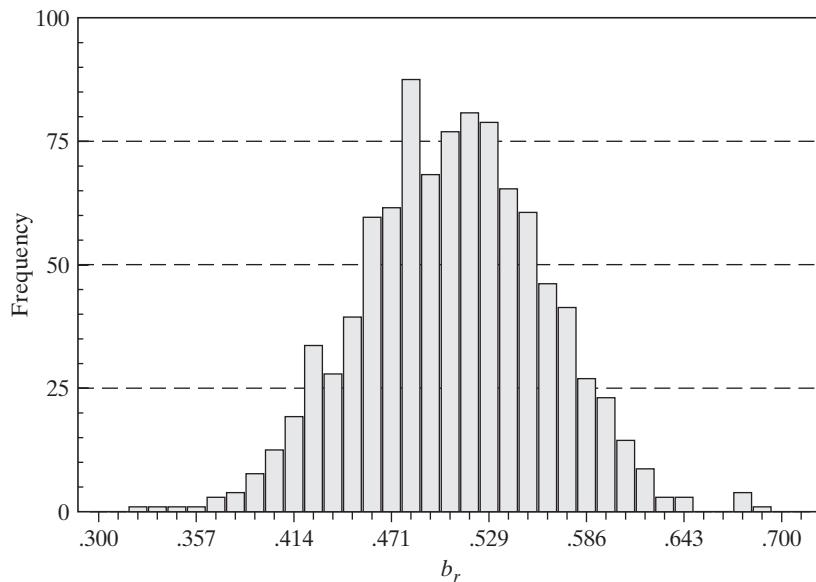
The preceding does not assert that no other competing estimator would ever be preferable to least squares. We have restricted attention to linear estimators. The preceding result precludes what might be an acceptably biased estimator. And, of course, the assumptions of the model might themselves not be valid. Although A5 and A6 are ultimately of minor consequence, the failure of any of the first four assumptions would make least squares much less attractive than we have suggested here.

### 4.3 FINITE SAMPLE PROPERTIES OF LEAST SQUARES

An “estimator” is a strategy, or formula for using the sample data that are drawn from a population. The “properties” of that estimator are a description of how that estimator can be expected to behave when it is applied to a sample of data. To consider an example, the concept of unbiasedness implies that “on average” an estimator (strategy) will correctly estimate the parameter in question; it will not be systematically too high or too low. It seems less than obvious how one could know this if they were only going to draw a single sample of data from the population and analyze that one sample. The argument adopted in classical econometrics is provided by the sampling properties of the estimation strategy. A conceptual experiment lies behind the description. One imagines “repeated sampling” from the population and characterizes the behavior of the “sample of samples.” The underlying statistical theory of the estimator provides the basis of the description. Example 4.1 illustrates.

#### *Example 4.1 The Sampling Distribution of a Least Squares Estimator*

The following sampling experiment shows the nature of a sampling distribution and the implication of unbiasedness. We drew two samples of 10,000 random draws on variables  $w_i$  and  $x_i$  from the standard normal population (mean zero, variance 1). We generated a set of  $\varepsilon_i$ 's equal to  $0.5w_i$  and then  $y_i = 0.5 + 0.5x_i + \varepsilon_i$ . We take this to be our population. We then drew 1,000 random samples of 100 observations on  $(y_i, x_i)$  from this population, and with each one, computed the least squares slope, using at replication  $r$ ,  $b_r = [\sum_{j=1}^{100}(x_{ir} - \bar{x}_r)y_{ir}] / [\sum_{j=1}^{100}(x_{ir} - \bar{x}_r)^2]$ . The histogram in Figure 4.1 shows the result of the experiment. Note that the distribution of slopes has a mean roughly equal to the “true value” of 0.5, and it has a substantial variance, reflecting the fact that the regression slope, like any other statistic computed from the sample, is a random variable. The concept of unbiasedness



**FIGURE 4.1** Histogram for Sampled Least Squares Regression Slopes.

relates to the central tendency of this distribution of values obtained in repeated sampling from the population. The shape of the histogram also suggests the normal distribution of the estimator that we will show theoretically in Section 4.3.8 (The experiment should be replicable with any regression program that provides a random number generator and a means of drawing a random sample of observations from a master data set.)

#### 4.3.1 UNBIASED ESTIMATION

The least squares estimator is unbiased in every sample. To show this, write

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}. \quad (4-4)$$

Now, take expectations, iterating over  $\mathbf{X}$ ;

$$E[\mathbf{b} | \mathbf{X}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} | \mathbf{X}].$$

By Assumption A3, the second term is  $\mathbf{0}$ , so

$$E[\mathbf{b} | \mathbf{X}] = \boldsymbol{\beta}. \quad (4-5)$$

Therefore,

$$E[\mathbf{b}] = E_{\mathbf{X}}\{E[\mathbf{b} | \mathbf{X}]\} = E_{\mathbf{X}}[\boldsymbol{\beta}] = \boldsymbol{\beta}. \quad (4-6)$$

The interpretation of this result is that for any particular set of observations,  $\mathbf{X}$ , the least squares estimator has expectation  $\boldsymbol{\beta}$ . Therefore, when we average this over the possible values of  $\mathbf{X}$ , we find the unconditional mean is  $\boldsymbol{\beta}$  as well.

## 56 PART I ♦ The Linear Regression Model

You might have noticed that in this section we have done the analysis conditioning on  $\mathbf{X}$ —that is, conditioning on the entire sample, while in Section 4.2 we have conditioned  $y_i$  on  $\mathbf{x}_i$ . (The sharp-eyed reader will also have noticed that in Table 4.1, in assumption A3, we have conditioned  $E[\varepsilon_i | \cdot]$  on  $\mathbf{x}_j$ , that is, on all  $i$  and  $j$ , which is, once again, on  $\mathbf{X}$ , not just  $\mathbf{x}_i$ . In Section 4.2, we have suggested a way to view the least squares estimator in the context of the joint distribution of a random variable,  $y$ , and a random vector,  $\mathbf{x}$ . For the purpose of the discussion, this would be most appropriate if our data were going to be a cross section of independent observations. In this context, as shown in Section 4.2.2, the least squares estimator emerges as the sample counterpart to the slope vector of the minimum mean squared error predictor,  $\boldsymbol{\gamma}$ , which is a feature of the population. In Section 4.3, we make a transition to an understanding of the process that is generating our observed sample of data. The statement that  $E[\mathbf{b}|\mathbf{X}] = \boldsymbol{\beta}$  is best understood from a Bayesian perspective; for the data that we have observed, we can expect certain behavior of the statistics that we compute, such as the least squares slope vector,  $\mathbf{b}$ . Much of the rest of this chapter, indeed much of the rest of this book, will examine the behavior of statistics as we consider whether what we learn from them in a particular sample can reasonably be extended to other samples if they were drawn under similar circumstances from the same population, or whether what we learn from a sample can be inferred to the full population. Thus, it is useful to think of the conditioning operation in  $E[\mathbf{b}|\mathbf{X}]$  in both of these ways at the same time, from the purely statistical viewpoint of deducing the properties of an estimator and from the methodological perspective of deciding how much can be learned about a broader population from a particular finite sample of data.

### 4.3.2 BIAS CAUSED BY OMISSION OF RELEVANT VARIABLES

The analysis has been based on the assumption that the correct specification of the regression model is known to be

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \quad (4-7)$$

There are numerous types of **specification errors** that one might make in constructing the regression model. The most common ones are the **omission of relevant variables** and the **inclusion of superfluous (irrelevant) variables**.

Suppose that a correctly specified regression model would be

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}, \quad (4-8)$$

where the two parts of  $\mathbf{X}$  have  $K_1$  and  $K_2$  columns, respectively. If we regress  $\mathbf{y}$  on  $\mathbf{X}_1$  without including  $\mathbf{X}_2$ , then the estimator is

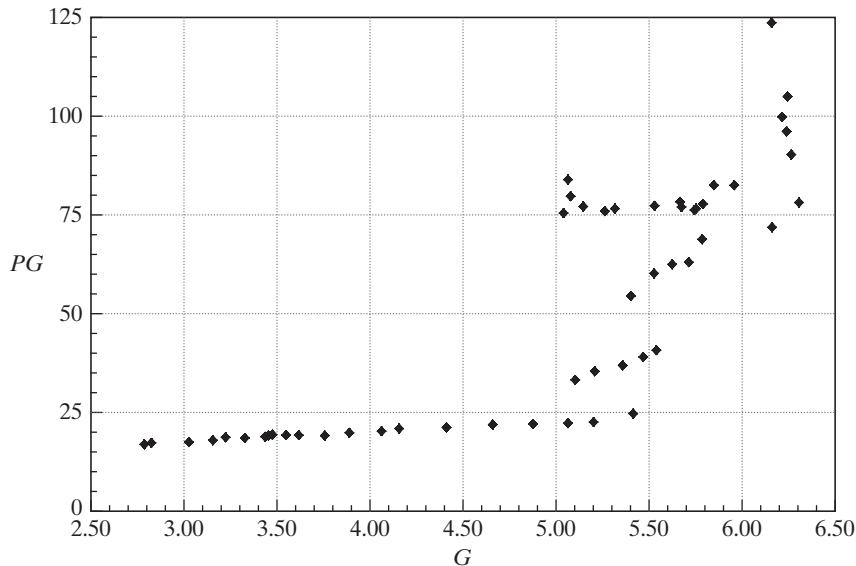
$$\mathbf{b}_1 = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{y} = \boldsymbol{\beta}_1 + (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{X}_2 \boldsymbol{\beta}_2 + (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \boldsymbol{\varepsilon}. \quad (4-9)$$

Taking the expectation, we see that unless  $\mathbf{X}'_1 \mathbf{X}_2 = \mathbf{0}$  or  $\boldsymbol{\beta}_2 = \mathbf{0}$ ,  $\mathbf{b}_1$  is biased. The well-known result is the **omitted variable formula**:

$$E[\mathbf{b}_1 | \mathbf{X}] = \boldsymbol{\beta}_1 + \mathbf{P}_{1.2} \boldsymbol{\beta}_2, \quad (4-10)$$

where

$$\mathbf{P}_{1.2} = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{X}_2. \quad (4-11)$$



## 58 PART I ♦ The Linear Regression Model

In this development, it is straightforward to deduce the directions of bias when there is a single included variable and one omitted variable. It is important to note, however, that if more than one variable is included, then the terms in the omitted variable formula involve multiple regression coefficients, which themselves have the signs of partial, not simple, correlations. For example, in the demand equation of the previous example, if the price of a closely related product had been included as well, then the simple correlation between price and income would be insufficient to determine the direction of the bias in the price elasticity. What would be required is the sign of the correlation between price and income net of the effect of the other price. This requirement might not be obvious, and it would become even less so as more regressors were added to the equation.

### 4.3.3 INCLUSION OF IRRELEVANT VARIABLES

If the regression model is correctly given by

$$\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{\epsilon} \quad (4-12)$$

and we estimate it as if (4-8) were correct (i.e., we include some extra variables), then it might seem that the same sorts of problems considered earlier would arise. In fact, this case is not true. We can view the omission of a set of relevant variables as equivalent to imposing an incorrect restriction on (4-8). In particular, omitting  $\mathbf{X}_2$  is equivalent to *incorrectly* estimating (4-8) subject to the restriction  $\boldsymbol{\beta}_2 = \mathbf{0}$ . Incorrectly imposing a restriction produces a biased estimator. Another way to view this error is to note that it amounts to incorporating incorrect information in our estimation. Suppose, however, that our error is simply a failure to use some information that is *correct*.

The inclusion of the irrelevant variables  $\mathbf{X}_2$  in the regression is equivalent to failing to impose  $\boldsymbol{\beta}_2 = \mathbf{0}$  on (4-8) in estimation. But (4-8) is not incorrect; it simply fails to incorporate  $\boldsymbol{\beta}_2 = \mathbf{0}$ . Therefore, we do not need to prove formally that the least squares estimator of  $\boldsymbol{\beta}$  in (4-8) is unbiased *even given* the restriction; we have already proved it. We can assert on the basis of all our earlier results that

$$E[\mathbf{b} | \mathbf{X}] = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{bmatrix}. \quad (4-13)$$

Then where is the problem? It would seem that one would generally want to “overfit” the model. From a theoretical standpoint, the difficulty with this view is that the failure to use correct information is always costly. In this instance, the cost will be reduced precision of the estimates. As we will show in Section 4.7.1, the covariance matrix in the short regression (omitting  $\mathbf{X}_2$ ) is never larger than the covariance matrix for the estimator obtained in the presence of the superfluous variables.<sup>2</sup> Consider a single-variable comparison. If  $\mathbf{x}_2$  is highly correlated with  $\mathbf{x}_1$ , then incorrectly including  $\mathbf{x}_2$  in the regression will greatly inflate the variance of the estimator of  $\boldsymbol{\beta}_1$ .

### 4.3.4 THE VARIANCE OF THE LEAST SQUARES ESTIMATOR

If the regressors can be treated as nonstochastic, as they would be in an experimental situation in which the analyst chooses the values in  $\mathbf{X}$ , then the **sampling variance**

<sup>2</sup>There is no loss if  $\mathbf{X}_1' \mathbf{X}_2 = \mathbf{0}$ , which makes sense in terms of the information about  $\mathbf{X}_1$  contained in  $\mathbf{X}_2$  (here, none). This situation is not likely to occur in practice, however.

## CHAPTER 4 ♦ The Least Squares Estimator 59

of the least squares estimator can be derived by treating  $\mathbf{X}$  as a matrix of constants. Alternatively, we can allow  $\mathbf{X}$  to be stochastic, do the analysis conditionally on the observed  $\mathbf{X}$ , then consider averaging over  $\mathbf{X}$  as we did in obtaining (4-6) from (4-5). Using (4-4) again, we have

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}. \quad (4-14)$$

Since we can write  $\mathbf{b} = \mathbf{A}\boldsymbol{\varepsilon}$ , where  $\mathbf{A}$  is  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ ,  $\mathbf{b}$  is a linear function of the disturbances, which by the definition we will use makes it a **linear estimator**. As we have seen, the expected value of the second term in (4-14) is  $\mathbf{0}$ . Therefore, *regardless of the distribution of  $\boldsymbol{\varepsilon}$ , under our other assumptions,  $\mathbf{b}$  is a linear, unbiased estimator of  $\boldsymbol{\beta}$ .* The conditional covariance matrix of the least squares slope estimator is

$$\begin{aligned} \text{Var}[\mathbf{b} | \mathbf{X}] &= E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})' | \mathbf{X}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} | \mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \quad (4-15)$$

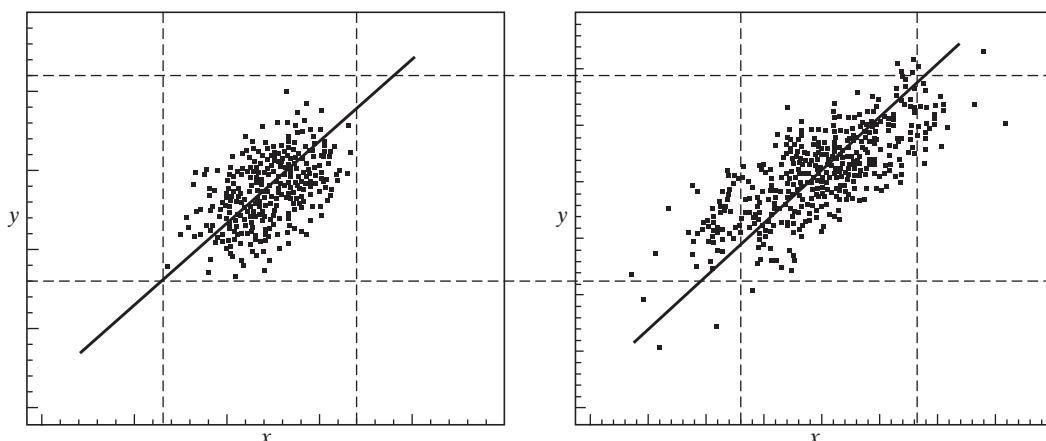
**Example 4.3 Sampling Variance in the Two-Variable Regression Model**

Suppose that  $\mathbf{X}$  contains only a constant term (column of 1s) and a single regressor  $x$ . The lower-right element of  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$  is

$$\text{Var}[b | \mathbf{x}] = \text{Var}[b - \beta | \mathbf{x}] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Note, in particular, the denominator of the variance of  $b$ . The greater the variation in  $x$ , the smaller this variance. For example, consider the problem of estimating the slopes of the two regressions in Figure 4.3. A more precise result will be obtained for the data in the right-hand panel of the figure.

**FIGURE 4.3** Effect of Increased Variation in  $x$  Given the Same Conditional and Overall Variation in  $y$ .



## 60 PART I ♦ The Linear Regression Model

### 4.3.5 THE GAUSS—MARKOV THEOREM

We will now obtain a general result for the class of linear unbiased estimators of  $\beta$ .

#### THEOREM 4.2 Gauss–Markov Theorem

*In the linear regression model with regressor matrix  $\mathbf{X}$ , the least squares estimator  $\mathbf{b}$  is the minimum variance linear unbiased estimator of  $\beta$ . For any vector of constants  $\mathbf{w}$ , the minimum variance linear unbiased estimator of  $\mathbf{w}'\beta$  in the regression model is  $\mathbf{w}'\mathbf{b}$ , where  $\mathbf{b}$  is the least squares estimator.*

Note that the theorem makes no use of Assumption A6, normality of the distribution of the disturbances. Only A1 to A4 are necessary. A direct approach to proving this important theorem would be to define the class of linear and unbiased estimators ( $\mathbf{b}_L = \mathbf{Cy}$  such that  $E[\mathbf{b}_L | \mathbf{X}] = \beta$ ) and then find the member of that class that has the smallest variance. We will use an indirect method instead. We have already established that  $\mathbf{b}$  is a linear unbiased estimator. We will now consider other linear unbiased estimators of  $\beta$  and show that any other such estimator has a larger variance.

Let  $\mathbf{b}_0 = \mathbf{Cy}$  be another linear unbiased estimator of  $\beta$ , where  $\mathbf{C}$  is a  $K \times n$  matrix. If  $\mathbf{b}_0$  is unbiased, then

$$E[\mathbf{Cy} | \mathbf{X}] = E[(\mathbf{CX}\beta + \mathbf{Ce}) | \mathbf{X}] = \beta,$$

which implies that  $\mathbf{CX} = \mathbf{I}$ . There are many candidates. For example, consider using just the first  $K$  (or, any  $K$ ) linearly independent rows of  $\mathbf{X}$ . Then  $\mathbf{C} = [\mathbf{X}_0^{-1} : \mathbf{0}]$ , where  $\mathbf{X}_0^{-1}$  is the inverse of the matrix formed from the  $K$  rows of  $\mathbf{X}$ . The covariance matrix of  $\mathbf{b}_0$  can be found by replacing  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  with  $\mathbf{C}$  in (4-14); the result is  $\text{Var}[\mathbf{b}_0 | \mathbf{X}] = \sigma^2 \mathbf{CC}'$ . Now let  $\mathbf{D} = \mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  so  $\mathbf{Dy} = \mathbf{b}_0 - \mathbf{b}$ . Then,

$$\text{Var}[\mathbf{b}_0 | \mathbf{X}] = \sigma^2 [(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')'].$$

We know that  $\mathbf{CX} = \mathbf{I} = \mathbf{DX} + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})$ , so  $\mathbf{DX}$  must equal  $\mathbf{0}$ . Therefore,

$$\text{Var}[\mathbf{b}_0 | \mathbf{X}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} + \sigma^2 \mathbf{DD}' = \text{Var}[\mathbf{b} | \mathbf{X}] + \sigma^2 \mathbf{DD}'.$$

Since a quadratic form in  $\mathbf{DD}'$  is  $\mathbf{q}'\mathbf{DD}'\mathbf{q} = \mathbf{z}'\mathbf{z} \geq 0$ , the conditional covariance matrix of  $\mathbf{b}_0$  equals that of  $\mathbf{b}$  plus a nonnegative definite matrix. Therefore, every quadratic form in  $\text{Var}[\mathbf{b}_0 | \mathbf{X}]$  is larger than the corresponding quadratic form in  $\text{Var}[\mathbf{b} | \mathbf{X}]$ , which establishes the first result.

The proof of the second statement follows from the previous derivation, since the variance of  $\mathbf{w}'\mathbf{b}$  is a quadratic form in  $\text{Var}[\mathbf{b} | \mathbf{X}]$ , and likewise for any  $\mathbf{b}_0$  and proves that each individual slope estimator  $b_k$  is the best linear unbiased estimator of  $\beta_k$ . (Let  $\mathbf{w}$  be all zeros except for a one in the  $k$ th position.) The theorem is much broader than this, however, since the result also applies to every other linear combination of the elements of  $\beta$ .

### 4.3.6 THE IMPLICATIONS OF STOCHASTIC REGRESSORS

The preceding analysis is done conditionally on the observed data. A convenient method of obtaining the unconditional statistical properties of  $\mathbf{b}$  is to obtain the desired results conditioned on  $\mathbf{X}$  first and then find the unconditional result by “averaging” (e.g., by

CHAPTER 4 ♦ The Least Squares Estimator **61**

integrating over) the conditional distributions. The crux of the argument is that if we can establish unbiasedness conditionally on an arbitrary  $\mathbf{X}$ , then we can average over  $\mathbf{X}$ 's to obtain an unconditional result. We have already used this approach to show the unconditional unbiasedness of  $\mathbf{b}$  in Section 4.3.1, so we now turn to the conditional variance.

The conditional variance of  $\mathbf{b}$  is

$$\text{Var}[\mathbf{b} | \mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

For the exact variance, we use the decomposition of variance of (B-69):

$$\text{Var}[\mathbf{b}] = E_{\mathbf{X}}[\text{Var}[\mathbf{b} | \mathbf{X}]] + \text{Var}_{\mathbf{X}}[E[\mathbf{b} | \mathbf{X}]].$$

The second term is zero since  $E[\mathbf{b} | \mathbf{X}] = \boldsymbol{\beta}$  for all  $\mathbf{X}$ , so

$$\text{Var}[\mathbf{b}] = E_{\mathbf{X}}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}] = \sigma^2 E_{\mathbf{X}}[(\mathbf{X}'\mathbf{X})^{-1}].$$

Our earlier conclusion is altered slightly. We must replace  $(\mathbf{X}'\mathbf{X})^{-1}$  with its expected value to get the appropriate covariance matrix, which brings a subtle change in the interpretation of these results. The unconditional variance of  $\mathbf{b}$  can only be described in terms of the average behavior of  $\mathbf{X}$ , so to proceed further, it would be necessary to make some assumptions about the variances and covariances of the regressors. We will return to this subject in Section 4.4.

We showed in Section 4.3.5 that

$$\text{Var}[\mathbf{b} | \mathbf{X}] \leq \text{Var}[\mathbf{b}_0 | \mathbf{X}]$$

for any linear and unbiased  $\mathbf{b}_0 \neq \mathbf{b}$  and for the specific  $\mathbf{X}$  in our sample. But if this inequality holds for every particular  $\mathbf{X}$ , then it must hold for

$$\text{Var}[\mathbf{b}] = E_{\mathbf{X}}[\text{Var}[\mathbf{b} | \mathbf{X}]].$$

That is, if it holds for every particular  $\mathbf{X}$ , then it must hold over the average value(s) of  $\mathbf{X}$ .

The conclusion, therefore, is that the important results we have obtained thus far for the least squares estimator, unbiasedness, and the Gauss–Markov theorem hold whether or not we condition on the particular sample in hand or consider, instead, sampling broadly from the population.

**THEOREM 4.3 Gauss–Markov Theorem (Concluded)**

*In the linear regression model, the least squares estimator  $\mathbf{b}$  is the minimum variance linear unbiased estimator of  $\boldsymbol{\beta}$  whether  $\mathbf{X}$  is stochastic or nonstochastic, so long as the other assumptions of the model continue to hold.*

**4.3.7 ESTIMATING THE VARIANCE OF THE LEAST SQUARES ESTIMATOR**

If we wish to test hypotheses about  $\boldsymbol{\beta}$  or to form confidence intervals, then we will require a sample estimate of the covariance matrix  $\text{Var}[\mathbf{b} | \mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ . The population

## 62 PART I ♦ The Linear Regression Model

parameter  $\sigma^2$  remains to be estimated. Since  $\sigma^2$  is the expected value of  $\varepsilon_i^2$  and  $e_i$  is an estimate of  $\varepsilon_i$ , by analogy,

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n e_i^2$$

would seem to be a natural estimator. But the least squares residuals are imperfect estimates of their population counterparts;  $e_i = y_i - \mathbf{x}'_i \boldsymbol{\beta} = \varepsilon_i - \mathbf{x}'_i (\boldsymbol{\beta} - \boldsymbol{\beta})$ . The estimator is distorted (as might be expected) because  $\boldsymbol{\beta}$  is not observed directly. The expected square on the right-hand side involves a second term that might not have expected value zero.

The least squares residuals are

$$\mathbf{e} = \mathbf{My} = \mathbf{M}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = \mathbf{M}\boldsymbol{\varepsilon},$$

as  $\mathbf{MX} = \mathbf{0}$ . [See (3-15).] An estimator of  $\sigma^2$  will be based on the sum of squared residuals:

$$\mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}. \quad (4-16)$$

The expected value of this quadratic form is

$$E[\mathbf{e}'\mathbf{e} | \mathbf{X}] = E[\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} | \mathbf{X}].$$

The scalar  $\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$  is a  $1 \times 1$  matrix, so it is equal to its trace. By using the result on cyclic permutations (A-94),

$$E[\text{tr}(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}) | \mathbf{X}] = E[\text{tr}(\mathbf{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') | \mathbf{X}].$$

Since  $\mathbf{M}$  is a function of  $\mathbf{X}$ , the result is

$$\text{tr}(\mathbf{M}E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}]) = \text{tr}(\mathbf{M}\sigma^2\mathbf{I}) = \sigma^2\text{tr}(\mathbf{M}).$$

The trace of  $\mathbf{M}$  is

$$\text{tr}[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \text{tr}(\mathbf{I}_n) - \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = \text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{I}_K) = n - K.$$

Therefore,

$$E[\mathbf{e}'\mathbf{e} | \mathbf{X}] = (n - K)\sigma^2,$$

so the natural estimator is biased toward zero, although the bias becomes smaller as the sample size increases. An unbiased estimator of  $\sigma^2$  is

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n - K}. \quad (4-17)$$

The estimator is unbiased unconditionally as well, since  $E[s^2] = E_{\mathbf{X}}\{E[s^2 | \mathbf{X}]\} = E_{\mathbf{X}}[\sigma^2] = \sigma^2$ . The **standard error of the regression** is  $s$ , the square root of  $s^2$ . With  $s^2$ , we can then compute

$$\text{Est. Var}[\mathbf{b} | \mathbf{X}] = s^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Henceforth, we shall use the notation  $\text{Est. Var}[\cdot]$  to indicate a sample estimate of the sampling variance of an estimator. The square root of the  $k$ th diagonal element of this matrix,  $\{[s^2(\mathbf{X}'\mathbf{X})^{-1}]_{kk}\}^{1/2}$ , is the **standard error** of the estimator  $b_k$ , which is often denoted simply “the standard error of  $b_k$ .”

CHAPTER 4 ♦ The Least Squares Estimator **63****4.3.8 THE NORMALITY ASSUMPTION**

To this point, our specification and analysis of the regression model are **semiparametric** (see Section 12.3). We have not used Assumption A6 (see Table 4.1), normality of  $\epsilon$ , in any of our results. The assumption is useful for constructing statistics for forming confidence intervals. In (4-4),  $\mathbf{b}$  is a linear function of the disturbance vector  $\epsilon$ . If we assume that  $\epsilon$  has a multivariate normal distribution, then we may use the results of Section B.10.2 and the mean vector and covariance matrix derived earlier to state that

$$\mathbf{b} | \mathbf{X} \sim N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]. \quad (4-18)$$

This specifies a multivariate normal distribution, so each element of  $\mathbf{b} | \mathbf{X}$  is normally distributed:



$$b_k | \mathbf{X} \sim N[\beta_k, \sigma^2(\mathbf{X}'\mathbf{X})_{kk}^{-1}]. \quad (4-19)$$

So found evidence of this result in Figure 4.1 in Example 4.1.

The distribution of  $\mathbf{b}$  is conditioned on  $\mathbf{X}$ . The normal distribution of  $\mathbf{b}$  in a finite sample is a consequence of our specific assumption of normally distributed disturbances. Without this assumption, and without some alternative specific assumption about the distribution of  $\epsilon$ , we will not be able to make any definite statement about the exact distribution of  $\mathbf{b}$ , conditional or otherwise. In an interesting result that we will explore at length in Section 4.4, we *will* be able to obtain an approximate normal distribution for  $\mathbf{b}$ , with or without assuming normally distributed disturbances and whether the regressors are stochastic or not.

**4.4 LARGE SAMPLE PROPERTIES OF THE LEAST SQUARES ESTIMATOR**

Using only assumptions A1 through A4 of the classical model listed in Table 4.1, we have established the following exact **finite-sample properties** for the least squares estimators  $\mathbf{b}$  and  $s^2$  of the unknown parameters  $\boldsymbol{\beta}$  and  $\sigma^2$ :

- $E[\mathbf{b}|\mathbf{X}] = E[\mathbf{b}] = \boldsymbol{\beta}$ —the least squares coefficient estimator is unbiased
- $E[s^2|\mathbf{X}] = E[s^2] = \sigma^2$ —the disturbance variance estimator is unbiased
- $\text{Var}[\mathbf{b}|\mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$  and  $\text{Var}[\mathbf{b}] = \sigma^2 E[(\mathbf{X}'\mathbf{X})^{-1}]$
- Gauss–Markov theorem: The MVALUE of  $\mathbf{w}'\boldsymbol{\beta}$  is  $\mathbf{w}'\mathbf{b}$  for any vector of constants,  $\mathbf{w}$ .

For this basic model, it is also straightforward to derive the large-sample, or asymptotic properties of the least squares estimator. The normality assumption, A6, becomes inessential at this point, and will be discarded save for discussions of maximum likelihood estimation in Section 4.4.6 and in Chapter 14.

**4.4.1 CONSISTENCY OF THE LEAST SQUARES ESTIMATOR OF  $\boldsymbol{\beta}$** 

Unbiasedness is a useful starting point for assessing the virtues of an estimator. It assures the analyst that their estimator will not persistently miss its target, either systematically too high or too low. However, as a guide to estimation strategy, it has two shortcomings. First, save for the least squares slope estimator we are discussing in this chapter, it is

## 64 PART I ♦ The Linear Regression Model

relatively rare for an econometric estimator to be unbiased. In nearly all cases beyond the multiple regression model, the best one can hope for is that the estimator improves in the sense suggested by unbiasedness as more information (data) is brought to bear on the study. As such, we will need a broader set of tools to guide the econometric inquiry. Second, the property of unbiasedness does not, in fact, imply that more information is better than less in terms of estimation of parameters. The sample means of random samples of 2, 100, and 10,000 are all unbiased estimators of a population mean—by this criterion all are equally desirable. Logically, one would hope that a larger sample is better than a smaller one in some sense that we are about to define (and, by extension, an extremely large sample should be much better, or even perfect). The property of **consistency** improves on unbiasedness in both of these directions.

To begin, we leave the data generating mechanism for  $\mathbf{X}$  unspecified— $\mathbf{X}$  may be any mixture of constants and random variables generated independently of the process that generates  $\boldsymbol{\varepsilon}$ . We do make two crucial assumptions. The first is a modification of Assumption A5 in Table 4.1;

**A5a.**  $(\mathbf{x}_i, \varepsilon_i) i = 1, \dots, n$  is a sequence of *independent* observations.

The second concerns the behavior of the data in large samples;

$$\text{plim}_{n \rightarrow \infty} \frac{\mathbf{X}'\mathbf{X}}{n} = \mathbf{Q}, \quad \text{a positive definite matrix.} \quad (4-20)$$

[We will return to (4-20) shortly.] The least squares estimator may be written

$$\mathbf{b} = \boldsymbol{\beta} + \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \left( \frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n} \right). \quad (4-21)$$

If  $\mathbf{Q}^{-1}$  exists, then

$$\text{plim } \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1} \text{plim} \left( \frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n} \right)$$

because the inverse is a continuous function of the original matrix. (We have invoked Theorem D.14.) We require the probability limit of the last term. Let

$$\frac{1}{n} \mathbf{X}' \boldsymbol{\varepsilon} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \varepsilon_i = \frac{1}{n} \sum_{i=1}^n \mathbf{w}_i = \bar{\mathbf{w}}. \quad (4-22)$$

Then

$$\text{plim } \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1} \text{plim } \bar{\mathbf{w}}.$$

From the exogeneity Assumption A3, we have  $E[\mathbf{w}_i] = E_{\mathbf{x}}[E[\mathbf{w}_i | \mathbf{x}_i]] = E_{\mathbf{x}}[\mathbf{x}_i E[\varepsilon_i | \mathbf{x}_i]] = \mathbf{0}$ , so the exact expectation is  $E[\bar{\mathbf{w}}] = \mathbf{0}$ . For any element in  $\mathbf{x}_i$  that is nonstochastic, the zero expectations follow from the marginal distribution of  $\varepsilon_i$ . We now consider the variance. By (B-70),  $\text{Var}[\bar{\mathbf{w}}] = E[\text{Var}[\bar{\mathbf{w}} | \mathbf{X}]] + \text{Var}[E[\bar{\mathbf{w}} | \mathbf{X}]]$ . The second term is zero because  $E[\varepsilon_i | \mathbf{x}_i] = 0$ . To obtain the first, we use  $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}] = \sigma^2 \mathbf{I}$ , so

$$\text{Var}[\bar{\mathbf{w}} | \mathbf{X}] = E[\bar{\mathbf{w}}\bar{\mathbf{w}}' | \mathbf{X}] = \frac{1}{n} \mathbf{X}' E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}] \mathbf{X} \frac{1}{n} = \left( \frac{\sigma^2}{n} \right) \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right).$$

**TABLE 4.2** Grenander Conditions for Well-Behaved Data

**G1.** For each column of  $\mathbf{X}$ ,  $\mathbf{x}_k$ , if  $d_{nk}^2 = \mathbf{x}'_k \mathbf{x}_k$ , then  $\lim_{n \rightarrow \infty} d_{nk}^2 = +\infty$ . Hence,  $\mathbf{x}_k$  does not degenerate to a sequence of zeros. Sums of squares will continue to grow as the sample size increases. No variable will degenerate to a sequence of zeros.

**G2.**  $\lim_{n \rightarrow \infty} x_{ik}^2/d_{nk}^2 = 0$  for all  $i = 1, \dots, n$ . This condition implies that no single observation will ever dominate  $\mathbf{x}'_k \mathbf{x}_k$ , and as  $n \rightarrow \infty$ , individual observations will become less important.

**G3.** Let  $\mathbf{R}_n$  be the sample correlation matrix of the columns of  $\mathbf{X}$ , excluding the constant term if there is one. Then  $\lim_{n \rightarrow \infty} \mathbf{R}_n = \mathbf{C}$ , a positive definite matrix. This condition implies that the full rank condition will always be met. We have already assumed that  $\mathbf{X}$  has full rank in a finite sample, so this assumption ensures that the condition will never be violated.

Therefore,

$$\text{Var}[\bar{\mathbf{w}}] = \left( \frac{\sigma^2}{n} \right) E\left( \frac{\mathbf{X}' \mathbf{X}}{n} \right).$$

The variance will collapse to zero if the expectation in parentheses is (or converges to) a constant matrix, so that the leading scalar will dominate the product as  $n$  increases. Assumption (4-20) should be sufficient. (Theoretically, the expectation could diverge while the probability limit does not, but this case would not be relevant for practical purposes.) It then follows that

$$\lim_{n \rightarrow \infty} \text{Var}[\bar{\mathbf{w}}] = 0 \cdot \mathbf{Q} = \mathbf{0}. \quad (4-23)$$

Since the mean of  $\bar{\mathbf{w}}$  is identically zero and its variance converges to zero,  $\bar{\mathbf{w}}$  converges in mean square to zero, so  $\text{plim } \bar{\mathbf{w}} = \mathbf{0}$ . Therefore,

$$\text{plim} \frac{\mathbf{X}' \boldsymbol{\varepsilon}}{n} = \mathbf{0}, \quad (4-24)$$

so

$$\text{plim } \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1} \cdot \mathbf{0} = \boldsymbol{\beta}. \quad (4-25)$$

This result establishes that under Assumptions A1–A4 and the additional assumption (4-20),  $\mathbf{b}$  is a **consistent estimator** of  $\boldsymbol{\beta}$  in the linear regression model.

Time-series settings that involve time trends, polynomial time series, and trending variables often pose cases in which the preceding assumptions are too restrictive. A somewhat weaker set of assumptions about  $\mathbf{X}$  that is broad enough to include most of these is the **Grenander conditions** listed in Table 4.2.<sup>3</sup> The conditions ensure that the data matrix is “well behaved” in large samples. The assumptions are very weak and likely to be satisfied by almost any data set encountered in practice.<sup>4</sup>

#### 4.4.2 ASYMPTOTIC NORMALITY OF THE LEAST SQUARES ESTIMATOR

As a guide to estimation, consistency is an improvement over unbiasedness. Since we are in the process of relaxing the more restrictive assumptions of the model, including A6, normality of the disturbances, we will also lose the normal distribution of the

<sup>3</sup>Judge et al. (1985, p. 162).

<sup>4</sup>White (2001) continues this line of analysis.

## 66 PART I ♦ The Linear Regression Model

estimator that will enable us to form confidence intervals in Section 4.5. It seems that the more general model we have built here has come at a cost. In this section, we will find that normality of the disturbances is not necessary for establishing the distributional results we need to allow statistical inference including confidence intervals and testing hypotheses. Under generally reasonable assumptions about the process that generates the sample data, large sample distributions will provide a reliable foundation for statistical inference in the regression model (and more generally, as we develop more elaborate estimators later in the book).

To derive the asymptotic distribution of the least squares estimator, we shall use the results of Section D.3. We will make use of some basic central limit theorems, so in addition to Assumption A3 (uncorrelatedness), we will assume that observations are *independent*. It follows from (4-21) that

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) = \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \left( \frac{1}{\sqrt{n}} \right) \mathbf{X}'\boldsymbol{\varepsilon}. \quad (4-26)$$

Since the inverse matrix is a continuous function of the original matrix,  $\text{plim}(\mathbf{X}'\mathbf{X}/n)^{-1} = \mathbf{Q}^{-1}$ . Therefore, if the limiting distribution of the random vector in (4-26) exists, then that limiting distribution is the same as that of

$$\left[ \text{plim} \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \right] \left( \frac{1}{\sqrt{n}} \right) \mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{Q}^{-1} \left( \frac{1}{\sqrt{n}} \right) \mathbf{X}'\boldsymbol{\varepsilon}. \quad (4-27)$$

Thus, we must establish the limiting distribution of

$$\left( \frac{1}{\sqrt{n}} \right) \mathbf{X}'\boldsymbol{\varepsilon} = \sqrt{n}(\bar{\mathbf{w}} - E[\bar{\mathbf{w}}]), \quad (4-28)$$

where  $E[\bar{\mathbf{w}}] = \mathbf{0}$ . [See (4-22).] We can use the multivariate Lindeberg–Feller version of the central limit theorem (D.19.A) to obtain the limiting distribution of  $\sqrt{n}\bar{\mathbf{w}}$ .<sup>5</sup> Using that formulation,  $\bar{\mathbf{w}}$  is the average of  $n$  independent random vectors  $\mathbf{w}_i = \mathbf{x}_i\varepsilon_i$ , with means  $\mathbf{0}$  and variances

$$\text{Var}[\mathbf{x}_i\varepsilon_i] = \sigma^2 E[\mathbf{x}_i\mathbf{x}'_i] = \sigma^2 \mathbf{Q}_i. \quad (4-29)$$

The variance of  $\sqrt{n}\bar{\mathbf{w}}$  is

$$\sigma^2 \bar{\mathbf{Q}}_n = \sigma^2 \left( \frac{1}{n} \right) [\mathbf{Q}_1 + \mathbf{Q}_2 + \cdots + \mathbf{Q}_n]. \quad (4-30)$$

As long as the sum is not dominated by any particular term and the regressors are well behaved, which in this case means that (4-20) holds,

$$\lim_{n \rightarrow \infty} \sigma^2 \bar{\mathbf{Q}}_n = \sigma^2 \mathbf{Q}. \quad (4-31)$$

Therefore, we may apply the Lindeberg–Feller central limit theorem to the vector  $\sqrt{n}\bar{\mathbf{w}}$ , as we did in Section D.3 for the univariate case  $\sqrt{n}\bar{x}$ . We now have the elements we need for a formal result. If  $[\mathbf{x}_i\varepsilon_i]$ ,  $i = 1, \dots, n$  are independent vectors distributed with

---

<sup>5</sup>Note that the Lindeberg–Levy version does not apply because  $\text{Var}[\mathbf{w}_i]$  is not necessarily constant.

CHAPTER 4 ♦ The Least Squares Estimator **67**

mean  $\mathbf{0}$  and variance  $\sigma^2 \mathbf{Q}_i < \infty$ , and if (4-20) holds, then

$$\left( \frac{1}{\sqrt{n}} \right) \mathbf{X}' \boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{0}, \sigma^2 \mathbf{Q}] \quad (4-32)$$

It then follows that

$$\mathbf{Q}^{-1} \left( \frac{1}{\sqrt{n}} \right) \mathbf{X}' \boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{Q}^{-1} \mathbf{0}, \mathbf{Q}^{-1} (\sigma^2 \mathbf{Q}) \mathbf{Q}^{-1}] \quad (4-33)$$

Combining terms,

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N[\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}] \quad (4-34)$$

Using the technique of Section D.3, we obtain the **asymptotic distribution** of  $\mathbf{b}$ :

**THEOREM 4.4 Asymptotic Distribution of  $\mathbf{b}$  with Independent Observations**

If  $\{\varepsilon_i\}$  are independently distributed with mean zero and finite variance  $\sigma^2$  and  $x_{ik}$  is such that the Grenander conditions are met, then

$$\mathbf{b} \xrightarrow{a} N \left[ \boldsymbol{\beta}, \frac{\sigma^2}{n} \mathbf{Q}^{-1} \right] \quad (4-35)$$

In practice, it is necessary to estimate  $(1/n)\mathbf{Q}^{-1}$  with  $(\mathbf{X}'\mathbf{X})^{-1}$  and  $\sigma^2$  with  $\mathbf{e}'\mathbf{e}/(n - K)$ .

If  $\boldsymbol{\varepsilon}$  is normally distributed, then result (4-18), normality of  $\mathbf{b}/\mathbf{X}$ , holds in *every* sample, so it holds asymptotically as well. The important implication of this derivation is that *if the regressors are well behaved and observations are independent*, then the **asymptotic normality** of the least squares estimator does not depend on normality of the disturbances; it is a consequence of the central limit theorem. We will consider other, more general cases in the sections to follow.

#### 4.4.3 CONSISTENCY OF $s^2$ AND THE ESTIMATOR OF ASY. VAR[ $\mathbf{b}$ ]

To complete the derivation of the asymptotic properties of  $\mathbf{b}$ , we will require an estimator of Asy. Var[ $\mathbf{b}$ ] =  $(\sigma^2/n)\mathbf{Q}^{-1}$ .<sup>6</sup> With (4-20), it is sufficient to restrict attention to  $s^2$ , so the purpose here is to assess the consistency of  $s^2$  as an estimator of  $\sigma^2$ . Expanding

$$s^2 = \frac{1}{n - K} \boldsymbol{\varepsilon}' \mathbf{M} \boldsymbol{\varepsilon}$$

produces

$$s^2 = \frac{1}{n - K} [\boldsymbol{\varepsilon}' \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\varepsilon}] = \frac{n}{n - k} \left[ \frac{\boldsymbol{\varepsilon}' \boldsymbol{\varepsilon}}{n} - \left( \frac{\boldsymbol{\varepsilon}' \mathbf{X}}{n} \right) \left( \frac{\mathbf{X}' \mathbf{X}}{n} \right)^{-1} \left( \frac{\mathbf{X}' \boldsymbol{\varepsilon}}{n} \right) \right].$$

The leading constant clearly converges to 1. We can apply (4-20), (4-24) (twice), and the product rule for **probability limits** (Theorem D.14) to assert that the second term

<sup>6</sup>See McCallum (1973) for some useful commentary on deriving the asymptotic covariance matrix of the least squares estimator.

## 68 PART I ♦ The Linear Regression Model

in the brackets converges to 0. That leaves

$$\bar{\varepsilon^2} = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2.$$

This is a narrow case in which the random variables  $\varepsilon_i^2$  are independent with the same finite mean  $\sigma^2$ , so not much is required to get the mean to converge almost surely to  $\sigma^2 = E[\varepsilon_i^2]$ . By the Markov theorem (D.8), what is needed is for  $E[|\varepsilon_i^2|^{1+\delta}]$  to be finite, so the minimal assumption thus far is that  $\varepsilon_i$  have finite moments up to slightly greater than 2. Indeed, if we further assume that every  $\varepsilon_i$  has the same distribution, then by the Khinchine theorem (D.5) or the corollary to D8, finite moments (of  $\varepsilon_i$ ) up to 2 is sufficient. **Mean square convergence** would require  $E[\varepsilon_i^4] = \phi_\varepsilon < \infty$ . Then the terms in the sum are independent, with mean  $\sigma^2$  and variance  $\phi_\varepsilon - \sigma^4$ . So, under fairly weak conditions, the first term in brackets converges in probability to  $\sigma^2$ , which gives our result,

$$\text{plim } s^2 = \sigma^2,$$

and, by the product rule,

$$\text{plim } s^2(\mathbf{X}'\mathbf{X}/n)^{-1} = \sigma^2 \mathbf{Q}^{-1}.$$

The appropriate *estimator* of the asymptotic covariance matrix of  $\mathbf{b}$  is

$$\text{Est. Asy. Var}[\mathbf{b}] = s^2(\mathbf{X}'\mathbf{X})^{-1}.$$

### 4.4.4 ASYMPTOTIC DISTRIBUTION OF A FUNCTION OF $\mathbf{b}$ : THE DELTA METHOD

We can extend Theorem D.22 to functions of the least squares estimator. Let  $\mathbf{f}(\mathbf{b})$  be a set of  $J$  continuous, linear, or nonlinear and continuously differentiable functions of the least squares estimator, and let

$$\mathbf{C}(\mathbf{b}) = \frac{\partial \mathbf{f}(\mathbf{b})}{\partial \mathbf{b}'},$$

where  $\mathbf{C}$  is the  $J \times K$  matrix whose  $j$ th row is the vector of derivatives of the  $j$ th function with respect to  $\mathbf{b}'$ . By the Slutsky theorem (D.12),

$$\text{plim } \mathbf{f}(\mathbf{b}) = \mathbf{f}(\boldsymbol{\beta})$$

and

$$\text{plim } \mathbf{C}(\mathbf{b}) = \frac{\partial \mathbf{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = \boldsymbol{\Gamma}.$$

Using a linear Taylor series approach, we expand this set of functions in the approximation

$$\mathbf{f}(\mathbf{b}) = \mathbf{f}(\boldsymbol{\beta}) + \boldsymbol{\Gamma} \times (\mathbf{b} - \boldsymbol{\beta}) + \text{higher-order terms.}$$

The higher-order terms become negligible in large samples if  $\text{plim } \mathbf{b} = \boldsymbol{\beta}$ . Then, the asymptotic distribution of the function on the left-hand side is the same as that on the right. Thus, the mean of the asymptotic distribution is  $\text{plim } \mathbf{f}(\mathbf{b}) = \mathbf{f}(\boldsymbol{\beta})$ , and the asymptotic covariance matrix is  $\{\boldsymbol{\Gamma}[\text{Asy. Var}(\mathbf{b} - \boldsymbol{\beta})]\boldsymbol{\Gamma}'\}$ , which gives us the following theorem:

**THEOREM 4.5 Asymptotic Distribution of a Function of  $\mathbf{b}$** 

If  $\mathbf{f}(\mathbf{b})$  is a set of continuous and continuously differentiable functions of  $\mathbf{b}$  such that  $\mathbf{\Gamma} = \partial\mathbf{f}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}'$  and if Theorem 4.4 holds, then

$$\mathbf{f}(\mathbf{b}) \xrightarrow{a} N \left[ \mathbf{f}(\boldsymbol{\beta}), \mathbf{\Gamma} \left( \frac{\sigma^2}{n} \mathbf{Q}^{-1} \right) \mathbf{\Gamma}' \right]. \quad (4-36)$$

In practice, the estimator of the asymptotic covariance matrix would be

$$\text{Est. Asy. Var}[\mathbf{f}(\mathbf{b})] = \mathbf{C}[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{C}'.$$

If any of the functions are nonlinear, then the property of unbiasedness that holds for  $\mathbf{b}$  may not carry over to  $\mathbf{f}(\mathbf{b})$ . Nonetheless, it follows from (4-25) that  $\mathbf{f}(\mathbf{b})$  is a consistent estimator of  $\mathbf{f}(\boldsymbol{\beta})$ , and the asymptotic covariance matrix is readily available.

**Example 4.4 Nonlinear Functions of Parameters: The Delta Method**

A dynamic version of the demand for gasoline model in Example 2.3 would be used to separate the short- and long-term impacts of changes in income and prices. The model would be

$$\begin{aligned} \ln(G/Pop)_t &= \beta_1 + \beta_2 \ln P_{G,t} + \beta_3 \ln(\text{Income}/Pop)_t + \beta_4 \ln P_{nc,t} \\ &\quad + \beta_5 \ln P_{uc,t} + \gamma \ln(G/Pop)_{t-1} + \varepsilon_t, \end{aligned}$$

where  $P_{nc}$  and  $P_{uc}$  are price indexes for new and used cars. In this model, the short-run price and income elasticities are  $\beta_2$  and  $\beta_3$ . The long-run elasticities are  $\phi_2 = \beta_2/(1 - \gamma)$  and  $\phi_3 = \beta_3/(1 - \gamma)$ , respectively. (See Section 21.3 for development of this model.) To estimate the long-run elasticities, we will estimate the parameters by least squares and then compute these two nonlinear functions of the estimates. We can use the delta method to estimate the standard errors.

Least squares estimates of the model parameters with standard errors and  $t$  ratios are given in Table 4.3. The estimated short-run elasticities are the estimates given in the table. The two estimated long-run elasticities are  $f_2 = b_2/(1 - c) = -0.069532/(1 - 0.830971) = -0.411358$  and  $f_3 = 0.164047/(1 - 0.830971) = 0.970522$ . To compute the estimates of the standard errors, we need the partial derivatives of these functions with respect to the six parameters in the model:

$$\begin{aligned} \mathbf{g}'_2 &= \partial\phi_2/\partial\boldsymbol{\beta}' = [0, 1/(1 - \gamma), 0, 0, 0, \beta_2/(1 - \gamma)^2] = [0, 5.91613, 0, 0, 0, -2.43365], \\ \mathbf{g}'_3 &= \partial\phi_3/\partial\boldsymbol{\beta}' = [0, 0, 1/(1 - \gamma), 0, 0, \beta_3/(1 - \gamma)^2] = [0, 0, 5.91613, 0, 0, 5.74174]. \end{aligned}$$

Using (4-36), we can now compute the estimates of the asymptotic variances for the two estimated long-run elasticities by computing  $\mathbf{g}'_2[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{g}_2$  and  $\mathbf{g}'_3[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{g}_3$ . The results are 0.023194 and 0.0263692, respectively. The two asymptotic standard errors are the square roots, 0.152296 and 0.162386.

**4.4.5 ASYMPTOTIC EFFICIENCY**

We have not established any large-sample counterpart to the Gauss–Markov theorem. That is, it remains to establish whether the large-sample properties of the least squares estimator are optimal by any measure. The Gauss–Markov theorem establishes finite

## 70 PART I ♦ The Linear Regression Model

**TABLE 4.3** Regression Results for a Demand Equation

Sum of squared residuals:	0.0127352		
Standard error of the regression:	0.0168227		
$R^2$ based on 51 observations	0.9951081		
Variable	Coefficient	Standard Error	t Ratio
Constant	-3.123195	0.99583	-3.136
$\ln P_G$	-0.069532	0.01973	-4.720
$\ln \text{Income}/\text{Pop}$	0.164047	0.05503	2.981
$\ln P_{nc}$	-0.178395	0.05517	-3.233
$\ln P_{uc}$	0.127009	0.03577	3.551
last period $\ln G/\text{Pop}$	0.830971	0.04576	18.158

*Estimated Covariance Matrix for  $b$  ( $e - n = \text{times } 10^{-n}$ )*

Constant	$\ln P_G$	$\ln(\text{Income}/\text{Pop})$	$\ln P_{nc}$	$\ln P_{uc}$	$\ln(G/\text{Pop})_{t-1}$
0.99168					
-0.0012088	0.00021705				
-0.052602	1.62165e-5	0.0030279			
0.0051016	-0.00021705	-0.00024708	0.0030440		
0.0091672	-4.0551e-5	-0.00060624	-0.0016782	0.0012795	
0.043915	-0.0001109	-0.0021881	0.00068116	8.57001e-5	0.0020943

sample conditions under which least squares is optimal. The requirements that the estimator be linear and unbiased limit the theorem's generality, however. One of the main purposes of the analysis in this chapter is to broaden the class of estimators in the linear regression model to those which might be biased, but which are consistent. Ultimately, we shall also be interested in nonlinear estimators. These cases extend beyond the reach of the Gauss–Markov theorem. To make any progress in this direction, we will require an alternative estimation criterion.

### DEFINITION 4.1 Asymptotic Efficiency

An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed, and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.

We can compare estimators based on their asymptotic variances. The complication in comparing two consistent estimators is that both converge to the true parameter as the sample size increases. Moreover, it usually happens (as in our example 4.5), that they converge at the same rate—that is, in both cases, the asymptotic variance of the two estimators are of the same order, such as  $O(1/n)$ . In such a situation, we can sometimes compare the asymptotic variances for the same  $n$  to resolve the ranking. The least absolute deviations estimator as an alternative to least squares provides an example.

## CHAPTER 4 ♦ The Least Squares Estimator 71

**Example 4.5 Least Squares vs. Least Absolute Deviations—A Monte Carlo Study**

We noted earlier (Section 4.2) that while it enjoys several virtues, least squares is not the only available estimator for the parameters of the linear regression model. Least absolute deviations (LAD) is an alternative. (The LAD estimator is considered in more detail in Section 7.3.1.) The LAD estimator is obtained as

$\mathbf{b}_{\text{LAD}} = \text{the minimizer of } \sum_{i=1}^n |y_i - \mathbf{x}'_i \mathbf{b}_0|,$   
in contrast to the least squares estimator,

$\mathbf{b}_{\text{LS}} = \text{the minimizer of } \sum_{i=1}^n (y_i - \mathbf{x}'_i \mathbf{b}_0)^2.$

Suppose the regression model is defined by

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i,$$

where the distribution of  $\varepsilon_i$  has conditional mean zero, constant variance  $\sigma^2$ , and conditional median zero as well—the distribution is symmetric—and  $\text{plim}(1/n)\mathbf{X}'\varepsilon = \mathbf{0}$ . That is, all the usual regression assumptions, but with the normality assumption replaced by symmetry of the distribution. Then, under our assumptions,  $\mathbf{b}_{\text{LS}}$  is a consistent and asymptotically normally distributed estimator with asymptotic covariance matrix given in Theorem 4.4, which we will call  $\sigma^2 \mathbf{A}$ . As Koenker and Bassett (1978, 1982), Huber (1987), Rogers (1993), and Koenker (2005) have discussed, under these assumptions,  $\mathbf{b}_{\text{LAD}}$  is also consistent. A good estimator of the asymptotic variance of  $\mathbf{b}_{\text{LAD}}$  would be  $(1/2)^2 [1/f(0)]^2 \mathbf{A}$  where  $f(0)$  is the density of  $\varepsilon$  at its median, zero. This means that we can compare these two estimators based on their asymptotic variances. The ratio of the asymptotic variance of the  $k$ th element of  $\mathbf{b}_{\text{LAD}}$  to the corresponding element of  $\mathbf{b}_{\text{LS}}$  would be

$$q_k = \text{Var}(b_{k,\text{LAD}}) / \text{Var}(b_{k,\text{LS}}) = (1/2)^2 (1/\sigma^2) [1/f(0)]^2.$$

If  $\varepsilon$  did actually have a normal distribution with mean (and median) zero, then

$$f(\varepsilon) = (2\pi\sigma^2)^{-1/2} \exp(-\varepsilon^2/(2\sigma^2))$$

so  $f(0) = (2\pi\sigma^2)^{-1/2}$  and for this special case  $q_k = \pi/2$ . Thus, if the disturbances are normally distributed, then LAD will be asymptotically less efficient by a factor of  $\pi/2 = 1.573$ .

The usefulness of the LAD estimator arises precisely in cases in which we cannot assume normally distributed disturbances. Then it becomes unclear which is the better estimator. It has been found in a long body search that the advantage of the LAD estimator is most likely to appear in small samples when the distribution of  $\varepsilon$  has thicker tails than the normal—that is, when outlying values of  $y_i$  are more likely. As the sample size grows larger, one can expect the LS estimator to regain its superiority. We will explore this aspect of the estimator in a small **Monte Carlo study**.

Examples 2.6 and 3.4 note an intriguing feature of the fine art market. At least in some settings, large paintings sell for more at auction than small ones. Appendix Table F4.1 contains the sale prices, widths, and heights of 430 Monet paintings. These paintings sold at auction for prices ranging from \$10,000 up to as much as \$33 million. A linear regression of the log of the price on a constant term, the log of the surface area, and the aspect ratio produces the results in the top line of Table 4.4. This is the focal point of our analysis. In order to study the different behaviors of the LS and LAD estimators, we will do the following Monte Carlo study.<sup>7</sup> We will draw without replacement 100 samples of  $R$  observations from the 430. For each of the 100 samples, we will compute  $\mathbf{b}_{\text{LS},r}$  and  $\mathbf{b}_{\text{LAD},r}$ . We then compute the average of

<sup>7</sup>Being a Monte Carlo study that uses a random number generator, there is a question of replicability. The study was done with NLOGIT and is replicable. The program can be found on the web site for the text. The qualitative results, if not the precise numerical values, can be reproduced with other programs that allow random sampling from a data set.

## 72 PART I ♦ The Linear Regression Model

**TABLE 4.4** Estimated Equations for Art Prices

<i>Full Sample</i>	<i>Constant</i>		<i>Log Area</i>		<i>Aspect Ratio</i>	
	<i>Mean</i>	<i>Standard Deviation</i>	<i>Mean</i>	<i>Standard Deviation</i>	<i>Mean</i>	<i>Standard Deviation</i>
<b>LS</b>	-8.42653	0.61184	1.33372	0.09072	-0.16537	0.12753
<b>LAD</b>	-7.62436	0.89055	1.20404	0.13626	-0.21260	0.13628
<b>R = 10</b>						
<b>LS</b>	-9.39384	6.82900	1.40481	1.00545	0.39446	2.14847
<b>LAD</b>	-8.97714	10.24781	1.34197	1.48038	0.35842	3.04773
<b>R = 50</b>						
<b>LS</b>	-8.73099	2.12135	1.36735	0.30025	-0.06594	0.52222
<b>LAD</b>	-8.91671	2.51491	1.38489	0.36299	-0.06129	0.63205
<b>R = 100</b>						
<b>LS</b>	-8.36163	1.32083	1.32758	0.17836	-0.17357	0.28977
<b>LAD</b>	-8.05195	1.54190	1.27340	0.21808	-0.20700	0.29465

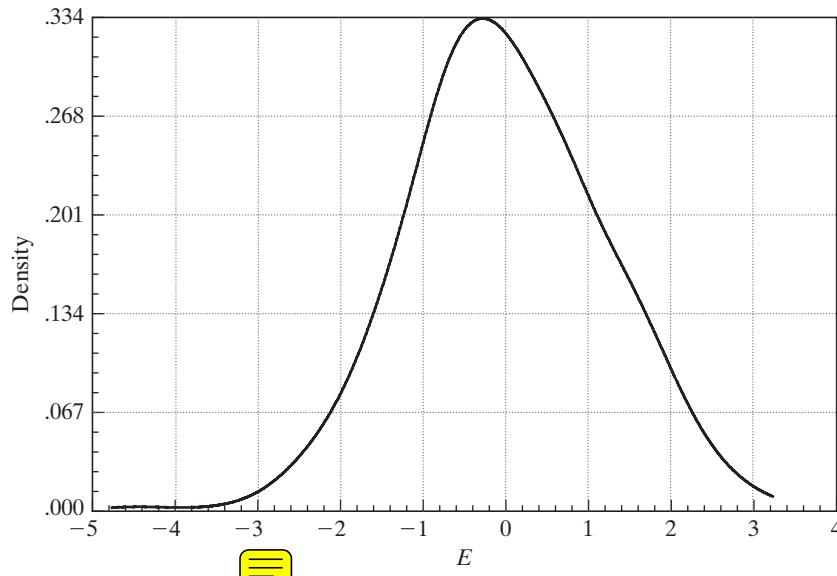
the 100 vectors and the sample variance of the 100 observations.<sup>8</sup> The sampling variability of the 100 sets of results corresponds to the notion of “variation in repeated samples.” For this experiment, we will do this for  $R = 10$ , 50, and 100. The overall sample size is fairly large, so it is reasonable to take the full sample results as at least approximately the “true parameters.” The standard errors reported for the full sample LAD estimator are computed using **bootstrapping**. Briefly, the procedure is carried out by drawing  $B$ —we used  $B = 100$ —samples of  $n$  (430) observations *with replacement*, from the full sample of  $n$  observations. The estimated variance of the LAD estimator is then obtained by computing the mean squared deviation of these  $B$  estimates around the full sample LAD estimates (not the mean of the  $B$  estimates). This procedure is discussed in detail in Section 15.4.

If the assumptions underlying our regression model are correct, we should observe the following:

1. Since both estimators are consistent, the averages should resemble the preceding main results, the more so as  $R$  increases.
2. As  $R$  increases, the sampling variance of the estimators should decline.
3. We should observe generally that the standard deviations of the LAD estimates are larger than the corresponding values for the LS estimator.
4. When  $R$  is small, the LAD estimator should compare more favorably to the LS estimator, but as  $R$  gets larger, a advantage of the LS estimator should become apparent.

A kernel density estimate for the distribution of the least squares residuals appears in Figure 4.4. There is a bit of skewness in the distribution, so a main assumption underlying our experiment may be violated to some degree. Results of the experiments are shown in Table 4.4. The force of the asymptotic results can be seen most clearly in the column for the coefficient on log Area. The decline of the standard deviation as  $R$  increases is evidence of the consistency of both estimators. In each pair of results (LS, LAD), we can also see that the estimated standard deviation of the LAD estimator is greater by a factor of about 1.2 to 1.4, which is also to be expected. Based on the normal distribution, we would have expected this ratio to be  $\sqrt{1.573} = 1.254$ .

<sup>8</sup>Note that the sample size  $R$  is not a negligible fraction of the population size, 430 for each replication. However, this does not call for a finite population correction of the variances in Table 4.4. We are not computing the variance of a sample of  $R$  observations drawn from a population of 430 paintings. We are computing the variance of a sample of  $R$  statistics each computed from a different subsample of the full population. There are a bit less than  $10^{20}$  different samples of 10 observations we can draw. The number of different samples of 50 or 100 is essentially infinite.



**FIGURE 4.4** Kernel Density Estimator for Least Squares Residuals.

#### 4.4.6 MAXIMUM LIKELIHOOD ESTIMATION

We have motivated the least squares estimator in two ways: First, we obtained Theorem 4.1 which states that the least squares estimator mimics the coefficients in the minimum mean squared error predictor of  $y$  in the joint distribution of  $y$  and  $\mathbf{x}$ . Second, Theorem 4.2, the Gauss–Markov theorem, states that the least squares estimator is the *minimum variance linear unbiased estimator* of  $\beta$  under the assumptions of the model. Neither of these results relies on Assumption A6, normality of the distribution of  $\varepsilon$ . A natural question at this point would be, what is the role of this assumption? There are two. First, the assumption of normality will produce the basis for determining the appropriate endpoints for confidence intervals in Sections 4.5 and 4.6. But, we found in Section 4.4.2 that based on the central limit theorem, we could base inference on the asymptotic normal distribution of  $\mathbf{b}$ , even if the disturbances were not normally distributed. That would seem to make the normality assumption no longer necessary, which is largely true but for a second result.

If the disturbances are normally distributed, then the least squares estimator is also the **maximum likelihood estimator (MLE)**. We will examine maximum likelihood estimation in detail in Chapter 14, so we will describe it only briefly at this point. The end result is that by virtue of being an MLE, least squares is *asymptotically efficient among consistent and asymptotically normally distributed estimators*. This is a large sample counterpart to the Gauss–Markov theorem (known formally as the Cramér–Rao bound). What the two theorems have in common is that they identify the least squares estimator as the most efficient estimator in the assumed class of estimators. They differ in the class of estimators assumed:

Gauss–Markov: Linear and unbiased estimators

ML: Based on normally distributed disturbances, consistent and asymptotically normally distributed estimators

## 74 PART I ♦ The Linear Regression Model

These are not “nested.” Notice, for example, that the MLE result does not require unbiasedness or linearity. Gauss–Markov does not require normality or consistency. The Gauss–Markov theorem is a finite sample result while the Cramér–Rao bound is an asymptotic (large-sample) property. The important aspect of the development concerns the efficiency property. Efficiency, in turn, relates to the question of how best to use the sample data for statistical inference. In general, it is difficult to establish that an estimator is efficient without being specific about the candidates. The Gauss–Markov theorem is a powerful result for the linear regression model. However, it has no counterpart in any other modeling context, so once we leave the linear model, we will require different tools for comparing estimators. The principle of maximum likelihood allows the analyst to assert asymptotic efficiency for the estimator, but only for the specific distribution assumed. Example 4.6 establishes that  $\mathbf{b}$  is the MLE in the regression model with normally distributed disturbances. Example 4.7 then considers a case in which the regression disturbances are not normally distributed and, consequently,  $\mathbf{b}$  is less efficient than the MLE.

### **Example 4.6 MLE with Normally Distributed Disturbances**

With normally distributed disturbances,  $y_i | \mathbf{x}_i$  is normally distributed with mean  $\mathbf{x}_i' \boldsymbol{\beta}$  and variance  $\sigma^2$ , so the density of  $y_i | \mathbf{x}_i$  is

$$f(y_i | \mathbf{x}_i) = \frac{\exp\left[-\frac{1}{2}(y_i - \mathbf{x}_i' \boldsymbol{\beta})^2\right]}{\sqrt{2\pi\sigma^2}}$$

The log likelihood for a sample of  $n$  independent observations is equal to the log of the joint density of the observed random variables. For a random sample, the joint density would be the product, so the log likelihood, given the data, which is written  $\ln L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X})$  would be the sum of the logs of the densities. This would be (after a bit of manipulation)

$$\ln L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) = -(n/2)[\ln \sigma^2 + \ln 2\pi + (1/\sigma^2) \frac{1}{n} \sum_i (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2].$$

The values of  $\boldsymbol{\beta}$  and  $\sigma^2$  that maximize this function are the maximum likelihood estimators of  $\boldsymbol{\beta}$  and  $\sigma^2$ . As we will explore further in Chapter 14, the functions of the data that maximize this function with respect to  $\boldsymbol{\beta}$  and  $\sigma^2$  are the least squares coefficient vector,  $\mathbf{b}$ , and the mean squared residual,  $\mathbf{e}'\mathbf{e}/n$ . Once again, we leave for Chapter 14 a derivation of the following result,

$$\text{Asy.Var}[\hat{\boldsymbol{\beta}}_{ML}] = -E[\partial^2 \ln L / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}']^{-1} = \sigma^2 E[(\mathbf{X}'\mathbf{X})^{-1}],$$

which is exactly what appears in Section 4.3.6. This shows that the least squares estimator is the maximum likelihood estimator. It is consistent, asymptotically (and exactly) normally distributed, and, under the assumption of normality, by virtue of Theorem 14.4, asymptotically efficient.

It is important to note that the properties of an MLE depend on the specific distribution assumed for the observed random variable. If some nonnormal distribution is specified for  $\varepsilon$  and it emerges that  $\mathbf{b}$  is not the MLE, then least squares may not be efficient. The following example illustrates.

### **Example 4.7 The Gamma Regression Model**

Greene (1980a) considers estimation in a regression model with an asymmetrically distributed disturbance,

$$y = (\alpha + \sigma\sqrt{P}) + \mathbf{x}'\boldsymbol{\beta} + (\varepsilon - \sigma\sqrt{P}) = \alpha^* + \mathbf{x}'\boldsymbol{\beta} + \varepsilon^*,$$

## CHAPTER 4 ♦ The Least Squares Estimator 75

where  $\varepsilon$  has the gamma distribution in Section B.4.5 [see (B-39)] and  $\sigma = \sqrt{P}/\lambda$  is the standard deviation of the disturbance. In this model, the covariance matrix of the least squares estimator of the slope coefficients (not including the constant term) is

$$\text{Asy. Var}[\mathbf{b} | \mathbf{X}] = \sigma^2 (\mathbf{X}' \mathbf{M}^0 \mathbf{X})^{-1},$$

whereas for the maximum likelihood estimator (which is not the least squares estimator),<sup>9</sup>

$$\text{Asy. Var}[\hat{\beta}_{ML}] \approx [1 - (2/P)]\sigma^2 (\mathbf{X}' \mathbf{M}^0 \mathbf{X})^{-1}.$$

But for the asymmetry parameter, this result would be the same as for the least squares estimator. We conclude that the estimator that accounts for the asymmetric disturbance distribution is more efficient asymptotically.

Another example that is somewhat similar to the model in Example 4.7 is the stochastic frontier model developed in Chapter 18. In these two cases in particular, the distribution of the disturbance is asymmetric. The maximum likelihood estimators are computed in a way that specifically accounts for this while the least squares estimator treats observations above and below the regression line symmetrically. That difference is the source of the asymptotic advantage of the MLE for these two models.

## 4.5 INTERVAL ESTIMATION

The objective of interval estimation is to present the best estimate of a parameter with an explicit expression of the uncertainty attached to that estimate. A general approach, for estimation of a parameter  $\theta$ , would be

$$\hat{\theta} \pm \text{sampling variability.} \quad (4-37)$$

(We are assuming that the interval of interest would be symmetric around  $\hat{\theta}$ .) Following the logic that the range of the sampling variability should convey the degree of (un)certainty, we consider the logical extremes. We can be absolutely (100 percent) certain that the true value of the parameter we are estimating lies in the range  $\hat{\theta} \pm \infty$ . Of course, this is not particularly informative. At the other extreme, we should place no certainty (0 percent) on the range  $\hat{\theta} \pm 0$ . The probability that our estimate precisely hits the true parameter value should be considered zero. The point is to choose a value of  $\alpha - 0.05$  or 0.01 is conventional—such that we can attach the desired confidence (probability),  $100(1 - \alpha)$  percent, to the interval in (4-37). We consider how to find that range and then apply the procedure to three familiar problems, interval estimation for one of the regression parameters, estimating a function of the parameters and predicting the value of the dependent variable in the regression using a specific setting of the independent variables. For this purpose, we depart from Assumption A6 that the disturbances are normally distributed. We will then relax that assumption and rely instead on the asymptotic normality of the estimator.

---

<sup>9</sup>The matrix  $\mathbf{M}^0$  produces data in the form of deviations from sample means. (See Section A.2.8.) In Greene's model,  $P$  must be greater than 2.

## 76 PART I ♦ The Linear Regression Model

### 4.5.1 FORMING A CONFIDENCE INTERVAL FOR A COEFFICIENT

From (4-18), we have that  $\mathbf{b}|\mathbf{X} \sim N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$ . It follows that for any particular element of  $\mathbf{b}$ , say  $b_k$ ,

$$b_k \sim N[\beta_k, \sigma^2 S^{kk}]$$

where  $S^{kk}$  denotes the  $k$ th diagonal element of  $(\mathbf{X}'\mathbf{X})^{-1}$ . By standardizing the variable, we find

$$z_k = \frac{b_k - \beta_k}{\sqrt{\sigma^2 S^{kk}}} \quad (4-38)$$

has a standard normal distribution. Note that  $z_k$ , which is a function of  $b_k$ ,  $\beta_k$ ,  $\sigma^2$  and  $S^{kk}$ , nonetheless has a distribution that involves none of the model parameters or the data;  $z_k$  is a **pivotal statistic**. Using our conventional 95 percent confidence level, we know that  $\text{Prob}[-1.96 \leq z_k \leq 1.96] = 0.95$ . By a simple manipulation, we find that

$$\text{Prob}\left[b_k - 1.96\sqrt{\sigma^2 S^{kk}} \leq \beta_k \leq b_k + 1.96\sqrt{\sigma^2 S^{kk}}\right] = 0.95. \quad (4-39)$$

Note that this is a statement about the probability that the random interval  $b_k \pm$  the sampling variability contains  $\beta_k$ , not the probability that  $\beta_k$  lies in the specified interval. If we wish to use some other level of confidence, not 95 percent, then the 1.96 in (4-39) is replaced by the appropriate  $z_{(1-\alpha/2)}$ . (We are using the notation  $z_{(1-\alpha/2)}$  to denote the value of  $z$  such that for the standard normal variable  $z$ ,  $\text{Prob}[z \leq z_{(1-\alpha/2)}] = 1 - \alpha/2$ . Thus,  $z_{0.975} = 1.96$ , which corresponds to  $\alpha = 0.05$ .)

We would have our desired confidence interval in (4-39), save for the complication that  $\sigma^2$  is not known, so the interval is not operational. It would seem natural to use  $s^2$  from the regression. This is, indeed, an appropriate approach. The quantity

$$\frac{(n - K)s^2}{\sigma^2} = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)' \mathbf{M} \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \quad (4-40)$$

is an idempotent quadratic form in a standard normal vector,  $(\boldsymbol{\varepsilon}/\sigma)$ . Therefore, it has a chi-squared distribution with degrees of freedom equal to the  $\text{rank}(\mathbf{M}) = \text{trace}(\mathbf{M}) = n - K$ . (See Section B11.4 for the proof of this result.) The chi-squared variable in (4-40) is independent of the standard normal variable in (38). To prove this, it suffices to show that

$$\left(\frac{\mathbf{b} - \boldsymbol{\beta}}{\sigma}\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X} \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)$$

is independent of  $(n - K)s^2/\sigma^2$ . In Section B11.7 (Theorem B.12), we found that a sufficient condition for the independence of a linear form  $\mathbf{Lx}$  and a idempotent quadratic form  $\mathbf{x}'\mathbf{A}\mathbf{x}$  in a standard normal vector  $\mathbf{x}$  is that  $\mathbf{LA} = \mathbf{0}$ . Letting  $\mathbf{L} = \frac{1}{\sigma}(\mathbf{b} - \boldsymbol{\beta})$  and  $\mathbf{A} = \frac{1}{\sigma^2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$ , we find that the requirement here would be that  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M} = \mathbf{0}$ . It does, as seen in (3-15). The general result is central in the derivation of many test statistics in regression analysis.

**THEOREM 4.6 Independence of  $\mathbf{b}$  and  $s^2$** 

If  $\boldsymbol{\epsilon}$  is normally distributed, then the least squares coefficient estimator  $\mathbf{b}$  is statistically independent of the residual vector  $\mathbf{e}$  and therefore, all functions of  $\mathbf{e}$ , including  $s^2$ .

Therefore, the ratio

$$t_k = \frac{(b_k - \beta_k) / \sqrt{\sigma^2 S^{kk}}}{\sqrt{[(n - K)s^2/\sigma^2]/(n - K)}} = \frac{b_k - \beta_k}{\sqrt{s^2 S^{kk}}} \quad (4-41)$$

has a  $t$  distribution with  $(n - K)$  degrees of freedom.<sup>10</sup> We can use  $t_k$  to test hypotheses or form confidence intervals about the individual elements of  $\beta$ .

The result in (4-41) differs from (38) in the use of  $s^2$  instead of  $\sigma^2$ , and in the pivotal distribution,  $t$  with  $(n - K)$  degrees of freedom, rather than standard normal. It follows that a confidence interval for  $\beta_k$  can be formed using

$$\text{Prob}\left[b_k - t_{(1-\alpha/2),[n-K]} \sqrt{s^2 S^{kk}} \leq \beta_k \leq b_k + t_{(1-\alpha/2),[n-K]} \sqrt{s^2 S^{kk}}\right] = 1 - \alpha, \quad (4-42)$$

where  $t_{(1-\alpha/2),[n-K]}$  is the appropriate critical value from the  $t$  distribution. Here, the distribution of the pivotal statistic depends on the sample size through  $(n - K)$ , but, once again, not on the parameters or the data. The practical advantage of (4-42) is that it does not involve any unknown parameters. A confidence interval for  $\beta_k$  can be based on (4-42).

**Example 4.8 Confidence Interval for the Income Elasticity of Demand for Gasoline**

Using the gasoline market data discussed in Examples 4.2 and 4.4, we estimated the following demand equation using the 52 observations:

$$\ln(G/Pop) = \beta_1 + \beta_2 \ln P_G + \beta_3 \ln(\text{Income}/Pop) + \beta_4 \ln P_{nc} + \beta_5 \ln P_{uc} + \varepsilon.$$

Least squares estimates of the model parameters with standard errors and  $t$  ratios are given in Table 4.5.

**TABLE 4.5 Regression Results for a Demand Equation**

Sum of squared residuals:	0.120871		
Standard error of the regression:	0.050712		
$R^2$ based on 52 observations	0.958443		
Variable	Coefficient	Standard Error	t Ratio
Constant	-21.21109	0.75322	-28.160
$\ln P_G$	-0.021206	0.04377	-0.485
$\ln \text{Income}/Pop$	1.095874	0.07771	14.102
$\ln P_{nc}$	-0.373612	0.15707	-2.379
$\ln P_{uc}$	0.02003	0.10330	0.194

<sup>10</sup>See (B-36) in Section B.4.2. It is the ratio of a standard normal variable to the square root of a chi-squared variable divided by its degrees of freedom.

## 78 PART I ♦ The Linear Regression Model

To form a confidence interval for the income elasticity, we need the critical value from the  $t$  distribution with  $n - K = 52 - 5 = 47$  degrees of freedom. The 95 percent critical value is 2.012. Therefore a 95 percent confidence interval for  $\beta_3$  is  $1.095874 \pm 2.012 (0.07771) = [0.9395, 1.2522]$ .

### 4.5.2 CONFIDENCE INTERVALS BASED ON LARGE SAMPLES

If the disturbances are not normally distributed, then the development in the previous section, which departs from this assumption, is not usable. But, the large sample results in Section 4.4 provide an alternative approach. Based on the development that we used to obtain Theorem 4.4 and (4-35), we have that the limiting distribution of the statistic

$$z_n = \frac{\sqrt{n}(b_{\beta_k} - \beta_k)}{\sqrt{\frac{\sigma^2}{n} Q^{kk}}}$$

is standard normal, where  $\mathbf{Q} = [\text{plim}(\mathbf{X}'\mathbf{X}/n)]^{-1}$  and  $Q^{kk}$  is the  $k$ th diagonal element of  $\mathbf{Q}$ . Based on the Slutsky theorem (D.16), we may replace  $\sigma^2$  with a consistent estimator,  $s^2$  and obtain a statistic with the same limiting distribution. And, of course, we estimate  $\mathbf{Q}$  with  $(\mathbf{X}'\mathbf{X}/n)^{-1}$ . This gives us precisely (4-41), which states that under the assumptions in Section 4.4, the “ $t$ ” statistic in (4-41) converges to standard normal even if the disturbances are not normally distributed. The implication would be that to employ the asymptotic distribution of  $\mathbf{b}$ , we should use (4-42) to compute the confidence interval but use the critical values from the standard normal table (e.g., 1.96) rather than from the  $t$  distribution. In practical terms, if the degrees of freedom in (4-42) are moderately large, say greater than 100, then the  $t$  distribution will be indistinguishable from the standard normal, and this large sample result would apply in any event. For smaller sample sizes, however, in the interest of conservatism, one might be advised to use the critical values from the  $t$  table rather than the standard normal, even in the absence of the normality assumption. In the application in Example 4.8, based on a sample of 52 observations, we formed a confidence interval for the income elasticity of demand using the critical value of 2.012 from the  $t$  table with 47 degrees of freedom. If we chose to base the interval on the asymptotic normal distribution, rather than the standard normal, we would use the 95 percent critical value of 1.96. One might think this is a bit optimistic, however, and retain the value 2.012, again, in the interest of conservatism.

**Example 4.9 Confidence Interval Based on the Asymptotic Distribution**  
In Example 4.4, we analyzed a dynamic form of the demand equation for gasoline,

$$\ln(G/\text{Pop})_t = \beta_1 + \beta_2 \ln P_{G,t} + \beta_3 \ln(\text{Income}/\text{Pop}) + \cdots + \gamma \ln(G/\text{POP})_{t-1} + \varepsilon_t.$$

In this model, the long-run price and income elasticities are  $\theta_P = \beta_2/(1-\gamma)$  and  $\theta_I = \beta_3/(1-\gamma)$ . We computed estimates of these two nonlinear functions using the least squares and the delta method, Theorem 4.5. The point estimates were  $-0.411358$  and  $0.970522$ , respectively. The estimated asymptotic standard errors were  $0.152296$  and  $0.162386$ . In order to form confidence intervals for  $\theta_P$  and  $\theta_I$ , we would generally use the asymptotic distribution, not the finite-sample distribution. Thus, the two confidence intervals are

$$\hat{\theta}_P = -0.411358 \pm 1.96(0.152296) = [-0.709858, -0.112858]$$

and

$$\hat{\theta}_I = 0.970523 \pm 1.96(0.162386) = [0.652246, 1.288800].$$

## CHAPTER 4 ♦ The Least Squares Estimator 79

In a sample of 51 observations, one might argue that using the critical value for the limiting normal distribution might be a bit optimistic. If so, using the critical value for the  $t$  distribution with  $51 - 6 = 45$  degrees of freedom would give a slightly wider interval. For example, for the income elasticity the interval would be  $0.970523 \pm 2.014(0.162386) = [0.643460, 1.297585]$ . We do note this is a practical adjustment. The statistic based on the asymptotic standard error does not actually have a  $t$  distribution with 45 degrees of freedom.

#### 4.5.3 CONFIDENCE INTERVAL FOR A LINEAR COMBINATION OF COEFFICIENTS: THE OAXACA DECOMPOSITION

With normally distributed disturbances, the least squares coefficient estimator,  $\hat{\beta}$ , is normally distributed with mean  $\beta$  and covariance matrix  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ . In Example 4.8, we showed how to use this result to form a confidence interval for one of the elements of  $\beta$ . By extending those results, we can show how to form a confidence interval for a linear function of the parameters. **Oaxaca's** (1973) and **Blinder's** (1973) **decomposition**<sup>11</sup> provides a frequently used application.<sup>11</sup>

Let  $\mathbf{w}$  denote a  $K \times 1$  vector of known constants. Then, the linear combination  $c = \mathbf{w}'\hat{\beta}$  is normally distributed with mean  $\gamma = \mathbf{w}'\beta$  and variance  $\sigma_c^2 = \mathbf{w}'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{w}$ , which we estimate with  $s_c^2 = \mathbf{w}'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{w}$ . With these in hand, we can use the earlier results to form a confidence interval for  $\gamma$ :

$$\text{Prob}[c - t_{(1-\alpha/2),[n-k]}s_c \leq \gamma \leq c + t_{(1-\alpha/2),[n-k]}s_c] = 1 - \alpha. \quad (4-43)$$

This general result can be used, for example, for the sum of the coefficients or for a difference.

Consider, then, Oaxaca's (1973) application. In a study of labor supply, separate wage regressions are fit for samples of  $n_m$  men and  $n_f$  women. The underlying regression models are

$$\ln \text{wage}_{m,i} = \mathbf{x}'_{m,i}\beta_m + \varepsilon_{m,i}, \quad i = 1, \dots, n_m$$

and

$$\ln \text{wage}_{f,j} = \mathbf{x}'_{f,j}\beta_f + \varepsilon_{f,j}, \quad j = 1, \dots, n_f.$$

The regressor vectors include sociodemographic variables, such as age, and human capital variables, such as education and experience. We are interested in comparing these two regressions, particularly to see if they suggest wage discrimination. Oaxaca suggested a comparison of the regression functions. For any two vectors of characteristics,

$$\begin{aligned} E[\ln \text{wage}_{m,i} | \mathbf{x}_{m,i}] - E[\ln \text{wage}_{f,j} | \mathbf{x}_{f,j}] &= \mathbf{x}'_{m,i}\beta_m - \mathbf{x}'_{f,j}\beta_f \\ &= \mathbf{x}'_{m,i}\beta_m - \mathbf{x}'_{m,i}\beta_f + \mathbf{x}'_{m,i}\beta_f - \mathbf{x}'_{f,j}\beta_f \\ &= \mathbf{x}'_{m,i}(\beta_m - \beta_f) + (\mathbf{x}_{m,i} - \mathbf{x}_{f,j})'\beta_f. \end{aligned}$$

The second term in this decomposition is identified with differences in human capital that would explain wage differences naturally, assuming that labor markets respond to these differences in ways that we would expect. The first term shows the differential in log wages that is attributable to differences unexplainable by human capital; holding these factors constant at  $\mathbf{x}_m$  makes the first term attributable to other factors. Oaxaca

<sup>11</sup>See Bourgignon et al. (2002) for an extensive application.

## 80 PART I ♦ The Linear Regression Model

suggested that this decomposition be computed at the means of the two regressor vectors,  $\bar{\mathbf{x}}_m$  and  $\bar{\mathbf{x}}_f$ , and the least squares coefficient vectors,  $\mathbf{b}_m$  and  $\mathbf{b}_f$ . If the regressions contain constant terms, then this process will be equivalent to analyzing  $\ln y_m - \ln y_f$ .

We are interested in forming a confidence interval for the first term, which will require two applications of our result. We will treat the two vectors of sample means as known vectors. Assuming that we have two independent sets of observations, our two estimators,  $\mathbf{b}_m$  and  $\mathbf{b}_f$ , are independent with means  $\beta_m$  and  $\beta_f$  and covariance matrices  $\sigma_m^2(\mathbf{X}'_m\mathbf{X}_m)^{-1}$  and  $\sigma_f^2(\mathbf{X}'_f\mathbf{X}_f)^{-1}$ . The covariance matrix of the difference is the sum of these two matrices. We are forming a confidence interval for  $\bar{\mathbf{x}}'_m \mathbf{d}$  where  $\mathbf{d} = \mathbf{b}_m - \mathbf{b}_f$ . The estimated covariance matrix is

$$\text{Est. Var}[\mathbf{d}] = s_m^2(\mathbf{X}'_m\mathbf{X}_m)^{-1} + s_f^2(\mathbf{X}'_f\mathbf{X}_f)^{-1}. \quad (4-44)$$

Now, we can apply the result above. We can also form a confidence interval for the second term; just define  $\mathbf{w} = \bar{\mathbf{x}}_m - \bar{\mathbf{x}}_f$  and apply the earlier result to  $\mathbf{w}'\mathbf{b}_f$ .

## 4.6 PREDICTION AND FORECASTING

After the estimation of the model parameters, a common use of regression modeling is for prediction of the dependent variable. We make a distinction between “prediction” and “forecasting” most easily based on the difference between cross section and time-series modeling. **Prediction** (which would apply to either case) involves using the regression model to compute fitted (predicted) values of the dependent variable, either within the sample or for observations outside the sample. The same set of results will apply to cross sections, panels, and time series. We consider these methods first. **Forecasting**, while largely the same exercise, explicitly gives a role to “time” and often involves lagged dependent variables and disturbances that are correlated with their past values. This exercise usually involves predicting future outcomes. An important difference between predicting and forecasting (as defined here) is that for predicting, we are usually examining a “scenario” of our own design. Thus, in the example below in which we are predicting the prices of Monet paintings, we might be interested in predicting the price of a hypothetical painting of a certain size and aspect ratio, or one that actually exists in the sample. In the time-series context, we will often try to forecast an event such as real investment next year, not based on a hypothetical economy but based on our best estimate of what economic conditions will be next year. We will use the term **ex post prediction** (or **ex post forecast**) for the cases in which the data used in the regression equation to make the prediction are either observed or constructed experimentally by the analyst. This would be the first case considered here. An **ex ante forecast** (in the time-series context) will be one that requires the analyst to forecast the independent variables first before it is possible to forecast the dependent variable. In an exercise for this chapter, real investment is forecasted using a regression model that contains real GDP and the consumer price index. In order to forecast real investment, we must first forecast real GDP and the price index. Ex ante forecasting is considered briefly here and again in Chapter 20.

## CHAPTER 4 ♦ The Least Squares Estimator 81

## 4.6.1 PREDICTION INTERVALS

Suppose that we wish to predict the value of  $y^0$  associated with a regressor vector  $\mathbf{x}^0$ . The actual value would be

$$y^0 = \mathbf{x}^{0\prime} \boldsymbol{\beta} + \varepsilon^0.$$

It follows from the Gauss–Markov theorem that

$$\hat{y}^0 = \mathbf{x}^{0\prime} \mathbf{b} \quad (4-45)$$

is the minimum variance linear unbiased estimator of  $E[y^0|\mathbf{x}^0] = \mathbf{x}^{0\prime} \boldsymbol{\beta}$ . The **prediction error** is

$$e^0 = \hat{y}^0 - y^0 = (\mathbf{b} - \boldsymbol{\beta})' \mathbf{x}^0 + \varepsilon^0.$$

The **prediction variance** of this estimator is

$$\text{Var}[e^0|\mathbf{X}, \mathbf{x}^0] = \sigma^2 + \text{Var}[(\mathbf{b} - \boldsymbol{\beta})' \mathbf{x}^0|\mathbf{X}, \mathbf{x}^0] = \sigma^2 + \mathbf{x}^{0\prime} [\sigma^2 (\mathbf{X}' \mathbf{X})^{-1}] \mathbf{x}^0. \quad (4-46)$$

If the regression contains a constant term, then an equivalent expression is

$$\text{Var}[e^0|\mathbf{X}, \mathbf{x}^0] = \sigma^2 \left[ 1 + \frac{1}{n} + \sum_{j=1}^{K-1} \sum_{k=1}^{K-1} (x_j^0 - \bar{x}_j) (x_k^0 - \bar{x}_k) (\mathbf{Z}' \mathbf{M}^0 \mathbf{Z})^{jk} \right], \quad (4-47)$$

where  $\mathbf{Z}$  is the  $K - 1$  columns of  $\mathbf{X}$  not including the constant,  $\mathbf{Z}' \mathbf{M}^0 \mathbf{Z}$  is the matrix of sums of squares and products for the columns of  $\mathbf{X}$  in deviations from their means [see (3-21)] and the “ $jk$ ” superscript indicates the  $jk$  element of the inverse of the matrix. This result suggests that the width of a confidence interval (i.e., a **prediction interval**) depends on the distance of the elements of  $\mathbf{x}^0$  from the center of the data. Intuitively, this idea makes sense; the farther the forecasted point is from the center of our experience, the greater is the degree of uncertainty. Figure 4.5 shows the effect for the bivariate case. Note that the prediction variance is composed of three parts. The second and third become progressively smaller as we accumulate more data (i.e., as  $n$  increases). But, the first term,  $\sigma^2$  is constant, which implies that no matter how much data we have, we can never predict perfectly.

The prediction variance can be estimated by using  $s^2$  in place of  $\sigma^2$ . A confidence (prediction) interval for  $y^0$  would then be formed using

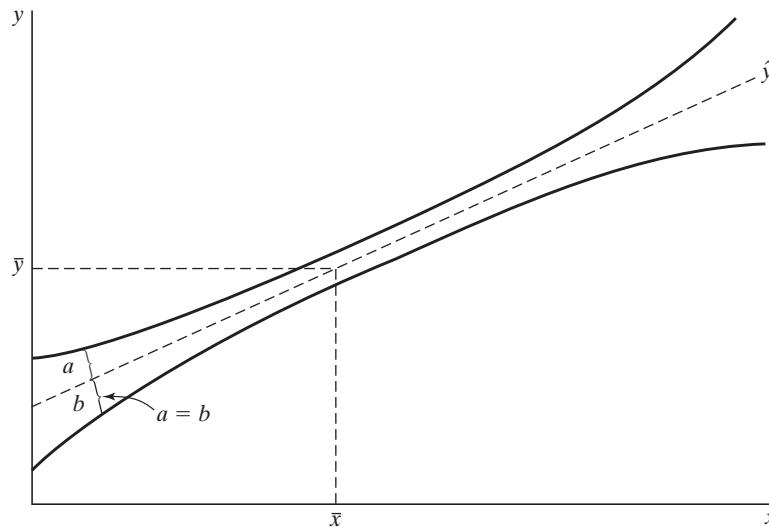
$$\text{prediction interval} = \hat{y}^0 \pm t_{(1-\alpha/2), [n-K]} se(e^0) \quad (4-48)$$

where  $t_{(1-\alpha/2), [n-K]}$  is the appropriate critical value for  $100(1 - \alpha)$  percent significance from the  $t$  table for  $n - K$  degrees of freedom and  $se(e^0)$  is the square root of the prediction variance.

4.6.2 PREDICTING  $y$  WHEN THE REGRESSION MODEL DESCRIBES  $\log y$ 

It is common to use the regression model to describe a function of the dependent variable, rather than the variable, itself. In Example 4.5 we model the sale prices of Monet paintings using

$$\ln Price = \beta_1 + \beta_2 \ln Area + \beta_3 \ln AspectRatio + \varepsilon$$

**82 PART I ♦ The Linear Regression Model**

**FIGURE 4.5** Prediction Intervals.

(area) is width times height of the painting and aspect ratio is the height divided by the width. The log form is convenient in that the coefficient provides the elasticity of the dependent variable with respect to the independent variable, that is, in this model,  $\beta_2 = \partial E[\ln Price | \ln Area, AspectRatio] / \partial \ln Area$ . However, the equation in this form is less interesting for prediction purposes than one that predicts the price, itself. The natural approach for a predictor of the form

$$\ln y^0 = \mathbf{x}^0' \mathbf{b}$$

would be to use

$$\hat{y}^0 = \exp(\mathbf{x}^0' \mathbf{b}).$$

The problem is that  $E[y|\mathbf{x}^0]$  is not equal to  $\exp(E[\ln y|\mathbf{x}^0])$ . The appropriate conditional mean function would be

$$\begin{aligned} E[y|\mathbf{x}^0] &= E[\exp(\mathbf{x}^0' \boldsymbol{\beta} + \varepsilon^0)|\mathbf{x}^0] \\ &= \exp(\mathbf{x}^0' \boldsymbol{\beta}) E[\exp(\varepsilon^0)|\mathbf{x}^0]. \end{aligned}$$

The second term is not  $\exp(E[\varepsilon^0|\mathbf{x}^0]) = 1$  in general. The precise result if  $\varepsilon^0|\mathbf{x}^0$  is normally distributed with mean zero and variance  $\sigma^2$  is  $E[\exp(\varepsilon^0)|\mathbf{x}^0] = \exp(\sigma^2/2)$ . (See Section B.4.4.) The implication for normally distributed disturbances would be that an appropriate predictor for the conditional mean would be

$$\hat{y}^0 = \exp(\mathbf{x}^0' \mathbf{b} + s^2/2) > \exp(\mathbf{x}^0' \mathbf{b}), \quad (4-49)$$

which would seem to imply that the naïve predictor would systematically underpredict  $y$ . However, this is not necessarily the appropriate interpretation of this result. The inequality implies that the naïve predictor will systematically underestimate the conditional mean function, not necessarily the realizations of the variable itself. The pertinent

## CHAPTER 4 ♦ The Least Squares Estimator 83

question is whether the conditional mean function is the desired predictor for the exponent of the dependent variable in the log regression. The conditional median might be more interesting, particularly for a financial variable such as income, expenditure, or the price of a painting. If the distribution of the variable in the log regression is symmetrically distributed (as they are when the disturbances are normally distributed), then the exponent will be asymmetrically distributed with a long tail in the positive direction, and the mean will exceed the median, possibly vastly so. In such cases, the median is often a preferred estimator of the center of a distribution. For estimating the median, rather than the mean, we would revert to the original naïve predictor,  $\hat{y}^0 = \exp(\mathbf{x}'\mathbf{b})$ .

Given the preceding, we consider estimating  $E[\exp(y)|\mathbf{x}^0]$ . If we wish to avoid the normality assumption, then it remains to determine what one should use for  $E[\exp(\varepsilon^0)|\mathbf{x}^0]$ . Duan (1983) suggested the consistent estimator (assuming that the expectation is a constant, that is, that the regression is homoscedastic),

$$\hat{E}[\exp(\varepsilon^0)|\mathbf{x}^0] = h^0 = \frac{1}{n} \sum_{i=1}^n \exp(e_i), \quad (4-50)$$

where  $e_i$  is a least squares residual in the original log form regression. Then, Duan's **smearing estimator** for prediction of  $y^0$  is

$$\hat{y}^0 = h^0 \exp(\mathbf{x}'\mathbf{b}).$$

#### 4.6.3 PREDICTION INTERVAL FOR $y$ WHEN THE REGRESSION MODEL DESCRIBES LOG $y$

We obtained a prediction interval in (4-48) for  $\ln y|\mathbf{x}^0$  in the loglinear model  $\ln y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$ ,

$$[\ln \hat{y}_{LOWER}^0, \ln \hat{y}_{UPPER}^0] = \left[ \mathbf{x}'\mathbf{b} - t_{(1-\alpha/2), [n-K]} se(e^0), \mathbf{x}'\mathbf{b} + t_{(1-\alpha/2), [n-K]} se(e^0) \right].$$

For a given choice of  $\alpha$ , say, 0.05, these values give the .025 and .975 quantiles of the distribution of  $\ln y|\mathbf{x}^0$ . If we wish specifically to estimate these quantiles of the distribution of  $y|\mathbf{x}^0$ , not  $\ln y|\mathbf{x}^0$ , then we would use;

$$[\hat{y}_{LOWER}^0, \hat{y}_{UPPER}^0] = \{ \exp[\mathbf{x}'\mathbf{b} - t_{(1-\alpha/2), [n-K]} se(e^0)], \exp[\mathbf{x}'\mathbf{b} + t_{(1-\alpha/2), [n-K]} se(e^0)] \}. \quad (4-51)$$

This follows from the result that if  $\text{Prob}[\ln y \leq \ln L] = 1 - \alpha/2$ , then  $\text{Prob}[y \leq L] = 1 - \alpha/2$ . The result is that the natural estimator is the right one for estimating the specific quantiles of the distribution of the original variable. However, if the objective is to find an interval estimator for  $y|\mathbf{x}^0$  that is as narrow as possible, then this approach is not optimal. If the distribution of  $y$  is asymmetric, as it would be for a loglinear model with normally distributed disturbances, then the naïve interval estimator is longer than necessary. Figure 4.6 shows why. We suppose that  $(L, U)$  in the figure is the prediction interval formed by (4-51). Then, the probabilities to the left of  $L$  and to the right of  $U$  each equal . Consider alternatives  $L_0 = 0$  and  $U_0$  instead. As we have constructed the figure, area (probability) between  $L_0$  and  $L$  equals the area between  $U_0$  and  $U$ . But, because the density is so much higher at  $L$ , the distance  $(0, U_0)$ , the dashed interval, is visibly shorter than that between  $(L, U)$ . The sum of the two tail probabilities is still equal to  $\alpha$ , so this provides a shorter prediction interval. We could improve on (4-51) by using, instead,  $(0, U_0)$  where  $U_0$  is simply  $\exp[\mathbf{x}'\mathbf{b} + t_{(1-\alpha), [n-K]} se(e^0)]$  (i.e., we put the

## 84 PART I ♦ The Linear Regression Model

entire tail area to the right of the upper value). However, while this is an improvement, it goes too far, as we now demonstrate.

Consider finding directly the shortest prediction interval. We treat this as an optimization problem:

$$\text{Minimize}(L, U) : I = U - L \text{ subject to } F(L) + [1 - F(U)] = \alpha,$$

where  $F$  is the cdf of the random variable  $y$  (not  $\ln y$ ). That is, we seek the shortest interval for which the two tail probabilities sum to our desired  $\alpha$  (usually 0.05). Formulate this as a Lagrangean problem,

$$\text{Minimize}(L, U, \lambda) : I^* = U - L + \lambda[F(L) + (1 - F(U)) - \alpha].$$

The solutions are found by equating the three partial derivatives to zero:

$$\partial I^*/\partial L = -1 + \lambda f(L) = 0,$$

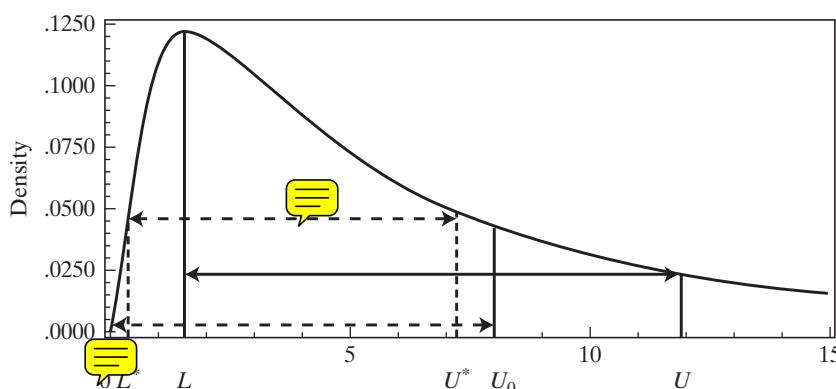
$$\partial I^*/\partial U = 1 - \lambda f(U) = 0,$$

$$\partial I^*/\partial \lambda = F(L) + [1 - F(U)] - \alpha = 0,$$

where  $f(L) = F'(L)$  and  $f(U) = F'(U)$  are the derivatives of the cdf, which are the densities of the random variable at  $L$  and  $U$ , respectively. The third equation enforces the restriction that the two tail areas sum to  $\alpha$  but does not force them to be equal. By adding the first two equations, we find that  $\lambda[f(L) - f(U)] = 0$ , which, if  $\lambda$  is not zero, means that the solution is obtained by locating  $(L^*, U^*)$  such that the tail areas sum to  $\alpha$  and the densities are equal. Looking again at Figure 4.6, we can see that the solution we would seek is  $(L^*, U^*)$  where  $0 < L^* < L$  and  $U^* < U_0$ . This is the shortest interval, and it is shorter than both  $[0, U_0]$  and  $[L, U]$ .

This derivation would apply for any distribution, symmetric or otherwise. For a symmetric distribution, however, we would obviously return to the symmetric interval in (4-51). It provides the correct solution for when the distribution is asymmetric. In Bayesian analysis, the counterpart when we examine the distribution of a parameter

**FIGURE 4.6** Lognorms Distrution for Prices of Monet Paintings.



## CHAPTER 4 ♦ The Least Squares Estimator 85

conditioned on the data, is the **highest posterior density interval**. (See Section 16.4.2.) For practical application, this computation requires a specific assumption for the distribution of  $y|\mathbf{x}^0$ , such as lognormal. Typically, we would use the smearing estimator specifically to avoid the distributional assumption. There also is no simple formula to use to locate this interval, even for the lognormal distribution. A crude grid search would probably be best, though each computation is very simple. What this derivation does establish is that one can do substantially better than the naïve interval estimator, for example using  $[0, U_0]$ .

**Example 4.10 Pricing Art**

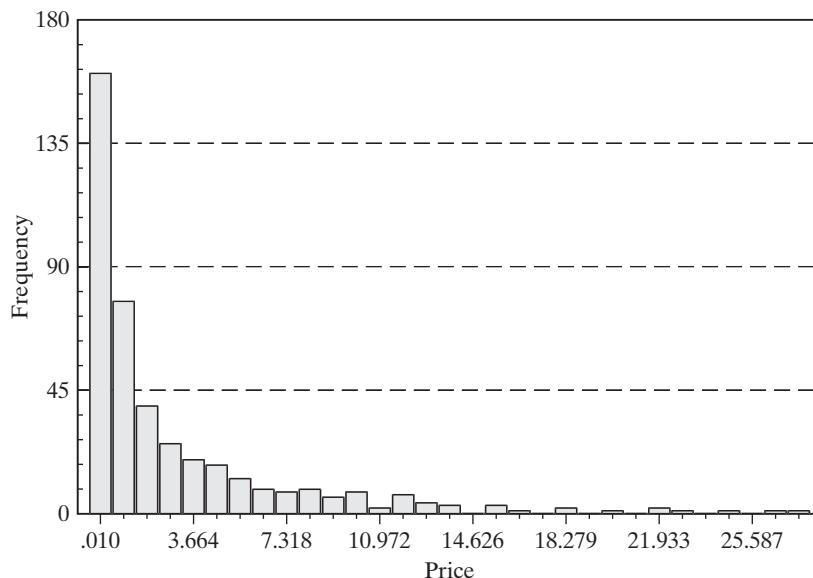
In Example 4.5, we suggested an intriguing feature of the market for Monet paintings, that larger paintings sold at auction for more than than smaller ones. In this example, we will examine that proposition empirically. Table F4.1 contains data on 430 auction prices for Monet paintings, with data on the dimensions of the paintings and several other variables that we will examine in later examples. Figure 4.7 shows a histogram for the sample of sale prices (in \$million). Figure 4.8 shows a histogram for the logs of the prices.

Results of the linear regression of  $\ln P$  on  $\ln \text{Area}$  (height times width) and Aspect Ratio (height divided by width) are given in Table 4.6.

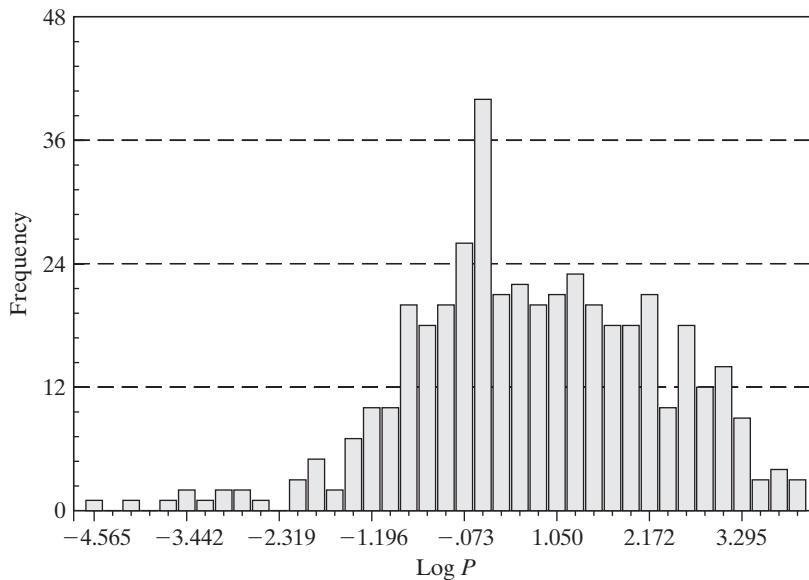
We consider using the regression model to predict the price of the painting, a 1903 painting of Charing Cross Bridge that sold for \$3,522,500. The painting is 25.6" high and 31.9" wide. (This is observation 60 in the sample.) The log area equals  $\ln(25.6 \times 31.9) = 6.705198$  and the aspect ratio equals  $25.6/31.9 = 0.802508$ . The prediction for the log of the price would be

$$\ln P|\mathbf{x}^0 = -8.42653 + 1.33372(6.705198) - 0.16537(0.802508) = 0.383636.$$

**FIGURE 4.7** Histogram for Sale Prices of 430 Monet Paintings (\$million).



## 86 PART I ♦ The Linear Regression Model



**FIGURE 4.8** Histogram of Logs of Auction Prices for Monet Paintings.

**TABLE 4.6** Estimated Equation for Log Price

Mean of log Price	.33274
Sum of squared residuals	519.17235
Standard error of regression	1.10266
R-squared	.33620
Adjusted R-squared	.33309
Number of observations	430

Variable	Coefficient	Standard Error	t	Mean of X
Constant	-8.42653	.61183	-13.77	1.00000
LOGAREA	1.33372	.09072	14.70	6.68007
ASPECT	-.16537	.12753	-1.30	0.90759
Estimated	Asymptotic Constant	Covariance LogArea	Matrix AspectRatio	
Constant	.37434	-.05429	-.00974	
LogArea	-.05429	.00823	-.00075	
AspectRatio	-.00974	-.00075	.01626	

Note that the mean log price is 0.33274, so this painting is expected to sell for roughly 5 percent more than the average painting, based on its dimensions. The estimate of the prediction variance is computed using (4-47);  $s_p = 1.104027$ . The sample is large enough to use the critical value from the standard normal table, 1.96, for a 95 percent confidence

## CHAPTER 4 ♦ The Least Squares Estimator 87

interval. A prediction interval for the log of the price is therefore

$$0.383636 \pm 1.96(1.104027) = [-1.780258, 2.547529].$$



For predicting the price, the naïve predictor would be  $\exp(0.383636) = \$1.476411M$ , which is far under the actual sale price of \$3.5225M. To compute the smearing estimator, we require the mean of the exponents of the residuals, which is 1.813045. The revised point estimate for the price would thus be  $1.813045 \times 1.47641 = \$2.660844M$ —this is better, but still fairly far off. This particular painting seems to have sold for relatively more than history (the data) would have predicted.

To compute an interval estimate for the price, we begin with the naïve prediction by simply exponentiating the lower and upper values for the log price, which gives a prediction interval for 95 percent confidence of  $[\$0.168595M, \$12.77503M]$ . Using the method suggested in Section 4.6.3, however, we are able to narrow this interval to  $[0.021261, 9.02]$ , a range of \$9M compared to the range based on the simple calculation of \$12.2M. The interval divides the .05 tail probability into 0.00063 on the left and .04937 on the right. The search algorithm is outlined next.

#### Grid Search Algorithm for Optimal Prediction Interval [LO, UO]

$$\mathbf{x}^0 = (1, \log(25.6 \times 31.9), 25.6/31.9)';$$

$$\hat{\mu}^0 = \exp(\mathbf{x}^0' \mathbf{b}), \hat{\sigma}_p^0 = \sqrt{s^2 + \mathbf{x}^0' [s^2 (\mathbf{X}' \mathbf{X})^{-1}] \mathbf{x}^0};$$

Confidence interval for  $\log P|\mathbf{x}^0$ : [Lower, Upper] =  $[\hat{\mu}^0 - 1.96\hat{\sigma}_p^0, \hat{\mu}^0 + 1.96\hat{\sigma}_p^0]$ ;

Naïve confidence interval for Price| $\mathbf{x}^0$ : L1 =  $\exp(\text{Lower})$ ; U1 =  $\exp(\text{Upper})$ ;

Initial value of L was .168595, LO = this value;

Grid search for optimal interval, decrement by  $\Delta = .005$  (chosen ad hoc);

Decrement LO and compute companion UO until densities match;

(\*) LO = LO -  $\Delta$  = new value of LO;

$$f(\text{LO}) = \left[ \text{LO} \hat{\sigma}_p^0 \sqrt{2\pi} \right]^{-1} \exp \left[ -\frac{1}{2} ((\ln \text{LO} - \hat{\mu}^0) / \hat{\sigma}_p^0)^2 \right];$$

$F(\text{LO}) = \Phi((\ln(\text{LO}) - \hat{\mu}^0) / \hat{\sigma}_p^0)$  = left tail probability;

UO =  $\exp(\hat{\sigma}_p^0 \Phi^{-1} [F(\text{LO}) + .95] + \hat{\mu}^0)$  = next value of UO;

$$f(\text{UO}) = \left[ \text{UO} \hat{\sigma}_p^0 \sqrt{2\pi} \right]^{-1} \exp \left[ -\frac{1}{2} ((\ln \text{UO} - \hat{\mu}^0) / \hat{\sigma}_p^0)^2 \right];$$

$1 - F(\text{UO}) = 1 - \Phi((\ln(\text{UO}) - \hat{\mu}^0) / \hat{\sigma}_p^0)$  = right tail probability;

Compare  $f(\text{LO})$  to  $f(\text{UO})$ . If not equal, return to (\*). If equal, exit.

#### 4.6.4 FORECASTING

The preceding discussion assumes that  $\mathbf{x}^0$  is known with certainty, ex post, or has been forecast perfectly, ex ante. If  $\mathbf{x}^0$  must, itself, be forecast (an ex ante forecast), then the formula for the forecast variance in (4-46) would have to be modified to incorporate the uncertainty in forecasting  $\mathbf{x}^0$ . This would be analogous to the term  $\sigma^2$  in the prediction variance that accounts for the implicit prediction of  $\varepsilon^0$ . This will vastly complicate the computation. Most authors view it as simply intractable. Beginning with Feldstein (1971), derivation of firm analytical results for the correct forecast variance for this case remain to be derived except for simple special cases. The one qualitative result that seems certain is that (4-46) will underestimate the true variance. McCullough (1996)

## 88 PART I ♦ The Linear Regression Model

presents an alternative approach to computing appropriate forecast standard errors based on the method of bootstrapping. (See Chapter 15.)

Various measures have been proposed for assessing the predictive accuracy of forecasting models.<sup>12</sup> Most of these measures are designed to evaluate ex post forecasts, that is, forecasts for which the independent variables do not themselves have to be forecast. Two measures that are based on the residuals from the forecasts are the **root mean squared error**,

$$\text{RMSE} = \sqrt{\frac{1}{n^0} \sum_i (y_i - \hat{y}_i)^2},$$

and the **mean absolute error**,

$$\text{MAE} = \frac{1}{n^0} \sum_i |y_i - \hat{y}_i|,$$

where  $n^0$  is the number of periods being forecasted. (Note that both of these, as well as the following measures, below are backward looking in that they are computed using the observed data on the independent variable.) These statistics have an obvious scaling problem—multiplying values of the dependent variable by any scalar multiplies the measure by that scalar as well. Several measures that are scale free are based on the **Theil  $U$  statistic**:<sup>13</sup>

$$U = \sqrt{\frac{(1/n^0) \sum_i (y_i - \hat{y}_i)^2}{(1/n^0) \sum_i y_i^2}}.$$

This measure is related to  $R^2$  but is not bounded by zero and one. Large values indicate a poor forecasting performance. An alternative is to compute the measure in terms of the changes in  $y$ :

$$U_\Delta = \sqrt{\frac{(1/n^0) \sum_i (\Delta y_i - \Delta \hat{y}_i)^2}{(1/n^0) \sum_i (\Delta y_i)^2}}$$

where  $\Delta y_i = y_i - y_{i-1}$  and  $\Delta \hat{y}_i = \hat{y}_i - \hat{y}_{i-1}$ , or, in percentage changes,  $\Delta y_i = (y_i - y_{i-1})/y_{i-1}$  and  $\Delta \hat{y}_i = (\hat{y}_i - \hat{y}_{i-1})/y_{i-1}$ . These measures will reflect the model's ability to track turning points in the data.

## 4.7 DATA PROBLEMS

The analysis to this point has assumed that the data in hand,  $\mathbf{X}$  and  $\mathbf{y}$ , are well measured and correspond to the assumptions of the model in Table 2.1 and to the variables described by the underlying theory. At this point, we consider several ways that “real-world” observed nonexperimental data fail to meet the assumptions. Failure of the assumptions generally has implications for the performance of the estimators of the

<sup>12</sup>See Theil (1961) and Fair (1984).

<sup>13</sup>Theil (1961).

CHAPTER 4 ♦ The Least Squares Estimator **89**

model parameters—unfortunately, none of them good. The cases we will examine are

- Multicollinearity: Although the full rank assumption, A2, is met, it almost fails. (“Almost” is a matter of degree, and sometimes a matter of interpretation.) Multicollinearity leads to imprecision in the estimator, though not to any systematic biases in estimation.
- Missing values: Gaps in  $\mathbf{X}$  and/or  $\mathbf{y}$  can be harmless. In many cases, the analyst can (and should) simply ignore them, and just use the complete data in the sample. In other cases, when the data are missing for reasons that are related to the outcome being studied, ignoring the problem can lead to inconsistency of the estimators.
- Measurement error: Data often correspond only imperfectly to the theoretical construct that appears in the model—individual data on income and education are familiar examples. Measurement error is never benign. The least harmful case is measurement error in the dependent variable. In this case, at least under probably reasonable assumptions, the implication is to degrade the fit of the model to the data compared to the (unfortunately hypothetical) case in which the data are accurately measured. Measurement error in the regressors is malignant—it produces systematic biases in estimation that are difficult to remedy.

#### 4.7.1 MULTICOLLINEARITY

The Gauss–Markov theorem states that among all linear unbiased estimators, the least squares estimator has the smallest variance. Although this result is useful, it does not assure us that the least squares estimator has a small variance in any absolute sense. Consider, for example, a model that contains two explanatory variables and a constant. For either slope coefficient,

$$\text{Var}[b_k | \mathbf{X}] = \frac{\sigma^2}{(1 - r_{12}^2) \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2} = \frac{\sigma^2}{(1 - r_{12}^2) S_{kk}}, \quad k = 1, 2. \quad (4-52)$$

If the two variables are perfectly correlated, then the variance is infinite. The case of an exact linear relationship among the regressors is a serious failure of the assumptions of the model, not of the data. The more common case is one in which the variables are highly, but not perfectly, correlated. In this instance, the regression model retains all its assumed properties, although potentially severe statistical problems arise. The problem faced by applied researchers when regressors are highly, although not perfectly, correlated include the following symptoms:

- Small changes in the data produce wide swings in the parameter estimates.
- Coefficients may have very high standard errors and low significance levels even though they are jointly significant and the  $R^2$  for the regression is quite high.
- Coefficients may have the “wrong” sign or implausible magnitudes.

For convenience, define the data matrix,  $\mathbf{X}$ , to contain a constant and  $K - 1$  other variables measured in deviations from their means. Let  $\mathbf{x}_k$  denote the  $k$ th variable, and let  $\mathbf{X}_{(k)}$  denote all the other variables (including the constant term). Then, in the inverse

## 90 PART I ♦ The Linear Regression Model

matrix,  $(\mathbf{X}'\mathbf{X})^{-1}$ , the  $k$ th diagonal element is

$$\begin{aligned} (\mathbf{x}'_k \mathbf{M}_{(k)} \mathbf{x}_k)^{-1} &= [\mathbf{x}'_k \mathbf{x}_k - \mathbf{x}'_k \mathbf{X}_{(k)} (\mathbf{X}'_{(k)} \mathbf{X}_{(k)})^{-1} \mathbf{X}'_{(k)} \mathbf{x}_k]^{-1} \\ &= \left[ \mathbf{x}'_k \mathbf{x}_k \left( 1 - \frac{\mathbf{x}'_k \mathbf{X}_{(k)} (\mathbf{X}'_{(k)} \mathbf{X}_{(k)})^{-1} \mathbf{X}'_{(k)} \mathbf{x}_k}{\mathbf{x}'_k \mathbf{x}_k} \right) \right]^{-1} \\ &= \frac{1}{(1 - R_{k.}^2) S_{kk}}, \end{aligned} \quad (4-53)$$

where  $R_{k.}^2$  is the  $R^2$  in the regression of  $x_k$  on all the other variables. In the multiple regression model, the variance of the  $k$ th least squares coefficient estimator is  $\sigma^2$  times this ratio. It then follows that the more highly correlated a variable is with the other variables in the model (collectively), the greater its variance will be. In the most extreme case, in which  $\mathbf{x}_k$  can be written as a linear combination of the other variables so that  $R_{k.}^2 = 1$ , the variance becomes infinite. The result

$$\text{Var}[b_k | \mathbf{X}] = \frac{\sigma^2}{(1 - R_{k.}^2) \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}, \quad (4-54)$$

shows the three ingredients of the precision of the  $k$ th least squares coefficient estimator:

- Other things being equal, the greater the correlation of  $x_k$  with the other variables, the higher the variance will be, due to multicollinearity.
- Other things being equal, the greater the variation in  $x_k$ , the lower the variance will be. This result is shown in Figure 4.3.
- Other things being equal, the better the overall fit of the regression, the lower the variance will be. This result would follow from a lower value of  $\sigma^2$ . We have yet to develop this implication, but it can be suggested by Figure 4.3 by imagining the identical figure in the right panel but with all the points moved closer to the regression line.

Since nonexperimental data will never be orthogonal ( $R_{k.}^2 = 0$ ), to some extent multicollinearity will always be present. When is multicollinearity a problem? That is, when are the variances of our estimates so adversely affected by this intercorrelation that we should be “concerned”? Some computer packages report a **variance inflation factor** (VIF),  $1/(1 - R_{k.}^2)$ , for each coefficient in a regression as a diagnostic statistic. As can be seen, the VIF for a variable shows the increase in  $\text{Var}[b_k]$  that can be attributable to the fact that this variable is not orthogonal to the other variables in the model. Another measure that is specifically directed at  $\mathbf{X}$  is the **condition number** of  $\mathbf{X}'\mathbf{X}$ , which is the square root of the ratio of the largest characteristic root of  $\mathbf{X}'\mathbf{X}$  (after scaling each column so that it has unit length) to the smallest. Values in excess of 20 are suggested as indicative of a problem [Belsley, Kuh, and Welsch (1980)]. (The condition number for the Longley data of Example 4.11 is over 15,000!)

### Example 4.11 Multicollinearity in the Longley Data

The data in Appendix Table F4.2 were assembled by J. Longley (1967) for the purpose of assessing the accuracy of least squares computations by computer programs. (These data are still widely used for that purpose.) The Longley data are notorious for severe multicollinearity. Note, for example, the last year of the data set. The last observation does not appear to

**TABLE 4.7** Longley Results: Dependent Variable is Employment

	<i>1947–1961</i>	<i>Variance Inflation</i>	<i>1947–1962</i>
Constant	1,459,415		1,169,087
Year	−721.756	143.4638	−576.464
GNP deflator	−181.123	75.6716	−19.7681
GNP	0.0910678	132.467	0.0643940
Armed Forces	−0.0749370	1.55319	−0.0101453

be unusual. But, the results in Table 4.7 show the dramatic effect of dropping this single observation from a regression of employment on a constant and the other variables. The last coefficient rises by 600 percent, and the third rises by 800 percent.

Several strategies have been proposed for finding and coping with multicollinearity.<sup>14</sup> Under the view that a multicollinearity “problem” arises because of a shortage of information, one suggestion is to obtain more data. One might argue that if analysts had such additional information available at the outset, they ought to have used it before reaching this juncture. More information need not mean more observations, however. The obvious practical remedy (and surely the most frequently used) is to drop variables suspected of causing the problem from the regression—that is, to impose on the regression an assumption, possibly erroneous, that the “problem” variable does not appear in the model. In doing so, one encounters the problems of specification that we will discuss in Section 4.7.2. If the variable that is dropped actually belongs in the model (in the sense that its coefficient,  $\beta_k$ , is not zero), then estimates of the remaining coefficients will be biased, possibly severely so. On the other hand, overfitting—that is, trying to estimate a model that is too large—is a common error, and dropping variables from an excessively specified model might have some virtue.

Using diagnostic tools to “detect” multicollinearity could be viewed as an attempt to distinguish a bad model from bad data. But, in fact, the problem only stems from a prior opinion with which the data seem to be in conflict. A finding that suggests multicollinearity is adversely affecting the estimates seems to suggest that but for this effect, all the coefficients would be statistically significant and of the right sign. Of course, this situation need not be the case. If the data suggest that a variable is unimportant in a model, then, the theory notwithstanding, the researcher ultimately has to decide how strong the commitment is to that theory. Suggested “remedies” for multicollinearity might well amount to attempts to force the theory on the data.

#### 4.7.2 PRETEST ESTIMATION

As a response to what appears to be a “multicollinearity problem,” it is often difficult to resist the temptation to drop what appears to be an offending variable from the regression, if it seems to be the one causing the problem. This “strategy” creates a subtle dilemma for the analyst. Consider the partitioned multiple regression

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

<sup>14</sup>See Hill and Adkins (2001) for a description of the standard set of tools for diagnosing collinearity.

## 92 PART I ♦ The Linear Regression Model

If we regress  $\mathbf{y}$  only on  $\mathbf{X}_1$ , the estimator is biased;

$$E[\mathbf{b}_1|\mathbf{X}] = \boldsymbol{\beta}_1 + \mathbf{P}_{1.2}\boldsymbol{\beta}_2.$$

The covariance matrix of this estimator is

$$\text{Var}[\mathbf{b}_1|\mathbf{X}] = \sigma^2(\mathbf{X}'_1\mathbf{X}_1)^{-1}.$$

(Keep in mind, this variance is around the  $E[\mathbf{b}_1|\mathbf{X}]$ , not around  $\boldsymbol{\beta}_1$ .) If  $\boldsymbol{\beta}_2$  is not actually zero, then in the multiple regression of  $\mathbf{y}$  on  $(\mathbf{X}_1, \mathbf{X}_2)$ , the variance of  $\mathbf{b}_{1.2}$  around its mean,  $\boldsymbol{\beta}_1$  would be

$$\text{Var}[\mathbf{b}_{1.2}|\mathbf{X}] = \sigma^2(\mathbf{X}'_1\mathbf{M}_2\mathbf{X}_1)^{-1}$$

where

$$\mathbf{M}_2 = \mathbf{I} - \mathbf{X}_2(\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{X}'_2,$$

or

$$\text{Var}[\mathbf{b}_{1.2}|\mathbf{X}] = \sigma^2[\mathbf{X}'_1\mathbf{X}_1 - \mathbf{X}'_1\mathbf{X}_2(\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{X}'_2\mathbf{X}_1]^{-1}.$$

We compare the two covariance matrices. It is simpler to compare the inverses. [See result (A-120).] Thus,

$$\{\text{Var}[\mathbf{b}_1|\mathbf{X}]\}^{-1} - \{\text{Var}[\mathbf{b}_{1.2}|\mathbf{X}]\}^{-1} = (1/\sigma^2)\mathbf{X}'_1\mathbf{X}_2(\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{X}'_2\mathbf{X}_1,$$

which is a nonnegative definite matrix. The implication is that the variance of  $\mathbf{b}_1$  is not larger than the variance of  $\mathbf{b}_{1.2}$  (since its inverse is at least as large). It follows that although  $\mathbf{b}_1$  is biased, its variance is never larger than the variance of the unbiased estimator. In any realistic case (i.e., if  $\mathbf{X}'_1\mathbf{X}_2$  is not zero), in fact it will be smaller. We get a useful comparison from a simple regression with two variables measured as deviations from their means. Then,  $\text{Var}[\mathbf{b}_1|\mathbf{X}] = \sigma^2/S_{11}$  where  $S_{11} = \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2$  and  $\text{Var}[\mathbf{b}_{1.2}|\mathbf{X}] = \sigma^2/[S_{11}(1 - r_{12}^2)]$  where  $r_{12}^2$  is the squared correlation between  $x_1$  and  $x_2$ .

The result in the preceding paragraph poses a bit of a dilemma for applied researchers. The situation arises frequently in the search for a model specification. Faced with a variable that a researcher suspects should be in the model, but that is causing a problem of multicollinearity, the analyst faces a choice of omitting the relevant variable or including it and estimating its (and all the other variables') coefficient imprecisely. This presents a choice between two estimators,  $b_1$  and  $b_{1.2}$ . In fact, what researchers usually do actually creates a third estimator. It is common to include the problem variable provisionally. If its  $t$  ratio is sufficiently large, it is retained; otherwise it is discarded. This third estimator is called a **pretest estimator**. What is known about pretest estimators is not encouraging. Certainly they are biased. How badly depends on the unknown parameters. Analytical results suggest that the pretest estimator is the least precise of the three when the researcher is most likely to use it. [See Judge et al. (1985).] The conclusion to be drawn is that as a general rule, the methodology leans away from estimation strategies that include ad hoc remedies for multicollinearity.

### 4.7.3 PRINCIPAL COMPONENTS

A device that has been suggested for “reducing” multicollinearity [see, e.g., Gurmu, Rilstone, and Stern (1999)] is to use a small number, say  $L$ , of **principal components**

## CHAPTER 4 ♦ The Least Squares Estimator 93

constructed as linear combinations of the  $K$  original variables. [See Johnson and Wichern (2005, Chapter 8).] (The mechanics are illustrated in Example 4.12.) The argument against using this approach is that if the original specification in the form  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$  were correct, then it is unclear what one is estimating when one regresses  $\mathbf{y}$  on some small set of linear combinations of the columns of  $\mathbf{X}$ . For a set of  $L < K$  principal components, if we regress  $\mathbf{y}$  on  $\mathbf{Z} = \mathbf{X}\mathbf{C}_L$  to obtain  $\mathbf{d}$ , it follows that  $E[\mathbf{d}] = \boldsymbol{\delta} = \mathbf{C}_L'\boldsymbol{\beta}$ . (The proof is considered in the exercises.) In an economic context, if  $\boldsymbol{\beta}$  has an interpretation, then it is unlikely that  $\boldsymbol{\delta}$  will. (E.g., how do we interpret the price elasticity minus twice the income elasticity?)

This orthodox interpretation cautions the analyst about mechanical devices for coping with multicollinearity that produce uninterpretable mixtures of the coefficients. But, there are also situations in which the model is built on a platform that might well involve a mixture of some measured variables. For example, one might be interested in a regression model that contains “ability,” ambiguously defined. As a measured counterpart, the analyst might have in hand standardized scores on a set of tests, none of which individually has any particular meaning in the context of the model. In this case, a mixture of the measured test scores might serve as one’s preferred proxy for the underlying variable. The study in Example 4.12 describes another natural example.

**Example 4.12 Predicting Movie Success**

Predicting the box office success of movies is a favorite exercise for econometricians. [See, e.g., Litman (1983), Ravid (1999), De Vany (2003), De Vany and Walls (1999, 2002, 2003), and Simonoff and Sparrow (2000).] The traditional predicting equation takes the form

$$\text{Box Office Receipts} = f(\text{Budget, Genre, MPAA Rating, Star Power, Sequel, etc.}) + \boldsymbol{\varepsilon}.$$

Coefficients of determination on the order of .4 are fairly common. Notwithstanding the relative power of such models, the common wisdom in Hollywood is “nobody knows.” There is tremendous randomness in movie success, and few really believe they can forecast it with any reliability.<sup>15</sup> Versaci (2009) added a new element to the model, “Internet buzz.” Internet buzz is vaguely defined to be Internet traffic and interest on familiar web sites such as RottenTomatoes.com, ImDB.com, Fandango.com, and traileraddict.com. None of these by itself defines Internet buzz. But, collectively, activity on these web sites, say three weeks before a movie’s opening, might be a useful predictor of upcoming success. Versaci’s data set (Table F4.3) contains data for 62 movies released in 2009, including four Internet buzz variables, all measured three weeks prior to the release of the movie:

- $buzz_1$  = number of Internet views of movie trailer at traileraddict.com
- $buzz_2$  = number of message board comments about the movie at ComingSoon.net
- $buzz_3$  = total number of “can’t wait” (for release) plus “don’t care” votes at Fandango.com
- $buzz_4$  = percentage of Fandango votes that are “can’t wait”

We have aggregated these into a single principal component as follows. We first computed the logs of  $buzz_1 - buzz_3$  to remove the scale effects. We then standardized the four variables, so  $z_k$  contains the original variable minus its mean,  $\bar{z}_k$ , then divided by its standard deviation,  $s_k$ . Let  $\mathbf{Z}$  denote the resulting  $62 \times 4$  matrix  $(z_1, z_2, z_3, z_4)$ . Then  $\mathbf{V} = (1/61)\mathbf{Z}'\mathbf{Z}$  is the sample correlation matrix. Let  $\mathbf{c}_1$  be the characteristic vector of  $\mathbf{V}$

<sup>15</sup>The assertion that “nobody knows” will be tested on a newly formed (April 2010) futures exchange where investors can place early bets on movie success (and producers can hedge their own bets). See <http://www.cantorexchange.com/> for discussion. The real money exchange was created by Cantor Fitzgerald, Inc. after they purchased the popular culture web site *Hollywood Stock Exchange*.

## 94 PART I ♦ The Linear Regression Model

**TABLE 4.8** Regression Results for Movie Success

Variable	Internet Buzz Model			Traditional Model		
	Coefficient	Std.Error	t	Coefficient	Std.Error	t
				22.30215	.58883	35.66514
Constant	15.4002	.64273	23.96	13.5768	.68825	19.73
ACTION	-.86932	.29333	-2.96	-.30682	.34401	-.89
COMEDY	-.01622	.25608	-.06	-.03845	.32061	-.12
ANIMATED	-.83324	.43022	-1.94	-.82032	.53869	-1.52
HORROR	.37460	.37109	1.01	1.02644	.44008	2.33
G	.38440	.55315	.69	.25242	.69196	.36
PG	.53359	.29976	1.78	.32970	.37243	.89
PG13	.21505	.21885	.98	.07176	.27206	.26
LOGBUDGT	.26088	.18529	1.41	.70914	.20812	3.41
SEQUEL	.27505	.27313	1.01	.64368	.33143	1.94
STARPOWR	.00433	.01285	.34	.00648	.01608	.40
BUZZ	.42906	.07839	5.47			

associated with the largest characteristic root. The first principal component (the one that explains most of the variation of the four variables) is  $\mathbf{Z}\mathbf{c}_1$ . (The roots are 2.4142, 0.7742, 0.4522, 0.3585 so the first principal component explains  $2.4142/4$  or 60.3 percent of the variation. Table 4.8 shows the regression results for the sample of 62 2009 movies. It appears that Internet buzz adds substantially to the predictive power of the regression. The  $R^2$  of the regression nearly doubles, from .34 to .58 when Internet buzz is added to the model. As we will discuss in Chapter 5, buzz is also a highly “significant” predictor of success.

### 4.7.4 MISSING VALUES AND DATA IMPUTATION

It is common for data sets to have gaps, for a variety of reasons. Perhaps the most frequent occurrence of this problem is in survey data, in which respondents may simply fail to respond to the questions. In a time series, the data may be missing because they do not exist at the frequency we wish to observe them; for example, the model may specify monthly relationships, but some variables are observed only quarterly. In panel data sets, the gaps in the data may arise because of **attrition** from the study. This is particularly common in health and medical research, when individuals choose to leave the study—possibly because of the success or failure of the treatment that is being studied.

There are several possible cases to consider, depending on why the data are missing. The data may be simply unavailable, for reasons unknown to the analyst and unrelated to the completeness or the values of the other observations in the sample. This is the most benign situation. If this is the case, then the complete observations in the sample constitute a usable data set, and the only issue is what possibly helpful information could be salvaged from the incomplete observations. Griliches (1986) calls this the **ignorable case** in that, for purposes of estimation, if we are not concerned with efficiency, then we may simply delete the incomplete observations and ignore the problem. Rubin (1976, 1987) and Little and Rubin (1987, 2002) label this case **missing completely at random**, or MCAR. A second case, which has attracted a great deal of attention in

## CHAPTER 4 ♦ The Least Squares Estimator 95

the econometrics literature, is that in which the gaps in the data set are not benign but are systematically related to the phenomenon being modeled. This case happens most often in surveys when the data are “self-selected” or “self-reported.”<sup>16</sup> For example, if a survey were designed to study expenditure patterns and if high-income individuals tended to withhold information about their income, then the gaps in the data set would represent more than just missing information. The clinical trial case is another instance. In this (worst) case, the complete observations would be qualitatively different from a sample taken at random from the full population. The missing data in this situation are termed **not missing at random**, or NMAR. We treat this second case in Chapter 18 with the subject of **sample selection**, so we shall defer our discussion until later.

The intermediate case is that in which there is information about the missing data contained in the complete observations that can be used to improve inference about the model. The incomplete observations in this **missing at random** (MAR) case are also ignorable, in the sense that unlike the NMAR case, simply using the complete data does not induce any biases in the analysis, as long as the underlying process that produces the missingness in the data does not share parameters with the model that is being estimated, which seems likely. [See Allison (2002).] This case is unlikely, of course, if “missingness” is based on the values of the dependent variable in a regression. Ignoring the incomplete observations when they are MAR but not MCAR, does ignore information that is in the sample and therefore sacrifices some efficiency. Researchers have used a variety of **data imputation** methods to fill gaps in data sets. The (by far) simplest case occurs when the gaps occur in the data on the regressors. For the case of missing data on the regressors, it helps to consider the simple regression and multiple regression cases separately. In the first case,  $\mathbf{X}$  has two columns: the column of 1s for the constant and a column with some blanks where the missing data would be if we had them. The **zero-order method** of replacing each missing  $x$  with  $\bar{x}$  based on the observed data results in no changes and is equivalent to dropping the incomplete data. (See Exercise 7 in Chapter 3.) However, the  $R^2$  will be lower. An alternative, **modified zero-order regression** fills the second column of  $\mathbf{X}$  with zeros and adds a variable that takes the value one for missing observations and zero for complete ones.<sup>17</sup> We leave this as an exercise to show that this is algebraically identical to simply filling the gaps with  $\bar{x}$ . There is also the possibility of computing fitted values for the missing  $x$ 's by a regression of  $x$  on  $y$  in the complete data. The sampling properties of the resulting estimator are largely unknown, but what evidence there is suggests that this is not a beneficial way to proceed.<sup>18</sup>

These same methods can be used when there are multiple regressors. Once again, it is tempting to replace missing values of  $\mathbf{x}_k$  with simple means of complete observations or with the predictions from linear regressions based on other variables in the model for which data are available when  $\mathbf{x}_k$  is missing. In most cases in this setting, a general characterization can be based on the principle that for any missing observation, the

<sup>16</sup>The vast surveys of Americans' opinions about sex by Ann Landers (1984, *passim*) and Shere Hite (1987) constitute two celebrated studies that were surely tainted by a heavy dose of self-selection bias. The latter was pilloried in numerous publications for purporting to represent the population at large instead of the opinions of those strongly enough inclined to respond to the survey. The former was presented with much greater modesty.

<sup>17</sup>See Maddala (1977a, p. 202).

<sup>18</sup>Afifi and Elashoff (1966, 1967) and Haitovsky (1968). Griliches (1986) considers a number of other possibilities.

## 96 PART I ♦ The Linear Regression Model

“true” unobserved  $x_{ik}$  is being replaced by an erroneous proxy that we might view as  $\hat{x}_{ik} = x_{ik} + u_{ik}$ , that is, in the framework of **measurement error**. Generally, the least squares estimator is biased (and inconsistent) in the presence of measurement error such as this. (We will explore the issue in Chapter 8.) A question does remain: Is the bias likely to be reasonably small? As intuition should suggest, it depends on two features of the data: (a) how good the prediction of  $x_{ik}$  is in the sense of how large the variance of the measurement error,  $u_{ik}$ , is compared to that of the actual data,  $x_{ik}$ , and (b) how large a proportion of the sample the analyst is filling.

The regression method replaces each missing value on an  $\mathbf{x}_k$  with a single prediction from a linear regression of  $\mathbf{x}_k$  on other exogenous variables—in essence, replacing the missing  $x_{ik}$  with an estimate of it based on the regression model. In a Bayesian setting, some applications that involve unobservable variables (such as our example for a binary choice model in Chapter 17) use a technique called **data augmentation** to treat the unobserved data as unknown “parameters” to be estimated with the structural parameters, such as  $\beta$  in our regression model. Building on this logic researchers, for example, Rubin (1987) and Allison (2002) have suggested taking a similar approach in classical estimation settings. The technique involves a data imputation step that is similar to what was suggested earlier, but with an extension that recognizes the variability in the estimation of the regression model used to compute the predictions. To illustrate, we consider the case in which the independent variable,  $\mathbf{x}_k$  is drawn in principle from a normal population, so it is a continuously distributed variable with a mean, a variance, and a joint distribution with other variables in the model. Formally, an imputation step would involve the following calculations:

1. Using as much information (complete data) as the sample will provide, linearly regress  $\mathbf{x}_k$  on other variables in the model (and/or outside it, if other information is available),  $\mathbf{Z}_k$ , and obtain the coefficient vector  $\mathbf{d}_k$  with associated asymptotic covariance matrix  $\mathbf{A}_k$  and estimated disturbance variance  $s_k^2$ .
2. For purposes of the imputation, we draw an observation from the estimated asymptotic normal distribution of  $\mathbf{d}_k$ , that is  $\mathbf{d}_{k,m} = \mathbf{d}_k + \mathbf{v}_k$  where  $\mathbf{v}_k$  is a vector of random draws from the normal distribution with mean zero and covariance matrix  $\mathbf{A}_k$ .
3. For each missing observation in  $\mathbf{x}_k$  that we wish to impute, we compute,  $x_{i,k,m} = \mathbf{d}'_{k,m} \mathbf{z}_{i,k} + s_{k,m} u_{i,k}$  where  $s_{k,m}$  is  $s_k$  divided by a random draw from the chi-squared distribution with degrees of freedom equal to the number of degrees of freedom in the imputation regression.

At this point, the iteration is the same as considered earlier, where the missing values are imputed using a regression, albeit, a much more elaborate procedure. The regression is then computed using the complete data and the imputed data for the missing observations, to produce coefficient vector  $\mathbf{b}_m$  and estimated covariance matrix,  $\mathbf{V}_m$ . This constitutes a single round. The technique of **multiple imputation** involves repeating this set of steps  $M$  times. The estimators of the parameter vector and the appropriate asymptotic covariance matrix are

$$\hat{\beta} = \bar{\mathbf{b}} = \frac{1}{M} \sum_{m=1}^M \mathbf{b}_m,$$

$$\hat{\mathbf{V}} = \bar{\mathbf{V}} + \mathbf{B} = \frac{1}{M} \sum_{m=1}^M \mathbf{V}_m + \left(1 + \frac{1}{M}\right) \left(\frac{1}{M-1}\right) \sum_{m=1}^M (\mathbf{b}_m - \bar{\mathbf{b}}) (\mathbf{b}_m - \bar{\mathbf{b}})'.$$

## CHAPTER 4 ♦ The Least Squares Estimator 97

Researchers differ on the effectiveness or appropriateness of multiple imputation. When all is said and done, the measurement error in the imputed values remains. It takes very strong assumptions to establish that the multiplicity of iterations will suffice to average away the effect of this error. Very elaborate techniques have been developed for the special case of joint normally distributed cross sections of regressors such as those suggested above. However, the typical application to survey data involves gaps due to nonresponse to qualitative questions with binary answers. The efficacy of the theory is much less well developed for imputation of binary, ordered, count or other qualitative variables.

The more manageable case is missing values of the dependent variable,  $y_i$ . Once again, it must be the case that  $y_i$  is at least MAR and that the mechanism that is determining presence in the sample does not share parameters with the model itself. Assuming the data on  $\mathbf{x}_i$  are complete for all observations, one might consider filling the gaps in the data on  $y_i$  by a two-step procedure: (1) estimate  $\beta$  with  $\mathbf{b}_c$  using the complete observations,  $\mathbf{X}_c$  and  $\mathbf{y}_c$ , then (2) fill the missing values,  $\mathbf{y}_m$ , with predictions,  $\hat{\mathbf{y}}_m = \mathbf{X}_m \mathbf{b}_c$ , and recompute the coefficients. We leave as an exercise (Exercise 17) to show that the second step estimator is exactly equal to the first. However, the variance estimator at the second step,  $s^2$ , must underestimate  $\sigma^2$ , intuitively because we are adding to the sample a set of observations that are fit perfectly. [See Cameron and Trivedi (2005, Chapter 27).] So, this is not a beneficial way to proceed. The flaw in the method comes back to the device used to impute the missing values for  $y_i$ . Recent suggestions that appear to provide some improvement involve using a randomized version,  $\hat{\mathbf{y}}_m = \mathbf{X}_m \mathbf{b}_c + \hat{\boldsymbol{\epsilon}}_m$ , where  $\hat{\boldsymbol{\epsilon}}_m$  are random draws from the (normal) population with zero mean and estimated variance  $s^2[\mathbf{I} + \mathbf{X}_m(\mathbf{X}'_c \mathbf{X}_c)^{-1} \mathbf{X}'_m]$ . (The estimated variance matrix corresponds to  $\mathbf{X}_m \mathbf{b}_c + \boldsymbol{\epsilon}_m$ .) This defines an iteration. After reestimating  $\beta$  with the augmented data, one can return to re-impute the augmented data with the new  $\hat{\beta}$ , then recompute  $\mathbf{b}$ , and so on. The process would continue until the estimated parameter vector stops changing. (A subtle point to be noted here: The same random draws should be used in each iteration. If not, there is no assurance that the iterations would ever converge.)

In general, not much is known about the properties of estimators based on using predicted values to fill missing values of  $y$ . Those results we do have are largely from simulation studies based on a particular data set or pattern of missing data. The results of these Monte Carlo studies are usually difficult to generalize. The overall conclusion seems to be that in a single-equation regression context, filling in missing values of  $y$  leads to biases in the estimator which are difficult to quantify. The only reasonably clear result is that imputations are more likely to be beneficial if the proportion of observations that are being filled is small—the smaller the better.

### 4.7.5 MEASUREMENT ERROR

There are any number of cases in which observed data are imperfect measures of their theoretical counterparts in the regression model. Examples include income, education, ability, health, “the interest rate,” output, capital, and so on. Mismeasurement of the variables in a model will generally produce adverse consequences for least squares estimation. Remedies are complicated and sometimes require heroic assumptions. In this section, we will provide a brief sketch of the issues. We defer to Section 8.5 a more

## 98 PART I ♦ The Linear Regression Model

detailed discussion of the problem of measurement error, the most common solution (instrumental variables estimation), and some applications.

It is convenient to distinguish between measurement error in the dependent variable and measurement error in the regressor(s). For the second case, it is also useful to consider the simple regression case and then extend it to the multiple regression model. Consider a model to describe expected income in a population,

$$I^* = \mathbf{x}'\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (4-55)$$

where  $I^*$  is the intended total income variable. Suppose the observed counterpart is  $I$ , earnings. How  $I$  relates to  $I^*$  is unclear; it is common to assume that the measurement error is additive, so  $I = I^* + w$ . Inserting the expression for  $I$  into (4-55) gives

$$\begin{aligned} I &= \mathbf{x}'\boldsymbol{\beta} + \boldsymbol{\varepsilon} + w \\ &= \mathbf{x}'\boldsymbol{\beta} + v, \end{aligned} \quad (4-56)$$

which appears to be a slightly more complicated regression, but otherwise similar to what we started with. As long as  $w$  and  $\mathbf{x}$  are uncorrelated, that is the case. If  $w$  is a homoscedastic zero mean error that is uncorrelated with  $\mathbf{x}$ , then the only difference between (4-55) and (4-56) is that the disturbance variance in (4-56) is  $\sigma_w^2 + \sigma_v^2 > \sigma_\varepsilon^2$ . Otherwise both are regressions and, evidently  $\boldsymbol{\beta}$  can be estimated consistently by least squares in either case. The cost of the measurement error is in the precision of the estimator, since the asymptotic variance of the estimator in (4-56) is  $(\sigma_v^2/n)[\text{plim}(\mathbf{X}'\mathbf{X}/n)]^{-1}$  while it is  $(\sigma_\varepsilon^2/n)[\text{plim}(\mathbf{X}'\mathbf{X}/n)]^{-1}$  if  $\boldsymbol{\beta}$  is estimated using (4-55). The measurement error also costs some fit. To see this, note that the  $R^2$  in the sample regression in (4-55) is

$$R_*^2 = 1 - (\mathbf{e}'\mathbf{e}/n)/(\mathbf{I}'\mathbf{M}^0\mathbf{I}^*/n).$$

The numerator converges to  $\sigma_\varepsilon^2$  while the denominator converges to the total variance of  $I^*$ , which would approach  $\sigma_\varepsilon^2 + \boldsymbol{\beta}'\mathbf{Q}\boldsymbol{\beta}$  where  $\mathbf{Q} = \text{plim}(\mathbf{X}'\mathbf{X}/n)$ . Therefore,

$$\text{plim} R_*^2 = \boldsymbol{\beta}'\mathbf{Q}\boldsymbol{\beta}/[\sigma_\varepsilon^2 + \boldsymbol{\beta}'\mathbf{Q}\boldsymbol{\beta}].$$

The counterpart for (4-56),  $R^2$ , differs only in that  $\sigma_\varepsilon^2$  is replaced by  $\sigma_v^2 > \sigma_\varepsilon^2$  in the denominator. It follows that

$$\text{plim} R_*^2 - \text{plim} R^2 > 0.$$

This implies that the fit of the regression in (4-56) will, at least broadly in expectation, be inferior to that in (4-55). (The preceding is an asymptotic approximation that might not hold in every finite sample.)

These results demonstrate the implications of measurement error in the dependent variable. We note, in passing, that if the measurement error is not additive, if it is correlated with  $\mathbf{x}$ , or if it has any other features such as heteroscedasticity, then the preceding results are lost, and nothing in general can be said about the consequence of the measurement error. Whether there is a “solution” is likewise an ambiguous question. The preceding explanation shows that it would be better to have the underlying variable if possible. In the absence, would it be preferable to use a proxy? Unfortunately,  $I$  is already a proxy, so unless there exists an available  $I'$  which has smaller measurement error variance, we have reached an impasse. On the other hand, it does seem that the outcome is fairly benign. The sample does not contain as much

## CHAPTER 4 ♦ The Least Squares Estimator 99

information as we might hope, but it does contain sufficient information consistently to estimate  $\beta$  and to do appropriate statistical inference based on the information we do have.

The more difficult case occurs when the measurement error appears in the independent variable(s). For simplicity, we retain the symbols  $I$  and  $I^*$  for our observed and theoretical variables. Consider a simple regression,

$$y = \beta_1 + \beta_2 I^* + \varepsilon,$$

where  $y$  is the perfectly measured dependent variable and the same measurement equation,  $I = I^* + w$  applies now to the independent variable. Inserting  $I$  into the equation and rearranging a bit, we obtain

$$\begin{aligned} y &= \beta_1 + \beta_2 I + (\varepsilon - \beta_2 w) \\ &= \beta_1 + \beta_2 I + v. \end{aligned} \tag{4-57}$$

It appears that we have obtained (4-56) once again. Unfortunately, this is not the case, because  $\text{Cov}[I, v] = \text{Cov}[I^* + w, \varepsilon - \beta_2 w] = -\beta_2 \sigma_w^2$ . Since the regressor in (4-57) is correlated with the disturbance, least squares regression in this case is inconsistent. There is a bit more that can be derived—this is pursued in Section 8.5, so we state it here without proof. In this case,

$$\text{plim } b_2 = \beta_2 [\sigma_*^2 / (\sigma_*^2 + \sigma_w^2)]$$

where  $\sigma_*^2$  is the marginal variance of  $I^*$ . The scale factor is less than one, so the least squares estimator is biased toward zero. The larger is the measurement error variance, the worse is the bias. (This is called **least squares attenuation**.) Now, suppose there are additional variables in the model;

$$y = \mathbf{x}' \boldsymbol{\beta} + \beta_2 T + \varepsilon.$$

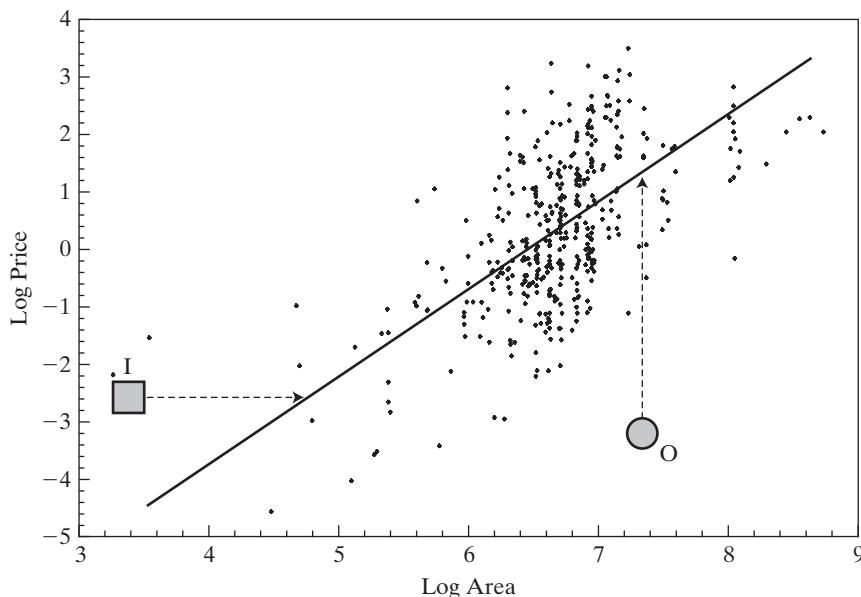
In this instance, almost no useful theoretical results are forthcoming. The following fairly general conclusions can be drawn—once again, proofs are deferred to Section 8.5:

1. The least squares estimator of  $\beta_2$  is still biased toward zero.
2. All the elements of the estimator of  $\boldsymbol{\beta}_1$  are biased, in unknown directions, even though the variables in  $\mathbf{x}$  are not measured with error.

Solutions to the “measurement error problem” come in two forms. If there is outside information on certain model parameters, then it is possible to deduce the scale factors (using the **method of moments**) and undo the bias. For the obvious example, in (4-57), if  $\sigma_w^2$  were known, then it would be possible to deduce  $\sigma_*^2$  from  $\text{Var}[I] = \sigma_*^2 + \sigma_w^2$  and thereby compute the necessary scale factor to undo the bias. This sort of information is generally not available. A second approach that has been used in many applications is the technique of instrumental variables. This is developed in detail for this setting in Section 8.5.

#### 4.7.6 OUTLIERS AND INFLUENTIAL OBSERVATIONS

Figure 4.9 shows a scatter plot of the data on sale prices of Monet paintings that were used in Example 4.10. Two points have been highlighted. The one marked “I” and noted with the square overlay shows the smallest painting in the data set. The circle marked

**100 PART I ♦ The Linear Regression Model**


**FIGURE 4.9** Log Price vs. Log Area for Monet Paintings.

“O” highlights a painting that fetched an unusually low price, at least in comparison to what the regression would have predicted. (It was not the least costly painting in the sample, but it was the one most poorly predicted by the regression.) Since least squares is based on squared deviations, the estimator is likely to be strongly influenced by extreme observations such as these, particularly if the sample is not very large.

An “influential observation” is one that is likely to have a substantial impact on the least squares regression coefficient(s). For a simple regression such as the one shown in Figure 4.9, Belsley, Kuh and Welsh (1980) defined an influence measure, for observation  $i$ ,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x}_n)^2}{\sum_{j=1}^n (x_j - \bar{x}_n)^2} \quad (4-58)$$

where  $\bar{x}_n$  and the summation in the denominator of the fraction are computed without this observation. (The measure derives from the difference between  $\mathbf{b}$  and  $\mathbf{b}_{(i)}$  where the latter is computed without the particular observation. We will return to this shortly.) It is suggested that an observation should be noted as influential if  $h_i > 2/n$ . The decision is whether to drop the observation or not. We should note, observations with high “leverage” are arguably not “outliers” (which remains to be defined), because the analysis is conditional on  $x_i$ . To underscore the point, referring to Figure 4.9, this observation would be marked even if it fell precisely on the regression line—the source of the influence is the numerator of the second term in  $h_i$ , which is unrelated to the distance of the point from the line. In our example, the “influential observation” happens to be the result of Monet’s decision to paint a small painting. The point is that in the absence of an underlying theory that explains (and justifies) the extreme values of  $x_i$ , eliminating

## CHAPTER 4 ♦ The Least Squares Estimator 101

such observations is an algebraic exercise that has the effect of forcing the regression line to be fitted with the values of  $x_i$  closest to the means.

The change in the linear regression coefficient vector in a multiple regression when an observation is added to the sample is

$$\mathbf{b} - \mathbf{b}_{(i)} = \Delta\mathbf{b} = \frac{1}{1 + \mathbf{x}'_i (\mathbf{X}'_{(i)} \mathbf{X}_{(i)})^{-1} \mathbf{x}_i} (\mathbf{X}'_{(i)} \mathbf{X}_{(i)})^{-1} \mathbf{x}_i (\mathbf{y}_i - \mathbf{x}'_i \mathbf{b}_{(i)}) \quad (4-59)$$

where  $\mathbf{b}$  is computed with observation  $i$  in the sample,  $\mathbf{b}_{(i)}$  is computed without observation  $i$  and  $\mathbf{X}_{(i)}$  does not include observation  $i$ . (See Exercise 6 in Chapter 3.) It is difficult to single out any particular feature of the observation that would drive this change. The influence measure,

$$\begin{aligned} h_{ii} &= \mathbf{x}'_i (\mathbf{X}'_{(i)} \mathbf{X}_{(i)})^{-1} \mathbf{x}_i \\ &= \frac{1}{n} + \sum_{j=1}^{K-1} \sum_{k=1}^{K-1} (x_{i,j} - \bar{x}_{n,j}) (x_{i,k} - \bar{x}_k) (\mathbf{Z}'_{(i)} \mathbf{M}^0 \mathbf{Z}_{(i)})^{jk}, \end{aligned} \quad (4-60)$$

has been used to flag influential observations. [See, once again, Belsley, Kuh and Welsh (1980) and Cook (1977).] In this instance, the selection criterion would be  $h_{ii} > 2(K-1)/n$ . Squared deviations of the elements of  $\mathbf{x}_i$  from the means of the variables appear in  $h_{ii}$ , so it is also operating on the difference of  $\mathbf{x}_i$  from the center of the data. (See the expression for the forecast variance in Section 4.6.1 for an application.)

In principle, an “outlier,” is an observation that appears to be outside the reach of the model, perhaps because it arises from a different data generating process. Point “O” in Figure 4.9 appears to be a candidate. Outliers could arise for several reasons. The simplest explanation would be actual data errors. Assuming the data are not erroneous, it then remains to define what constitutes an outlier. Unusual residuals are an obvious choice. But, since the distribution of the disturbances would anticipate a certain small percentage of extreme observations in any event, simply singling out observations with large residuals is actually a dubious exercise. On the other hand, one might suspect that the outlying observations are actually generated by a different population. “Studentized” residuals are constructed with this in mind by computing the regression coefficients and the residual variance without observation  $i$  for each observation in the sample and then standardizing the modified residuals. The  $i$ th studentized residual is

$$e(i) = \frac{e_i}{(1 - h_{ii})} \sqrt{\frac{\mathbf{e}' \mathbf{e} - e_i^2 / (1 - h_{ii})}{n - 1 - K}} \quad (4-61)$$

where  $\mathbf{e}$  is the residual vector for the full sample, based on  $\mathbf{b}$ , including  $e_i$  the residual for observation  $i$ . In principle, this residual has a  $t$  distribution with  $n - 1 - K$  degrees of freedom (or a standard normal distribution asymptotically). Observations with large studentized residuals, that is, greater than 2.0, would be singled out as outliers.

There are several complications that arise with isolating outlying observations in this fashion. First, there is no a priori assumption of which observations are from the alternative population, if this is the view. From a theoretical point of view, this would suggest a skepticism about the model specification. If the sample contains a substantial proportion of outliers, then the properties of the estimator based on the reduced sample are difficult to derive. In the following application, following, the procedure

**102 PART I ♦ The Linear Regression Model**
**TABLE 4.9** Estimated Equations for Log Price

Number of observations	430	410				
Mean of log Price	0.33274	.36043				
Sum of squared residuals	519.17235	383.17982				
Standard error of regression	1.10266	0.97030				
R-squared	0.33620	0.39170				
Adjusted R-squared	0.33309	0.38871				
	<i>Coefficient</i>	<i>Standard Error</i>	<i>t</i>			
Variable	<i>n</i> = 430	<i>n</i> = 410	<i>n</i> = 430	<i>n</i> = 410		
Constant	-8.42653	-8.67356	.61183	.57529	-13.77	-15.08
LOGAREA	1.33372	1.36982	.09072	.08472	14.70	16.17
ASPECT	-0.16537	-0.14383	.12753	.11412	-1.30	-1.26

deletes 4.7 percent of the sample (20 observations). Finally, it will usually occur that observations that were not outliers in the original sample will become “outliers” when the original set of outliers is removed. It is unclear how one should proceed at this point. (Using the Monet paintings data, the first round of studentizing the residuals removes 20 observations. After 16 iterations, the sample size stabilizes at 316 of the original 430 observations, a reduction of 26.5 percent.) Table 4.9 shows the original results (from Table 4.6) and the modified results with 20 outliers removed. Since 430 is a relatively large sample, the modest change in the results is to be expected.

It is difficult to draw a firm general conclusions from this exercise. It remains likely that in very small samples, some caution and close scrutiny of the data are called for. If it is suspected at the outset that a process prone to large observations is at work, it may be useful to consider a different estimator altogether, such as least absolute deviations, or even a different model specification that accounts for this possibility. For example, the idea that the sample may contain some observations that are generated by a different process lies behind the latent class model that is discussed in Chapters 14 and 18.

## 4.8 SUMMARY AND CONCLUSIONS

This chapter has examined a set of properties of the least squares estimator that will apply in all samples, including unbiasedness and efficiency among unbiased estimators. The formal assumptions of the linear model are pivotal in the results of this chapter. All of them are likely to be violated in more general settings than the one considered here. For example, in most cases examined later in the book, the estimator has a possible bias, but that bias diminishes with increasing sample sizes. For purposes of forming confidence intervals and testing hypotheses, the assumption of normality is narrow, so it was necessary to extend the model to allow nonnormal disturbances. These and other “large-sample” extensions of the linear model were considered in Section 4.4. The crucial results developed here were the consistency of the estimator and a method of obtaining an appropriate covariance matrix and large-sample distribution that provides the basis for forming confidence intervals and testing hypotheses. Statistical inference in

CHAPTER 4 ♦ The Least Squares Estimator **103**

the form of interval estimation for the model parameters and for values of the dependent variable was considered in Sections 4.5 and 4.6. This development will continue in Chapter 5 where we will consider hypothesis testing and model selection.

Finally, we considered some practical problems that arise when data are less than perfect for the estimation and analysis of the regression model, including multicollinearity, missing observations, measurement error, and outliers.

**Key Terms and Concepts**

- Assumptions
- Asymptotic covariance matrix
- Asymptotic distribution
- Asymptotic efficiency
- Asymptotic normality
- Asymptotic properties
- Attrition
- Bootstrap
- condition number
- Confidence interval
- Consistency
- Consistent estimator
- Data imputation
- Effect size
- Ergodic
- Estimator
- Ex ante forecast
- Ex post forecast
- Finite sample properties
- Gauss–Markov theorem
- Grenander conditions
- Highest posterior density interval
- Identification
- Ignorable case
- Inclusion of superfluous (irrelevant) variables
- Indicator
- Interval estimation
- Least squares attenuation
- Lindeberg–Feller central limit theorem
- Linear estimator
- Linear unbiased estimator
- Maximum likelihood estimator
- Mean absolute error
- Mean square convergence
- Mean squared error
- Measurement error
- Method of moments
- Minimum mean squared error
- Minimum variance linear unbiased estimator
- Missing at random
- Missing completely at random
- Missing observations
- Modified zero-order regression
- Monte Carlo study
- Multicollinearity
- Not missing at random
- Oaxaca's and Blinder's decomposition
- Omission of relevant variables
- Optimal linear predictor
- Orthogonal random variables
- Panel data
- Pivotal statistic
- Point estimation
- Prediction error
- Prediction interval
- Prediction variance
- Pretest estimator
- Principal components
- Probability limit
- Root mean squared error
- Sample selection
- Sampling distribution
- Sampling variance
- Semiparametric
- Smearing estimator
- Specification errors
- Standard error
- Standard error of the regression
- Stationary process
- Statistical properties
- Stochastic regressors
- Theil  $U$  statistic
- $t$  ratio
- Variance inflation factor
- Zero-order method

**Exercises**

1. Suppose that you have two independent unbiased estimators of the same parameter  $\theta$ , say  $\hat{\theta}_1$  and  $\hat{\theta}_2$ , with different variances  $v_1$  and  $v_2$ . What linear combination  $\hat{\theta} = c_1\hat{\theta}_1 + c_2\hat{\theta}_2$  is the minimum variance unbiased estimator of  $\theta$ ?
2. Consider the simple regression  $y_i = \beta x_i + \varepsilon_i$  where  $E[\varepsilon | x] = 0$  and  $E[\varepsilon^2 | x] = \sigma^2$ 
  - a. What is the minimum mean squared error linear estimator of  $\beta$ ? [Hint: Let the estimator be  $(\hat{\beta} = \mathbf{c}'\mathbf{y})$ . Choose  $\mathbf{c}$  to minimize  $\text{Var}(\hat{\beta}) + (E(\hat{\beta} - \beta))^2$ . The answer is a function of the unknown parameters.]

**104 PART I ♦ The Linear Regression Model**

- b. For the estimator in part a, show that ratio of the mean squared error of  $\hat{\beta}$  to that of the ordinary least squares estimator  $b$  is

$$\frac{\text{MSE}[\hat{\beta}]}{\text{MSE}[b]} = \frac{\tau^2}{(1 + \tau^2)}, \quad \text{where } \tau^2 = \frac{\beta^2}{[\sigma^2/\mathbf{x}'\mathbf{x}]}$$

Note that  $\tau$  is the square of the population analog to the “ $t$  ratio” for testing the hypothesis that  $\beta = 0$ , which is given in (4-14). How do you interpret the behavior of this ratio as  $\tau \rightarrow \infty$ ?

3. Suppose that the classical regression model applies but that the true value of the constant is zero. Compare the variance of the least squares slope estimator computed without a constant term with that of the estimator computed with an unnecessary constant term.
4. Suppose that the regression model is  $y_i = \alpha + \beta x_i + \varepsilon_i$ , where the disturbances  $\varepsilon_i$  have  $f(\varepsilon_i) = (1/\lambda) \exp(-\varepsilon_i/\lambda)$ ,  $\varepsilon_i \geq 0$ . This model is rather peculiar in that all the disturbances are assumed to be nonnegative. Note that the disturbances have  $E[\varepsilon_i | x_i] = \lambda$  and  $\text{Var}[\varepsilon_i | x_i] = \lambda^2$ . Show that the least squares slope is unbiased but that the intercept is biased.
5. Prove that the least squares intercept estimator in the classical regression model is the minimum variance linear unbiased estimator.
6. As a profit-maximizing monopolist, you face the demand curve  $Q = \alpha + \beta P + \varepsilon$ . In the past, you have set the following prices and sold the accompanying quantities:

<b><math>Q</math></b>	3	3	7	6	10	15	16	13	9	15	9	15	12	18	21
<b><math>P</math></b>	18	16	17	12	15	15	4	13	11	6	8	10	7	7	7

Suppose that your marginal cost is 10. Based on the least squares regression, compute a 95 percent confidence interval for the expected value of the profit-maximizing output.

7. The following sample moments for  $x = [1, x_1, x_2, x_3]$  were computed from 100 observations produced using a random number generator:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 100 & 123 & 96 & 109 \\ 123 & 252 & 125 & 189 \\ 96 & 125 & 167 & 146 \\ 109 & 189 & 146 & 168 \end{bmatrix}, \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 460 \\ 810 \\ 615 \\ 712 \end{bmatrix}, \quad \mathbf{y}'\mathbf{y} = 3924.$$

The true model underlying these data is  $y = x_1 + x_2 + x_3 + \varepsilon$ .

- a. Compute the simple correlations among the regressors.
- b. Compute the ordinary least squares coefficients in the regression of  $y$  on a constant  $x_1$ ,  $x_2$ , and  $x_3$ .
- c. Compute the ordinary least squares coefficients in the regression of  $y$  on a constant  $x_1$  and  $x_2$ , on a constant  $x_1$  and  $x_3$ , and on a constant  $x_2$  and  $x_3$ .
- d. Compute the variance inflation factor associated with each variable.
- e. The regressors are obviously collinear. Which is the problem variable?
8. Consider the multiple regression of  $y$  on  $K$  variables  $\mathbf{X}$  and an additional variable  $\mathbf{z}$ . Prove that under the assumptions A1 through A6 of the classical regression model, the true variance of the least squares estimator of the slopes on  $\mathbf{X}$  is larger when  $\mathbf{z}$

## CHAPTER 4 ♦ The Least Squares Estimator 105

is included in the regression than when it is not. Does the same hold for the sample estimate of this covariance matrix? Why or why not? Assume that  $\mathbf{X}$  and  $\mathbf{z}$  are nonstochastic and that the coefficient on  $\mathbf{z}$  is nonzero.

9. For the classical normal regression model  $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$  with no constant term and  $K$  regressors, assuming that the true value of  $\beta$  is zero, what is the exact expected value of  $F[K, n - K] = (R^2/K)/[(1 - R^2)/(n - K)]$ ?
10. Prove that  $E[\mathbf{b}'\mathbf{b}] = \beta'\beta + \sigma^2 \sum_{k=1}^K (1/\lambda_k)$  where  $\mathbf{b}$  is the ordinary least squares estimator and  $\lambda_k$  is a characteristic root of  $\mathbf{X}'\mathbf{X}$ .
11. For the classical normal regression model  $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$  with no constant term and  $K$  regressors, what is  $\text{plim } F[K, n - K] = \text{plim } \frac{R^2/K}{(1-R^2)/(n-K)}$ , assuming that the true value of  $\beta$  is zero?
12. Let  $e_i$  be the  $i$ th residual in the ordinary least squares regression of  $\mathbf{y}$  on  $\mathbf{X}$  in the classical regression model, and let  $\varepsilon_i$  be the corresponding true disturbance. Prove that  $\text{plim}(e_i - \varepsilon_i) = 0$ .
13. For the simple regression model  $y_i = \mu + \varepsilon_i$ ,  $\varepsilon_i \sim N[0, \sigma^2]$ , prove that the sample mean is consistent and asymptotically normally distributed. Now consider the alternative estimator  $\hat{\mu} = \sum_i w_i y_i$ ,  $w_i = \frac{i}{n(n+1)/2} = \frac{i}{\sum_i i}$ . Note that  $\sum_i w_i = 1$ . Prove that this is a consistent estimator of  $\mu$  and obtain its asymptotic variance. [Hint:  $\sum_i i^2 = n(n+1)(2n+1)/6$ .]
14. Consider a data set consisting of  $n$  observations,  $n_c$  complete and  $n_m$  incomplete for which the dependent variable,  $y_i$ , is missing. Data on the independent variables,  $\mathbf{x}_i$ , are complete for all  $n$  observations,  $\mathbf{X}_c$  and  $\mathbf{X}_m$ . We wish to use the data to estimate the parameters of the linear regression model  $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$ . Consider the following the imputation strategy: Step 1: Linearly regress  $\mathbf{y}_c$  on  $\mathbf{X}_c$  and compute  $\mathbf{b}_c$ . Step 2: Use  $\mathbf{X}_m$  to predict the missing  $\mathbf{y}_m$  with  $\mathbf{X}_m\mathbf{b}_c$ . Then regress the full sample of observations,  $(\mathbf{y}_c, \mathbf{X}_m\mathbf{b}_c)$ , on the full sample of regressors,  $(\mathbf{X}_c, \mathbf{X}_m)$ .
  - a. Show that the first and second step least squares coefficient vectors are identical.
  - b. Is the second step coefficient estimator unbiased?
  - c. Show that the sum of squared residuals is the same at both steps.
  - d. Show that the second step estimator of  $\sigma^2$  is biased downward.
15. In (4-13), we find that when superfluous variables  $\mathbf{X}_2$  are added to the regression of  $\mathbf{y}$  on  $\mathbf{X}_1$  the least squares coefficient estimator is an unbiased estimator of the true parameter vector,  $\beta = (\beta'_1, \mathbf{0}')'$ . Show that in this long regression,  $\mathbf{e}'\mathbf{e}/(n - K_1 - K_2)$  is also unbiased as estimator of  $\sigma^2$ .
16. In Section 4.7.3, we consider regressing  $\mathbf{y}$  on a set of principal components, rather than the original data. For simplicity, assume that  $\mathbf{X}$  does not contain a constant term, and that the  $K$  variables are measured in deviations from the means and are “standardized” by dividing by the respective standard deviations. We consider regression of  $\mathbf{y}$  on  $L$  principal components,  $\mathbf{Z} = \mathbf{X}\mathbf{C}_L$ , where  $L < K$ . Let  $\mathbf{d}$  denote the coefficient vector. The regression model is  $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$ . In the discussion, it is claimed that  $E[\mathbf{d}] = \mathbf{C}'_L \beta$ . Prove the claim.
17. Example 4.9 presents a regression model that is used to predict the auction prices of Monet paintings. The most expensive painting in the sample sold for \$33.0135M ( $\log = 17.3124$ ). The height and width of this painting were 35" and 39.4", respectively. Use these data and the model to form prediction intervals for the log of the price and then the price for this painting.

## 106 PART I ♦ The Linear Regression Model

### Applications

1. Data on U.S. gasoline consumption for the years 1953 to 2004 are given in Table F2.2. Note, the consumption data appear as total expenditure. To obtain the per capita quantity variable, divide GASEXP by GASP times Pop. The other variables do not need transformation.
  - a. Compute the multiple regression of per capita consumption of gasoline on per capita income, the price of gasoline, all the other prices and a time trend. Report all results. Do the signs of the estimates agree with your expectations?
  - b. Test the hypothesis that at least in regard to demand for gasoline, consumers do not differentiate between changes in the prices of new and used cars.
  - c. Estimate the own price elasticity of demand, the income elasticity, and the cross-price elasticity with respect to changes in the price of public transportation. Do the computations at the 2004 point in the data.
  - d. Reestimate the regression in logarithms so that the coefficients are direct estimates of the elasticities. (Do not use the log of the time trend.) How do your estimates compare with the results in the previous question? Which specification do you prefer?
  - e. Compute the simple correlations of the price variables. Would you conclude that multicollinearity is a “problem” for the regression in part a or part d?
  - f. Notice that the price index for gasoline is normalized to 100 in 2000, whereas the other price indices are anchored at 1983 (roughly). If you were to renormalize the indices so that they were all 100.00 in 2004, then how would the results of the regression in part a change? How would the results of the regression in part d change?
  - g. This exercise is based on the model that you estimated in part d. We are interested in investigating the change in the gasoline market that occurred in 1973. First, compute the average values of log of per capita gasoline consumption in the years 1953–1973 and 1974–2004 and report the values and the difference. If we divide the sample into these two groups of observations, then we can decompose the change in the expected value of the log of consumption into a change attributable to change in the regressors and a change attributable to a change in the model coefficients, as shown in Section 4.5.3. Using the Oaxaca–Blinder approach described there, compute the decomposition by partitioning the sample and computing separate regressions. Using your results, compute a confidence interval for the part of the change that can be attributed to structural change in the market, that is, change in the regression coefficients.
2. Christensen and Greene (1976) estimated a generalized Cobb–Douglas cost function for electricity generation of the form

$$\ln C = \alpha + \beta \ln Q + \gamma \left[ \frac{1}{2} (\ln Q)^2 \right] + \delta_k \ln P_k + \delta_l \ln P_l + \delta_f \ln P_f + \varepsilon.$$

$P_k$ ,  $P_l$ , and  $P_f$  indicate unit prices of capital, labor, and fuel, respectively,  $Q$  is output and  $C$  is total cost. To conform to the underlying theory of production, it is necessary to impose the restriction that the cost function be homogeneous of degree one in the three prices. This is done with the restriction  $\delta_k + \delta_l + \delta_f = 1$ , or  $\delta_f = 1 - \delta_k - \delta_l$ .

**CHAPTER 4 ♦ The Least Squares Estimator 107**

Inserting this result in the cost function and rearranging produces the estimating equation,

$$\ln(C/P_f) = \alpha + \beta \ln Q + \gamma \left[ \frac{1}{2} (\ln Q)^2 \right] + \delta_k \ln(P_k/P_f) + \delta_l \ln(P_l/P_f) + \varepsilon.$$

The purpose of the generalization was to produce a U-shaped average total cost curve. [See Example 6.6 for discussion of Nerlove's (1963) predecessor to this study.] We are interested in the **efficient scale**, which is the output at which the cost curve reaches its minimum. That is the point at which  $(\partial \ln C / \partial \ln Q)|_{Q=Q^*} = 1$  or  $Q^* = \exp[(1 - \beta)/\gamma]$ .

- a. Data on 158 firms extracted from Christensen and Greene's study are given in Table F4.4. Using all 158 observations, compute the estimates of the parameters in the cost function and the estimate of the asymptotic covariance matrix.
- b. Note that the cost function does not provide a direct estimate of  $\delta_f$ . Compute this estimate from your regression results, and estimate the asymptotic standard error.
- c. Compute an estimate of  $Q^*$  using your regression results and then form a confidence interval for the estimated efficient scale.
- d. Examine the raw data and determine where in the sample the efficient scale lies. That is, determine how many firms in the sample have reached this scale, and whether, in your opinion, this scale is large in relation to the sizes of firms in the sample. Christensen and Greene approached this question by computing the proportion of total output in the sample that was produced by firms that had not yet reached efficient scale. (*Note:* there is some double counting in the data set—more than 20 of the largest “firms” in the sample we are using for this exercise are holding companies and power pools that are aggregates of other firms in the sample. We will ignore that complication for the purpose of our numerical exercise.)

## 5

## HYPOTHESIS TESTS AND MODEL SELECTION

---

### 5.1 INTRODUCTION

The linear regression model is used for three major purposes: estimation and prediction, which were the subjects of the previous chapter, and hypothesis testing. In this chapter, we will examine some applications of hypothesis tests using the linear regression model. We begin with the methodological and statistical theory. Some of this theory was developed in Chapter 4 (including the idea of a pivotal statistic in Section 4.5.1) and in Appendix C.7. In Section 5.2, we will extend the methodology to hypothesis testing based on the regression model. After the theory is developed, Sections 5.3–5.7 will examine some applications in regression modeling. This development will be concerned with the implications of restrictions on the parameters of the model, such as whether a variable is ‘relevant’ (i.e., has a nonzero coefficient) or whether the regression model itself is supported by the data (i.e., whether the data seem consistent with the hypothesis that all of the coefficients are zero). We will primarily be concerned with linear restrictions in this discussion. We will turn to nonlinear restrictions near the end of the development in Section 5.7. Section 5.8 considers some broader types of hypotheses, such as choosing between two competing models, such as whether a linear or a loglinear model is better suited to the data. In each of the cases so far, the testing procedure attempts to resolve a competition between two theories for the data; in Sections 5.2–5.7 between a narrow model and a broader one and in Section 5.8, between two arguably equal models. Section 5.9 illustrates a particular **specification test**, which is essentially a test of a proposition such as “the model is correct” vs. “the model is inadequate.” This test pits the theory of the model against “some other unstated theory.” Finally, Section 5.10 presents some general principles and elements of a strategy of model testing and selection.

### 5.2 HYPOTHESIS TESTING METHODOLOGY

We begin the analysis with the regression model as a statement of a proposition,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \quad (5-1)$$

To consider a specific application, Example 4.6 depicted the auction prices of paintings

$$\ln Price = \beta_1 + \beta_2 \ln Size + \beta_3 AspectRatio + \varepsilon. \quad (5-2)$$

Some questions might be raised about the “model” in (5-2), fundamentally, about the variables. It seems natural that fine art enthusiasts would be concerned about aspect ratio, which is an element of the aesthetic quality of a painting. But, the idea that size should

CHAPTER 5 ♦ Hypothesis Tests and Model Selection **109**

be an element of the price is counterintuitive, particularly weighed against the surprisingly small sizes of some of the world's most iconic paintings such as the *Mona Lisa* (30" high and 21" wide) or Dali's *Persistence of Memory* (only 9.5" high and 13" wide). A skeptic might question the presence of *InSize* in the equation, or, equivalently, the nonzero coefficient,  $\beta_2$ . To settle the issue, the relevant empirical question is whether the equation specified appears to be consistent with the data—that is, the observed sale prices of paintings. In order to proceed, the obvious approach for the analyst would be to fit the regression first and then examine the estimate of  $\beta_2$ . The “test” at this point, is whether  $b_2$  in the least squares regression is zero or not. Recognizing that the least squares slope is a random variable that will never be exactly zero even if  $\beta_2$  really is, we would soften the question to be whether the sample estimate seems to be close enough to zero for us to conclude that its population counterpart is actually zero, that is, that the nonzero value we observe is nothing more than noise that is due to sampling variability. Remaining to be answered are questions including; How close to zero is close enough to reach this conclusion? What metric is to be used? How certain can we be that we have reached the right conclusion? (Not absolutely, of course.) How likely is it that our decision rule, whatever we choose, will lead us to the wrong conclusion? This section will formalize these ideas. After developing the methodology in detail, we will construct a number of numerical examples.

### 5.2.1 RESTRICTIONS AND HYPOTHESES

The approach we will take is to formulate a hypothesis as a restriction on a model. Thus, in the classical methodology considered here, the model is a general statement and a hypothesis is a proposition that narrows that statement. In the art example in (5-2), while the narrower statement is (5-2) with the additional statement that  $\beta_2 = 0$ —without comment on  $\beta_1$  or  $\beta_3$ . We define the **null hypothesis** as the statement that narrows the model and the **alternative hypothesis** as the broader one. In the example, the broader model allows the equation to contain both *InSize* and *AspectRatio*—it admits the possibility that either coefficient might be zero but does not insist upon it. The null hypothesis insists that  $\beta_2 = 0$  while it also makes no comment about  $\beta_1$  or  $\beta_3$ . The formal notation used to frame this hypothesis would be

$$\begin{aligned} \ln Price &= \beta_1 + \beta_2 \ln Size + \beta_3 AspectRatio + \varepsilon, \\ H_0: \beta_2 &= 0, \\ H_1: \beta_2 &\neq 0. \end{aligned} \tag{5-3}$$

Note that the null and alternative hypotheses, together, are exclusive and exhaustive. There is no third possibility; either one or the other of them is true, not both.

The analysis from this point on will be to measure the null hypothesis against the data. The data might persuade the econometrician to reject the null hypothesis. It would seem appropriate at that point to “accept” the alternative. However, in the interest of maintaining flexibility in the methodology, that is, an openness to new information, the appropriate conclusion here will be either to reject the null hypothesis or not to reject it. Not rejecting the null hypothesis is not equivalent to “accepting” it—though the language might suggest so. By accepting the null hypothesis, we would implicitly be closing off further investigation. Thus, the traditional, classical methodology leaves open the possibility that further evidence might still change the conclusion. Our testing

## 110 PART I ♦ The Linear Regression Model

methodology will be constructed so as either to

Reject  $H_0$ : The data are inconsistent with the hypothesis with a reasonable degree of certainty.

Do not reject  $H_0$ : The data appear to be consistent with the null hypothesis.

### 5.2.2 NESTED MODELS

The general approach to testing a hypothesis is to formulate a statistical model that contains the hypothesis as a restriction on its parameters. A theory is said to have **testable implications** if it implies some testable restrictions on the model. Consider, for example, a model of investment,  $I_t$ ,

$$\ln I_t = \beta_1 + \beta_2 i_t + \beta_3 \Delta p_t + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t, \quad (5-4)$$

which states that investors are sensitive to nominal interest rates,  $i_t$ , the rate of inflation,  $\Delta p_t$ , (the log of) real output,  $\ln Y_t$ , and other factors that trend upward through time, embodied in the time trend,  $t$ . An alternative theory states that “investors care about real interest rates.” The alternative model is

$$\ln I_t = \beta_1 + \beta_2 (i_t - \Delta p_t) + \beta_3 \Delta p_t + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t. \quad (5-5)$$

Although this new model does embody the theory, the equation still contains both nominal interest and inflation. The theory has no testable implication for our model. But, consider the stronger hypothesis, “investors care *only* about real interest rates.” The resulting equation,

$$\ln I_t = \beta_1 + \beta_2 (i_t - \Delta p_t) + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t, \quad (5-6)$$

is now restricted; in the context of (5-4), the implication is that  $\beta_2 + \beta_3 = 0$ . The stronger statement implies something specific about the parameters in the equation that may or may not be supported by the empirical evidence.

The description of testable implications in the preceding paragraph suggests (correctly) that testable restrictions will imply that only some of the possible models contained in the original specification will be “valid”; that is, consistent with the theory. In the example given earlier, (5-4) specifies a model in which there are five unrestricted parameters ( $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ ). But, (5-6) shows that only some values are consistent with the theory, that is, those for which  $\beta_3 = -\beta_2$ . This subset of values is contained within the unrestricted set. In this way, the models are said to be **nested**. Consider a different hypothesis, “investors do not care about inflation.” In this case, the smaller set of coefficients is  $(\beta_1, \beta_2, 0, \beta_4, \beta_5)$ . Once again, the restrictions imply a valid **parameter space** that is “smaller” (has fewer dimensions) than the unrestricted one. The general result is that the hypothesis specified by the restricted model is contained within the unrestricted model.

Now, consider an alternative pair of models: Model<sub>0</sub>: “Investors care only about inflation”; Model<sub>1</sub>: “Investors care only about the nominal interest rate.” In this case, the two parameter vectors are  $(\beta_1, 0, \beta_3, \beta_4, \beta_5)$  by Model<sub>0</sub> and  $(\beta_1, \beta_2, 0, \beta_4, \beta_5)$  by Model<sub>1</sub>. In this case, the two specifications are both subsets of the unrestricted model, but neither model is obtained as a restriction on the other. They have the same number of parameters; they just contain different variables. These two models are **nonnested**. For the present, we are concerned only with nested models. Nonnested models are considered in Section 5.8.

## CHAPTER 5 ♦ Hypothesis Tests and Model Selection 111

**5.2.3 TESTING PROCEDURES—NEYMAN–PEARSON METHODOLOGY**

In the example in (5-2), intuition suggests a testing approach based on measuring the data against the hypothesis. The essential methodology suggested by the work of Neyman and Pearson (1933) provides a reliable guide to testing hypotheses in the setting we are considering in this chapter. Broadly, the analyst follows the logic, “What type of data will lead me to reject the hypothesis?” Given the way the hypothesis is posed in Section 5.2.1, the question is equivalent to asking what sorts of data will support the model. The data that one can observe are divided into a **rejection region** and an **acceptance region**. The testing procedure will then be reduced to a simple up or down examination of the statistical evidence. Once it is determined what the rejection region is, if the observed data appear in that region, the null hypothesis is rejected. To see how this operates in practice, consider, once again, the hypothesis about size in the art price equation. Our test is of the hypothesis that  $\beta_2$  equals zero. We will compute the least squares slope. We will decide in advance how far the estimate of  $\beta_2$  must be from zero to lead to rejection of the null hypothesis. Once the rule is laid out, the test, itself, is mechanical. In particular, for this case,  $b_2$  is “far” from zero if  $b_2 > \beta_2^{0+}$  or  $b_2 < \beta_2^{0-}$ . If either case occurs, the hypothesis is rejected. The crucial element is that the rule is decided upon in advance.

**5.2.4 SIZE, POWER, AND CONSISTENCY OF A TEST**

Since the testing procedure is determined in advance and the estimated coefficient(s) in the regression are random, there are two ways the Neyman–Pearson method can make an error. To put this in a numerical context, the sample regression corresponding to (5-2) appears in Table 4.6. The estimate of the coefficient on  $\ln Area$  is 1.33372 with an estimated standard error of 0.09072. Suppose the rule to be used to test is decided arbitrarily (at this point—we will formalize it shortly) to be: If  $b_2$  is greater than +1.0 or less than -1.0, then we will reject the hypothesis that the coefficient is zero (and conclude that art buyers really do care about the sizes of paintings). So, based on this rule, we will, in fact, reject the hypothesis. However, since  $b_2$  is a random variable, there are the following possible errors:

Type I error:  $\beta_2 = 0$ , but we reject the hypothesis.

The null hypothesis is incorrectly rejected.

Type II error:  $\beta_2 \neq 0$ , but we do not reject the hypothesis.

The null hypothesis is incorrectly retained.

The probability of a Type I error is called the **size of the test**. The size of a test is the probability that the test will incorrectly reject the null hypothesis. As will emerge later, the analyst determines this in advance. One minus the probability of a Type II error is called the **power of a test**. The power of a test is the probability that it will correctly reject a false null hypothesis. The power of a test depends on the alternative. It is not under the control of the analyst. To consider the  example once again, we are going to reject the hypothesis if  $|b_2| > 1$ . If  $\beta_2$  is actually 1.5, based on the results we've seen, we are quite likely to find a value of  $b_2$  that is greater than 1.0. On the other hand, if  $\beta_2$  is only 0.3, then it does not appear likely that we will observe a sample value greater than 1.0. Thus, again, the power of a test depends on the actual parameters that underlie the data. The idea of power of a test relates to its ability to find what it is looking for.

## 112 PART I ♦ The Linear Regression Model

A test procedure is **consistent** if its power goes to 1.0 as the sample size grows to infinity. This quality is easy to see, again, in the context of a single parameter, such as the one being considered here. Since least squares is consistent, it follows that as the sample size grows, we will be able to learn the exact value of  $\beta_2$ , so we will know if it is zero or not. Thus, for this example, it is clear that as the sample size grows, we will know with certainty if we should reject the hypothesis. For most of our work in this text, we can use the following guide: A testing procedure about the parameters in a model is consistent if it is based on a consistent estimator of those parameters. Since nearly all our work in this book is based on consistent estimators and save for the latter sections of this chapter, where our tests will be about the parameters in nested models, our tests will be consistent.

### 5.2.5 A METHODOLOGICAL DILEMMA: BAYESIAN VS. CLASSICAL TESTING

As we noted earlier, the Neyman–Pearson testing methodology we will employ here is an all-or-nothing proposition. We will determine the testing rule(s) in advance, gather the data, and either reject or not reject the null hypothesis. There is no middle ground. This presents the researcher with two uncomfortable dilemmas. First, the testing outcome, that is, the sample data might be uncomfortably close to the boundary of the rejection region. Consider our example. If we have decided in advance to reject the null hypothesis if  $b_2 > 1.00$ , and the sample value is 0.9999, it will be difficult to resist the urge to reject the null hypothesis anyway, particularly if we entered the analysis with a strongly held belief that the null hypothesis is incorrect. (I.e., intuition notwithstanding, I am convinced that art buyers really do care about size.) Second, the methodology we have laid out here has no way of incorporating other studies. To continue our example, if I were the tenth analyst to study the art market, and the previous nine had decisively rejected the hypothesis that  $\beta_2 = 0$ , I will find it very difficult not to reject that hypothesis even if my evidence suggests, based on my testing procedure, that I should.

This dilemma is built into the classical testing methodology. There is a middle ground. The Bayesian methodology that we will discuss in Chapter 15 does not face this dilemma because Bayesian analysts never reach a firm conclusion. They merely update their priors. Thus, the first case noted, in which the observed data are close to the boundary of the rejection region, the analyst will merely be updating the prior with somewhat slightly less persuasive evidence than might be hoped for. But, the methodology is comfortable with this. For the second instance, we have a case in which there is a wealth of prior evidence in favor of rejecting  $H_0$ . It will take a powerful tenth body of evidence to overturn the previous nine conclusions. The results of the tenth study (the posterior results) will incorporate not only the current evidence, but the wealth of prior data as well.

## 5.3 TWO APPROACHES TO TESTING HYPOTHESES

The **general linear hypothesis** is a set of  $J$  restrictions on the linear regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

## CHAPTER 5 ♦ Hypothesis Tests and Model Selection 113

The restrictions are written

$$\begin{aligned} r_{11}\beta_1 + r_{12}\beta_2 + \cdots + r_{1K}\beta_K &= q_1 \\ r_{21}\beta_1 + r_{22}\beta_2 + \cdots + r_{2K}\beta_K &= q_2 \\ &\dots \\ r_{J1}\beta_1 + r_{J2}\beta_2 + \cdots + r_{JK}\beta_K &= q_J. \end{aligned} \tag{5-7}$$

The simplest case is a single restriction on one coefficient, such as

$$\beta_k = 0.$$

The more general case can be written in the matrix form,

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{q}. \tag{5-8}$$

Each row of  $\mathbf{R}$  is the coefficients in one of the restrictions. Typically,  $\mathbf{R}$  will have only a few rows and numerous zeros in each row. Some examples would be as follows:

1. One of the coefficients is zero,  $\beta_j = 0$ ,

$$\mathbf{R} = [0 \ 0 \ \cdots \ 1 \ 0 \ \cdots \ 0] \text{ and } \mathbf{q} = 0.$$

2. Two of the coefficients are equal,  $\beta_k = \beta_j$ ,

$$\mathbf{R} = [0 \ 0 \ 1 \ \cdots \ -1 \ \cdots \ 0] \text{ and } \mathbf{q} = 0.$$

3. A set of the coefficients sum to one,  $\beta_2 + \beta_3 + \beta_4 = 1$ ,

$$\mathbf{R} = [0 \ 1 \ 1 \ 1 \ 0 \ \cdots] \text{ and } \mathbf{q} = 1.$$

4. A subset of the coefficients are all zero,  $\beta_1 = 0$ ,  $\beta_2 = 0$ , and  $\beta_3 = 0$ ,

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} = [\mathbf{I} \ \mathbf{0}] \text{ and } \mathbf{q} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

5. Several linear restrictions,  $\beta_2 + \beta_3 = 1$ ,  $\beta_4 + \beta_6 = 0$ , and  $\beta_5 + \beta_6 = 0$ ,

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \text{ and } \mathbf{q} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

6. All the coefficients in the model except the constant term are zero,

$$\mathbf{R} = [\mathbf{0} : \mathbf{I}_{K-1}] \text{ and } \mathbf{q} = \mathbf{0}.$$

The matrix  $\mathbf{R}$  has  $K$  columns to be conformable with  $\boldsymbol{\beta}$ ,  $J$  rows for a total of  $J$  restrictions, and *full row rank*, so  $J$  must be less than or equal to  $K$ . The rows of  $\mathbf{R}$  must be linearly independent. Although it does not violate the condition, the case of  $J = K$  must also be ruled out. If the  $K$  coefficients satisfy  $J = K$  restrictions, then  $\mathbf{R}$  is square and nonsingular and  $\boldsymbol{\beta} = \mathbf{R}^{-1}\mathbf{q}$ . There is no estimation or inference problem. The restriction  $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$  imposes  $J$  restrictions on  $K$  otherwise free parameters. Hence, with the restrictions imposed, there are, in principle, only  $K - J$  free parameters remaining.

We will want to extend the methods to nonlinear restrictions. In a following example, below, the hypothesis takes the form  $H_0: \beta_j/\beta_k = \beta_l/\beta_m$ . The **general nonlinear**

## 114 PART I ♦ The Linear Regression Model

**hypothesis** involves a set of  $J$  possibly nonlinear restrictions,

$$\mathbf{c}(\boldsymbol{\beta}) = \mathbf{q}, \quad (5-9)$$

where  $\mathbf{c}(\boldsymbol{\beta})$  is a set of  $J$  nonlinear functions of  $\boldsymbol{\beta}$ . The linear hypothesis is a special case. The counterpart to our requirements for the linear case are that, once again,  $J$  be strictly less than  $K$ , and the matrix of derivatives,

$$\mathbf{G}(\boldsymbol{\beta}) = \partial \mathbf{c}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}', \quad (5-10)$$

have full row rank. This means that the restrictions are **functionally independent**. In the linear case,  $\mathbf{G}(\boldsymbol{\beta})$  is the matrix of constants,  $\mathbf{R}$  that we saw earlier and functional independence is equivalent to linear independence. We will consider nonlinear restrictions in detail in Section 5.7. For the present, we will restrict attention to the general linear hypothesis.

The hypothesis implied by the restrictions is written

$$H_0: \mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0},$$

$$H_1: \mathbf{R}\boldsymbol{\beta} - \mathbf{q} \neq \mathbf{0}.$$

We will consider two approaches to testing the hypothesis, Wald tests and fit based tests. The hypothesis characterizes the population. If the hypothesis is correct, then the sample statistics should mimic that description. To continue our earlier example, the hypothesis states that a certain coefficient in a regression model equals zero. If the hypothesis is correct, then the least squares coefficient should be close to zero, at least within sampling variability. The tests will proceed as follows:

- Wald tests: The hypothesis states that  $\mathbf{R}\boldsymbol{\beta} - \mathbf{q}$  equals  $\mathbf{0}$ . The least squares estimator,  $\mathbf{b}$ , is an unbiased and consistent estimator of  $\boldsymbol{\beta}$ . If the hypothesis is correct, then the **sample discrepancy**,  $\mathbf{R}\mathbf{b} - \mathbf{q}$  should be close to zero. For the example of a single coefficient, if the hypothesis that  $\beta_k$  equals zero is correct, then  $b_k$  should be close to zero. The Wald test measures how close  $\mathbf{R}\mathbf{b} - \mathbf{q}$  is to zero.
- Fit based tests: We obtain the best possible fit—highest  $R^2$ —by using least squares without imposing the restrictions. We proved this in Chapter 3. We will show here that the sum of squares will never decrease when we impose the restrictions—except for an unlikely special case, it will increase. For example, when we impose  $\beta_k = 0$  by leaving  $x_k$  out of the model, we should expect  $R^2$  to fall. The empirical device to use for testing the hypothesis will be a measure of how much  $R^2$  falls when we impose the restrictions.

### AN IMPORTANT ASSUMPTION

To develop the test statistics in this section, we will assume normally distributed disturbances. As we saw in Chapter 4, with this assumption, we will be able to obtain the exact distributions of the test statistics. In Section 5.6, we will consider the implications of relaxing this assumption and develop an alternative set of results that allows us to proceed without it.

## 5.4 WALD TESTS BASED ON THE DISTANCE MEASURE

The **Wald test** is the most commonly used procedure. It is often called a “significance test.” The operating principle of the procedure is to fit the regression without the restrictions, and then assess whether the results appear, within sampling variability, to agree with the hypothesis.

### 5.4.1 TESTING A HYPOTHESIS ABOUT A COEFFICIENT

The simplest case is a test of the value of a single coefficient. Consider, once again, our art market example in Section 5.2. The null hypothesis is

$$H_0: \beta_2 = \beta_2^0,$$

where  $\beta_2^0$  is the hypothesized value of the coefficient, in this case, zero. The **Wald distance** of a coefficient estimate from a hypothesized value is the linear distance, measured in standard deviation units. Thus, for this case, the distance of  $b_k$  from  $\beta_k^0$  would be

$$W_k = \frac{b_k - \beta_k^0}{\sqrt{\sigma^2 S^{kk}}}. \quad (5-11)$$

As we saw in (4-38),  $W_k$  (which we called  $z_k$  before) has a standard normal distribution assuming that  $E[b_k] = \beta_k^0$ . Note that if  $E[b_k]$  is not equal to  $\beta_k^0$ , then  $W_k$  still has a normal distribution, but the mean is not zero. In particular, if  $E[b_k]$  is  $\beta_k^1$  which is different from  $\beta_k^0$ , then

$$E\{W_k | E[b_k] = \beta_k^1\} = \frac{\beta_k^1 - \beta_k^0}{\sqrt{\sigma^2 S^{kk}}}. \quad (5-12)$$

(E.g., if the hypothesis is that  $\beta_k = \beta_k^0 = 0$ , and  $\beta_k$  does not equal zero, then the expected of  $W_k = b_k / \sqrt{\sigma^2 S^{kk}}$  will equal  $\beta_k^1 / \sqrt{\sigma^2 S^{kk}}$ , which is not zero.) For purposes of using  $W_k$  to test the hypothesis, our interpretation is that if  $\beta_k$  does equal  $\beta_k^0$ , then  $b_k$  will be close to  $\beta_k^0$ , with the distance measured in standard error units. Therefore, the logic of the test, to this point, will be to conclude that  $H_0$  is incorrect—should be rejected—if  $W_k$  is “large.”

Before we determine a benchmark for large, we note that the Wald measure suggested here is not usable because  $\sigma^2$  is not known. It was estimated by  $s^2$ . Once again, invoking our results from Chapter 4, if we compute  $W_k$  using the sample estimate of  $\sigma^2$ , we obtain

$$t_k = \frac{b_k - \beta_k^0}{\sqrt{s^2 S^{kk}}} \quad (5-13)$$

Assuming that  $\beta_k$  does indeed equal  $\beta_k^0$ , that is, “under the assumption of the null hypothesis,” then  $t_k$  has a *t* distribution with  $n - K$  degrees of freedom. [See (4-41).] We can now construct the testing procedure. The test is carried out by determining in advance the desired confidence with which we would like to draw the conclusion—the standard value is 95 percent. Based on (5-13), we can say that

$$\text{Prob}\{-t_{(1-\alpha/2),[n-K]}^* < t_k < +t_{(1-\alpha/2),[n-K]}^*\}$$

## 116 PART I ♦ The Linear Regression Model

where  $t^*_{(1-\alpha/2),[n-K]}$  is the appropriate value from the  $t$  table (in Appendix G of this book). By this construction, finding a sample value of  $t_k$  that falls outside this range is unlikely. Our test procedure states that it is so unlikely that we would conclude that it could not happen if the hypothesis were correct, so the hypothesis must be incorrect.

A common test is the hypothesis that a parameter equals zero—equivalently, this is a test of the relevance of a variable in the regression. To construct the test statistic, we set  $\beta_k^0$  to zero in (5-13) to obtain the standard “ $t$  ratio,”

$$t_k = \frac{b_k}{s_{bk}}.$$

This statistic is reported in the regression results in several of our earlier examples, such as 4.10 where the regression results for the model in (5-2) appear. This statistic is usually labeled the  **$t$  ratio** for the estimator  $b_k$ . If  $|b_k|/s_{bk} > t_{(1-\alpha/2),[n-K]}$ , where  $t_{(1-\alpha/2),[n-K]}$  is the  $100(1 - \alpha/2)$  percent critical value from the  $t$  distribution with  $(n - K)$  degrees of freedom, then the null hypothesis that the coefficient is zero is rejected and the coefficient (actually, the associated variable) is said to be “statistically significant.” The value of 1.96, which would apply for the 95 percent significance level in a large sample, is often used as a benchmark value when a table of critical values is not immediately available. The  $t$  ratio for the test of the hypothesis that a coefficient equals zero is a standard part of the regression output of most computer programs.

Another view of the testing procedure is useful. Also based on (4-39) and (5-13), we formed a confidence interval for  $\beta_k$  as  $b_k \pm t^* s_k$ . We may view this interval as the set of plausible values of  $\beta_k$  with a confidence level of  $100(1 - \alpha)$  percent, where we choose  $\alpha$ , typically 5 percent. The confidence interval provides a convenient tool for testing a hypothesis about  $\beta_k$ , since we may simply ask whether the hypothesized value,  $\beta_k^0$  is contained in this range of plausible values.

### Example 5.1 Art Appreciation

Regression results for the model in (5-3) based on a sample of 430 sales of Monet paintings appear in Table 4.6 in Example 4.10. The estimated coefficient on  $\ln(\text{Area})$  is 1.33372 with an estimated standard error of 0.09072. The distance of the estimated coefficient from zero is  $1.33372/0.09072 = 14.70$ . Since this is far larger than the 95 percent critical value of 1.96, we reject the hypothesis that  $\beta_2$  equals zero; evidently buyers of Monet paintings do care about size. In contrast, the coefficient on  $\text{AspectRatio}$  is  $-0.16537$  with an estimated standard error of 0.12753, so the associated  $t$  ratio for the test of  $H_0: \beta_3 = 0$  is only  $-1.30$ . Since this is well under 1.96, we conclude that art buyers (of Monet paintings) do not care about the aspect ratio of the paintings. As a final consideration, we examine another (equally bemusing) hypothesis, whether auction prices are inelastic  $H_0: \beta_2 \leq 1$  or elastic  $H_1: \beta_2 > 1$  with respect to area. This is a **one-sided test**. Using our Neyman–Pearson guideline for formulating the test, we will reject the null hypothesis if the estimated coefficient is sufficiently larger than 1.0 (and not if it is less than or equal to 1.0). To maintain a test of size 0.05, we will then place all of the area for the critical region (the rejection region) to the right of 1.0; the critical value from the table is 1.645. The test statistic is  $(1.33372 - 1.0)/0.09072 = 3.679 > 1.645$ . Thus, we will reject this null hypothesis as well.

### Example 5.2 Earnings Equation

Appendix Table F5.1 contains 753 observations used in Mroz's (1987) study of the labor supply behavior of married women. We will use these data at several points in this example. Of the 753 individuals in the sample, 428 were participants in the formal labor market. For these individuals, we will fit a semilog earnings equation of the form suggested in Example 2.2;

$$\ln(\text{earnings}) = \beta_1 + \beta_2 \text{age} + \beta_3 \text{age}^2 + \beta_4 \text{education} + \beta_5 \text{kids} + \varepsilon,$$

## CHAPTER 5 ♦ Hypothesis Tests and Model Selection 117

**TABLE 5.1** Regression Results for an Earnings Equation

Sum of squared residuals:	599.4582			
Standard error of the regression:	1.19044			
$R^2$ based on 428 observations	0.040995			
Variable	Coefficient	Standard Error	t Ratio	
Constant	3.24009	1.7674	1.833	
Age	0.20056	0.08386	2.392	
Age <sup>2</sup>	-0.0023147	0.00098688	-2.345	
Education	0.067472	0.025248	2.672	
Kids	-0.35119	0.14753	-2.380	
<i>Estimated Covariance Matrix for b (e - n = times 10<sup>-n</sup>)</i>				
Constant	Age	Age <sup>2</sup>	Education	Kids
3.12381				
-0.14409	0.0070325			
0.0016617	-8.23237e-5	9.73928e-7		
-0.0092609	5.08549e-5	-4.96761e-7	0.00063729	
0.026749	-0.0026412	3.84102e-5	-5.46193e-5	0.021766

where *earnings* is *hourly wage times hours worked*, *education* is measured in years of schooling, and *kids* is a binary variable which equals one if there are children under 18 in the household. (See the data description in Appendix F for details.) Regression results are shown in Table 5.1. There are 428 observations and 5 parameters, so the *t* statistics have  $(428 - 5) = 423$  degrees of freedom. For 95 percent significance levels, the standard normal value of 1.96 is appropriate when the degrees of freedom are this large. By this measure, all variables are statistically significant and signs are consistent with expectations. It will be interesting to investigate whether the effect of *kids* is on the wage or hours, or both. We interpret the schooling variable to imply that an additional year of schooling is associated with a 6.7 percent increase in earnings. The quadratic age profile suggests that for a given education level and family size, earnings rise to the peak at  $-b_2/(2b_3)$  which is about 43 years of age, at which point they begin to decline. Some points to note: (1) Our selection of only those individuals who had positive hours worked is not an innocent sample selection mechanism. Since individuals chose whether or not to be in the labor force, it is likely (almost certain) that earnings potential was a significant factor, along with some other aspects we will consider in Chapter 18.

(2) The earnings equation is a mixture of a labor supply equation—hours worked by the individual—and a labor demand outcome—the wage is, presumably, an accepted offer. As such, it is unclear what the precise nature of this equation is. Presumably, it is a hash of the equations of an elaborate structural equation system. (See Example 1.1 for discussion.)

**5.4.2 THE F STATISTIC AND THE LEAST SQUARES DISCREPANCY**

We now consider testing a set of  $J$  linear restrictions stated in the **null hypothesis**

$$H_0 : \mathbf{R}\beta - \mathbf{q} = \mathbf{0}$$

against the **alternative hypothesis**,

$$H_1 : \mathbf{R}\beta - \mathbf{q} \neq \mathbf{0}.$$

Given the least squares estimator  $\mathbf{b}$ , our interest centers on the **discrepancy vector**  $\mathbf{R}\mathbf{b} - \mathbf{q} = \mathbf{m}$ . It is unlikely that  $\mathbf{m}$  will be exactly  $\mathbf{0}$ . The statistical question is whether

## 118 PART I ♦ The Linear Regression Model

the deviation of  $\mathbf{m}$  from  $\mathbf{0}$  can be attributed to sampling error or whether it is significant. Since  $\mathbf{b}$  is normally distributed [see (4-18)] and  $\mathbf{m}$  is a linear function of  $\mathbf{b}$ ,  $\mathbf{m}$  is also normally distributed. If the null hypothesis is true, then  $\mathbf{R}\beta - \mathbf{q} = \mathbf{0}$  and  $\mathbf{m}$  has mean vector

$$E[\mathbf{m} | \mathbf{X}] = \mathbf{R}E[\mathbf{b} | \mathbf{X}] - \mathbf{q} = \mathbf{R}\beta - \mathbf{q} = \mathbf{0}.$$

and covariance matrix

$$\text{Var}[\mathbf{m} | \mathbf{X}] = \text{Var}[\mathbf{R}\mathbf{b} - \mathbf{q} | \mathbf{X}] = \mathbf{R}\{\text{Var}[\mathbf{b} | \mathbf{X}]\}\mathbf{R}' = \sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'.$$

We can base a test of  $H_0$  on the **Wald criterion**. Conditioned on  $\mathbf{X}$ , we find:

$$\begin{aligned} W &= \mathbf{m}'\{\text{Var}[\mathbf{m} | \mathbf{X}]\}^{-1}\mathbf{m}. \\ &= (\mathbf{R}\mathbf{b} - \mathbf{q})'[\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}) \\ &= \frac{(\mathbf{R}\mathbf{b} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q})}{\sigma^2} \\ &\sim \chi^2[J]. \end{aligned} \tag{5-14}$$

The statistic  $W$  has a chi-squared distribution with  $J$  degrees of freedom if the hypothesis is correct.<sup>1</sup> Intuitively, the larger  $\mathbf{m}$  is—that is, the worse the failure of least squares to satisfy the restrictions—the larger the chi-squared statistic. Therefore, a large chi-squared value will weigh against the hypothesis.

The chi-squared statistic in (5-14) is not usable because of the unknown  $\sigma^2$ . By using  $s^2$  instead of  $\sigma^2$  and dividing the result by  $J$ , we obtain a usable  $F$  statistic with  $J$  and  $n - K$  degrees of freedom. Making the substitution in (5-14), dividing by  $J$ , and multiplying and dividing by  $n - K$ , we obtain

$$\begin{aligned} F &= \frac{W \sigma^2}{J s^2} \\ &= \left( \frac{(\mathbf{R}\mathbf{b} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q})}{\sigma^2} \right) \left( \frac{1}{J} \right) \left( \frac{\sigma^2}{s^2} \right) \left( \frac{(n - K)}{(n - K)} \right) \\ &= \frac{(\mathbf{R}\mathbf{b} - \mathbf{q})'[\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q})/J}{[(n - K)s^2/\sigma^2]/(n - K)}. \end{aligned} \tag{5-15}$$

If  $\mathbf{R}\beta = \mathbf{q}$ , that is, if the null hypothesis is true, then  $\mathbf{R}\mathbf{b} - \mathbf{q} = \mathbf{R}\mathbf{b} - \mathbf{R}\beta = \mathbf{R}(\mathbf{b} - \beta) = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}$ . [See (4-4).] Let  $\mathbf{C} = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']$  since

$$\frac{\mathbf{R}(\mathbf{b} - \beta)}{\sigma} = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(\frac{\boldsymbol{\epsilon}}{\sigma}\right) = \mathbf{D}\left(\frac{\boldsymbol{\epsilon}}{\sigma}\right),$$

the numerator of  $F$  equals  $[(\boldsymbol{\epsilon}/\sigma)' \mathbf{T}(\boldsymbol{\epsilon}/\sigma)]/J$  where  $\mathbf{T} = \mathbf{D}'\mathbf{C}^{-1}\mathbf{D}$ . The numerator is  $W/J$  from (5-14) and is distributed as  $1/J$  times a chi-squared  $[J]$ , as we showed earlier. We found in (4-16) that  $s^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon}/(n - K) = \boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}/(n - K)$  where  $\mathbf{M}$  is an idempotent matrix. Therefore, the denominator of  $F$  equals  $[(\boldsymbol{\epsilon}/\sigma)' \mathbf{M}(\boldsymbol{\epsilon}/\sigma)]/(n - K)$ . This statistic is distributed as  $1/(n - K)$  times a chi-squared  $[n - K]$ . Therefore, the  $F$  statistic is the ratio of two chi-squared variables each divided by its degrees of freedom. Since  $\mathbf{M}(\boldsymbol{\epsilon}/\sigma)$  and

<sup>1</sup>This calculation is an application of the “full rank quadratic form” of Section B.11.6. Note that although the chi-squared distribution is conditioned on  $\mathbf{X}$ , it is also free of  $\mathbf{X}$ .

## CHAPTER 5 ♦ Hypothesis Tests and Model Selection 119

$\mathbf{T}(\boldsymbol{\varepsilon}/\sigma)$  are both normally distributed and their covariance  $\mathbf{TM}$  is  $\mathbf{0}$ , the vectors of the quadratic forms are independent. The numerator and denominator of  $F$  are functions of independent random vectors and are therefore independent. This completes the proof of the  $F$  distribution. [See (B-35).] Canceling the two appearances of  $\sigma^2$  in (5-15) leaves the  $F$  statistic for testing a linear hypothesis:

$$F[J, n - K | \mathbf{X}] = \frac{(\mathbf{R}\mathbf{b} - \mathbf{q})' \{ \mathbf{R}[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}' \}^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q})}{J}. \quad (5-16)$$

For testing one linear restriction of the form

$$H_0 : r_1\beta_1 + r_2\beta_2 + \cdots + r_K\beta_K = \mathbf{r}'\boldsymbol{\beta} = q \quad \text{[Text Box]}$$

(usually, some of the  $r$ 's will be zero), the  $F$  statistic is

$$F[1, n - K] = \frac{(\sum_j r_j b_j - q)^2}{\sum_j \sum_k r_j r_k \text{Est. Cov}[b_j, b_k]}.$$

If the hypothesis is that the  $j$ th coefficient is equal to a particular value, then  $\mathbf{R}$  has a single row with a 1 in the  $j$ th position and 0s elsewhere,  $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$  is the  $j$ th diagonal element of the inverse matrix, and  $\mathbf{R}\mathbf{b} - \mathbf{q}$  is  $(b_j - q)$ . The  $F$  statistic is then

$$F[1, n - K] = \frac{(b_j - q)^2}{\text{Est. Var}[b_j]}.$$

Consider an alternative approach. The sample estimate of  $\mathbf{r}'\boldsymbol{\beta}$  is

$$r_1 b_1 + r_2 b_2 + \cdots + r_K b_K = \mathbf{r}'\mathbf{b} = \hat{q}.$$

If  $\hat{q}$  differs significantly from  $q$ , then we conclude that the sample data are not consistent with the hypothesis. It is natural to base the test on

$$t = \frac{\hat{q} - q}{\text{se}(\hat{q})}. \quad (5-17)$$

We require an estimate of the standard error of  $\hat{q}$ . Since  $\hat{q}$  is a linear function of  $\mathbf{b}$  and we have an estimate of the covariance matrix of  $\mathbf{b}$ ,  $s^2(\mathbf{X}'\mathbf{X})^{-1}$ , we can estimate the variance of  $\hat{q}$  with

$$\text{Est. Var}[\hat{q} | \mathbf{X}] = \mathbf{r}'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{r}.$$

The denominator of  $t$  is the square root of this quantity. In words,  $t$  is the distance in standard error units between the hypothesized function of the true coefficients and the same function of our estimates of them. If the hypothesis is true, then our estimates should reflect that, at least within the range of sampling variability. Thus, if the absolute value of the preceding  $t$  ratio is larger than the appropriate critical value, then doubt is cast on the hypothesis.

There is a useful relationship between the statistics in (5-16) and (5-17). We can write the square of the  $t$  statistic as

$$t^2 = \frac{(\hat{q} - q)^2}{\text{Var}(\hat{q} - q | \mathbf{X})} = \frac{(\mathbf{r}'\mathbf{b} - q)\{\mathbf{r}'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{r}\}^{-1}(\mathbf{r}'\mathbf{b} - q)}{1}.$$

It follows, therefore, that for testing a single restriction, the  $t$  statistic is the square root of the  $F$  statistic that would be used to test that hypothesis.

## 120 PART I ♦ The Linear Regression Model

### Example 5.3 Restricted Investment Equation

Section 5.2.2 suggested a theory about the behavior of investors: They care only about real interest rates. If investors were only interested in the real rate of interest, then equal increases in interest rates and the rate of inflation would have no independent effect on investment. The null hypothesis is

$$H_0: \beta_2 + \beta_3 = 0.$$

Estimates of the parameters of equations (5-4) and (5-6) using 1950.1 to 2000.4 quarterly data on real investment, real GDP, an interest rate (the 90-day T-bill rate), and inflation measured by the change in the log of the CPI given in Appendix Table F5.2 are presented in Table 5.2. (One observation is lost in computing the change in the CPI.)

To form the appropriate test statistic, we require the standard error of  $\hat{q} = b_2 + b_3$ , which is

$$se(\hat{q}) = [0.00319^2 + 0.00234^2 + 2(-3.718 \times 10^{-6})]^{1/2} = 0.002866.$$

The  $t$  ratio for the test is therefore

$$t = \frac{-0.00860 + 0.00331}{0.002866} = -1.845.$$

Using the 95 percent critical value from  $t$  [203-5] = 1.96 (the standard normal value), we conclude that the sum of the two coefficients is not significantly different from zero, so the hypothesis should not be rejected.

There will usually be more than one way to formulate a restriction in a regression model. One convenient way to parameterize a constraint is to set it up in such a way that the standard test statistics produced by the regression can be used without further computation to test the hypothesis. In the preceding example, we could write the regression model as specified in (5-5). Then an equivalent way to test  $H_0$  would be to fit the investment equation with both the real interest rate and the rate of inflation as regressors and to test our theory by simply testing the hypothesis that  $\beta_3$  equals zero, using the standard  $t$  statistic that is routinely computed. When the regression is computed this way,  $b_3 = -0.00529$  and the estimated standard error is 0.00287, resulting in a  $t$  ratio of  $-1.844()$ . (Exercise: Suppose that the nominal interest rate, rather than the rate of inflation, were included as the extra regressor. What do you think the coefficient and its standard error would be?)

Finally, consider a test of the joint hypothesis

$$\beta_2 + \beta_3 = 0 \quad (\text{investors consider the real interest rate}),$$

$$\beta_4 = 1 \quad (\text{the marginal propensity to invest equals 1}),$$

$$\beta_5 = 0 \quad (\text{there is no time trend}).$$

**TABLE 5.2** Estimated Investment Equations (Estimated standard errors in parentheses)

	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$
<b>Model (5-4)</b>	-9.135 (1.366)	-0.00860 (0.00319)	0.00331 (0.00234)	1.930 (0.183)	-0.00566 (0.00149)
	$s = 0.08618, R^2 = 0.979753, \mathbf{e}'\mathbf{e} = 1.47052,$ $\text{Est. Cov}[b_2, b_3] = -3.718e-6$				
<b>Model (5-6)</b>	-7.907 (1.201)	-0.00443 (0.00227)	0.00443 (0.00227)	1.764 (0.161)	-0.00440 (0.00133)
	$s = 0.8670, R^2 = 0.979405, \mathbf{e}'\mathbf{e} = 1.49578$				

CHAPTER 5 ♦ Hypothesis Tests and Model Selection **121**

Then,

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{q} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{R}\mathbf{b} - \mathbf{q} = \begin{bmatrix} -0.0053 \\ 0.9302 \\ -0.0057 \end{bmatrix}.$$

Inserting these values in  $F$  yields  $F = 109.84$ . The 5 percent critical value for  $F[3, 198]$  is 2.65. We conclude, therefore, that these data are not consistent with the hypothesis. The result gives no indication as to which of the restrictions is most influential in the rejection of the hypothesis. If the three restrictions are tested one at a time, the  $t$  statistics in (5-17) are  $-1.844$ ,  $5.076$ , and  $-3.803$ . Based on the individual test statistics, therefore, we would expect both the second and third hypotheses to be rejected.

## 5.5 TESTING RESTRICTIONS USING THE FIT OF THE REGRESSION

A different approach to hypothesis testing focuses on the fit of the regression. Recall that the least squares vector  $\mathbf{b}$  was chosen to minimize the sum of squared deviations,  $\mathbf{e}'\mathbf{e}$ . Since  $R^2$  equals  $1 - \mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$  and  $\mathbf{y}'\mathbf{M}^0\mathbf{y}$  is a constant that does not involve  $\mathbf{b}$ , it follows that  $\mathbf{b}$  is chosen to maximize  $R^2$ . One might ask whether choosing some other value for the slopes of the regression leads to a significant loss of fit. For example, in the investment equation (5-4), one might be interested in whether assuming the hypothesis (that investors care only about real interest rates) leads to a substantially worse fit than leaving the model unrestricted. To develop the test statistic, we first examine the computation of the least squares estimator subject to a set of restrictions. We will then construct a test statistic that is based on comparing the  $R^2$ 's from the two regressions.

### 5.5.1 THE RESTRICTED LEAST SQUARES ESTIMATOR

Suppose that we explicitly impose the restrictions of the general linear hypothesis in the regression. The restricted least squares estimator is obtained as the solution to

$$\text{Minimize}_{\mathbf{b}_0} S(\mathbf{b}_0) = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'(\mathbf{y} - \mathbf{X}\mathbf{b}_0) \quad \text{subject to } \mathbf{R}\mathbf{b}_0 = \mathbf{q}. \quad (5-18)$$

A Lagrangean function for this problem can be written

$$L^*(\mathbf{b}_0, \lambda) = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'(\mathbf{y} - \mathbf{X}\mathbf{b}_0) + 2\lambda'(\mathbf{R}\mathbf{b}_0 - \mathbf{q}).^2 \quad (5-19)$$

The solutions  $\mathbf{b}_*$  and  $\lambda_*$  will satisfy the necessary conditions

$$\begin{aligned} \frac{\partial L^*}{\partial \mathbf{b}_*} &= -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}_*) + 2\mathbf{R}'\lambda_* = \mathbf{0} \\ \frac{\partial L^*}{\partial \lambda_*} &= 2(\mathbf{R}\mathbf{b}_* - \mathbf{q}) = \mathbf{0}. \end{aligned} \quad (5-20)$$

Dividing through by 2 and expanding terms produces the partitioned matrix equation

$$\begin{aligned} \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{R}' \\ \mathbf{R} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b}_* \\ \lambda_* \end{bmatrix} &= \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{q} \end{bmatrix} \\ \text{or} \end{aligned} \quad (5-21)$$

$$\mathbf{A}\mathbf{d}_* = \mathbf{v}.$$

<sup>2</sup>Since  $\lambda$  is not restricted, we can formulate the constraints in terms of  $2\lambda$ . The convenience of the scaling shows up in (5-20).

## 122 PART I ♦ The Linear Regression Model

Assuming that the partitioned matrix in brackets is nonsingular, the restricted least squares estimator is the upper part of the solution

$$\mathbf{d}_* = \mathbf{A}^{-1}\mathbf{v}. \quad (5-22)$$

If, in addition,  $\mathbf{X}'\mathbf{X}$  is nonsingular, then explicit solutions for  $\mathbf{b}_*$  and  $\lambda_*$  may be obtained by using the formula for the partitioned inverse (A-74),<sup>3</sup>

$$\begin{aligned}\mathbf{b}_* &= \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}) \\ &= \mathbf{b} - \mathbf{Cm}\end{aligned}$$

and

$$\lambda_* = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}).$$

Greene and Seaks (1991) show that the covariance matrix for  $\mathbf{b}_*$  is simply  $\sigma^2$  times the upper left block of  $\mathbf{A}^{-1}$ . Once again, in the usual case in which  $\mathbf{X}'\mathbf{X}$  is nonsingular, an explicit formulation may be obtained:

$$\text{Var}[\mathbf{b}_* | \mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}. \quad (5-24)$$

Thus,

$$\text{Var}[\mathbf{b}_* | \mathbf{X}] = \text{Var}[\mathbf{b} | \mathbf{X}] - \text{a nonnegative definite matrix.}$$

One way to interpret this reduction in variance is as the value of the information contained in the restrictions.

Note that the explicit solution for  $\lambda_*$  involves the discrepancy vector  $\mathbf{R}\mathbf{b} - \mathbf{q}$ . If the unrestricted least squares estimator satisfies the restriction, the Lagrangean multipliers will equal zero and  $\mathbf{b}_*$  will equal  $\mathbf{b}$ . Of course, this is unlikely. The constrained solution  $\mathbf{b}_*$  is equal to the unconstrained solution  $\mathbf{b}$  minus a term that accounts for the failure of the unrestricted solution to satisfy the constraints.

### 5.5.2 THE LOSS OF FIT FROM RESTRICTED LEAST SQUARES

To develop a test based on the restricted least squares estimator, we consider a single coefficient first and then turn to the general case of  $J$  linear restrictions. Consider the change in the fit of a multiple regression when a variable  $z$  is added to a model that already contains  $K - 1$  variables,  $\mathbf{x}$ . We showed in Section 3.5 (Theorem 3.6) (3-29) that the effect on the fit would be given by

$$R_{\mathbf{Xz}}^2 = R_{\mathbf{X}}^2 + (1 - R_{\mathbf{X}}^2)r_{yz}^{*2}, \quad (5-25)$$

where  $R_{\mathbf{Xz}}^2$  is the new  $R^2$  after  $z$  is added,  $R_{\mathbf{X}}^2$  is the original  $R^2$  and  $r_{yz}^{*2}$  is the partial correlation between  $y$  and  $z$ , controlling for  $\mathbf{x}$ . So, as we knew, the fit improves (or, at the least, does not deteriorate). In deriving the partial correlation coefficient between  $y$  and  $z$  in (3-22) we obtained the convenient result

$$r_{yz}^{*2} = \frac{t_z^2}{t_z^2 + (n - K)}, \quad (5-26)$$

<sup>3</sup>The general solution given for  $\mathbf{d}_*$  may be usable even if  $\mathbf{X}'\mathbf{X}$  is singular. Suppose, for example, that  $\mathbf{X}'\mathbf{X}$  is  $4 \times 4$  with rank 3. Then  $\mathbf{X}'\mathbf{X}$  is singular. But if there is a parametric restriction on  $\beta$ , then the  $5 \times 5$  matrix in brackets may still have rank 5. This formulation and a number of related results are given in Greene and Seaks (1991).

## CHAPTER 5 ♦ Hypothesis Tests and Model Selection 123

where  $t_z^2$  is the square of the  $t$  ratio for testing the hypothesis that the coefficient on  $z$  is zero in the *multiple* regression of  $\mathbf{y}$  on  $\mathbf{X}$  and  $\mathbf{z}$ . If we solve (5-25) for  $r_{yz}^{*2}$  and (5-26) for  $t_z^2$  and then insert the first solution in the second, then we obtain the result

$$t_z^2 = \frac{(R_{\mathbf{Xz}}^2 - R_{\mathbf{X}}^2)/1}{(1 - R_{\mathbf{Xz}}^2)/(n - K)}. \quad (5-27)$$

We saw at the end of Section 5.4.2 that for a single restriction, such as  $\beta_z = 0$ ,

$$F[1, n - K] = t^2[n - K],$$

which gives us our result. That is, in (5-27), we see that the squared  $t$  statistic (i.e., the  $F$  statistic) is computed using the change in the  $R^2$ . By interpreting the preceding as the result of *removing*  $z$  from the regression, we see that we have proved a result for the case of testing whether a single slope is zero. But the preceding result is general. The test statistic for a single linear restriction is the square of the  $t$  ratio in (5-17). By this construction, we see that for a single restriction,  $F$  is a measure of the loss of fit that results from imposing that restriction. To obtain this result, we will proceed to the general case of  $J$  linear restrictions, which will include one restriction as a special case.

The fit of the restricted least squares coefficients cannot be better than that of the unrestricted solution. Let  $\mathbf{e}_*$  equal  $\mathbf{y} - \mathbf{X}\mathbf{b}_*$ . Then, using a familiar device,

$$\mathbf{e}_* = \mathbf{y} - \mathbf{X}\mathbf{b} - \mathbf{X}(\mathbf{b}_* - \mathbf{b}) = \mathbf{e} - \mathbf{X}(\mathbf{b}_* - \mathbf{b}).$$

The new sum of squared deviations is

$$\mathbf{e}'_* \mathbf{e}_* = \mathbf{e}' \mathbf{e} + (\mathbf{b}_* - \mathbf{b})' \mathbf{X}' \mathbf{X}(\mathbf{b}_* - \mathbf{b}) \geq \mathbf{e}' \mathbf{e}.$$

(The middle term in the expression involves  $\mathbf{X}'\mathbf{e}$ , which is zero.) The loss of fit is

$$\mathbf{e}'_* \mathbf{e}_* - \mathbf{e}' \mathbf{e} = (\mathbf{R}\mathbf{b} - \mathbf{q})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q}). \quad (5-28)$$

This expression appears in the numerator of the  $F$  statistic in (5-7). Inserting the remaining parts, we obtain

$$F[J, n - K] = \frac{(\mathbf{e}'_* \mathbf{e}_* - \mathbf{e}' \mathbf{e})/J}{\mathbf{e}' \mathbf{e}/(n - K)}. \quad (5-29)$$

Finally, by dividing both numerator and denominator of  $F$  by  $\sum_i (y_i - \bar{y})^2$ , we obtain the general result:

$$F[J, n - K] = \frac{(R^2 - R_*^2)/J}{(1 - R^2)/(n - K)}. \quad (5-30)$$

This form has some intuitive appeal in that the difference in the fits of the two models is directly incorporated in the test statistic. As an example of this approach, consider the joint test that all the slopes in the model are zero. This is the overall  $F$  ratio that will be discussed in Section 5.5.3, where  $R_*^2 = 0$ .

For imposing a set of **exclusion restrictions** such as  $\beta_k = 0$  for one or more coefficients, the obvious approach is simply to omit the variables from the regression and base the test on the sums of squared residuals for the restricted and unrestricted regressions. The  $F$  statistic for testing the hypothesis that a subset, say  $\beta_2$ , of the coefficients are all zero is constructed using  $\mathbf{R} = (\mathbf{0} : \mathbf{I})$ ,  $\mathbf{q} = \mathbf{0}$ , and  $J = K_2$  = the number of elements in  $\beta_2$ . The matrix  $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'$  is the  $K_2 \times K_2$  lower right block of the full inverse matrix.

## 124 PART I ♦ The Linear Regression Model

Using our earlier results for partitioned inverses and the results of Section 3.3, we have

$$\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' = (\mathbf{X}'_2\mathbf{M}_1\mathbf{X}_2)^{-1}$$

and

$$\mathbf{R}\mathbf{b} - \mathbf{q} = \mathbf{b}_2.$$

Inserting these in (5-28) gives the loss of fit that results when we drop a subset of the variables from the regression:

$$\mathbf{e}'_*\mathbf{e}_* - \mathbf{e}'\mathbf{e} = \mathbf{b}'_2\mathbf{X}'_2\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2.$$

The procedure for computing the appropriate  $F$  statistic amounts simply to comparing the sums of squared deviations from the “short” and “long” regressions, which we saw earlier.

### Example 5.4 Production Function

The data in Appendix Table F5.3 have been used in several studies of production functions.<sup>4</sup> Least squares regression of log output (value added) on a constant and the logs of labor and capital produce the estimates of a Cobb–Douglas production function shown in Table 5.3. We will construct several hypothesis tests based on these results. A generalization of the Cobb–Douglas model is the *translog* model,<sup>5</sup> which is

$$\ln Y = \beta_1 + \beta_2 \ln L + \beta_3 \ln K + \beta_4 \left( \frac{1}{2} \ln^2 L \right) + \beta_5 \left( \frac{1}{2} \ln^2 K \right) + \beta_6 \ln L \ln K + \varepsilon.$$

As we shall analyze further in Chapter 10, this model differs from the Cobb–Douglas model in that it relaxes the Cobb–Douglas’s assumption of a unitary elasticity of substitution. The Cobb–Douglas model is obtained by the restriction  $\beta_4 = \beta_5 = \beta_6 = 0$ . The results for the two regressions are given in Table 5.3. The  $F$  statistic for the hypothesis of a Cobb–Douglas model is

$$F[3, 21] = \frac{(0.85163 - 0.67993)/3}{0.67993/21} = 1.768.$$

The critical value from the  $F$  table is 3.07, so we would not reject the hypothesis that a Cobb–Douglas model is appropriate.

The hypothesis of constant returns to scale is often tested in studies of production. This hypothesis is equivalent to a restriction that the two coefficients of the Cobb–Douglas production function sum to 1. For the preceding data,

$$F[1, 24] = \frac{(0.6030 + 0.3757 - 1)^2}{0.01586 + 0.00728 - 2(0.00961)} = 0.1157,$$

which is substantially less than the 95 percent critical value of 4.26. We would not reject the hypothesis; the data are consistent with the hypothesis of constant returns to scale. The equivalent test for the translog model would be  $\beta_2 + \beta_3 = 1$  and  $\beta_4 + \beta_5 + 2\beta_6 = 0$ . The  $F$  statistic with 2 and 21 degrees of freedom is 1.8991, which is less than the critical value of 3.47. Once again, the hypothesis is not rejected.

In most cases encountered in practice, it is possible to incorporate the restrictions of a hypothesis directly on the regression and estimate a restricted model.<sup>6</sup> For example, to

<sup>4</sup>The data are statewide observations on SIC 33, the primary metals industry. They were originally constructed by Hildebrand and Liu (1957) and have subsequently been used by a number of authors, notably Aigner, Lovell, and Schmidt (1977). The 28th data point used in the original study is incomplete; we have used only the remaining 27.



<sup>5</sup>Berndt and Christensen (1973). See Example 2.4 and Section 10.4.2 for discussion.

<sup>6</sup>This case is not true when the restrictions are nonlinear. We consider this issue in Chapter 7.

CHAPTER 5 ♦ Hypothesis Tests and Model Selection **125****TABLE 5.3** Estimated Production Functions

	<i>Translog</i>		<i>Cobb-Douglas</i>			
<i>Variable</i>	<i>Coefficient</i>	<i>Standard Error</i>	<i>t Ratio</i>	<i>Coefficient</i>	<i>Standard Error</i>	<i>t Ratio</i>
Sum of squared residuals	0.67993			0.85163		
Standard error of regression	0.17994			0.18837		
<i>R</i> -squared	0.95486			0.94346		
Adjusted <i>R</i> -squared	0.94411			0.93875		
Number of observations	27			27		
<i>Constant</i>	0.944196	2.911	0.324	1.171	0.3268	3.582
$\ln L$	3.61364	1.548	2.334	0.6030	0.1260	4.787
$\ln K$	-1.89311	1.016	-1.863	0.3757	0.0853	4.402
$\frac{1}{2} \ln^2 L$	-0.96405	0.7074	-1.363			
$\frac{1}{2} \ln^2 K$	0.08529	0.2926	0.291			
$\ln L \times \ln K$	0.31239	0.4389	0.712			

<i>Estimated Covariance Matrix for Translog (Cobb-Douglas) Coefficient Estimates</i>						
	<i>Constant</i>	$\ln L$	$\ln K$	$\frac{1}{2} \ln^2 L$	$\frac{1}{2} \ln^2 K$	$\ln L \ln K$
<b><i>Constant</i></b>	8.472 (0.1068)					
<b><math>\ln L</math></b>	-2.388 (-0.01984)	2.397 (0.01586)				
<b><math>\ln K</math></b>	-0.3313 (0.001189)	-1.231 (-0.00961)	1.033 (0.00728)			
<b><math>\frac{1}{2} \ln^2 L</math></b>	-0.08760	-0.6658	0.5231	0.5004		
<b><math>\frac{1}{2} \ln^2 K</math></b>	-0.2332	0.03477	0.02637	0.1467	0.08562	
<b><math>\ln L \ln K</math></b>	0.3635	0.1831	-0.2255	-0.2880	-0.1160	0.1927

impose the constraint  $\beta_2 = 1$  on the Cobb–Douglas model, we would write

$$\ln Y = \beta_1 + 1.0 \ln L + \beta_3 \ln K + \varepsilon$$

or

$$\ln Y - \ln L = \beta_1 + \beta_3 \ln K + \varepsilon.$$

Thus, the restricted model is estimated by regressing  $\ln Y - \ln L$  on a constant and  $\ln K$ . Some care is needed if this regression is to be used to compute an *F* statistic. If the *F* statistic is computed using the sum of squared residuals [see (5-29)], then no problem will arise. If (5-30) is used instead, however, then it may be necessary to account for the restricted regression having a different dependent variable from the unrestricted one. In the preceding regression, the dependent variable in the unrestricted regression is  $\ln Y$ , whereas in the restricted regression, it is  $\ln Y - \ln L$ . The  $R^2$  from the restricted regression is only 0.26979, which would imply an *F* statistic of 285.96, whereas the correct value is 9.935. If we compute the appropriate  $R^2_*$  using the correct denominator, however, then its value is 0.92006 and the correct *F* value results.

Note that the coefficient on  $\ln K$  is negative in the translog model. We might conclude that the estimated output elasticity with respect to capital now has the wrong sign. This conclusion would be incorrect, however; in the translog model, the capital elasticity of output is

$$\frac{\partial \ln Y}{\partial \ln K} = \beta_3 + \beta_5 \ln K + \beta_6 \ln L.$$

## 126 PART I ♦ The Linear Regression Model

If we insert the coefficient estimates and the mean values for  $\ln K$  and  $\ln L$  (not the logs of the means) of 7.44592 and 5.7637, respectively, then the result is 0.5425, which is quite in line with our expectations and is fairly close to the value of 0.3757 obtained for the Cobb-Douglas model. The estimated standard error for this linear combination of the least squares estimates is computed as the square root of

$$\text{Est. Var}[b_3 + b_5 \bar{\ln} K + b_6 \bar{\ln} L] = \mathbf{w}'(\text{Est. Var}[\mathbf{b}])\mathbf{w},$$

where

$$\mathbf{w} = (0, 0, 1, 0, \bar{\ln} K, \bar{\ln} L)'$$

and  $\mathbf{b}$  is the full  $6 \times 1$  least squares coefficient vector. This value is 0.1122, which is reasonably close to the earlier estimate of 0.0853.

### 5.5.3 TESTING THE SIGNIFICANCE OF THE REGRESSION

A question that is usually of interest is whether the regression equation as a whole is significant. This test is a joint test of the hypotheses that *all* the coefficients except the constant term are zero. If all the slopes are zero, then the multiple correlation coefficient,  $R^2$ , is zero as well, so we can base a test of this hypothesis on the value of  $R^2$ . The central result needed to carry out the test is given in (5-30). This is the special case with  $R_*^2 = 0$ , so the  $F$  statistic, which is usually reported with multiple regression results is

$$F[K - 1, n - K] = \frac{R^2/(K - 1)}{(1 - R^2)/(n - K)}.$$

If the hypothesis that  $\beta_2 = \mathbf{0}$  (the part of  $\beta$  not including the constant) is true and the disturbances are normally distributed, then this statistic has an  $F$  distribution with  $K-1$  and  $n-K$  degrees of freedom. Large values of  $F$  give evidence against the validity of the hypothesis. Note that a large  $F$  is induced by a large value of  $R^2$ . The logic of the test is that the  $F$  statistic is a measure of the loss of fit (namely, all of  $R^2$ ) that results when we impose the restriction that all the slopes are zero. If  $F$  is large, then the hypothesis is rejected.

#### **Example 5.5 F Test for the Earnings Equation**

The  $F$  ratio for testing the hypothesis that the four slopes in the earnings equation in Example 5.2 are all zero is

$$F[4, 423] = \frac{0.040995/(5 - 1)}{(1 - 0.040995)/(428 - 5)} = 4.521,$$

which is far larger than the 95 percent critical value of 2.39. We conclude that the data are inconsistent with the hypothesis that all the slopes in the earnings equation are zero. We might have expected the preceding result, given the substantial  $t$  ratios presented earlier. But this case need not always be true. Examples can be constructed in which the individual coefficients are statistically significant, while jointly they are not. This case can be regarded as pathological, but the opposite one, in which none of the coefficients is significantly different from zero while  $R^2$  is highly significant, is relatively common. The problem is that the interaction among the variables may serve to obscure their individual contribution to the fit of the regression, whereas their joint effect may still be significant.

### 5.5.4 SOLVING OUT THE RESTRICTIONS AND A CAUTION ABOUT USING $R^2$

In principle, one can usually solve out the restrictions imposed by a linear hypothesis. Algebraically, we would begin by partitioning  $\mathbf{R}$  into two groups of columns, one with

## CHAPTER 5 ♦ Hypothesis Tests and Model Selection 127

$J$  and one with  $K - J$ , so that the first set are linearly independent. (There are many ways to do so; any one will do for the present.) Then, with  $\beta$  likewise partitioned and its elements reordered in whatever way is needed, we may write

$$\mathbf{R}\beta = \mathbf{R}_1\beta_1 + \mathbf{R}_2\beta_2 = \mathbf{q}.$$

If the  $J$  columns of  $\mathbf{R}_1$  are independent, then

$$\beta_1 = \mathbf{R}_1^{-1}[\mathbf{q} - \mathbf{R}_2\beta_2].$$

This suggests that one might estimate the restricted model directly using a transformed equation, rather than use the rather cumbersome restricted estimator shown in (5-23). A simple example illustrates. Consider imposing constant returns to scale on a two input production function,

$$\ln y = \beta_1 + \beta_2 \ln x_1 + \beta_3 \ln x_2 + \varepsilon.$$

The hypothesis of linear homogeneity is  $\beta_2 + \beta_3 = 1$  or  $\beta_3 = 1 - \beta_2$ . Simply building the restriction into the model produces

$$\ln y = \beta_1 + \beta_2 \ln x_1 + (1 - \beta_2) \ln x_2 + \varepsilon$$

or

$$\ln y = \ln x_2 + \beta_1 + \beta_2(\ln x_1 - \ln x_2) + \varepsilon.$$

One can obtain the restricted least squares estimates by linear regression of  $(\ln y - \ln x_2)$  on a constant and  $(\ln x_1 - \ln x_2)$ . However, the test statistic for the hypothesis cannot be tested using the familiar result in (5-30), because the denominators in the two  $R^2$ 's are different. The statistic in (5-30) could even be negative. The appropriate approach would be to use the equivalent, but appropriate computation based on the sum of squared residuals in . The general result from this example is that one must be careful in using (5-30) and that the dependent variable in the two regressions must be the same.

## 5.6 NONNORMAL DISTURBANCES AND LARGE-SAMPLE TESTS

We now consider the relation between the sample test statistics and the data in  $\mathbf{X}$ . First, consider the conventional  $t$  statistic in (4-41) for testing  $H_0 : \beta_k = \beta_k^0$ ,

$$t|\mathbf{X} = \frac{b_k - \beta_k^0}{\sqrt{s^2(\mathbf{X}'\mathbf{X})_{kk}^{-1}}}.$$

Conditional on  $\mathbf{X}$ , if  $\beta_k = \beta_k^0$  (i.e., under  $H_0$ ), then  $t|\mathbf{X}$  has a  $t$  distribution with  $(n - K)$  degrees of freedom. What interests us, however, is the marginal, that is, the unconditional distribution of  $t$ . As we saw,  $\mathbf{b}$  is only normally distributed conditionally on  $\mathbf{X}$ ; the marginal distribution may not be normal because it depends on  $\mathbf{X}$  (through the conditional variance). Similarly, because of the presence of  $\mathbf{X}$ , the denominator of the  $t$  statistic is not the square root of a chi-squared variable divided by its degrees of freedom, again, except conditional on this  $\mathbf{X}$ . But, because the distributions of  $(b_k - \beta_k)/\sqrt{s^2(\mathbf{X}'\mathbf{X})_{kk}^{-1}}|\mathbf{X}$  and  $[(n - K)s_2/\sigma^2]|\mathbf{X}$  are still independent  $N[0, 1]$  and

## 128 PART I ♦ The Linear Regression Model

$\chi^2[n - K]$ , respectively, which do not involve  $\mathbf{X}$ , we have the surprising result that, regardless of the distribution of  $\mathbf{X}$ , or even of whether  $\mathbf{X}$  is stochastic or nonstochastic, the marginal distributions of  $t$  is still  $t$ , even though the final distribution of  $b_k$  may be nonnormal. This intriguing result follows because  $f(t | \mathbf{X})$  is not a function of  $\mathbf{X}$ . The same reasoning can be used to deduce that the usual  $F$  ratio used for testing linear restrictions, discussed in the previous section, is valid whether  $\mathbf{X}$  is stochastic or not. This result is very powerful. The implication is that *if the disturbances are normally distributed, then we may carry out tests and construct confidence intervals for the parameters without making any changes in our procedures, regardless of whether the regressors are stochastic, nonstochastic, or some mix of the two.*

The distributions of these statistics do follow from the normality assumption for  $\boldsymbol{\epsilon}$ , but they do not depend on  $\mathbf{X}$ . Without the normality assumption, however, the exact distributions of these statistics depend on the data and the parameters and are not  $F$ ,  $t$ , and chi-squared. At least at first blush, it would seem that we need either a new set of critical values for the tests or perhaps a new set of test statistics. In this section, we will examine results that will generalize the familiar procedures. These large-sample results suggest that although the usual  $t$  and  $F$  statistics are still usable, in the more general case without the special assumption of normality, they are viewed as approximations whose quality improves as the sample size increases. By using the results of Section D.3 (on asymptotic distributions) and some large-sample results for the least squares estimator, we can construct a set of usable inference procedures based on already familiar computations.

Assuming the data are well behaved, the *asymptotic* distribution of the least squares coefficient estimator,  $\mathbf{b}$ , is given by

$$\mathbf{b} \xrightarrow{a} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n} \mathbf{Q}^{-1}\right] \quad \text{where } \mathbf{Q} = \text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right). \quad (5-31)$$

The interpretation is that, absent normality of  $\boldsymbol{\epsilon}$ , as the sample size,  $n$ , grows, the normal distribution becomes an increasingly better approximation to the true, though at this point unknown, distribution of  $\mathbf{b}$ . As  $n$  increases, the distribution of  $\sqrt{n}(\mathbf{b} - \boldsymbol{\beta})$  converges exactly to a normal distribution, which is how we obtain the preceding finite-sample approximation. This result is based on the central limit theorem and does not require normally distributed disturbances. The second result we will need concerns the estimator of  $\sigma^2$ :

$$\text{plim } s^2 = \sigma^2, \quad \text{where } s^2 = \mathbf{e}'\mathbf{e}/(n - K).$$

With these in place, we can obtain some large-sample results for our test statistics that suggest how to proceed in a finite sample with nonnormal disturbances.

The sample statistic for testing the hypothesis that one of the coefficients,  $\beta_k$  equals a particular value,  $\beta_k^0$  is

$$t_k = \frac{\sqrt{n}(b_k - \beta_k^0)}{\sqrt{s^2(\mathbf{X}'\mathbf{X}/n)_{kk}}}.$$

(Note that two occurrences of  $\sqrt{n}$  cancel to produce our familiar result.) Under the null hypothesis, with normally distributed disturbances,  $t_k$  is exactly distributed as  $t$  with  $n - K$  degrees of freedom. [See Theorem 4.4 and the beginning of this section.] The

CHAPTER 5 ♦ Hypothesis Tests and Model Selection **129**

exact distribution of this statistic is unknown, however, if  $\epsilon$  is not normally distributed. From the preceding results, we find that the denominator of  $t_k$  converges to  $\sqrt{\sigma^2 \mathbf{Q}_{kk}^{-1}}$ . Hence, if  $t_k$  has a limiting distribution, then it is the same as that of the statistic that has this latter quantity in the denominator. (See point 3 Theorem D.16.) That is, the large-sample distribution of  $t_k$  is the same as that of

$$\tau_k = \frac{\sqrt{n}(b_k - \beta_k^0)}{\sqrt{\sigma^2 \mathbf{Q}_{kk}^{-1}}}.$$

But  $\tau_k = (b_k - E[b_k]) / (\text{Asy. Var}[b_k])^{1/2}$  from the asymptotic normal distribution (under the hypothesis  $\beta_k = \beta_k^0$ ), so it follows that  $\tau_k$  has a standard normal asymptotic distribution, and this result is the large-sample distribution of our  $t$  statistic. Thus, as a large-sample approximation, we will use the standard normal distribution to approximate the true distribution of the test statistic  $t_k$  and use the critical values from the standard normal distribution for testing hypotheses.

The result in the preceding paragraph is valid only in large samples. For moderately sized samples, it provides only a suggestion that the  $t$  distribution may be a reasonable approximation. The appropriate critical values only *converge* to those from the standard normal, and generally *from above*, although we cannot be sure of this. In the interest of conservatism—that is, in controlling the probability of a Type I error—one should generally use the critical value from the  $t$  distribution even in the absence of normality. Consider, for example, using the standard normal critical value of 1.96 for a two-tailed test of a hypothesis based on 25 degrees of freedom. The nominal size of this test is 0.05. The actual size of the test, however, is the true, but unknown, probability that  $|t_k| > 1.96$ , which is 0.0612 if the  $t[25]$  distribution is correct, and some other value if the disturbances are not normally distributed. The end result is that the standard  $t$  test retains a large sample validity. Little can be said about the true size of a test based on the  $t$  distribution unless one makes some other equally narrow assumption about  $\epsilon$ , but the  $t$  distribution is generally used as a reliable approximation.

We will use the same approach to analyze the  $F$  statistic for testing a set of  $J$  linear restrictions. Step 1 will be to show that with normally distributed disturbances,  $JF$  converges to a chi-squared variable as the sample size increases. We will then show that this result is actually independent of the normality of the disturbances; it relies on the central limit theorem. Finally, we consider, as before, the appropriate critical values to use for this test statistic, which only has large sample validity.

The  $F$  statistic for testing the validity of  $J$  linear restrictions,  $\mathbf{R}\beta - \mathbf{q} = \mathbf{0}$ , is given in (5-6). With normally distributed disturbances and under the null hypothesis, the exact distribution of this statistic is  $F[J, n - K]$ . To see how  $F$  behaves more generally, divide the numerator and denominator in (5-16) by  $\sigma^2$  and rearrange the fraction slightly, so

$$F = \frac{(\mathbf{R}\mathbf{b} - \mathbf{q})' \{ \mathbf{R}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}] \mathbf{R}' \}^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q})}{J(s^2/\sigma^2)}. \quad (5-32)$$

Since  $\text{plim } s^2 = \sigma^2$ , and  $\text{plim}(\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}$ , the denominator of  $F$  converges to  $J$  and the bracketed term in the numerator will behave the same as  $(\sigma^2/n)\mathbf{R}\mathbf{Q}^{-1}\mathbf{R}'$ . (See Theorem D16.3.) Hence, regardless of what this distribution is, if  $F$  has a limiting distribution,

## 130 PART I ♦ The Linear Regression Model

then it is the same as the limiting distribution of

$$\begin{aligned} W^* &= \frac{1}{J}(\mathbf{R}\mathbf{b} - \mathbf{q})'[\mathbf{R}(\sigma^2/n)\mathbf{Q}^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}) \\ &= \frac{1}{J}(\mathbf{R}\mathbf{b} - \mathbf{q})'\{\text{Asy. Var}[\mathbf{R}\mathbf{b} - \mathbf{q}]\}^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}). \end{aligned}$$

This expression is  $(1/J)$  times a **Wald statistic**, based on the asymptotic distribution. The large-sample distribution of  $W^*$  will be that of  $(1/J)$  times a chi-squared with  $J$  degrees of freedom. It follows that with normally distributed disturbances,  $JF$  converges to a chi-squared variate with  $J$  degrees of freedom. The proof is instructive. [See White (2001, 9.76).]

### THEOREM 5.1 Limiting Distribution of the Wald Statistic

If  $\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N[\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}]$  and if  $H_0 : \mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$  is true, then

$$W = (\mathbf{R}\mathbf{b} - \mathbf{q})'[\mathbf{R}s^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}) = JF \xrightarrow{d} \chi^2[J].$$

**Proof:** Since  $\mathbf{R}$  is a matrix of constants and  $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$ ,

$$\sqrt{n}\mathbf{R}(\mathbf{b} - \boldsymbol{\beta}) = \sqrt{n}(\mathbf{R}\mathbf{b} - \mathbf{q}) \xrightarrow{d} N[\mathbf{0}, \mathbf{R}(\sigma^2 \mathbf{Q}^{-1})\mathbf{R}']. \quad (1)$$

For convenience, write this equation as

$$\mathbf{z} \xrightarrow{d} N[\mathbf{0}, \mathbf{P}]. \quad (2)$$

In Section A.6.11, we define the inverse square root of a positive definite matrix  $\mathbf{P}$  as another matrix, say  $\mathbf{T}$ , such that  $\mathbf{T}^2 = \mathbf{P}^{-1}$ , and denote  $\mathbf{T}$  as  $\mathbf{P}^{-1/2}$ . Then, by the same reasoning as in (1) and (2),

$$\text{if } \mathbf{z} \xrightarrow{d} N[\mathbf{0}, \mathbf{P}], \text{ then } \mathbf{P}^{-1/2}\mathbf{z} \xrightarrow{d} N[\mathbf{0}, \mathbf{P}^{-1/2}\mathbf{P}\mathbf{P}^{-1/2}] = N[\mathbf{0}, \mathbf{I}]. \quad (3)$$

We now invoke Theorem D.21 for the limiting distribution of a function of a random variable. The sum of squares of uncorrelated (i.e., independent) standard normal variables is distributed as chi-squared. Thus, the limiting distribution of

$$(\mathbf{P}^{-1/2}\mathbf{z})'(\mathbf{P}^{-1/2}\mathbf{z}) = \mathbf{z}'\mathbf{P}^{-1}\mathbf{z} \xrightarrow{d} \chi^2(J). \quad (4)$$

Reassembling the parts from before, we have shown that the limiting distribution of

$$n(\mathbf{R}\mathbf{b} - \mathbf{q})'[\mathbf{R}(\sigma^2 \mathbf{Q}^{-1})\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}) \quad (5)$$

is chi-squared, with  $J$  degrees of freedom. Note the similarity of this result to the results of Section B.11.6. Finally, if

$$\text{plim } s^2 \left( \frac{1}{n} \mathbf{X}'\mathbf{X} \right)^{-1} = \sigma^2 \mathbf{Q}^{-1}, \quad (6)$$

then the statistic obtained by replacing  $\sigma^2 \mathbf{Q}^{-1}$  by  $s^2(\mathbf{X}'\mathbf{X}/n)^{-1}$  in (5) has the same limiting distribution. The  $n$ 's cancel, and we are left with the same Wald statistic we looked at before. This step completes the proof.

CHAPTER 5 ♦ Hypothesis Tests and Model Selection **131**

The appropriate critical values for the  $F$  test of the restrictions  $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$  converge from above to  $1/J$  times those for a chi-squared test based on the Wald statistic (see the Appendix tables). For example, for testing  $J = 5$  restrictions, the critical value from the chi-squared table (Appendix Table G.4) for 95 percent significance is 11.07. The critical values from the  $F$  table (Appendix Table G.5) are  $3.33 = 16.65/5$  for  $n - K = 10$ ,  $2.60 = 13.00/5$  for  $n - K = 25$ ,  $2.40 = 12.00/5$  for  $n - K = 50$ ,  $2.31 = 11.55/5$  for  $n - K = 100$ , and  $2.214 = 11.07/5$  for large  $n - K$ . Thus, with normally distributed disturbances, as  $n$  gets large, the  $F$  test can be carried out by referring  $JF$  to the critical values from the chi-squared table.

The crucial result for our purposes here is that the distribution of the Wald statistic is built up from the distribution of  $\mathbf{b}$ , which is asymptotically normal even without normally distributed disturbances. The implication is that an appropriate large sample test statistic is chi-squared =  $JF$ . Once again, this implication relies on the central limit theorem, not on normally distributed disturbances. Now, what is the appropriate approach for a small or moderately sized sample? As we saw earlier, the critical values for the  $F$  distribution converge from above to  $(1/J)$  times those for the preceding chi-squared distribution. As before, one cannot say that this will always be true in every case for every possible configuration of the data and parameters. Without some special configuration of the data and parameters, however, one can expect it to occur generally. The implication is that absent some additional firm characterization of the model, the  $F$  statistic, with the critical values from the  $F$  table, remains a conservative approach that becomes more accurate as the sample size increases.

Exercise 7 at the end of this chapter suggests another approach to testing that has validity in large samples, a **Lagrange multiplier test**. The vector of Lagrange multipliers in (5-23) is  $[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q})$ , that is, a multiple of the least squares discrepancy vector. In principle, a test of the hypothesis that  $\lambda_*$  equals zero should be equivalent to a test of the null hypothesis. Since the leading matrix has full rank, this can only equal zero if the discrepancy equals zero. A Wald test of the hypothesis that  $\lambda_* = \mathbf{0}$  is indeed a valid way to proceed. The large sample distribution of the Wald statistic would be chi-squared with  $J$  degrees of freedom. (The procedure is considered in Exercise 7.) For a set of exclusion restrictions,  $\boldsymbol{\beta}_2 = \mathbf{0}$ , there is a simple way to carry out this test. The chi-squared statistic, in this case with  $K_2$  degrees of freedom can be computed as  $nR^2$  in the regression of  $\mathbf{e}_*$  (the residuals in the short regression) on the full set of independent variables.

## 5.7 TESTING NONLINEAR RESTRICTIONS

The preceding discussion has relied heavily on the linearity of the regression model. When we analyze nonlinear functions of the parameters and nonlinear regression models, most of these exact distributional results no longer hold.

The general problem is that of testing a hypothesis that involves a nonlinear function of the regression coefficients:

$$H_0: c(\boldsymbol{\beta}) = q.$$

We shall look first at the case of a single restriction. The more general one, in which  $\mathbf{c}(\boldsymbol{\beta}) = \mathbf{q}$  is a set of restrictions, is a simple extension. The counterpart to the test statistic

## 132 PART I ♦ The Linear Regression Model

we used earlier would be

$$z = \frac{c(\hat{\beta}) - q}{\text{estimated standard error}} \quad (5-33)$$

or its square, which in the preceding were distributed as  $t[n - K]$  and  $F[1, n - K]$ , respectively. The discrepancy in the numerator presents no difficulty. Obtaining an estimate of the sampling variance of  $c(\hat{\beta}) - q$ , however, involves the variance of a nonlinear function of  $\hat{\beta}$ .

The results we need for this computation are presented in Sections 4.4.4, B.10.3, and D.3.1. A linear Taylor series approximation to  $c(\hat{\beta})$  around the true parameter vector  $\beta$  is

$$c(\hat{\beta}) \approx c(\beta) + \left( \frac{\partial c(\beta)}{\partial \beta} \right)' (\hat{\beta} - \beta). \quad (5-34)$$

We must rely on consistency rather than unbiasedness here, since, in general, the expected value of a nonlinear function is not equal to the function of the expected value. If  $\text{plim } \hat{\beta} = \beta$ , then we are justified in using  $c(\hat{\beta})$  as an estimate of  $c(\beta)$ . (The relevant result is the Slutsky theorem.) Assuming that our use of this approximation is appropriate, the variance of the nonlinear function is approximately equal to the variance of the right-hand side, which is, then,

$$\text{Var}[c(\hat{\beta})] \approx \left( \frac{\partial c(\beta)}{\partial \beta} \right)' \text{Var}[\hat{\beta}] \left( \frac{\partial c(\beta)}{\partial \beta} \right). \quad (5-35)$$

The derivatives in the expression for the variance are functions of the unknown parameters. Since these are being estimated, we use our sample estimates in computing the derivatives. To estimate the variance of the estimator, we can use  $s^2(\mathbf{X}'\mathbf{X})^{-1}$ . Finally, we rely on Theorem D.22 in Section D.3.1 and use the standard normal distribution instead of the  $t$  distribution for the test statistic. Using  $\mathbf{g}(\hat{\beta})$  to estimate  $\mathbf{g}(\beta) = \partial c(\beta)/\partial \beta$ , we can now test a hypothesis in the same fashion we did earlier.

### Example 5.6 A Long-Run Marginal Propensity to Consume

A consumption function that has different short- and long-run marginal propensities to consume can be written in the form

$$\ln C_t = \alpha + \beta \ln Y_t + \gamma \ln C_{t-1} + \varepsilon_t,$$

which is a **distributed lag** model. In this model, the short-run marginal propensity to consume (MPC) (elasticity, since the variables are in logs) is  $\beta$ , and the long-run MPC is  $\delta = \beta/(1 - \gamma)$ . Consider testing the hypothesis that  $\delta = 1$ .

Quarterly data on aggregate U.S. consumption and disposable personal income for the years 1950 to 2000 are given in Appendix Table F5.2. The estimated equation based on these data is

$$\begin{aligned} \ln C_t &= 0.003142 + 0.07495 \ln Y_t + 0.9246 \ln C_{t-1} + \varepsilon_t, & R^2 &= 0.999712, & s &= 0.00874 \\ (0.01055) & (0.02873) & (0.02859) & & & \end{aligned}$$

Estimated standard errors are shown in parentheses. We will also require  $\text{Est. Cov}[b, c] = -0.0008207$ . The estimate of the long-run MPC is  $d = b/(1 - c) = 0.07495/(1 - 0.9246) = 0.99403$ . To compute the estimated variance of  $d$ , we will require

$$g_b = \frac{\partial d}{\partial b} = \frac{1}{1 - c} = 13.2626, \quad g_c = \frac{\partial d}{\partial c} = \frac{b}{(1 - c)^2} = 13.1834.$$

CHAPTER 5 ♦ Hypothesis Tests and Model Selection **133**

The estimated asymptotic variance of  $d$  is

$$\begin{aligned}\text{Est. Asy. Var}[d] &= g_b^2 \text{Est. Asy. Var}[b] + g_c^2 \text{Est. Asy. Var}[c] + 2g_b g_c \text{Est. Asy. Cov}[b, c] \\ &= 13.2626^2 \times 0.02873^2 + 13.1834^2 \times 0.02859^2 \\ &\quad + 2(13.2626)(13.1834)(-0.0008207) = 0.0002585.\end{aligned}$$

The square root is 0.016078. To test the hypothesis that the long-run MPC is greater than or equal to 1, we would use

$$z = \frac{0.99403 - 1}{0.016078} = -0.37131.$$

Because we are using a large sample approximation, we refer to a standard normal table instead of the  $t$  distribution. The hypothesis that  $\gamma = 1$  is not rejected.

You may have noticed that we could have tested this hypothesis with a linear restriction instead; if  $\delta = 1$ , then  $\beta = 1 - \gamma$ , or  $\beta + \gamma = 1$ . The estimate is  $q = b + c - 1 = -0.00045$ . The estimated standard error of this linear function is  $[0.02873^2 + 0.02859^2 - 2(0.0008207)]^{1/2} = 0.00118$ . The  $t$  ratio for this test is  $-0.38135$ , which is almost the same as before. Since the sample used here is fairly large, this is to be expected. However, there is nothing in the computations that ensures this outcome. In a smaller sample, we might have obtained a different answer. For example, using the last 11 years of the data, the  $t$  statistics for the two hypotheses are 7.652 and 5.681. The Wald test is not invariant to how the hypothesis is formulated. In a borderline case, we could have reached a different conclusion. This **lack of invariance** does not occur with the likelihood ratio or Lagrange multiplier tests discussed in Chapter 16.<sup>7</sup> On the other hand, both of these tests require an assumption of normality, whereas the Wald statistic does not. This illustrates one of the trade-offs between a more detailed specification and the power of the test procedures that are implied.

The generalization to more than one function of the parameters proceeds along similar lines. Let  $\mathbf{c}(\hat{\beta})$  be a set of  $J$  functions of the estimated parameter vector and let the  $J \times K$  matrix of derivatives of  $\mathbf{c}(\hat{\beta})$  be

$$\hat{\mathbf{G}} = \frac{\partial \mathbf{c}(\hat{\beta})}{\partial \hat{\beta}'}. \quad (5-36)$$

The estimate of the asymptotic covariance matrix of these functions is

$$\text{Est. Asy. Var}[\hat{\mathbf{c}}] = \hat{\mathbf{G}} \{ \text{Est. Asy. Var}[\hat{\beta}] \} \hat{\mathbf{G}}'. \quad (5-37)$$

The  $j$ th row of  $\hat{\mathbf{G}}$  is  $K$  derivatives of  $c_j$  with respect to the  $K$  elements of  $\hat{\beta}$ . For example, the covariance matrix for estimates of the short- and long-run marginal propensities to consume would be obtained using

$$\mathbf{G} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1/(1-\gamma) & \beta/(1-\gamma)^2 \end{bmatrix}.$$

The statistic for testing the  $J$  hypotheses  $\mathbf{c}(\beta) = \mathbf{q}$  is

$$W = (\hat{\mathbf{c}} - \mathbf{q})' \{ \text{Est. Asy. Var}[\hat{\mathbf{c}}] \}^{-1} (\hat{\mathbf{c}} - \mathbf{q}). \quad (5-38)$$

In large samples,  $W$  has a chi-squared distribution with degrees of freedom equal to the number of restrictions. Note that for a single restriction, this value is the square of the statistic in (5-33).

## 134 PART I ♦ The Linear Regression Model

### 5.8 CHOOSING BETWEEN NONNESTED MODELS

The classical testing procedures that we have been using have been shown to be most powerful for the types of hypotheses we have considered.<sup>7</sup> Although use of these procedures is clearly desirable, the requirement that we express the hypotheses in the form of restrictions on the model  $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$ ,

$$H_0 : \mathbf{R}\beta = \mathbf{q}$$

versus

$$H_1 : \mathbf{R}\beta \neq \mathbf{q},$$

can be limiting. Two common exceptions are the general problem of determining which of two possible sets of regressors is more appropriate and whether a linear or loglinear model is more appropriate for a given analysis. For the present, we are interested in comparing two competing linear models:

$$H_0 : \mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}_0 \quad (5-39a)$$

and

$$H_1 : \mathbf{y} = \mathbf{Z}\gamma + \boldsymbol{\varepsilon}_1. \quad (5-39b)$$

The classical procedures we have considered thus far provide no means of forming a preference for one model or the other. The general problem of testing nonnested hypotheses such as these has attracted an impressive amount of attention in the theoretical literature and has appeared in a wide variety of empirical applications.<sup>8</sup>

#### 5.8.1 TESTING NONNESTED HYPOTHESES

A useful distinction between hypothesis testing as discussed in the preceding chapters and model selection as considered here will turn on the asymmetry between the null and alternative hypotheses that is a part of the classical testing procedure.<sup>9</sup> Because, by construction, the classical procedures seek evidence in the sample to refute the “null” hypothesis, how one frames the null can be crucial to the outcome. Fortunately, the Neyman–Pearson methodology provides a prescription; the null is usually cast as the narrowest model in the set under consideration. On the other hand, the classical procedures never reach a sharp conclusion. Unless the significance level of the testing procedure is made so high as to exclude all alternatives, there will always remain the possibility of a Type 1 error. As such, the null hypothesis is never rejected with certainty, but only with a prespecified degree of confidence. Model selection tests, in contrast, give the competing hypotheses equal standing. There is no natural null hypothesis. However, the end of the process is a firm decision—in testing (5-39a, b), one of the models will be rejected and the other will be retained; the analysis will then proceed in

<sup>7</sup>See, for example, Stuart and Ord (1989, Chap. 27).

<sup>8</sup>Surveys on this subject are White (1982a, 1983), Gourieroux and Monfort (1994), McAleer (1995), and Pesaran and Weeks (2001). McAleer’s survey tabulates an array of applications, while Gourieroux and Monfort focus on the underlying theory.

<sup>9</sup>See Granger and Pesaran (2000) for discussion.

CHAPTER 5 ♦ Hypothesis Tests and Model Selection **135**

the framework of that one model and not the other. Indeed, it cannot proceed until one of the models is discarded. It is common, for example, in this new setting for the analyst first to test with one model cast as the null, then with the other. Unfortunately, given the way the tests are constructed, it can happen that both or neither model is rejected; in either case, further analysis is clearly warranted. As we shall see, the science is a bit inexact.

The earliest work on nonnested hypothesis testing, notably Cox (1961, 1962), was done in the framework of sample likelihoods and maximum likelihood procedures. Recent developments have been structured around a common pillar labeled the **encompassing principle** [Mizon and Richard (1986)]. In the large, the principle directs attention to the question of whether a maintained model can explain the features of its competitors, that is, whether the maintained model encompasses the alternative. Yet a third approach is based on forming a **comprehensive model** that contains both competitors as special cases. When possible, the test between models can be based, essentially, on classical (-like) testing procedures. We will examine tests that exemplify all three approaches.

### 5.8.2 AN ENCOMPASSING MODEL

The encompassing approach is one in which the ability of one model to explain features of another is tested. Model 0 “encompasses” Model 1 if the features of Model 1 can be explained by Model 0, but the reverse is not true.<sup>10</sup> Because  $H_0$  cannot be written as a restriction on  $H_1$ , none of the procedures we have considered thus far is appropriate. One possibility is an artificial nesting of the two models. Let  $\bar{\mathbf{X}}$  be the set of variables in  $\mathbf{X}$  that are not in  $\mathbf{Z}$ , define  $\bar{\mathbf{Z}}$  likewise with respect to  $\mathbf{Z}$ , and let  $\mathbf{W}$  be the variables that the models have in common. Then  $H_0$  and  $H_1$  could be combined in a “supermodel”:

$$\mathbf{y} = \bar{\mathbf{X}}\bar{\boldsymbol{\beta}} + \bar{\mathbf{Z}}\bar{\boldsymbol{\gamma}} + \mathbf{W}\boldsymbol{\delta} + \boldsymbol{\varepsilon}.$$

In principle,  $H_1$  is rejected if it is found that  $\bar{\boldsymbol{\gamma}} = \mathbf{0}$  by a conventional  $F$  test, whereas  $H_0$  is rejected if it is found that  $\bar{\boldsymbol{\beta}} = \mathbf{0}$ . There are two problems with this approach. First,  $\boldsymbol{\delta}$  remains a mixture of parts of  $\boldsymbol{\beta}$  and  $\boldsymbol{\gamma}$ , and it is not established by the  $F$  test that either of these parts is zero. Hence, this test does not really distinguish between  $H_0$  and  $H_1$ ; it distinguishes between  $H_1$  and a hybrid model. Second, this compound model may have an extremely large number of regressors. In a time-series setting, the problem of collinearity may be severe.

Consider an alternative approach. If  $H_0$  is correct, then  $\mathbf{y}$  will, apart from the random disturbance  $\boldsymbol{\varepsilon}$ , be fully explained by  $\mathbf{X}$ . Suppose we then attempt to estimate  $\boldsymbol{\gamma}$  by regression of  $\mathbf{y}$  on  $\mathbf{Z}$ . Whatever set of parameters is estimated by this regression, say,  $\mathbf{c}$ , if  $H_0$  is correct, then we should estimate exactly the same coefficient vector if we were to regress  $\mathbf{X}\boldsymbol{\beta}$  on  $\mathbf{Z}$ , since  $\boldsymbol{\varepsilon}_0$  is random noise under  $H_0$ . Because  $\boldsymbol{\beta}$  must be estimated, suppose that we use  $\mathbf{X}\mathbf{b}$  instead and compute  $\mathbf{c}_0$ . A test of the proposition that Model 0 “encompasses” Model 1 would be a test of the hypothesis that  $E[\mathbf{c} - \mathbf{c}_0] = \mathbf{0}$ . It is straightforward to show [see Davidson and MacKinnon (2004, pp. 671–672)] that the test can be carried out by using a standard  $F$  test to test the hypothesis that  $\boldsymbol{\gamma}_1 = \mathbf{0}$ .

<sup>10</sup>See Deaton (1982), Dastoor (1983), Gourieroux et al. (1983, 1995) and, especially, Mizon and Richard (1986).

## 136 PART I ♦ The Linear Regression Model

in the augmented regression,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}_1,$$

where  $\mathbf{Z}_1$  is the variables in  $\mathbf{Z}$  that are not in  $\mathbf{X}$ . (Of course, a line of manipulation reveals that  $\bar{\mathbf{Z}}$  and  $\mathbf{Z}_1$  are the same, so the tests are also.)

### 5.8.3 COMPREHENSIVE APPROACH—THE J TEST

The underpinnings of the comprehensive approach are tied to the density function as the characterization of the data generating process. Let  $f_0(y_i | \text{data}, \boldsymbol{\beta}_0)$  be the assumed density under Model 0 and define the alternative likewise as  $f_1(y_i | \text{data}, \boldsymbol{\beta}_1)$ . Then, a comprehensive model which subsumes both of these is

$$f_c(y_i | \text{data}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1) = \frac{[f_0(y_i | \text{data}, \boldsymbol{\beta}_0)]^{1-\lambda}[f_1(y_i | \text{data}, \boldsymbol{\beta}_1)]^\lambda}{\int_{\text{range of } y_i} [f_0(y_i | \text{data}, \boldsymbol{\beta}_0)]^{1-\lambda}[f_1(y_i | \text{data}, \boldsymbol{\beta}_1)]^\lambda dy_i}.$$

Estimation of the comprehensive model followed by a test of  $\lambda = 0$  or 1 is used to assess the validity of Model 0 or 1, respectively.<sup>11</sup>

The **J test** proposed by Davidson and MacKinnon (1981) can be shown [see Pesaran and Weeks (2001)] to be an application of this principle to the linear regression model. Their suggested alternative to the preceding compound model is

$$\mathbf{y} = (1 - \lambda)\mathbf{X}\boldsymbol{\beta} + \lambda(\mathbf{Z}\boldsymbol{\gamma}) + \boldsymbol{\varepsilon}.$$

In this model, a test of  $\lambda = 0$  would be a test against  $H_1$ . The problem is that  $\lambda$  cannot be separately estimated in this model; it would amount to a redundant scaling of the regression coefficients. Davidson and MacKinnon's *J* test consists of estimating  $\boldsymbol{\gamma}$  by a least squares regression of  $\mathbf{y}$  on  $\mathbf{Z}$  followed by a least squares regression of  $\mathbf{y}$  on  $\mathbf{X}$  and  $\mathbf{Z}\hat{\boldsymbol{\gamma}}$ , the fitted values in the first regression. A valid test, at least asymptotically, of  $H_1$  is to test  $H_0 : \lambda = 0$ . If  $H_0$  is true, then  $\text{plim } \hat{\lambda} = 0$ . Asymptotically, the ratio  $\hat{\lambda}/\text{se}(\hat{\lambda})$  (i.e., the usual *t* ratio) is distributed as standard normal and may be referred to the standard table to carry out the test. Unfortunately, in testing  $H_0$  versus  $H_1$  and vice versa, all four possibilities (reject both, neither, or either one of the two hypotheses) could occur. This issue, however, is a finite sample problem. Davidson and MacKinnon show that as  $n \rightarrow \infty$ , if  $H_1$  is true, then the probability that  $\hat{\lambda}$  will differ significantly from 0 approaches 1.

#### **Example 5.7 J Test for a Consumption Function**

Gaver and Geisel (1974) propose two forms of a consumption function:

$$H_0 : C_t = \beta_1 + \beta_2 Y_t + \beta_3 Y_{t-1} + \varepsilon_{0t}$$

and

$$H_1 : C_t = \gamma_1 + \gamma_2 Y_t + \gamma_3 C_{t-1} + \varepsilon_{1t}.$$

The first model states that consumption responds to changes in income over two periods, whereas the second states that the effects of changes in income on consumption persist for many periods. Quarterly data on aggregate U.S. real consumption and real disposable income are given in Appendix Table F5.2. Here we apply the *J* test to these data and the two proposed specifications. First, the two models are estimated separately (using observations

<sup>11</sup>Silva (2001) presents an application to the choice of probit or logit model for binary choice.

## CHAPTER 5 ♦ Hypothesis Tests and Model Selection 137

1950.2 through 2000.4). The least squares regression of  $C$  on a constant,  $Y$ , lagged  $Y$ , and the fitted values from the second model produces an estimate of  $\lambda$  of 1.0145 with a  $t$  ratio of 62.861. Thus,  $H_0$  should be rejected in favor of  $H_1$ . But reversing the roles of  $H_0$  and  $H_1$ , we obtain an estimate of  $\lambda$  of  $-10.677$  with a  $t$  ratio of  $-7.188$ . Thus,  $H_1$  is rejected as well.<sup>12</sup>

### 5.9 A SPECIFICATION TEST

The tests considered so far have evaluated nested models. The presumption is that one of the two models is correct. In Section 5.8, we broadened the range of models considered to allow two nonnested models. It is not assumed that either model is necessarily the true data generating process; the test attempts to ascertain which of two competing models is closer to the truth. Specification tests fall between these two approaches. The idea of a **specification test** is to consider a particular null model and alternatives that are not explicitly given in the form of restrictions on the regression equation. A useful way to consider some specification tests is as if the core model,  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$  is the null hypothesis and the alternative is a possibly unstated generalization of that model. Ramsey's (1969) **RESET test** is one such test which seeks to uncover nonlinearities in the functional form. One (admittedly ambiguous) way to frame the analysis is

$$H_0: \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$H_1: \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \text{higher order powers of } x_k \text{ and other terms} + \boldsymbol{\varepsilon}.$$

A straightforward approach would be to add squares, cubes, and cross products of the regressors to the equation and test down to  $H_0$  as a restriction on the larger model. Two complications are that this approach might be too specific about the form of the alternative hypothesis and, second, with a large number of variables in  $\mathbf{X}$ , it could become unwieldy. Ramsey's proposed solution is to add powers of  $\mathbf{x}_i'\boldsymbol{\beta}$  to the regression using the least squares predictions—typically, one would add the square and, perhaps the cube. This would require a two-step estimation procedure, since in order to add  $(\mathbf{x}_i'\mathbf{b})^2$  and  $(\mathbf{x}_i'\mathbf{b})^3$ , one needs the coefficients. The suggestion, then, is to fit the null model first, using least squares. Then, for the second step, the squares (and cubes) of the predicted values from this first-step regression are added to the equation and it is refit with the additional variables. A (large-sample) Wald test is then used to test the hypothesis of the null model.

As a general strategy, this sort of specification is designed to detect failures of the assumptions of the null model. The obvious virtue of such a test is that it provides much greater generality than a simple test of restrictions such as whether a coefficient is zero. But, that generality comes at considerable cost:

1. The test is nonconstructive. It gives no indication what the researcher should do next if the null model is rejected. This is a general feature of specification tests. Rejection of the null model does not imply any particular alternative.
2. Since the alternative hypothesis is unstated, it is unclear what the power of this test is against any specific alternative.
3. For this specific test (perhaps not for some other specification tests we will examine later), because  $\mathbf{x}_i'\mathbf{b}$  uses the same  $\mathbf{b}$  for every observation, the observations are

---

<sup>12</sup>For related discussion of this possibility, see McAleer, Fisher, and Volker (1982).

**138 PART I ♦ The Linear Regression Model**

correlated, while they are assumed to be uncorrelated in the original model. Because of the two-step nature of the estimator, it is not clear what is the appropriate covariance matrix to use for the  $\chi^2$  test. Two other complications emerge for this test. First, it is unclear what  $\gamma$  converges to, assuming it converges to a  $\chi^2$  limit. Second, variance of the difference between  $\mathbf{x}_i'\mathbf{b}$  and  $\mathbf{x}_i'\boldsymbol{\beta}$  is a function of  $x$ , so the second-step regression might be heteroscedastic. The implication is that neither the size nor the power of this test is necessarily what might be expected.

**Example 5.8 Size of a RESET Test**

To investigate the true size of the RESET test in a particular application, we carried out a Monte Carlo experiment. The results in Table 4.6 give the following estimates of equation (5-2):

$$\ln Price = -8.42653 + 1.33372 \ln Area - 0.16537 \text{Aspect Ratio} + e \text{ where } sd(e) = 1.10266.$$

We take the estimated right-hand side to be our population. We generated 5,000 samples of 430 (the original sample size), by reusing the regression coefficients and generating a new sample of disturbances for each replication. Thus, with each replication,  $r$ , we have a new sample of observations on  $\ln Price_{ir}$ , where the regression part is as above reused and a new set of disturbances is generated each time. With each sample, we computed the least squares coefficient, then the predictions. We then recomputed the least squares regression while adding the square and cube of the prediction to the regression. Finally, with each sample, we computed the chi-squared statistic, and rejected the null model if the chi-squared statistic is larger than 5.99, the 95th percentile of the chi-squared distribution with two degrees of freedom. The **nominal size** of this test is 0.05. Thus, in samples of 100, 500, 1,000, and 5,000, we should reject the null model 5, 25, 50, and 250 times. In our experiment, the computed chi-squared exceeded 5.99 8, 31, 65, and 259 times, respectively, which suggests that at least with sufficient replications, the test performs as might be expected. We then investigated the power of the test by adding 0.1 times the square of  $\ln Area$  to the predictions. It is not possible to deduce the exact power of the RESET test to detect this failure of the null model. In our experiment, with 1,000 replications, the null hypothesis is rejected 321 times. We conclude that the procedure does appear have power to detect this failure of the model assumptions.

**5.10 MODEL BUILDING—A GENERAL TO SIMPLE STRATEGY**

There has been a shift in the general approach to model building in the past 20 years or so, partly based on the results in the previous two sections. With an eye toward maintaining simplicity, model builders would generally begin with a small specification and gradually build up the model ultimately of interest by adding variables. But, based on the preceding results, we can surmise that just about any criterion that would be used to decide whether to add a variable to a current specification would be tainted by the biases caused by the incomplete specification at the early steps. Omitting variables from the equation seems generally to be the worse of the two errors. Thus, the **simple-to-general** approach to model building has little to recommend it. Building on the work of Hendry [e.g., (1995)] and aided by advances in estimation hardware and software, researchers are now more comfortable beginning their specification searches with large elaborate models involving many variables and perhaps long and complex lag structures. The attractive strategy is then to adopt a **general-to-simple**, downward reduction of the

CHAPTER 5 ♦ Hypothesis Tests and Model Selection **139**

model to the preferred specification. [This approach has been completely automated in Hendry's *PCGets*<sup>(c)</sup> computer program. See, e.g., Hendry and Kotzis (2001).] Of course, this must be tempered by two related considerations. In the "kitchen sink" regression, which contains every variable that might conceivably be relevant, the adoption of a fixed probability for the Type I error, say, 5 percent, ensures that in a big enough model, some variables will appear to be significant, even if "by accident." Second, the problems of pretest estimation and **stepwise model building** also pose some risk of ultimately misspecifying the model. To cite one unfortunately common example, the statistics involved often produce unexplainable lag structures in dynamic models with many lags of the dependent or independent variables.

### 5.10.1 MODEL SELECTION CRITERIA

The preceding discussion suggested some approaches to model selection based on nonnested hypothesis tests. Fit measures  testing procedures based on the sum of squared residuals, such as  $R^2$  and the Cox test, are useful when interest centers on the within-sample fit or within-sample prediction of the dependent variable. When the model building is directed toward forecasting, within-sample measures are not necessarily optimal. As we have seen,  $R^2$  cannot fall when variables are added to a model, so there is a built-in tendency to overfit the model. This criterion may point us away from the best forecasting model, because adding variables to a model may increase the variance of the forecast error (see Section 4.6) despite the improved fit to the data. With this thought in mind, the **adjusted  $R^2$** ,

$$\bar{R}^2 = 1 - \frac{n-1}{n-K}(1-R^2) = 1 - \frac{n-1}{n-K} \left( \frac{\mathbf{e}'\mathbf{e}}{\sum_{i=1}^n (y_i - \bar{y})^2} \right), \quad (5-40)$$

has been suggested as a fit measure that appropriately penalizes the loss of degrees of freedom that result from adding variables to the model. Note that  $\bar{R}^2$  may fall when a variable is added to a model if the sum of squares does not fall fast enough. (The applicable result appears in Theorem 3.7;  $\bar{R}^2$  does not rise when a variable is added to a model unless the  $t$  ratio associated with that variable exceeds one in absolute value.) The adjusted  $R^2$  has been found to be a preferable fit measure for assessing the fit of forecasting models. [See Diebold (2003), who argues that the simple  $R^2$  has a downward bias as a measure of the out-of-sample, one-step-ahead prediction error variance.]

The adjusted  $R^2$  penalizes the loss of degrees of freedom that occurs when a model is expanded. There is, however, some question about whether the penalty is sufficiently large to ensure that the criterion will necessarily lead the analyst to the correct model (assuming that it is among the ones considered) as the sample size increases. Two alternative fit measures that have been suggested are the **Akaike Information Criterion**,

$$AIC(K) = s_y^2(1 - R^2)e^{2K/n} \quad (5-41)$$

and the Schwarz or **Bayesian Information Criterion**,

$$BIC(K) = s_y^2(1 - R^2)n^{K/n}. \quad (5-42)$$

(There is no degrees of freedom correction in  $s_y^2$ .) Both measures improve (decline) as  $R^2$  increases (decreases), but, everything else constant, degrade as the model size increases. Like  $\bar{R}^2$ , these measures place a premium on achieving a given fit with a smaller

## 140 PART I ♦ The Linear Regression Model

number of parameters per observation,  $K/n$ . Logs are usually more convenient; the measures reported by most software are

$$\text{AIC}(K) = \ln\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{2K}{n} \quad (5-43)$$

$$\text{BIC}(K) = \ln\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{K \ln n}{n}. \quad (5-44)$$

Both **prediction criteria** have their virtues, and neither has an obvious advantage over the other. [See Diebold (2003).] The **Schwarz criterion**, with its heavier penalty for degrees of freedom lost, will lean toward a simpler model. All else given, simplicity does have some appeal.

### 5.10.2 MODEL SELECTION

The preceding has laid out a number of choices for **model selection**, but, at the same time, has posed some uncomfortable propositions. The pretest estimation aspects of specification search are based on the model builder's knowledge of "the truth" and the consequences of failing to use that knowledge. While the cautions about blind search for statistical significance are well taken, it does seem optimistic to assume that the correct model is likely to be known with hard certainty at the outset of the analysis. The bias documented in (4-10) is well worth the modeler's attention. But, in practical terms, knowing anything about the magnitude presumes that we know what variables are in  $\mathbf{X}_2$ , which need not be the case. While we can agree that the model builder will omit income from a demand equation at their peril, we could also have some sympathy for the analyst faced with finding the right specification for their forecasting model among dozens of choices. The tests for nonnested models would seem to free the modeler from having to claim that the specified set of models contain "the truth." But, a moment's thought should suggest that the cost of this is the possibly deflated power of these procedures to point toward that truth. The  $J$  test may provide a sharp choice between two alternatives, but it neglects the third possibility, that both models are wrong. Vuong's test does but, of course, it suffers from the fairly large inconclusive region, which is a symptom of its relatively low power against many alternatives. The upshot of all of this is that there remains much to be accomplished in the area of model selection. Recent commentary has provided suggestions from two perspective, classical and Bayesian.

### 5.10.3 CLASSICAL MODEL SELECTION

Hansen (2005) lists four shortcomings of the methodology we have considered here:

1. parametric vision
2. assuming a true data generating process
3. evaluation based on fit
4. ignoring model uncertainty

All four of these aspects have framed the analysis of the preceding sections. Hansen's view is that the analysis considered here is too narrow and stands in the way of progress in model discovery.

CHAPTER 5 ♦ Hypothesis Tests and Model Selection **141**

All the model selection procedures considered here are based on the likelihood function, which requires a specific distributional assumption. Hansen argues for a focus, instead, on semiparametric structures. For regression analysis, this points toward generalized method of moments estimators. Casualties of this reorientation will be distributionally based test statistics such as the Cox and Vuong statistics, and even the AIC and BIC measures, which are transformations of the likelihood function. However, alternatives have been proposed [e.g., by Hong, Preston, and Shum (2000)]. The second criticism is one we have addressed. The assumed “true” model can be a straight-jacket. Rather (he argues), we should view our specifications as approximations to the underlying true data generating process—this greatly widens the specification search, to one for a model which provides the best approximation. Of course, that now forces the question of what is “best.” So far, we have focused on the likelihood function, which in the classical regression can be viewed as an increasing function of  $R^2$ . The author argues for a more “focused” information criterion (FIC) that examines directly the parameters of interest, rather than the fit of the model to the data. Each of these suggestions seeks to improve the process of model selection based on familiar criteria, such as test statistics based on fit measures and on characteristics of the model.

A (perhaps *the*) crucial issue remaining is uncertainty about the model itself. The search for the correct model is likely to have the same kinds of impacts on statistical inference as the search for a specification given the form of the model (see Sections 4.3.2 and 4.3.3). Unfortunately, incorporation of this kind of uncertainty in statistical inference procedures remains an unsolved problem. Hansen suggests one potential route would be the Bayesian model averaging methods discussed next although he does express some skepticism about Bayesian methods in general.

#### 5.10.4 BAYESIAN MODEL AVERAGING

If we have doubts as to which of two models is appropriate, then we might well be convinced to concede that possibly neither one is really “the truth.” We have painted ourselves into a corner with our “left or right” approach to testing. The Bayesian approach to this question would treat it as a problem of comparing the two hypotheses rather than testing for the validity of one over the other. We enter our sampling experiment with a set of prior probabilities about the relative merits of the two hypotheses, which is summarized in a “prior odds ratio,”  $P_{01} = \text{Prob}[H_0]/\text{Prob}[H_1]$ . After gathering our data, we construct the Bayes factor, which summarizes the weight of the sample evidence in favor of one model or the other. After the data have been analyzed, we have our “posterior odds ratio,”  $P_{01} | \text{data} = \text{Bayes factor} \times P_{01}$ . The upshot is that ex post, neither model is discarded; we have merely revised our assessment of the comparative likelihood of the two in the face of the sample data. Of course, this still leaves the specification question open. Faced with a choice among models, how can we best use the information we have? Recent work on **Bayesian model averaging** [Hoeting et al. (1999)] has suggested an answer.

An application by Wright (2003) provides an interesting illustration. Recent advances such as Bayesian VARs have improved the forecasting performance of econometric models. Stock and Watson (2001, 2004) report that striking improvements in predictive performance of international inflation can be obtained by averaging a large

## 142 PART I ♦ The Linear Regression Model

number of forecasts from different models and sources. The result is remarkably consistent across subperiods and countries. Two ideas are suggested by this outcome. First, the idea of blending different models is very much in the spirit of Hansen's fourth point. Second, note that the focus of the improvement is not on the fit of the model (point 3), but its predictive ability. Stock and Watson suggested that simple equal-weighted averaging, while one could not readily explain why, seems to bring large improvements. Wright proposed Bayesian model averaging as a means of making the choice of the weights for the average more systematic and of gaining even greater predictive performance.

Leamer (1978) appears to be the first to propose Bayesian model averaging as a means of combining models. The idea has been studied more recently by Min and Zellner (1993) for output growth forecasting, Doppelhofer et al. (2000) for cross-country growth regressions, Koop and Potter (2004) for macroeconomic forecasts, and others. Assume that there are  $M$  models to be considered, indexed by  $m = 1, \dots, M$ . For simplicity, we will write the  $m$ th model in a simple form,  $f_m(\mathbf{y} | \mathbf{Z}, \boldsymbol{\theta}_m)$  where  $f(\cdot)$  is the density,  $\mathbf{y}$  and  $\mathbf{Z}$  are the data, and  $\boldsymbol{\theta}_m$  is the parameter vector for model  $m$ . Assume, as well, that model  $m^*$  is the true model, unknown to the analyst. The analyst has priors  $\pi_m$  over the probabilities that model  $m$  is the correct model, so  $\pi_m$  is the prior probability that  $m = m^*$ . The posterior probabilities for the models are

$$\Pi_m = \text{Prob}(m = m^* | \mathbf{y}, \mathbf{Z}) = \frac{P(\mathbf{y}, \mathbf{Z} | m)\pi_m}{\sum_{r=1}^M P(\mathbf{y}, \mathbf{Z} | r)\pi_r}, \quad (5-45)$$

where  $P(\mathbf{y}, \mathbf{Z} | m)$  is the marginal likelihood for the  $m$ th model,

$$P(\mathbf{y}, \mathbf{Z} | m) = \int_{\theta_m} P(\mathbf{y}, \mathbf{Z} | \theta_m, m)P(\theta_m)d\theta_m, \quad (5-46)$$

while  $P(\mathbf{y}, \mathbf{Z} | \theta_m, m)$  is the conditional (on  $\theta_m$ ) likelihood for the  $m$ th model and  $P(\theta_m)$  is the analyst's prior over the parameters of the  $m$ th model. This provides an alternative set of weights to the  $\Pi_m = 1/M$  suggested by Stock and Watson. Let  $\hat{\theta}_m$  denote the Bayesian estimate (posterior mean) of the parameters of model  $m$ . (See Chapter 16.) Each model provides an appropriate posterior forecast density,  $f^*(\mathbf{y} | \mathbf{Z}, \hat{\theta}_m, m)$ . The Bayesian model averaged forecast density would then be

$$\bar{f}^* = \sum_{m=1}^M f^*(\mathbf{y} | \mathbf{Z}, \hat{\theta}_m, m)\Pi_m. \quad (5-47)$$

A point forecast would be a similarly weighted average of the forecasts from the individual models.

### Example 5.9 Bayesian Averaging of Classical Estimates

Many researchers have expressed skepticism of Bayesian methods because of the apparent arbitrariness of the specifications of prior densities over unknown parameters. In the Bayesian model averaging setting, the analyst requires prior densities over not only the model probabilities,  $\pi_m$ , but also the model specific parameters,  $\theta_m$ . In their application, Doppelhofer, Miller, and Sala-i-Martin (2000) were interested in the appropriate set of regressors to include in a long-term macroeconomic (income) growth equation. With 32 candidates,  $M$  for their application was  $2^{32}$  (minus one if the zero regressors model is ignored), or roughly four billion. Forming this many priors would be optimistic in the extreme. The authors proposed a novel method of weighting a large subset (roughly 21 million) of the  $2^M$  possible (classical) least squares regressions. The weights are formed using a Bayesian procedure; however,

**CHAPTER 5 ♦ Hypothesis Tests and Model Selection 143**

the estimates that are weighted are the classical least squares estimates. While this saves considerable computational effort, it still requires the computation of millions of least squares coefficient vectors. [See Sala-i-Martin (1997).] The end result is a model with 12 independent variables.

### 5.11 SUMMARY AND CONCLUSIONS

This chapter has focused on two uses of the linear regression model, hypothesis testing, and basic prediction. The central result for testing hypotheses is the  $F$  statistic. The  $F$  ratio can be produced in two equivalent ways; first, by measuring the extent to which the unrestricted least squares estimate differs from what a hypothesis would predict, and second, by measuring the loss of fit that results from assuming that a hypothesis is correct. We then extended the  $F$  statistic to more general settings by examining its large-sample properties, which allow us to discard the assumption of normally distributed disturbances and by extending it to nonlinear restrictions.

This is the last of five chapters that we have devoted specifically to the methodology surrounding the most heavily used tool in econometrics, the classical linear regression model. We began in Chapter 2 with a statement of the regression model. Chapter 3 then described computation of the parameters by least squares—a purely algebraic exercise. Chapter 4 reinterpreted least squares as an estimator of an unknown parameter vector and described the finite sample and large-sample characteristics of the sampling distribution of the estimator. Chapter 5 was devoted to building and sharpening the regression model, with statistical results for testing hypotheses about the underlying population. In this chapter, we have examined some broad issues related to model specification and selection of a model among a set of competing alternatives. The concepts considered here are tied very closely to one of the pillars of the paradigm of econometrics; Underlying the model is a theoretical construction, a set of true behavioral relationships that constitute *the model*. It is only on this notion that the concepts of bias and biased estimation and model selection make any sense—“bias” as a concept can only be described with respect to some underlying “model” against which an estimator can be said to be biased. That is, there must be a yardstick. This concept is a central result in the analysis of specification, where we considered the implications of underfitting (omitting variables) and overfitting (including superfluous variables) the model. We concluded this chapter (and our discussion of the classical linear regression model) with an examination of procedures that are used to choose among competing model specifications.

#### Key Terms and Concepts

- Acceptance region
- Adjusted R-squared
- Akaike Information Criterion
- Alternative hypothesis
- Bayesian model averaging
- Bayesian Information Criterion
- Biased estimator
- Comprehensive model
- Consistent
- Distributed lag
- Discrepancy vector
- Encompassing principle
- Exclusion restrictions
- Ex post forecast
- Functionally independent
- General nonlinear hypothesis
- General-to-simple strategy
- Inclusion of superfluous variables
- $J$  test
- Lack of invariance

## 144 PART I ♦ The Linear Regression Model

- Lagrange multiplier test
- Linear restrictions
- Mean squared error
- Model selection
- Nested
- Nested models
- Nominal size
- Nonnested
- Nonnested models
- Nonnormality
- Null hypothesis
- One-sided test
- Parameter space
- Power of a test
- Prediction criterion
- Prediction interval
- Prediction variance
- Rejection region
- Restricted least squares
- Root mean squared error
- Sample discrepancy
- Schwarz criterion
- Simple-to-general
- Size of the test
- Specification test
- Stepwise model building
- $t$  ratio
- Testable implications
- Theil  $U$  statistic
- Wald criterion
- Wald distance
- Wald statistic
- Wald test

### Exercises

1. A multiple regression of  $y$  on a constant  $x_1$  and  $x_2$  produces the following results:  
 $\hat{y} = 4 + 0.4x_1 + 0.9x_2$ ,  $R^2 = 8/60$ ,  $\mathbf{e}'\mathbf{e} = 520$ ,  $n = 29$ ,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 29 & 0 & 0 \\ 0 & 50 & 10 \\ 0 & 10 & 80 \end{bmatrix}.$$

Test the hypothesis that the two slopes sum to 1.

2. Using the results in Exercise 1, test the hypothesis that the slope on  $x_1$  is 0 by running the restricted regression and comparing the two sums of squared deviations.
3. The regression model to be analyzed is  $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$ , where  $\mathbf{X}_1$  and  $\mathbf{X}_2$  have  $K_1$  and  $K_2$  columns, respectively. The restriction is  $\boldsymbol{\beta}_2 = \mathbf{0}$ .
  - a. Using (5-23), prove that the restricted estimator is simply  $[\mathbf{b}_{1*}, \mathbf{0}]$ , where  $\mathbf{b}_{1*}$  is the least squares coefficient vector in the regression of  $\mathbf{y}$  on  $\mathbf{X}_1$ .
  - b. Prove that if the restriction is  $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0$  for a nonzero  $\boldsymbol{\beta}_2^0$ , then the restricted estimator of  $\boldsymbol{\beta}_1$  is  $\mathbf{b}_{1*} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_2^0)$ .
4. The expression for the restricted coefficient vector in (5-23) may be written in the form  $\mathbf{b}_* = [\mathbf{I} - \mathbf{C}\mathbf{R}]\mathbf{b} + \mathbf{w}$ , where  $\mathbf{w}$  does not involve  $\mathbf{b}$ . What is  $\mathbf{C}$ ? Show that the covariance matrix of the restricted least squares estimator is

$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$$

and that this matrix may be written as

$$\text{Var}[\mathbf{b} | \mathbf{X}] \{ [\text{Var}(\mathbf{b} | \mathbf{X})]^{-1} - \mathbf{R}'[\text{Var}(\mathbf{R}\mathbf{b}) | \mathbf{X}]^{-1}\mathbf{R} \} \text{Var}[\mathbf{b} | \mathbf{X}].$$

5. Prove the result that the restricted least squares estimator never has a larger covariance matrix than the unrestricted least squares estimator.
6. Prove the result that the  $R^2$  associated with a restricted least squares estimator is never larger than that associated with the unrestricted least squares estimator. Conclude that imposing restrictions never improves the fit of the regression.
7. An alternative way to test the hypothesis  $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$  is to use a Wald test of the hypothesis that  $\boldsymbol{\lambda}_* = \mathbf{0}$ , where  $\boldsymbol{\lambda}_*$  is defined in (5-23). Prove that

$$\chi^2 = \boldsymbol{\lambda}'_* \{ \text{Est. Var}[\boldsymbol{\lambda}_*] \}^{-1} \boldsymbol{\lambda}_* = (n - K) \left[ \frac{\mathbf{e}'_* \mathbf{e}_*}{\mathbf{e}' \mathbf{e}} - 1 \right].$$

CHAPTER 5 ♦ Hypothesis Tests and Model Selection **145**

Note that the fraction in brackets is the ratio of two estimators of  $\sigma^2$ . By virtue of (5-28) and the preceding discussion, we know that this ratio is greater than 1. Finally, prove that this test statistic is equivalent to  $JF$ , where  $J$  is the number of restrictions being tested and  $F$  is the conventional  $F$  statistic given in (5-16). Formally, the Lagrange multiplier test requires that the variance estimator be based on the restricted sum of squares, not the unrestricted. Then, the test statistic would be  $LM = nJ/[(n - K)/F + J]$ . See Godfrey (1988).

8. Use the test statistic defined in Exercise 7 to test the hypothesis in Exercise 1.
9. Prove that under the hypothesis that  $\mathbf{R}\beta = \mathbf{q}$ , the estimator

$$s_*^2 = \frac{(\mathbf{y} - \mathbf{X}\mathbf{b}_*)'(\mathbf{y} - \mathbf{X}\mathbf{b}_*)}{n - K + J},$$

where  $J$  is the number of restrictions, is unbiased for  $\sigma^2$ .

10. Show that in the multiple regression of  $\mathbf{y}$  on a constant,  $\mathbf{x}_1$  and  $\mathbf{x}_2$  while imposing the restriction  $\beta_1 + \beta_2 = 1$  leads to the regression of  $\mathbf{y} - \mathbf{x}_1$  on a constant and  $\mathbf{x}_2 - \mathbf{x}_1$ .
11. Suppose the true regression model is given by (4-8). The result in (4-10) shows that if either  $\mathbf{P}_{1,2}$  is nonzero or  $\beta_2$  is nonzero, then regression of  $\mathbf{y}$  on  $\mathbf{X}_1$  alone produces a biased and inconsistent estimator of  $\beta_1$ . Suppose the objective is to forecast  $\mathbf{y}$ , not to estimate the parameters. Consider regression of  $\mathbf{y}$  on  $\mathbf{X}_1$  alone to estimate  $\beta_1$  with  $\mathbf{b}_1$  (which is biased). Is the forecast of  $\mathbf{y}$  computed using  $\mathbf{X}_1\mathbf{b}_1$  also biased? Assume that  $E[\mathbf{X}_2 | \mathbf{X}_1]$  is a linear function of  $\mathbf{X}_1$ . Discuss your findings generally. What are the implications for prediction when variables are omitted from a regression?
12. Compare the mean squared errors of  $b_1$  and  $b_{1,2}$  in Section 4.7.2. (*Hint:* The comparison depends on the data and the model parameters, but you can devise a compact expression for the two quantities.)
13. The log likelihood function for the linear regression model with normally distributed disturbances is shown in Example 4.6. Show that at the maximum likelihood estimators of  $\mathbf{b}$  for  $\beta$  and  $\mathbf{e}'\mathbf{e}/n$  for  $\sigma^2$ , the log likelihood is an increasing function of  $R^2$  for the model.
14. Show that the model of the alternative hypothesis in Example 5.7 can be written

$$H_1: C_t = \theta_1 + \theta_2 Y_t + \theta_3 Y_{t-1} + \sum_{s=2}^{\infty} \theta_{s+2} Y_{t-s} + \varepsilon_{it} + \sum_{s=1}^{\infty} \lambda_s \varepsilon_{t-s}.$$

As such, it does appear that  $H_0$  is a restriction on  $H_1$ . However, because there are an infinite number of constraints, this does not reduce the test to a standard test of restrictions. It does suggest the connections between the two formulations. (We will revisit models of this sort in Chapter 21.)

## **Applications**

1. The application in Chapter 3 used 15 of the 17,919 observations in Koop and Tobias's (2004) study of the relationship between wages and education, ability, and family characteristics. (See Appendix Table F3.2.) We will use the full data set for this exercise. The data may be downloaded from the *Journal of Applied Econometrics* data archive at <http://www.econ.queensu.ca/jae/12004-v19.7/koop-tobias/>. The

## 146 PART I ♦ The Linear Regression Model

data file is in two parts. The first file contains the panel of 17,919 observations on variables:

- Column 1; *Person id* (ranging from 1 to 2,178),
- Column 2; *Education*,
- Column 3; *Log of hourly wage*,
- Column 4; *Potential experience*,
- Column 5; *Time trend*.

Columns 2–5 contain time varying variables. The second part of the data set contains time invariant variables for the 2,178 households. These are

- Column 1; *Ability*,
- Column 2; *Mother's education*,
- Column 3; *Father's education*,
- Column 4; *Dummy variable for residence in a broken home*,
- Column 5; *Number of siblings*.

To create the data set for this exercise, it is necessary to merge these two data files. The  $i$ th observation in the second file will be replicated  $T_i$  times for the set of  $T_i$  observations in the first file. The *person id* variable indicates which rows must contain the data from the second file. (How this preparation is carried out will vary from one computer package to another.) (Note: We are not attempting to replicate Koop and Tobias's results here—we are only employing their interesting data set.) Let  $\mathbf{X}_1 = [\text{constant}, \text{education}, \text{experience}, \text{ability}]$  and let  $\mathbf{X}_2 = [\text{mother's education}, \text{father's education}, \text{broken home}, \text{number of siblings}]$ .

- a. Compute the full regression of *log wage* on  $\mathbf{X}_1$  and  $\mathbf{X}_2$  and report all results.
- b. Use an  $F$  test to test the hypothesis that all coefficients except the constant term are zero.
- c. Use an  $F$  statistic to test the joint hypothesis that the coefficients on the four household variables in  $\mathbf{X}_2$  are zero.
- d. Use a Wald test to carry out the test in part c.

2. The generalized Cobb–Douglas cost function examined in Application 2 in Chapter 4 is a special case of the **translog cost function**,

$$\begin{aligned} \ln C = & \alpha + \beta \ln Q + \delta_k \ln P_k + \delta_l \ln P_l + \delta_f \ln P_f \\ & + \phi_{kk} \left[ \frac{1}{2} (\ln P_k)^2 \right] + \phi_{ll} \left[ \frac{1}{2} (\ln P_l)^2 \right] + \phi_{ff} \left[ \frac{1}{2} (\ln P_f)^2 \right] \\ & + \phi_{kl} [\ln P_k][\ln P_l] + \phi_{kf} [\ln P_k][\ln P_f] + \phi_{lf} [\ln P_l][\ln P_f] \\ & + \gamma \left[ \frac{1}{2} (\ln Q)^2 \right] \\ & + \theta_{Qk} [\ln Q][\ln P_k] + \theta_{Ql} [\ln Q][\ln P_l] + \theta_{Qf} [\ln Q][\ln P_f] + \varepsilon. \end{aligned}$$

The theoretical requirement of linear homogeneity in the factor prices imposes the following restrictions:

$$\begin{array}{lcl} \delta_k + \delta_l + \delta_f = 1 & \phi_{kk} + \phi_{kl} + \phi_{kf} = 0 & \phi_{kl} + \phi_{ll} + \phi_{lf} = 0 \\ \phi_{kf} + \phi_{lf} + \phi_{ff} = 0 & \theta_{QK} + \theta_{QL} + \theta_{QF} = 0 & \end{array}$$

Note that although the underlying theory requires it, the model can be estimated (by least squares) without imposing the linear homogeneity restrictions. [Thus, one

## CHAPTER 5 ♦ Hypothesis Tests and Model Selection 147

could “test” the underlying theory by testing the validity of these restrictions. See Christensen, Jorgenson, and Lau (1975).] We will repeat this exercise in part b.

A number of additional restrictions were explored in Christensen and Greene’s (1976) study. The hypothesis of homotheticity of the production structure would add the additional restrictions

$$\theta_{Qk} = 0, \quad \theta_{Ql} = 0, \quad \theta_{Qf} = 0.$$

Homogeneity of the production structure adds the restriction  $\gamma = 0$ . The hypothesis that all elasticities of substitution in the production structure are equal to  $-1$  is imposed by the six restrictions  $\phi_{ij} = 0$  for all  $i$  and  $j$ .

We will use the data from the earlier application to test these restrictions. For the purposes of this exercise, denote by  $\beta_1, \dots, \beta_{15}$  the 15 parameters in the cost function above in the order that they appear in the model, starting in the first line and moving left to right and downward.

- a. Write out the **R** matrix and **q** vector in (5-8) that are needed to impose the restriction of linear homogeneity in prices.
- b. “Test” the theory of production using all 158 observations. Use an *F* test to test the restrictions of linear homogeneity. Note, you can use the general form of the *F* statistic in (5-16) to carry out the test. Christensen and Greene enforced the linear homogeneity restrictions by building them into the model. You can do this by dividing cost and the prices of capital and labor by the price of fuel. Terms with *f* subscripts fall out of the model, leaving an equation with 10 parameters. Compare the sums of squares for the two models to carry out the test. Of course, the test may be carried out either way and will produce the same result.
- c. Test the hypothesis homotheticity of the production structure under the assumption of linear homogeneity in prices.
- d. Test the hypothesis of the generalized Cobb–Douglas cost function in Chapter 4 against the more general translog model suggested here, once again (and henceforth) assuming linear homogeneity in the prices.
- e. The simple Cobb–Douglas function appears in the first line of the model above. Test the hypothesis of the Cobb–Douglas model against the alternative of the full translog model.
- f. Test the hypothesis of the generalized Cobb–Douglas model against the homothetic translog model. 
- g. Which of the several functional forms suggested here do you conclude is the most appropriate for these data?
3. The gasoline consumption model suggested in part d of Application 1 in Chapter 4 may be written as 

$$\ln(G/Pop) = \alpha + \beta_P \ln P_g + \beta_I \ln (Income/Pop) + \gamma_{nc} \ln P_{nc} + \gamma_{uc} \ln P_{uc} + \gamma_{pt} \ln P_{pt} + \tau_{year} + \delta_d \ln P_d + \delta_n \ln P_n + \delta_s \ln P_s + \varepsilon.$$

- a. Carry out a test of the hypothesis that the three aggregate price indices are not significant determinants of the demand for gasoline.
- b. Consider the hypothesis that the microelasticities are a constant proportion of the elasticity with respect to their corresponding aggregate. Thus, for some positive  $\theta$  (presumably between 0 and 1),  $\gamma_{nc} = \theta \delta_d$ ,  $\gamma_{uc} = \theta \delta_d$ ,  $\gamma_{pt} = \theta \delta_s$ . The first

**148 PART I ♦ The Linear Regression Model**

two imply the simple linear restriction  $\gamma_{nc} = \gamma_{uc}$ . By taking ratios, the first (or second) and third imply the nonlinear restriction

$$\frac{\gamma_{nc}}{\gamma_{pt}} = \frac{\delta_d}{\delta_s} \quad \text{or} \quad \gamma_{nc}\delta_s - \gamma_{pt}\delta_d = 0.$$

Describe in detail how you would test the validity of the restriction.

- c. Using the gasoline market data in Table F2.2, test the two restrictions suggested here, separately and jointly.
- 4. The  $J$  test in Example 5.7 is carried out using more than 50 years of data. It is optimistic to hope that the underlying structure of the economy did not change in 50 years. Does the result of the test carried out in Example 5.7 persist if it is based on data only from 1980 to 2000? Repeat the computation with this subset of the data.

## 6

# FUNCTIONAL FORM AND STRUCTURAL CHANGE

---

## 6.1 INTRODUCTION

This chapter will complete our analysis of the linear regression model. We begin by examining different aspects of the functional form of the regression model. Many different types of functions are *linear* by the definition in Section 2.3.1. By using different transformations of the dependent and independent variables, binary variables, and different arrangements of functions of variables, a wide variety of models can be constructed that are all estimable by linear least squares. Section 6.2 considers using binary variables to accommodate nonlinearities in the model. Section 6.3 broadens the class of models that are linear in the parameters. By using logarithms, quadratic terms, and interaction terms (products of variables), the regression model can accommodate a wide variety of functional forms in the data.

Section 6.4 examines the issue of specifying and testing for discrete change in the underlying process that generates the data, under the heading of **structural change**. In a time-series context, this relates to abrupt changes in the economic environment, such as major events in financial (e.g., the world financial crisis of 2007–2009) or commodity markets (such as the several upheavals in the oil market). In a cross section, we can modify the regression model to account for discrete differences across groups such as different preference structures or market experiences of men and women.

## 6.2 USING BINARY VARIABLES

One of the most useful devices in regression analysis is the **binary**, or **dummy variable**. A dummy variable takes the value one for some observations to indicate the presence of an effect or membership in a group and zero for the remaining observations. Binary variables are a convenient means of building discrete shifts of the function into a regression model.

### 6.2.1 BINARY VARIABLES IN REGRESSION

Dummy variables are usually used in regression equations that also contain other quantitative variables. In the earnings equation in Example 5.2, we included a variable *Kids* to indicate whether there were children in the household, under the assumption that for many married women, this fact is a significant consideration in labor supply behavior. The results shown in Example 6.1 appear to be consistent with this hypothesis.

## 150 PART I ♦ The Linear Regression Model

**TABLE 6.1** Estimated Earnings Equation

$\ln \text{earnings} = \beta_1 + \beta_2 \text{age} + \beta_3 \text{age}^2 + \beta_4 \text{education} + \beta_5 \text{kids} + \varepsilon$			
Sum of squared residuals:	599.4582		
Standard error of the regression:	1.19044		
$R^2$ based on 428 observations	0.040995		
Variable	Coefficient	Standard Error	t Ratio
Constant	3.24009	1.7674	1.833
Age	0.20056	0.08386	2.392
Age <sup>2</sup>	-0.0023147	0.00098688	-2.345
Education	0.067472	0.025248	2.672
Kids	-0.35119	0.14753	-2.380

### Example 6.1 Dummy Variable in an Earnings Equation

Table 6.1 following reproduces the estimated earnings equation in Example 5.2. The variable *Kids* is a dummy variable, which equals one if there are children under 18 in the household and zero otherwise. Since this is a **semilog equation**, the value of -0.35 for the coefficient is an extremely large effect, one which suggests that all other things equal, the earnings of women with children are nearly a third less than those without. This is a large difference, but one that would certainly merit closer scrutiny. Whether this effect results from different labor market effects that influence wages and not hours, or the reverse, remains to be seen. Second, having chosen a nonrandomly selected sample of those with only positive earnings to begin with, it is unclear whether the sampling mechanism has, itself, induced a bias in this coefficient.

Dummy variables are particularly useful in loglinear regressions. In a model of the form

$$\ln y = \beta_1 + \beta_2 x + \beta_3 d + \varepsilon,$$

the coefficient on the dummy variable,  $d$ , indicates a multiplicative shift of the function. The percentage change in  $E[y|x,d]$  associated with the change in  $d$  is

$$\begin{aligned} \% (\Delta E[y|x, d]/\Delta d) &= 100\% \left\{ \frac{E[y|x, d=1] - E[y|x, d=0]}{E[y|x, d=0]} \right\} \\ &= 100\% \left\{ \frac{\exp(\beta_1 + \beta_2 x + \beta_3) E[\exp(\varepsilon)] - \exp(\beta_1 + \beta_2 x) E[\exp(\varepsilon)]}{\exp(\beta_1 + \beta_2 x) E[\exp(\varepsilon)]} \right\} \\ &= 100\% [\exp(\beta_3) - 1]. \end{aligned}$$

### Example 6.2 Value of a Signature

In Example 4.10 we explored the relationship between (log of) sale price and surface area for 430 sales of Monet paintings. Regression results from the example are included in Table 6.2. The results suggest a strong relationship between area and price—the coefficient is 1.33372 indicating a highly elastic relationship and the  $t$  ratio of 14.70 suggests the relationship is highly significant. A variable (effect) that is clearly left out of the model is the effect of the artist's signature on the sale price. Of the 430 sales in the sample, 77 are for unsigned paintings. The results at the right of Table 6.2 include a dummy variable for whether the painting is signed or not. The results show an extremely strong effect. The regression results imply that

$$\begin{aligned} E[Price|Area, Aspect, Signature] &= \\ &\exp[-9.64 + 1.35 \ln Area - .08 \text{AspectRatio} + 1.23 \text{Signature} + .993^2/2]. \end{aligned}$$

CHAPTER 6 ♦ Functional Form and Structural Change **151****TABLE 6.2** Estimated Equations for Log Price

$\ln \text{price} = \beta_0 + \beta_1 \ln \text{Area} + \beta_2 \text{aspect ratio} + \beta_3 \text{signature} + \varepsilon$					
Mean of log Price	.33274				
Number of observations	430				
Sum of squared residuals	519.17235				420.16787
Standard error	1.10266				0.99313
R-squared	0.33620				0.46279
Adjusted R-squared	0.33309				0.45900
Variable	Coefficient	Standard Error	t	Coefficient	Standard Error
Constant	-8.42653	0.61183	-13.77	-9.64028	.56422
Ln area	1.33372	0.09072	14.70	1.34935	.08172
Aspect ratio	-0.16537	0.12753	-1.30	-0.07857	.11519
Signature	0.00000	0.00000	0.00	1.25541	.12530
					10.02

(See Section 4.6.) Computing this result for a painting of the same area and aspect ratio, we find the model predicts that the signature effect would be

$$100\% \times (\Delta E[\text{Price}]/\text{Price}) = 100\%[\exp(1.26) - 1] = 252\%.$$

The effect of a signature on an otherwise similar painting is to more than double the price. The estimated standard error for the signature coefficient is 0.1253. Using the delta method, we obtain an estimated standard error for  $[\exp(b_3) - 1]$  of the square root of  $[\exp(b_3)]^2 \times .1253^2$ , which is 0.4417. For the percentage difference of 252%, we have an estimated standard error of 44.17%.

Superficially, it is possible that the size effect we observed earlier could be explained by the presence of the signature. If the artist tended on average to sign only the larger paintings, then we would have an explanation for the counterintuitive effect of size. (This would be an example of the effect of multicollinearity of a sort.) For a regression with a continuous variable and a dummy variable, we can easily confirm or refute this proposition. The average size for the 77 sales of unsigned paintings is 1,228.69 square inches. The average size of the other 353 is 940.812 square inches. There does seem to be a substantial systematic difference between signed and unsigned paintings, but it goes in the other direction. We are left with significant findings of both a size and a signature effect in the auction prices of Monet paintings. *Aspect Ratio*, however, appears still to be inconsequential.

There is one remaining feature of this sample for us to explore. These 430 sales involved only 387 different paintings. Several sales involved repeat sales of the same painting. The assumption that observations are independent draws is violated, at least for some of them. We will examine this form of “clustering” in Chapter 11 in our treatment of panel data.

It is common for researchers to include a dummy variable in a regression to account for something that applies only to a single observation. For example, in time-series analyses, an occasional study includes a dummy variable that is one only in a single unusual year, such as the year of a major strike or a major policy event. (See, for example, the application to the German money demand function in Section 23.3.5.) It is easy to show (we consider this in the exercises) the very useful implication of this:

A dummy variable that takes the value one only for one observation has the effect of deleting that observation from computation of the least squares slopes and variance estimator (but not R-squared).

## 152 PART I ♦ The Linear Regression Model

### 6.2.2 SEVERAL CATEGORIES

When there are several categories, a set of binary variables is necessary. Correcting for seasonal factors in macroeconomic data is a common application. We could write a consumption function for quarterly data as

$$C_t = \beta_1 + \beta_2 x_t + \delta_1 D_{t1} + \delta_2 D_{t2} + \delta_3 D_{t3} + \varepsilon_t,$$

where  $x_t$  is disposable income. Note that only three of the four quarterly dummy variables are included in the model. If the fourth were included, then the four dummy variables would sum to one at every observation, which would reproduce the constant term—a case of perfect multicollinearity. This is known as the **dummy variable trap**. Thus, to avoid the dummy variable trap, we drop the dummy variable for the fourth quarter. (Depending on the application, it might be preferable to have four separate dummy variables and drop the overall constant.)<sup>1</sup> Any of the four quarters (or 12 months) can be used as the base period.

The preceding is a means of *deseasonalizing* the data. Consider the alternative formulation:

$$C_t = \beta x_t + \delta_1 D_{t1} + \delta_2 D_{t2} + \delta_3 D_{t3} + \delta_4 D_{t4} + \varepsilon_t. \quad (6-1)$$

Using the results from Section 3.3 on partitioned regression, we know that the preceding multiple regression is equivalent to first regressing  $C$  and  $x$  on the four dummy variables and then using the residuals from these regressions in the subsequent regression of deseasonalized consumption on deseasonalized income. Clearly, deseasonalizing in this fashion prior to computing the simple regression of consumption on income produces the same coefficient on income (and the same vector of residuals) as including the set of dummy variables in the regression.

#### **Example 6.3 Genre Effects on Movie Box Office Receipts**

Table 4.8 in Example 4.12 presents the results of the regression of log of box office receipts for 62 2009 movies on a number of variables including a set of dummy variables for genre: *Action*, *Comedy*, *Animated*, or *Horror*. The left out category is “any of the remaining 9 genres” in the standard set of 13 that is usually used in models such as this one. The four coefficients are  $-.869$ ,  $-.016$ ,  $-.833$ ,  $+.375$ , respectively. This suggests that, save for horror movies, these genres typically fare substantially worse at the box office than other types of movies. We note the use of  $b$  directly to estimate the percentage change for the category, as we did in example 6.1 when we interpreted the coefficient of  $-.35$  on *Kids* as indicative of a 35 percent change in income, is an approximation that works well when  $b$  is close to zero but deteriorates as it gets far from zero. Thus, the value of  $-.869$  above does not translate to an 87 percent difference between *Action* movies and other movies. Using the formula we used in Example 6.2, we find an estimated difference closer to  $[\exp(-.869) - 1]$  or about 58 percent.

### 6.2.3 SEVERAL GROUPINGS

The case in which several sets of dummy variables are needed is much the same as those we have already considered, with one important exception. Consider a model of statewide per capita expenditure on education  $y$  as a function of statewide per capita income  $x$ . Suppose that we have observations on all  $n = 50$  states for  $T = 10$  years.

---

<sup>1</sup>See Suits (1984) and Greene and Seaks (1991).

## CHAPTER 6 ♦ Functional Form and Structural Change 153

A regression model that allows the expected expenditure to change over time as well as across states would be

$$y_{it} = \alpha + \beta x_{it} + \delta_i + \theta_t + \varepsilon_{it}. \quad (6-2)$$

As before, it is necessary to drop one of the variables in each set of dummy variables to avoid the dummy variable trap. For our example, if a total of 50 state dummies and 10 time dummies is retained, a problem of “perfect multicollinearity” remains; the sums of the 50 state dummies and the 10 time dummies are the same, that is, 1. One of the variables in each of the sets (or the overall constant term and one of the variables in one of the sets) must be omitted.

### **Example 6.4 Analysis of Covariance**

The data in Appendix Table F6.1 were used in a study of efficiency in production of airline services in Greene (2007a). The airline industry has been a favorite subject of study [e.g., Schmidt and Sickles (1984); Sickles, Good, and Johnson (1986)], partly because of interest in this rapidly changing market in a period of deregulation and partly because of an abundance of large, high-quality data sets collected by the (no longer existent) Civil Aeronautics Board. The original data set consisted of 25 firms observed yearly for 15 years (1970 to 1984), a “balanced panel.” Several of the firms merged during this period and several others experienced strikes, which reduced the number of complete observations substantially. Omitting these and others because of missing data on some of the variables left a group of 10 full observations, from which we have selected six for the examples to follow. We will fit a cost equation of the form

$$\begin{aligned} \ln C_{i,t} = & \beta_1 + \beta_2 \ln Q_{i,t} + \beta_3 \ln^2 Q_{i,t} + \beta_4 \ln P_{fuel\ i,t} + \beta_5 \text{Loadfactor}_{i,t} \\ & + \sum_{t=1}^{14} \theta_t D_{i,t} + \sum_{i=1}^5 \delta_i F_{i,t} + \varepsilon_{i,t}. \end{aligned}$$

The dummy variables are  $D_{i,t}$  which is the year variable and  $F_{i,t}$  which is the firm variable. We have dropped the last one in each group. The estimated model for the full specification is

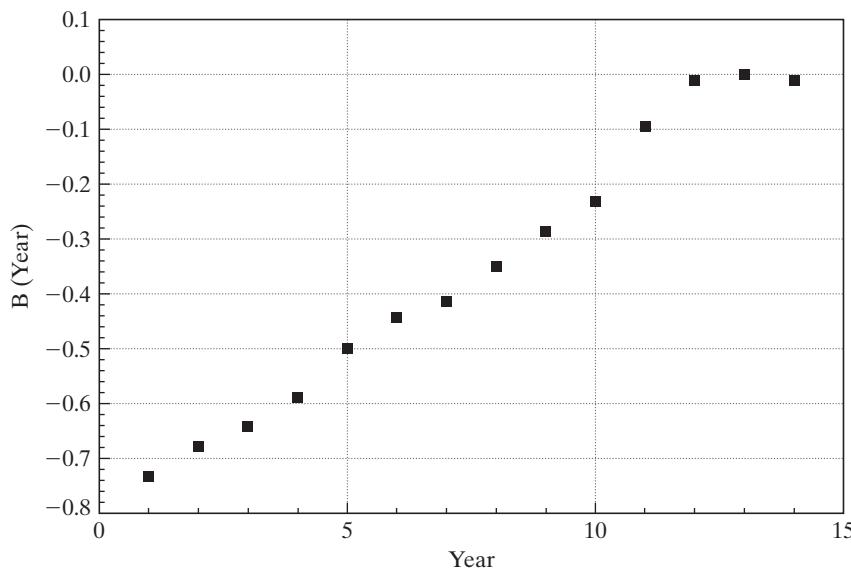
$$\begin{aligned} \ln C_{i,t} = & 13.56 + 0.8866 \ln Q_{i,t} + 0.0221 \ln^2 Q_{i,t} + 0.1281 \ln P_{fi,t} - 0.8855 LF_{i,t} \\ & + \text{time effects} + \text{firm effects} \end{aligned}$$

The year effects display a revealing pattern, as shown in Figure 6.1. This was a period of rapidly rising fuel prices, so the cost effects are to be expected. Since one year dummy variable is dropped, the effect shown is relative to this base year (1984).

We are interested in whether the firm effects, the time effects, both, or neither are statistically significant. Table 6.3 presents the sums of squares from the four regressions. The  $F$  statistic for the hypothesis that there are no firm-specific effects is 65.94, which is highly significant. The statistic for the time effects is only 2.61, which is larger than the critical value of 1.84, but perhaps less so than Figure 6.1 might have suggested. In the absence of the

**TABLE 6.3** *F* tests for Firm and Year Effects

Model	Sum of Squares	Restrictions	F	Deg.Fr.
Full model	0.17257	0	—	
Time effects only	1.03470	5	65.94	[5, 66]
Firm effects only	0.26815	14	2.61	[14, 66]
No effects	1.27492	19	22.19	[19, 66]

**154 PART I ♦ The Linear Regression Model**

**FIGURE 6.1** Estimated Year Dummy Variable Coefficients.

year-specific dummy variables, the year-specific effects are probably largely absorbed by the price of fuel.

**6.2.4 THRESHOLD EFFECTS AND CATEGORICAL VARIABLES**

In most applications, we use dummy variables to account for purely qualitative factors, such as membership in a group, or to represent a particular time period. There are cases, however, in which the dummy variable(s) represents levels of some underlying factor that might have been measured directly if this were possible. For example, education is a case in which we typically observe certain thresholds rather than, say, years of education. Suppose, for example, that our interest is in a regression of the form

$$\text{income} = \beta_1 + \beta_2 \text{age} + \text{effect of education} + \varepsilon.$$

The data on education might consist of the highest level of education attained, such as high school (*HS*), undergraduate (*B*), master's (*M*), or Ph.D. (*P*). An obviously unsatisfactory way to proceed is to use a variable *E* that is 0 for the first group, 1 for the second, 2 for the third, and 3 for the fourth. That is,  $\text{income} = \beta_1 + \beta_2 \text{age} + \beta_3 E + \varepsilon$ . The difficulty with this approach is that it assumes that the increment in income at each threshold is the same;  $\beta_3$  is the difference between income with a Ph.D. and a master's and between a master's and a bachelor's degree. This is unlikely and unduly restricts the regression. A more flexible model would use three (or four) binary variables, one for each level of education. Thus, we would write

$$\text{income} = \beta_1 + \beta_2 \text{age} + \delta_B B + \delta_M M + \delta_P P + \varepsilon.$$

## CHAPTER 6 ♦ Functional Form and Structural Change 155

The correspondence between the coefficients and income for a given age is

$$\begin{aligned}\text{High school: } E[\text{income} | \text{age}, \text{HS}] &= \beta_1 + \beta_2 \text{age}, \\ \text{Bachelor's: } E[\text{income} | \text{age}, \text{B}] &= \beta_1 + \beta_2 \text{age} + \delta_B, \\ \text{Master's: } E[\text{income} | \text{age}, \text{M}] &= \beta_1 + \beta_2 \text{age} + \delta_M, \\ \text{Ph.D.: } E[\text{income} | \text{age}, \text{P}] &= \beta_1 + \beta_2 \text{age} + \delta_P.\end{aligned}$$

The differences between, say,  $\delta_P$  and  $\delta_M$  and between  $\delta_M$  and  $\delta_B$  are of interest. Obviously, these are simple to compute. An alternative way to formulate the equation that reveals these differences directly is to redefine the dummy variables to be 1 if the individual has the degree, rather than whether the degree is the highest degree obtained. Thus, for someone with a Ph.D., all three binary variables are 1, and so on. By defining the variables in this fashion, the regression is now

$$\begin{aligned}\text{High school: } E[\text{income} | \text{age}, \text{HS}] &= \beta_1 + \beta_2 \text{age}, \\ \text{Bachelor's: } E[\text{income} | \text{age}, \text{B}] &= \beta_1 + \beta_2 \text{age} + \delta_B, \\ \text{Master's: } E[\text{income} | \text{age}, \text{M}] &= \beta_1 + \beta_2 \text{age} + \delta_B + \delta_M, \\ \text{Ph.D.: } E[\text{income} | \text{age}, \text{P}] &= \beta_1 + \beta_2 \text{age} + \delta_B + \delta_M + \delta_P.\end{aligned}$$

Instead of the difference between a Ph.D. and the base case, in this model  $\delta_P$  is the marginal value of the Ph.D. How equations with dummy variables are formulated is a matter of convenience. All the results can be obtained from a basic equation.

### 6.2.5 TREATMENT EFFECTS AND DIFFERENCE IN DIFFERENCES REGRESSION

Researchers in many fields have studied the effect of a **treatment** on some kind of **response**. Examples include the effect of going to college on lifetime income [Dale and Krueger (2002)], the effect of cash transfers on child health [Gertler (2004)], the effect of participation in job training programs on income [LaLonde (1986)] and pre-versus postregime shifts in macroeconomic models [Mankiw (2006)], to name but a few. These examples can be formulated in regression models involving a single dummy variable:



$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \delta D_i + \varepsilon_i,$$

where the shift parameter,  $\delta$ , measures the impact of the treatment or the policy change (conditioned on  $\mathbf{x}$ ) on the sampled individuals. In the simplest case of a comparison of one group to another,

$$y_i = \beta_1 + \beta_2 D_i + \varepsilon_i,$$

we will have  $b_1 = (\bar{y}|D_i = 0)$ , that is, the average outcome of those who did not experience the intervention, and  $b_2 = (\bar{y}|D_i = 1) - (\bar{y}|D_i = 0)$ , the difference in the means of the two groups. In the Dale and Krueger (2002) study, the model compared the incomes of students who attended elite colleges to those who did not. When the analysis is of an intervention that occurs over time, such as Krueger's (1999) analysis of the Tennessee STAR experiment in which school performance measures were observed before and after a policy dictated a change in class sizes, the treatment dummy

## 156 PART I ♦ The Linear Regression Model

variable will be a period indicator,  $D_t = 0$  in period 1 and 1 in period 2. The effect in  $\beta_2$  measures the change in the outcome variable, for example, school performance, pre- to postintervention;  $b_2 = \bar{y}_1 - \bar{y}_0$ .

The assumption that the treatment group does not change from period 1 to period 2 weakens this comparison. A strategy for strengthening the result is to include in the sample a group of **control observations** that do not receive the treatment. The change in the outcome for the **treatment group** can then be compared to the change for the **control group** under the presumption that the difference is due to the intervention. An intriguing application of this strategy is often used in clinical trials for health interventions to accommodate the **placebo effect**. The placebo “effect” is a controversial, but apparently tangible outcome in some clinical trials in which subjects “respond” to the treatment even when the treatment is a  intervention, such as a sugar or starch pill in a drug trial. [See Hróbjartsson and Peter C. Gøtzsche, 2001]. A broad template for assessment of the results of such a clinical trial is as follows: The subjects who receive the placebo are the controls. The outcome variable—level of cholesterol for example—is measured at the baseline for both groups. The treatment group receives the drug; the control group receives the placebo, and the outcome variable is measured posttreatment. The impact is measured by the difference in differences,

$$E = [(\bar{y}_{exit}|treatment) - (\bar{y}_{baseline}|treatment)] - [(\bar{y}_{exit}|placebo) - (\bar{y}_{baseline}|placebo)].$$

The presumption is that the difference in differences measurement is robust to the placebo effect *if it exists*. If there is no placebo effect, the result is even stronger (assuming there is a result).

An increasingly common social science application of treatment effect models with dummy variables is in the evaluation of the effects of discrete changes in policy.<sup>2</sup> A pioneering application is the study of the Manpower Development and Training Act (MDTA) by Ashenfelter and Card (1985). The simplest form of the model is one with a pre- and posttreatment observation on a group, where the outcome variable is  $y$ , with

$$y_{it} = \beta_1 + \beta_2 T_t + \beta_3 D_i + \beta_4 T_t \times D_i + \varepsilon, \quad t = 1, 2. \quad (6-3)$$

In this model,  $T_t$  is a dummy variable that is zero in the pretreatment period and one after the treatment and  $D_i$  equals one for those individuals who received the “treatment.” The change in the outcome variable for the “treated” individuals will be

$$(y_{i2}|D_i = 1) - (y_{i1}|D_i = 1) = (\beta_1 + \beta_2 + \beta_3 + \beta_4) - (\beta_1 + \beta_3) = \beta_2 + \beta_4.$$

For the controls, this is

$$(y_{i2}|D_i = 0) - (y_{i1}|D_i = 0) = (\beta_1 + \beta_2) - (\beta_1) = \beta_2.$$

The **difference in differences** is

$$[(y_{i2}|D_i = 1) - (y_{i1}|D_i = 1)] - [(y_{i2}|D_i = 0) - (y_{i1}|D_i = 0)] = \beta_4.$$

<sup>2</sup>Surveys of literatures on treatment effects, including use of D-i-D estimators, are provided by Imbens and Wooldridge (2009) and Millimet, Smith, and Vytlacil (2008).

## CHAPTER 6 ♦ Functional Form and Structural Change **157**

In the multiple regression of  $y_{it}$  on a constant,  $T$ ,  $D$  and  $TD$ , the least squares estimate of  $\beta_4$  will equal the difference in the changes in the means,

$$\begin{aligned} b_4 &= (\bar{y}|D = 1, \text{Period } 2) - (\bar{y}|D = 1, \text{Period } 1) \\ &\quad - (\bar{y}|D = 0, \text{Period } 2) - (\bar{y}|D = 0, \text{Period } 1) \\ &= \Delta\bar{y}|\text{treatment} - \Delta\bar{y}|\text{control}. \end{aligned}$$

The regression is called a difference in differences estimator in reference to this result.

When the treatment is the result of a policy change or event that occurs completely outside the context of the study, the analysis is often termed a **natural experiment**. Card's (1990) study of a major immigration into Miami in 1979 discussed in Example 6.5 is an application.

### **Example 6.5 A Natural Experiment: The Mariel Boatlift**

A sharp change in policy can constitute a natural experiment. An example studied by Card (1990) is the Mariel boatlift from Cuba to Miami (May–September 1980) which increased the Miami labor force by 7 percent. The author examined the impact of this abrupt change in labor market conditions on wages and employment for nonimmigrants. The model compared Miami to a similar city, Los Angeles. Let  $i$  denote an individual and  $D$  denote the “treatment,” which for an individual would be equivalent to “lived in a city that experienced the immigration.” For an individual in either Miami or Los Angeles, the outcome variable is

$$(Y_i) = 1 \text{ if they are unemployed and 0 if they are employed.}$$

Let  $c$  denote the city and let  $t$  denote the period, before (1979) or after (1981) the immigration. Then, the unemployment rate in city  $c$  at time  $t$  is  $E[y_{i,0}|c, t]$  if there is no immigration and it is  $E[y_{i,1}|c, t]$  if there is the immigration. These rates are assumed to be constants. Then,

$$\begin{aligned} E[y_{i,0}|c, t] &= \beta_t + \gamma_c && \text{without the immigration,} \\ E[y_{i,1}|c, t] &= \beta_t + \gamma_c + \delta && \text{with the immigration.} \end{aligned}$$

The effect of the immigration on the unemployment rate is measured by  $\delta$ . The natural experiment is that the immigration occurs in Miami and not in Los Angeles but is not a result of any action by the people in either city. Then,

$$\begin{aligned} E[y_i|M, 79] &= \beta_{79} + \gamma_M && \text{and } E[y_i|M, 81] = \beta_{81} + \gamma_M + \delta && \text{for Miami,} \\ E[y_i|L, 79] &= \beta_{79} + \gamma_L && \text{and } E[y_i|L, 81] = \beta_{81} + \gamma_L && \text{for Los Angeles.} \end{aligned}$$

It is assumed that unemployment growth in the two cities would be the same if there were no immigration. If neither city experienced the immigration, the change in the unemployment rate would be

$$\begin{aligned} E[y_{i,0}|M, 81] - E[y_{i,0}|M, 79] &= \beta_{81} - \beta_{79} && \text{for Miami,} \\ E[y_{i,0}|L, 81] - E[y_{i,0}|L, 79] &= \beta_{81} - \beta_{79} && \text{for Los Angeles.} \end{aligned}$$

If both cities were exposed to migration,

$$\begin{aligned} E[y_{i,1}|M, 81] - E[y_{i,1}|M, 79] &= \beta_{81} - \beta_{79} + \delta && \text{for Miami} \\ E[y_{i,1}|L, 81] - E[y_{i,1}|L, 79] &= \beta_{81} - \beta_{79} + \delta && \text{for Los Angeles.} \end{aligned}$$

Only Miami experienced the migration (the “treatment”). The difference in differences that quantifies the result of the experiment is

$$\{E[y_{i,1}|M, 81] - E[y_{i,1}|M, 79]\} - \{E[y_{i,1}|L, 81] - E[y_{i,1}|L, 79]\} = \delta.$$

## 158 PART I ♦ The Linear Regression Model

The author examined changes in employment rates and wages in the two cities over several years after the boatlift. The effects were surprisingly modest given the scale of the experiment in Miami.

One of the important issues in policy analysis concerns measurement of such treatment effects when the dummy variable results from an individual participation decision. In the clinical trial example given earlier, the control observations (it is assumed) do not know they are in the control group. The treatment assignment is exogenous to the experiment. In contrast, in Keueger and Dale's study, the assignment to the treatment group, attended the elite college, is completely voluntary and determined by the individual. A crucial aspect of the analysis in this case is to accommodate the almost certain outcome that the "treatment dummy" might be measuring the latent motivation and initiative of the participants rather than the effect of the program itself. That is the main appeal of the natural experiment approach—it more closely (possibly exactly) replicates the exogenous treatment assignment of a clinical trial.<sup>3</sup> We will examine some of these cases in Chapters 8 and 18.

### 6.3 NONLINEARITY IN THE VARIABLES

It is useful at this point to write the linear regression model in a very general form: Let  $\mathbf{z} = z_1, z_2, \dots, z_L$  be a set of  $L$  independent variables; let  $f_1, f_2, \dots, f_K$  be  $K$  linearly independent functions of  $\mathbf{z}$ ; let  $g(y)$  be an observable function of  $y$ ; and retain the usual assumptions about the disturbance. The linear regression model is

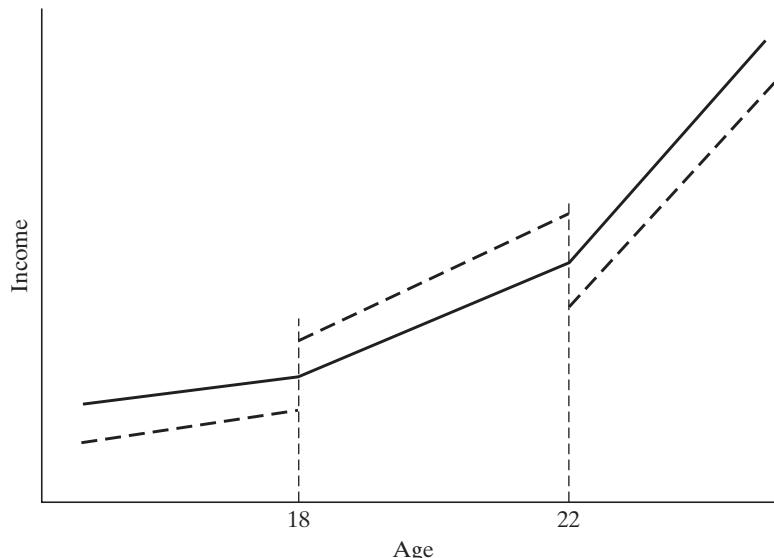
$$\begin{aligned} g(y) &= \beta_1 f_1(\mathbf{z}) + \beta_2 f_2(\mathbf{z}) + \cdots + \beta_K f_K(\mathbf{z}) + \varepsilon \\ &= \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \varepsilon \\ &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon. \end{aligned} \tag{6-4}$$

By using logarithms, exponentials, reciprocals, transcendental functions, polynomials, products, ratios, and so on, this "linear" model can be tailored to any number of situations.

#### 6.3.1 PIECEWISE LINEAR REGRESSION

If one is examining income data for a large cross section of individuals of varying ages in a population, then certain patterns with regard to some age thresholds will be clearly evident. In particular, throughout the range of values of age, income will be rising, but the slope might change at some distinct milestones, for example, at age 18, when the typical individual graduates from high school, and at age 22, when he or she graduates from college. The **time profile** of income for the typical individual in this population might appear as in Figure 6.2. Based on the discussion in the preceding paragraph, we could fit such a regression model just by dividing the sample into three subsamples. However, this would neglect the continuity of the proposed function. The result would appear more like the dotted figure than the continuous function we had in mind. Restricted

<sup>3</sup>See Angrist and Krueger (2001) and Angrist and Pischke (2010) for discussions of this approach.

CHAPTER 6 ♦ Functional Form and Structural Change **159****FIGURE 6.2** Spline Function.

regression and what is known as a **spline** function can be used to achieve the desired effect.<sup>4</sup>

The function we wish to estimate is

$$\begin{aligned} E[\text{income} | \text{age}] &= \alpha^0 + \beta^0 \text{age} && \text{if } \text{age} < 18, \\ & & & \alpha^1 + \beta^1 \text{age} && \text{if } \text{age} \geq 18 \text{ and } \text{age} < 22, \\ & & & \alpha^2 + \beta^2 \text{age} && \text{if } \text{age} \geq 22. \end{aligned}$$

The threshold values, 18 and 22, are called **knots**. Let

$$\begin{aligned} d_1 &= 1 && \text{if } \text{age} \geq t_1^*, \\ d_2 &= 1 && \text{if } \text{age} \geq t_2^*, \end{aligned}$$

where  $t_1^* = 18$  and  $t_2^* = 22$ . To combine all three equations, we use

$$\text{income} = \beta_1 + \beta_2 \text{age} + \gamma_1 d_1 + \delta_1 d_1 \text{age} + \gamma_2 d_2 + \delta_2 d_2 \text{age} + \varepsilon.$$

This relationship is the dashed function in Figure 6.2. The slopes in the three segments are  $\beta_2$ ,  $\beta_2 + \delta_1$ , and  $\beta_2 + \delta_1 + \delta_2$ . To make the function **piecewise continuous**, we require that the segments join at the knots—that is,

$$\beta_1 + \beta_2 t_1^* = (\beta_1 + \gamma_1) + (\beta_2 + \delta_1) t_1^*$$

and

$$(\beta_1 + \gamma_1) + (\beta_2 + \delta_1) t_2^* = (\beta_1 + \gamma_1 + \gamma_2) + (\beta_2 + \delta_1 + \delta_2) t_2^*.$$

---

<sup>4</sup>An important reference on this subject is Poirier (1974). An often-cited application appears in Garber and Poirier (1974).

## 160 PART I ♦ The Linear Regression Model

These are linear restrictions on the coefficients. Collecting terms, the first one is

$$\gamma_1 + \delta_1 t_1^* = 0 \quad \text{or} \quad \gamma_1 = -\delta_1 t_1^*.$$

Doing likewise for the second and inserting these in (6-3), we obtain

$$\text{income} = \beta_1 + \beta_2 \text{age} + \delta_1 d_1 (\text{age} - t_1^*) + \delta_2 d_2 (\text{age} - t_2^*) + \varepsilon.$$

Constrained least squares estimates are obtainable by multiple regression, using a constant and the variables

$$x_1 = \text{age},$$

$$x_2 = \text{age} - 18 \quad \text{if } \text{age} \geq 18 \text{ and 0 otherwise,}$$

and

$$x_3 = \text{age} - 22 \quad \text{if } \text{age} \geq 22 \text{ and 0 otherwise.}$$

We can test the hypothesis that the slope of the function is constant with the joint test of the two restrictions  $\delta_1 = 0$  and  $\delta_2 = 0$ .

### 6.3.2 FUNCTIONAL FORMS

A commonly used form of regression model is the **loglinear model**,

$$\ln y = \ln \alpha + \sum_k \beta_k \ln X_k + \varepsilon = \beta_1 + \sum_k \beta_k x_k + \varepsilon.$$

In this model, the coefficients are elasticities:

$$\left( \frac{\partial y}{\partial X_k} \right) \left( \frac{X_k}{y} \right) = \frac{\partial \ln y}{\partial \ln X_k} = \beta_k. \quad (6-5)$$

In the loglinear equation, measured changes are in proportional or percentage terms;  $\beta_k$  measures the percentage change in  $y$  associated with a 1 percent change in  $X_k$ . This removes the units of measurement of the variables from consideration in using the regression model. An alternative approach sometimes taken is to measure the variables and associated changes in standard deviation units. If the data are “standardized” before estimation using  $x_{ik}^* = (x_{ik} - \bar{x}_k)/s_k$  and likewise for  $y$ , then the least squares regression coefficients measure changes in standard deviation units rather than natural units or percentage terms. (Note that the constant term disappears from this regression.) It is not necessary actually to transform the data to produce these results; multiplying each least squares coefficient  $b_k$  in the original regression by  $s_k/s_y$  produces the same result.

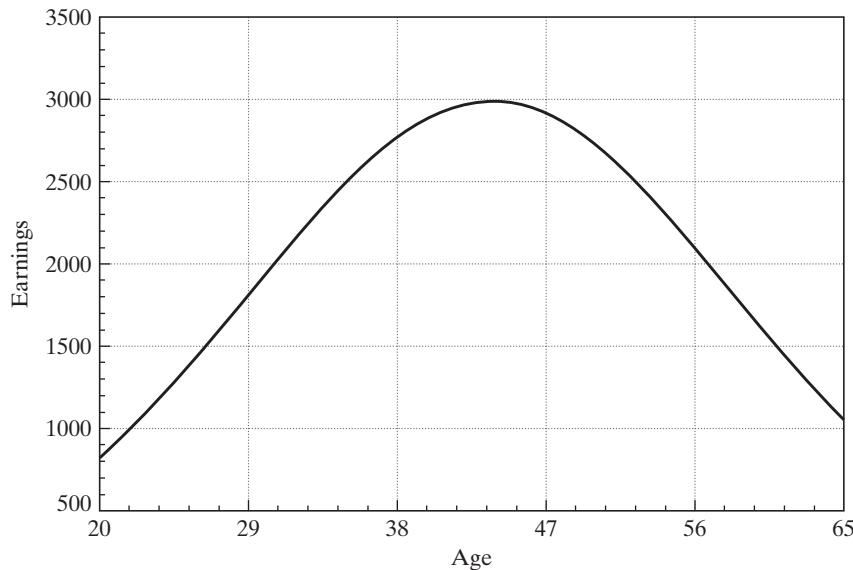
A hybrid of the linear and loglinear models is the semilog equation

$$\ln y = \beta_1 + \beta_2 x + \varepsilon. \quad (6-6)$$

We used this form in the investment equation in Section 5.2.2,

$$\ln I_t = \beta_1 + \beta_2 (i_t - \Delta p_t) + \beta_3 \Delta p_t + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t,$$

where the log of investment is modeled in the levels of the real interest rate, the price level, and a time trend. In a semilog equation with a time trend such as this one,  $d \ln I/dt = \beta_5$  is the average rate of growth of  $I$ . The estimated value of  $-0.00566$  in Table 5.2 suggests that over the full estimation period, after accounting for all other factors, the average rate of growth of investment was  $-0.566$  percent per year.



**FIGURE 6.3** Age-Earnings Profile.

The coefficients in the semilog model are partial- or semi-elasticities; in (6-6),  $\beta_2$  is  $\partial \ln y / \partial x$ . This is a natural form for models with dummy variables such as the earnings equation in Example 5.2. The coefficient on *Kids* of  $-0.35$  suggests that all else equal, earnings are approximately 35 percent less when there are children in the household.

The quadratic earnings equation in Example 6.1 shows another use of nonlinearities in the variables. Using the results in Example 6.1, we find that for a woman with 12 years of schooling and children in the household, the age-earnings profile appears as in Figure 6.3. This figure suggests an important question in this framework. It is tempting to conclude that Figure 6.3 shows the earnings trajectory of a person at different ages, but that is not what the data provide. The model is based on a cross section, and what it displays is the earnings of different people of different ages. How this profile relates to the expected earnings path of one individual is a different, and complicated question.

### 6.3.3 INTERACTION EFFECTS

Another useful formulation of the regression model is one with **interaction terms**. For example, a model relating braking distance  $D$  to speed  $S$  and road wetness  $W$  might be

$$D = \beta_1 + \beta_2 S + \beta_3 W + \beta_4 SW + \varepsilon.$$

In this model,

$$\frac{\partial E[D | S, W]}{\partial S} = \beta_2 + \beta_4 W,$$

which implies that the **marginal effect** of higher speed on braking distance is increased when the road is wetter (assuming that  $\beta_4$  is positive). If it is desired to form confidence intervals or test hypotheses about these marginal effects, then the necessary standard

## 162 PART I ♦ The Linear Regression Model

error is computed from

$$\text{Var}\left(\frac{\partial \hat{E}[D|S, W]}{\partial S}\right) = \text{Var}[\hat{\beta}_2] + W^2 \text{Var}[\hat{\beta}_4] + 2W \text{Cov}[\hat{\beta}_2, \hat{\beta}_4],$$

and similarly for  $\partial E[D|S, W]/\partial W$ . A value must be inserted for  $W$ . The sample mean is a natural choice, but for some purposes, a specific value, such as an extreme value of  $W$  in this example, might be preferred.

### 6.3.4 IDENTIFYING NONLINEARITY

If the functional form is not known a priori, then there are a few approaches that may help at least to identify any nonlinearity and provide some information about it from the sample. For example, if the suspected nonlinearity is with respect to a single regressor in the equation, then fitting a quadratic or cubic polynomial rather than a linear function may capture some of the nonlinearity. By choosing several ranges for the regressor in question and allowing the slope of the function to be different in each range, a piecewise linear approximation to the nonlinear function can be fit.

#### **Example 6.6 Functional Form for a Nonlinear Cost Function**

In a celebrated study of economies of scale in the U.S. electric power industry, Nerlove (1963) analyzed the production costs of 145 American electricity generating companies. This study produced several innovations in microeconomics. It was among the first major applications of statistical cost analysis. The theoretical development in Nerlove's study was the first to show how the fundamental theory of duality between production and cost functions could be used to frame an econometric model. Finally, Nerlove employed several useful techniques to sharpen his basic model.

The focus of the paper was economies of scale, typically modeled as a characteristic of the production function. He chose a Cobb-Douglas function to model output as a function of capital,  $K$ , labor,  $L$ , and fuel,  $F$ :

$$Q = \alpha_0 K^{\alpha_K} L^{\alpha_L} F^{\alpha_F} e^{\varepsilon_i},$$

where  $Q$  is output and  $\varepsilon_i$  embodies the unmeasured differences across firms. The economies of scale parameter is  $r = \alpha_K + \alpha_L + \alpha_F$ . The value 1 indicates constant returns to scale. In this study, Nerlove investigated the widely accepted assumption that producers in this industry enjoyed substantial economies of scale. The production model is loglinear, so assuming that other conditions of the classical regression model are met, the four parameters could be estimated by least squares. However, he argued that the three factors could not be treated as exogenous variables. For a firm that optimizes by choosing its factors of production, the demand for fuel would be  $F^* = F^*(Q, P_K, P_L, P_F)$  and likewise for labor and capital, so certainly the assumptions of the classical model are violated.

In the regulatory framework in place at the time, state commissions set rates and firms met the demand forthcoming at the regulated prices. Thus, it was argued that output (as well as the factor prices) could be viewed as exogenous to the firm and, based on an argument by Zellner, Kmenta, and Dreze (1966), Nerlove argued that at equilibrium, the deviation of costs from the long-run optimum  would be independent of output. (This has a testable implication which we will explore in Chapter 8.) Thus, the firm's objective was cost minimization subject to the constraint of the production function. This can be formulated as a Lagrangean problem,

$$\text{Min}_{K,L,F} P_K K + P_L L + P_F F + \lambda(Q - \alpha_0 K^{\alpha_K} L^{\alpha_L} F^{\alpha_F}).$$

The solution to this minimization problem is the three factor demands and the multiplier (which measures marginal cost). Inserted back into total costs, this produces an (intrinsically

CHAPTER 6 ♦ Functional Form and Structural Change **163****TABLE 6.4** Cobb–Douglas Cost Functions (standard errors in parentheses)

	$\log Q$	$\log P_L - \log P_F$	$\log P_K - \log P_F$	$R^2$
All firms	0.721 (0.0174)	0.593 (0.205)	-0.0085 (0.191)	0.932
Group 1	0.400	0.615	-0.081	0.513
Group 2	0.658	0.094	0.378	0.633
Group 3	0.938	0.402	0.250	0.573
Group 4	0.912	0.507	0.093	0.826
Group 5	1.044	0.603	-0.289	0.921

linear) loglinear cost function,

$$P_K K + P_L L + P_F F = C(Q, P_K, P_L, P_F) = r A Q^{1/r} P_K^{\alpha_K/r} P_L^{\alpha_L/r} P_F^{\alpha_F/r} e^{\varepsilon_i/r},$$

or

$$\ln C = \beta_1 + \beta_q \ln Q + \beta_K \ln P_K + \beta_L \ln P_L + \beta_F \ln P_F + u_i, \quad (6-7)$$

where  $\beta_q = 1/(\alpha_K + \alpha_L + \alpha_F)$  is now the parameter of interest and  $\beta_j = \alpha_j/r$ ,  $j = K, L, F$ . Thus, the duality between production and cost functions has been used to derive the estimating equation from first principles.

A complication remains. The cost parameters must sum to one;  $\beta_K + \beta_L + \beta_F = 1$ , so estimation must be done subject to this constraint.<sup>5</sup> This restriction can be imposed by regressing  $\ln(C/P_F)$  on a constant,  $\ln Q$ ,  $\ln(P_K/P_F)$ , and  $\ln(P_L/P_F)$ . This first set of results appears at the top of Table 6.4.<sup>6</sup>

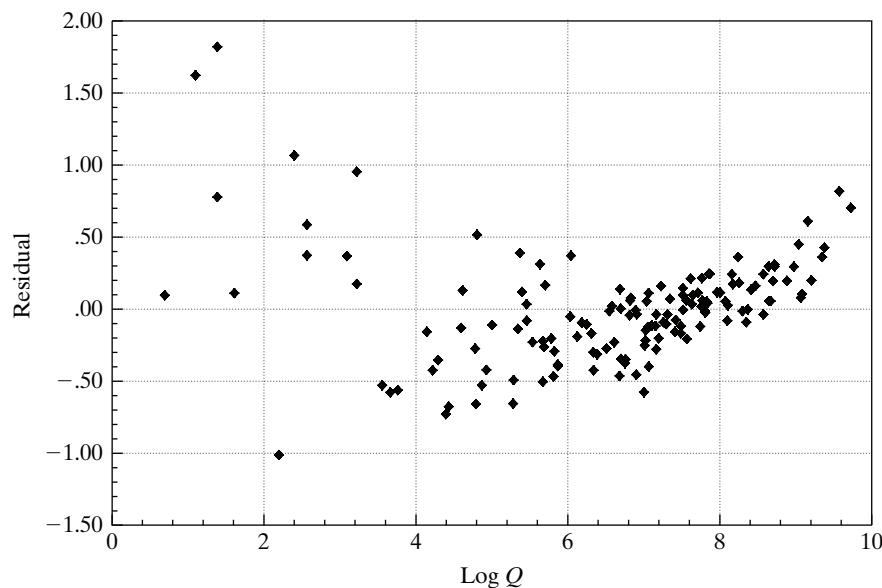
Initial estimates of the parameters of the cost function are shown in the top row of Table 6.4. The hypothesis of constant returns to scale can be firmly rejected. The  $t$  ratio is  $(0.721 - 1)/0.0174 = -16.03$ , so we conclude that this estimate is significantly less than 1 or, by implication,  $r$  is significantly greater than 1. Note that the coefficient on the capital price is negative. In theory, this should equal  $\alpha_K/r$ , which (unless the marginal product of capital is negative) should be positive. Nerlove attributed this to measurement error in the capital price variable. This seems plausible, but it carries with it the implication that the other coefficients are mismeasured as well. [Christensen and Greene (1976) estimator of this model with these data produced a positive estimate. See Section 10.4.2.]

The striking pattern of the residuals shown in Figure 6.4 and some thought about the implied form of the production function suggested that something was missing from the model.<sup>7</sup> In theory, the estimated model implies a continually declining average cost curve,

<sup>5</sup>In the context of the econometric model, the restriction has a testable implication by the definition in Chapter 5. But, the underlying economics require this restriction—it was used in deriving the cost function. Thus, it is unclear what is implied by a test of the restriction. Presumably, if the hypothesis of the restriction is rejected, the analysis should stop at that point, since without the restriction, the cost function is not a valid representation of the production function. We will encounter this conundrum again in another form in Chapter 10. Fortunately, in this instance, the hypothesis is not rejected. (It is in the application in Chapter 10.)

<sup>6</sup>Readers who attempt to replicate Nerlove's study should note that he used common (base 10) logs in his calculations, not natural logs. A practical tip: to convert a natural log to a common log, divide the former by  $\log_{10} e = 2.302585093$ . Also, however, although the first 145 rows of the data in Appendix Table F6.2 are accurately transcribed from the original study, the only regression listed in Table 6.3 that can be reproduced with these data is the first one. The results for Groups 1–5 in the table have been recomputed here and do not match Nerlove's results. Likewise, the results in Table 6.4 have been recomputed and do not match the original study.

<sup>7</sup>A Durbin–Watson test of correlation among the residuals (see Section 20.7) revealed to the author a substantial autocorrelation. Although normally used with time series data, the Durbin–Watson statistic and a test for “autocorrelation” can be a useful tool for determining the appropriate functional form in a cross-sectional model. To use this approach, it is necessary to sort the observations based on a variable of interest (output). Several clusters of residuals of the same sign suggested a need to reexamine the assumed functional form.

**164 PART I ♦ The Linear Regression Model**

**FIGURE 6.4** Residuals from Predicted Cost.

which in turn implies persistent economies of scale at all levels of output. This conflicts with the textbook notion of a U-shaped average cost curve and appears implausible for the data. Note the three clusters of residuals in the figure. Two approaches were used to extend the model.

By sorting the sample into five groups of 29 firms on the basis of output and fitting separate regressions to each group, Nerlove fit a piecewise loglinear model. The results are given in the lower rows of Table 6.4, where the firms in the successive groups are progressively larger. The results are persuasive that the (log)linear cost function is inadequate. The output coefficient that rises toward and then crosses 1.0 is consistent with a U-shaped cost curve as surmised earlier.

A second approach was to expand the cost function to include a quadratic term in log output. This approach corresponds to a much more general model and produced the results given in Table 6.5. Again, a simple  $t$  test strongly suggests that increased generality is called for;  $t = 0.051/0.00054 = 9.44$ . The output elasticity in this quadratic model is  $\beta_q + 2\gamma_{qq} \log Q$ .<sup>8</sup> There are economies of scale when this value is less than 1 and constant returns to scale when it equals 1. Using the two values given in the table (0.152 and 0.0052, respectively), we find that this function does, indeed, produce a U-shaped average cost curve with minimum at  $\ln Q = (1 - 0.152)/(2 \times 0.051) = 8.31$ , or  $Q = 4079$ , which is roughly in the middle of the range of outputs for Nerlove's sample of firms.

This study was updated by Christensen and Greene (1976). Using the same data but a more elaborate (translog) functional form and by simultaneously estimating the factor demands and the cost function, they found results broadly similar to Nerlove's. Their preferred functional form did suggest that Nerlove's generalized model in Table 6.5 did somewhat underestimate the range of outputs in which unit costs of production would continue to decline. They also redid the study using a sample of 123 firms from 1970 and found similar results.

<sup>8</sup>Nerlove inadvertently measured economies of scale from this function as  $1/(\beta_q + \delta \log Q)$ , where  $\beta_q$  and  $\delta$  are the coefficients on  $\log Q$  and  $\log^2 Q$ . The correct expression would have been  $1/[\partial \log C / \partial \log Q] = 1/[\beta_q + 2\delta \log Q]$ . This slip was periodically rediscovered in several later papers.

CHAPTER 6 ♦ Functional Form and Structural Change **165****TABLE 6.5** Log-Quadratic Cost Function (standard errors in parentheses)

	$\log Q$	$\log^2 Q$	$\log P_L - \log P_F$	$\log P_K - \log P_F$	$R^2$
All firms	0.152 (0.062)	0.051 (0.0054)	0.481 (0.161)	0.074 (0.150)	0.96

In the latter sample, however, it appeared that many firms had expanded rapidly enough to exhaust the available economies of scale. We will revisit the 1970 data set in a study of production costs in Chapters 10 and 18.

The preceding example illustrates three useful tools in identifying and dealing with unspecified nonlinearity: analysis of residuals, the use of piecewise linear regression, and the use of polynomials to approximate the unknown regression function.

### 6.3.5 INTRINSICALLY LINEAR MODELS

The loglinear model illustrates an intermediate case of a nonlinear regression model. The equation is **intrinsically linear**, however. By taking logs of  $Y_i = \alpha X_i^{\beta_2} e^{\varepsilon_i}$ , we obtain

$$\ln Y_i = \ln \alpha + \beta_2 \ln X_i + \varepsilon_i$$

or

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i.$$

Although this equation is linear in most respects, something has changed in that it is no longer linear in  $\alpha$ . Written in terms of  $\beta_1$ , we obtain a fully linear model. But that may not be the form of interest. Nothing is lost, of course, since  $\beta_1$  is just  $\ln \alpha$ . If  $\beta_1$  can be estimated, then an obvious estimator of  $\alpha$  is suggested,  $\hat{\alpha} = \exp(\beta_1)$ .

This fact leads us to a useful aspect of intrinsically linear models; they have an “invariance property.” Using the nonlinear least squares procedure described in the next chapter, we could estimate  $\alpha$  and  $\beta_2$  directly by minimizing the sum of squares function;

$$\text{Minimize with respect to } (\alpha, \beta_2) : S(\alpha, \beta_2) = \sum_{i=1}^n (\ln Y_i - \ln \alpha - \beta_2 \ln X_i)^2. \quad (6-8)$$

This is a complicated mathematical problem because of the appearance of the term  $\ln \alpha$ . However, the equivalent linear least squares problem,

$$\text{Minimize with respect to } (\beta_1, \beta_2) : S(\beta_1, \beta_2) = \sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2, \quad (6-9)$$

is simple to solve with the least squares estimator we have used up to this point. The invariance feature that applies is that the two sets of results will be numerically identical; we will get the identical result from estimating  $\alpha$  using (6-8) and from using  $\exp(\beta_1)$  from (6-9). By exploiting this result, we can broaden the definition of linearity and include some additional cases that might otherwise be quite complex.

**166 PART I ♦ The Linear Regression Model**
**TABLE 6.6** Estimates of the Regression in a Gamma Model: Least Squares versus Maximum Likelihood

	$\beta$		$\rho$	
	<i>Estimate</i>	<i>Standard Error</i>	<i>Estimate</i>	<i>Standard Error</i>
Least squares	-1.708	8.689	2.426	1.592
Maximum likelihood	-4.719	2.345	3.151	0.794

**DEFINITION 6.1 Intrinsic Linearity**

In the classical linear regression model, if the  $K$  parameters  $\beta_1, \beta_2, \dots, \beta_K$  can be written as  $K$  one-to-one, possibly nonlinear functions of a set of  $K$  underlying parameters  $\theta_1, \theta_2, \dots, \theta_K$ , then the model is intrinsically linear in  $\theta$ .

**Example 6.7 Intrinsically Linear Regression**

In Section 14.6.4, we will estimate by maximum likelihood the parameters of the model

$$f(y | \beta, x) = \frac{(\beta + x)^{-\rho}}{\Gamma(\rho)} y^{\rho-1} e^{-y/(\beta+x)}.$$

In this model,  $E[y|x] = (\beta\rho) + \rho x$ , which suggests another way that we might estimate the two parameters. This function is an intrinsically linear regression model,  $E[y|x] = \beta_1 + \beta_2 x$ , in which  $\beta_1 = \beta\rho$  and  $\beta_2 = \rho$ . We can estimate the parameters by least squares and then retrieve the estimate of  $\beta$  using  $b_1/b_2$ . Because this value is a nonlinear function of the estimated parameters, we use the delta method to estimate the standard error. Using the data from that example,<sup>9</sup> the least squares estimates of  $\beta_1$  and  $\beta_2$  (with standard errors in parentheses) are -4.1431 (23.734) and 2.4261 (1.5915). The estimated covariance is -36.979. The estimate of  $\beta$  is  $-4.1431/2.4261 = -1.7077$ . We estimate the sampling variance of  $\hat{\beta}$  with

$$\begin{aligned} \text{Est. Var}[\hat{\beta}] &= \left( \frac{\partial \hat{\beta}}{\partial b_1} \right)^2 \widehat{\text{Var}}[b_1] + \left( \frac{\partial \hat{\beta}}{\partial b_2} \right)^2 \widehat{\text{Var}}[b_2] + 2 \left( \frac{\partial \hat{\beta}}{\partial b_1} \right) \left( \frac{\partial \hat{\beta}}{\partial b_2} \right) \widehat{\text{Cov}}[b_1, b_2] \\ &= 8.6889^2. \end{aligned}$$

Table 6.6 compares the least squares and maximum likelihood estimates of the parameters. The lower standard errors for the maximum likelihood estimates result from the inefficient (equal) weighting given to the observations by the least squares procedure. The gamma distribution is highly skewed. In addition, we know from our results in Appendix C that this distribution is an exponential family. We found for the gamma distribution that the sufficient statistics for this density were  $\sum_i y_i$  and  $\sum_i \ln y_i$ . The least squares estimator does not use the second of these, whereas an efficient estimator will.

The emphasis in intrinsic linearity is on “one to one.” If the conditions are met, then the model can be estimated in terms of the functions  $\beta_1, \dots, \beta_K$ , and the underlying parameters derived after these are estimated. The one-to-one correspondence is an **identification condition**. If the condition is met, then the underlying parameters of the

<sup>9</sup>The data are given in Appendix Table FC.1.

## CHAPTER 6 ♦ Functional Form and Structural Change **167**

regression ( $\theta$ ) are said to be **exactly identified** in terms of the parameters of the linear model  $\beta$ . An excellent example is provided by Kmenta (1986, p. 515, and 1967).

**Example 6.8 CES Production Function**

The constant elasticity of substitution production function may be written

$$\ln y = \ln \gamma - \frac{\nu}{\rho} \ln[\delta K^{-\rho} + (1 - \delta)L^{-\rho}] + \varepsilon. \quad (6-10)$$

A Taylor series approximation to this function around the point  $\rho = 0$  is

$$\begin{aligned} \ln y &= \ln \gamma + \nu \delta \ln K + \nu(1 - \delta) \ln L + \rho \nu \delta(1 - \delta) \left\{ -\frac{1}{2} [\ln K - \ln L]^2 \right\} + \varepsilon' \\ &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon', \end{aligned} \quad (6-11)$$

where  $x_1 = 1$ ,  $x_2 = \ln K$ ,  $x_3 = \ln L$ ,  $x_4 = -\frac{1}{2} \ln^2(K/L)$ , and the transformations are

$$\begin{aligned} \beta_1 &= \ln \gamma, & \beta_2 &= \nu \delta, & \beta_3 &= \nu(1 - \delta), & \beta_4 &= \rho \nu \delta(1 - \delta), \\ \gamma &= e^{\beta_1}, & \delta &= \beta_2 / (\beta_2 + \beta_3), & \nu &= \beta_2 + \beta_3, & \rho &= \beta_4(\beta_2 + \beta_3) / (\beta_2 \beta_3). \end{aligned} \quad (6-12)$$

Estimates of  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$ , and  $\beta_4$  can be computed by least squares. The estimates of  $\gamma$ ,  $\delta$ ,  $\nu$ , and  $\rho$  obtained by the second row of (6-12) are the same as those we would obtain had we found the nonlinear least squares estimates of (6-11) directly. As Kmenta shows, however, they are not the same as the nonlinear least squares estimates of (6-10) due to the use of the Taylor series approximation to get to (6-11). We would use the delta method to construct the estimated asymptotic covariance matrix for the estimates of  $\theta' = [\gamma, \delta, \nu, \rho]$ . The derivatives matrix is

$$\mathbf{C} = \frac{\partial \theta}{\partial \beta'} = \begin{bmatrix} e^{\beta_1} & 0 & 0 & 0 \\ 0 & \beta_3 / (\beta_2 + \beta_3)^2 & -\beta_2 / (\beta_2 + \beta_3)^2 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & -\beta_3 \beta_4 / (\beta_2^2 \beta_3) & -\beta_2 \beta_4 / (\beta_2 \beta_3^2) & (\beta_2 + \beta_3) / (\beta_2 \beta_3) \end{bmatrix}.$$

The estimated covariance matrix for  $\hat{\theta}$  is  $\hat{\mathbf{C}} [s^2(\mathbf{X}'\mathbf{X})^{-1}]\hat{\mathbf{C}}'$ .

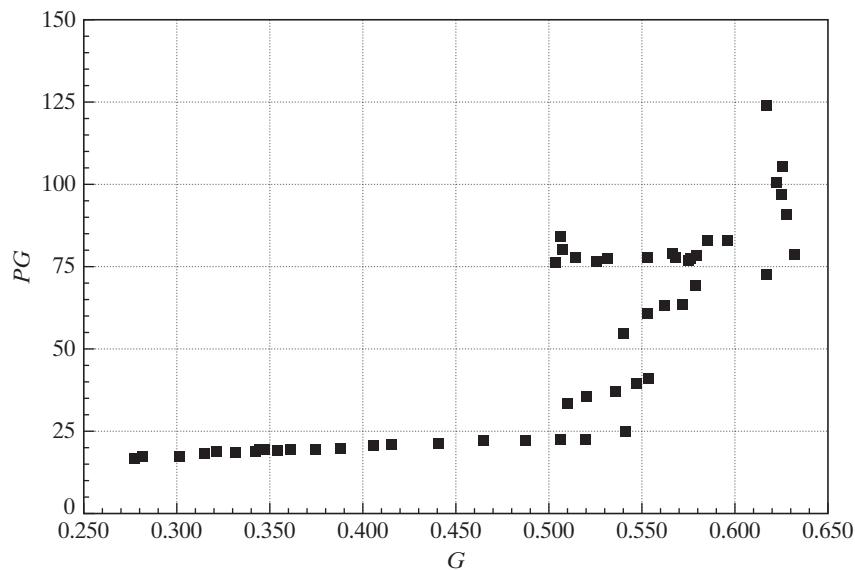
Not all models of the form

$$y_i = \beta_1(\theta)x_{i1} + \beta_2(\theta)x_{i2} + \cdots + \beta_K(\theta)x_{ik} + \varepsilon_i \quad (6-13)$$

are intrinsically linear. Recall that the condition that the functions be one to one (i.e., that the parameters be exactly identified) was required. For example,

$$y_i = \alpha + \beta x_{i1} + \gamma x_{i2} + \beta \gamma x_{i3} + \varepsilon_i$$

is nonlinear. The reason is that if we write it in the form of (6-13), we fail to account for the condition that  $\beta_4$  equals  $\beta_2 \beta_3$ , which is a **nonlinear restriction**. In this model, the three parameters  $\alpha$ ,  $\beta$ , and  $\gamma$  are **overidentified** in terms of the four parameters  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$ , and  $\beta_4$ . Unrestricted least squares estimates of  $\beta_2$ ,  $\beta_3$ , and  $\beta_4$  can be used to obtain two estimates of each of the underlying parameters, and there is no assurance that these will be the same. Models that are not intrinsically linear are treated in Chapter 7.

**168 PART I ♦ The Linear Regression Model**


**FIGURE 6.5** Gasoline Price and Per Capita Consumption, 1953–2004.

## 6.4 MODELING AND TESTING FOR A STRUCTURAL BREAK

One of the more common applications of the  $F$  test is in tests of **structural change**.<sup>10</sup> In specifying a regression model, we assume that its assumptions apply to all the observations in our sample. It is straightforward, however, to test the hypothesis that some or all of the regression coefficients are different in different subsets of the data. To analyze a number of examples, we will revisit the data on the U.S. gasoline market that we examined in Examples 2.3, 4.2, 4.4, 4.8 and 4.9. As Figure 6.5 suggests, this market behaved in predictable, unremarkable fashion prior to the oil shock of 1973 and was quite volatile thereafter. The large jumps in price in 1973 and 1980 are clearly visible, as is the much greater variability in consumption.<sup>11</sup> It seems unlikely that the same regression model would apply to both periods.

### 6.4.1 DIFFERENT PARAMETER VECTORS

The gasoline consumption data span two very different periods. Up to 1973, fuel was plentiful and world prices for gasoline had been stable or falling for at least two decades. The embargo of 1973 marked a transition in this market, marked by shortages, rising prices, and intermittent turmoil. It is possible that the entire relationship described by our regression model changed in 1974. To test this as a hypothesis, we could proceed as follows: Denote the first 21 years of the data in  $\mathbf{y}$  and  $\mathbf{X}$  as  $\mathbf{y}_1$  and  $\mathbf{X}_1$  and the remaining

<sup>10</sup>This test is often labeled a **Chow test**, in reference to Chow (1960).

<sup>11</sup>The observed data will doubtless reveal similar disruption in 2006.

## CHAPTER 6 ♦ Functional Form and Structural Change **169**

years as  $y_2$  and  $\mathbf{X}_2$ . An unrestricted regression that allows the coefficients to be different in the two periods is

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}. \quad (6-14)$$

Denoting the data matrices as  $\mathbf{y}$  and  $\mathbf{X}$ , we find that the unrestricted least squares estimator is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} \mathbf{X}'_1\mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}'_2\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'_1 y_1 \\ \mathbf{X}'_2 y_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}, \quad (6-15)$$

which is least squares applied to the two equations separately. Therefore, the total sum of squared residuals from this regression will be the sum of the two residual sums of squares from the two separate regressions:

$$\mathbf{e}'\mathbf{e} = \mathbf{e}'_1\mathbf{e}_1 + \mathbf{e}'_2\mathbf{e}_2.$$

The restricted coefficient vector can be obtained in two ways. Formally, the restriction  $\beta_1 = \beta_2$  is  $\mathbf{R}\beta = \mathbf{q}$ , where  $\mathbf{R} = [\mathbf{I} : -\mathbf{I}]$  and  $\mathbf{q} = \mathbf{0}$ . The general result given earlier can be applied directly. An easier way to proceed is to build the restriction directly into the model. If the two coefficient vectors are the same, then (6-14) may be written

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix},$$

and the restricted estimator can be obtained simply by stacking the data and estimating a single regression. The residual sum of squares from this restricted regression,  $\mathbf{e}'_*\mathbf{e}_*$ , then forms the basis for the test. The test statistic is then given in (5-16), where  $J$ , the number of restrictions, is the number of columns in  $\mathbf{X}_2$  and the denominator degrees of freedom is  $n_1 + n_2 - 2k$ .

### 6.4.2 INSUFFICIENT OBSERVATIONS

In some circumstances, the data series are not long enough to estimate one or the other of the separate regressions for a test of structural change. For example, one might surmise that consumers took a year or two to adjust to the turmoil of the two oil price shocks in 1973 and 1979, but that the market never actually fundamentally changed or that it only changed temporarily. We might consider the same test as before, but now only single out the four years 1974, 1975, 1980, and 1981 for special treatment. Because there are six coefficients to estimate but only four observations, it is not possible to fit the two separate models. Fisher (1970) has shown that in such a circumstance, a valid way to proceed is as follows:

1. Estimate the regression, using the full data set, and compute the restricted sum of squared residuals,  $\mathbf{e}'_*\mathbf{e}_*$ .
2. Use the longer (adequate) subperiod ( $n_1$  observations) to estimate the regression, and compute the unrestricted sum of squares,  $\mathbf{e}'_1\mathbf{e}_1$ . This latter computation is done assuming that with only  $n_2 < K$  observations, we could obtain a perfect fit and thus contribute zero to the sum of squares.

## 170 PART I ♦ The Linear Regression Model

3. The  $F$  statistic is then computed, using

$$F[n_2, n_1 - K] = \frac{(\mathbf{e}'_* \mathbf{e}_* - \mathbf{e}'_1 \mathbf{e}_1)/n_2}{\mathbf{e}'_1 \mathbf{e}_1/(n_1 - K)}. \quad (6-16)$$

Note that the numerator degrees of freedom is  $n_2$ , not  $K$ .<sup>12</sup> This test has been labeled the **Chow predictive test** because it is equivalent to extending the restricted model to the shorter subperiod and basing the test on the prediction errors of the model in this latter period.

### 6.4.3 CHANGE IN A SUBSET OF COEFFICIENTS

The general formulation previously suggested lends itself to many variations that allow a wide range of possible tests. Some important particular cases are suggested by our gasoline market data. One possible description of the market is that after the oil shock of 1973, Americans simply reduced their consumption of gasoline by a fixed proportion, but other relationships in the market, such as the income elasticity, remained unchanged. This case would translate to a simple shift downward of the loglinear regression model or a reduction only in the constant term. Thus, the unrestricted equation has separate coefficients in the two periods, while the restricted equation is a pooled regression with separate constant terms. The regressor matrices for these two cases would be of the form

$$\text{(unrestricted) } \mathbf{X}_U = \begin{bmatrix} \mathbf{i} & \mathbf{0} & \mathbf{W}_{\text{pre}73} & \mathbf{0} \\ \mathbf{0} & \mathbf{i} & \mathbf{0} & \mathbf{W}_{\text{post}73} \end{bmatrix}$$

and

$$\text{(restricted) } \mathbf{X}_R = \begin{bmatrix} \mathbf{i} & \mathbf{0} & \mathbf{W}_{\text{pre}73} \\ \mathbf{0} & \mathbf{i} & \mathbf{W}_{\text{post}73} \end{bmatrix}.$$

The first two columns of  $\mathbf{X}_U$  are dummy variables that indicate the subperiod in which the observation falls.

Another possibility is that the constant and one or more of the slope coefficients changed, but the remaining parameters remained the same. The results in Example 6.9 suggest that the constant term and the price and income elasticities changed much more than the cross-price elasticities and the time trend. The Chow test for this type of restriction looks very much like the one for the change in the constant term alone. Let  $\mathbf{Z}$  denote the variables whose coefficients are believed to have changed, and let  $\mathbf{W}$  denote the variables whose coefficients are thought to have remained constant. Then, the regressor matrix in the constrained regression would appear as

$$\mathbf{X} = \begin{bmatrix} \mathbf{i}_{\text{pre}} & \mathbf{Z}_{\text{pre}} & \mathbf{0} & \mathbf{0} & \mathbf{W}_{\text{pre}} \\ \mathbf{0} & \mathbf{0} & \mathbf{i}_{\text{post}} & \mathbf{Z}_{\text{post}} & \mathbf{W}_{\text{post}} \end{bmatrix}. \quad (6-17)$$

As before, the unrestricted coefficient vector is the combination of the two separate regressions.

---

<sup>12</sup>One way to view this is that only  $n_2 < K$  coefficients are needed to obtain this perfect fit.

#### 6.4.4 TESTS OF STRUCTURAL BREAK WITH UNEQUAL VARIANCES

An important assumption made in using the Chow test is that the disturbance variance is the same in both (or all) regressions. In the restricted model, if this is not true, the first  $n_1$  elements of  $\epsilon$  have variance  $\sigma_1^2$ , whereas the next  $n_2$  have variance  $\sigma_2^2$ , and so on. The restricted model is, therefore, heteroscedastic, and our results for the classical regression model no longer apply. As analyzed by Schmidt and Sickles (1977), Ohtani and Toyoda (1985), and Toyoda and Ohtani (1986), it is quite likely that the actual probability of a type I error will be larger than the significance level we have chosen. (That is, we shall regard as large an  $F$  statistic that is actually less than the *appropriate* but unknown critical value.) Precisely how severe this effect is going to be will depend on the data and the extent to which the variances differ, in ways that are not likely to be obvious.

If the sample size is reasonably large, then we have a test that is valid whether or not the disturbance variances are the same. Suppose that  $\hat{\theta}_1$  and  $\hat{\theta}_2$  are two consistent and asymptotically normally distributed estimators of a parameter based on independent samples,<sup>13</sup> with asymptotic covariance matrices  $\mathbf{V}_1$  and  $\mathbf{V}_2$ . Then, under the null hypothesis that the true parameters are the same,

$$\hat{\theta}_1 - \hat{\theta}_2 \text{ has mean } \mathbf{0} \text{ and asymptotic covariance matrix } \mathbf{V}_1 + \mathbf{V}_2.$$

Under the null hypothesis, the Wald statistic,

$$W = (\hat{\theta}_1 - \hat{\theta}_2)'(\hat{\mathbf{V}}_1 + \hat{\mathbf{V}}_2)^{-1}(\hat{\theta}_1 - \hat{\theta}_2), \quad (6-18)$$

has a limiting chi-squared distribution with  $K$  degrees of freedom. A test that the difference between the parameters is zero can be based on this statistic.<sup>14</sup> It is straightforward to apply this to our test of common parameter vectors in our regressions. Large values of the statistic lead us to reject the hypothesis.

In a small or moderately sized sample, the Wald test has the unfortunate property that the probability of a type I error is persistently larger than the critical level we use to carry it out. (That is, we shall too frequently reject the null hypothesis that the parameters are the same in the subsamples.) We should be using a larger critical value. Ohtani and Kobayashi (1986) have devised a “bounds” test that gives a partial remedy for the problem.<sup>15</sup>

It has been observed that the size of the **Wald test** may differ from what we have assumed, and that the deviation would be a function of the alternative hypothesis. There are two general settings in which a test of this sort might be of interest. For comparing two possibly different populations—such as the labor supply equations for men versus women—not much more can be said about the suggested statistic in the absence of specific information about the alternative hypothesis. But a great deal of work on this type of statistic has been done in the time-series context. In this instance, the nature of the alternative is rather more clearly defined.

<sup>13</sup>Without the required independence, this test and several similar ones will fail completely. The problem becomes a variant of the famous Behrens–Fisher problem.

<sup>14</sup>See Andrews and Fair (1988). The true size of this suggested test is uncertain. It depends on the nature of the alternative. If the variances are radically different, the assumed critical values might be somewhat unreliable.

<sup>15</sup>See also Kobayashi (1986). An alternative, somewhat more cumbersome test is proposed by Jayatissa (1977). Further discussion is given in Thursby (1982).

## 172 PART I ♦ The Linear Regression Model

### Example 6.9 Structural Break in the Gasoline Market

Figure 6.5 shows a plot of prices and quantities in the U.S. gasoline market from 1953 to 2004. The first 21 points are the layer at the bottom of the figure and suggest an orderly market. The remainder clearly reflect the subsequent turmoil in this market.

We will use the Chow tests described to examine this market. The model we will examine is the one suggested in Example 2.3, with the addition of a time trend:

$$\ln(G/Pop)_t = \beta_1 + \beta_2 \ln(Income/Pop)_t + \beta_3 \ln PG_t + \beta_4 \ln PNC_t + \beta_5 \ln PUC_t + \beta_6 t + \varepsilon_t.$$

The three prices in the equation are for G, new cars and used cars. *Income/Pop* is per capita Income, and *G/Pop* is per capita gasoline consumption. The time trend is computed as  $t = \text{Year} - 1952$ , so in the first period  $t = 1$ . Regression results for four functional forms are shown in Table 6.7. Using the data for the entire sample, 1953 to 2004, and for the two subperiods, 1953 to 1973 and 1974 to 2004, we obtain the three estimated regressions in the first and last two columns. The *F* statistic for testing the restriction that the coefficients in the two equations are the same is

$$F[6, 40] = \frac{(0.101997 - (0.00202244 + 0.007127899))/6}{(0.00202244 + 0.007127899)/(21 + 31 - 12)} = 67.645.$$

The tabled critical value is 2.336, so, consistent with our expectations, we would reject the hypothesis that the coefficient vectors are the same in the two periods. Using the full set of 52 observations to fit the model, the sum of squares is  $e^T e^* = 0.101997$ . When the  $n_2 = 4$  observations for 1974, 1975, 1980, and 1981 are removed from the sample, the sum of squares falls to  $e^T e^* = 0.0973936$ . The *F* statistic is 0.496. Because the tabled critical value for  $F[4, 48 - 6]$  is 2.594, we would not reject the hypothesis of stability. The conclusion to this point would be that although something has surely changed in the market, the hypothesis of a temporary disequilibrium seems not to be an adequate explanation.

An alternative way to compute this statistic might be more convenient. Consider the original arrangement, with all 52 observations. We now add to this regression four binary variables, Y1974, Y1975, Y1980, and Y1981. Each of these takes the value one in the single year indicated and zero in all 51 remaining years. We then compute the regression with the original six variables and these four additional dummy variables. The sum of squared residuals in this regression is 0.0973936 (precisely the same as when the four observations are deleted from the sample—see Exercise 7 in Chapter 3), so the *F* statistic for testing the joint hypothesis that the four coefficients are zero is

$$F[4, 42] = \frac{(0.101997 - 0.0973936)/4}{0.0973936/(52 - 6 - 4)} = 0.496$$

once again. (See Section 6.4.2 for discussion of this test.)

**TABLE 6.7** Gasoline Consumption Functions

Coefficients	1953–2004	Pooled	Preshock	Postshock
Constant	-26.6787	-24.9009	-22.1647	
Constant		-24.8167		-15.3283
$\ln Income/Pop$	1.6250	1.4562	0.8482	0.3739
$\ln PG$	-0.05392	-0.1132	-0.03227	-0.1240
$\ln PNC$	-0.08343	-0.1044	0.6988	-0.001146
$\ln PUC$	-0.08467	-0.08646	-0.2905	-0.02167
Year	-0.01393	-0.009232	0.01006	0.004492
$R^2$	0.9649	0.9683	0.9975	0.9529
Standard error	0.04709	0.04524	0.01161	0.01689
Sum of squares	0.101997	0.092082	0.00202244	0.007127899

## CHAPTER 6 ♦ Functional Form and Structural Change 173

The  $F$  statistic for testing the restriction that the coefficients in the two equations are the same apart from the constant term is based on the last three sets of results in the table:

$$F[5, 40] = \frac{(0.092082 - (0.00202244 + 0.007127899)) / 5}{(0.00202244 + 0.007127899) / (21 + 31 - 12)} = 72.506.$$

The tabled critical value is 2.449, so this hypothesis is rejected as well. The data suggest that the models for the two periods are systematically different, beyond a simple shift in the constant term.

The  $F$  ratio that results from estimating the model subject to the restriction that the two automobile price elasticities and the coefficient on the time trend are unchanged is

$$F[3, 40] = \frac{(0.01441975 - (0.00202244 + 0.007127899)) / 3}{(0.00202244 + 0.007127899) / (52 - 6 - 6)} = 7.678.$$

(The restricted regression is not shown.) The critical value from the  $F$  table is 2.839, so this hypothesis is rejected as well. Note, however, that this value is far smaller than those we obtained previously. This fact suggests that the bulk of the difference in the models across the two periods is, indeed, explained by the changes in the constant and the price and income elasticities.

The test statistic in (6-18) for the regression results in Table 6.7 gives a value of 502.34. The 5 percent critical value from the chi-squared table for six degrees of freedom is 12.59. So, on the basis of the Wald test, we would once again reject the hypothesis that the same coefficient vector applies in the two subperiods 1953 to 1973 and 1974 to 2004. We should note that the Wald statistic is valid only in large samples, and our samples of 21 and 31 observations hardly meet that standard. We have tested the hypothesis that the regression model for the gasoline market changed in 1973, and on the basis of the  $F$  test (Chow test) we strongly rejected the hypothesis of model stability.

### **Example 6.10 The World Health Report**

The 2000 version of the World Health Organization's (WHO) *World Health Report* contained a major country-by-country inventory of the world's health care systems. [World Health Organization (2000). See also <http://www.who.int/whr/en/>.] The book documented years of research and has thousands of pages of material. Among the most controversial and most publicly debated parts of the report was a single chapter that described a comparison of the delivery of health care by 191 countries—nearly all of the world's population. [Evans et al. (2000a,b). See, e.g., Hilts (2000) for reporting in the popular press.] The study examined the efficiency of health care delivery on two measures: the standard one that is widely studied, (disability adjusted) life expectancy (DALE), and an innovative new measure created by the authors that was a composite of five outcomes (COMP) and that accounted for efficiency and fairness in delivery. The regression-style modeling, which was done in the setting of a frontier model (see Chapter 18), related health care attainment to two major inputs, education and (per capita) health care expenditure. The residuals were analyzed to obtain the country comparisons.

The data in Appendix Table F6.3 were used by the researchers at the WHO for the study. (They used a panel of data for the years 1993 to 1997. We have extracted the 1997 data for this example.) The WHO data have been used by many researchers in subsequent analyses. [See, e.g., Hollingsworth and Wildman (2002), Gravelle et al. (2002), and Greene (2004).] The regression model used by the WHO contained DALE or COMP on the left-hand side and health care expenditure, education, and education squared on the right. Greene (2004) added a number of additional variables such as per capita GDP, a measure of the distribution of income, and World Bank measures of government effectiveness and democratization of the political structure.

Among the controversial aspects of the study was the fact that the model aggregated countries of vastly different characteristics. A second striking aspect of the results, suggested in Hilts (2000) and documented in Greene (2004), was that, in fact, the "efficient" countries in the study were the 30 relatively wealthy OECD members, while the rest of the world on average fared much more poorly. We will pursue that aspect here with respect to DALE. Analysis of COMP is left as an exercise. Table 6.8 presents estimates of the regression models for

**174 PART I ♦ The Linear Regression Model**
**TABLE 6.8** Regression Results for Life Expectancy

	<i>All Countries</i>	<i>OECD</i>		<i>Non-OECD</i>	
Constant	25.237	38.734	42.728	49.328	26.812
Health exp	0.00629	-0.00180	0.00268	0.00114	0.00955
Education	7.931	7.178	6.177	5.156	7.0433
Education <sup>2</sup>	-0.439	-0.426	-0.385	-0.329	-0.374
Gini coeff		-17.333		-5.762	-21.329
Tropic		-3.200		-3.298	-3.144
Pop. Dens.		-0.255e-4		0.000167	-0.425e-4
Public exp		-0.0137		-0.00993	-0.00939
PC GDP		0.000483		0.000108	0.000600
Democracy		1.629		-0.546	1.909
Govt. Eff.		0.748		1.224	0.786
R <sup>2</sup>	0.6824	0.7299	0.6483	0.7340	0.6133
Std. Err.	6.984	6.565	1.883	1.916	7.366
Sum of sq.	9121.795	7757.002	92.21064	69.74428	8518.750
N	191		30		161
GDP/Pop	6609.37		18199.07		4449.79
F test	4.524		0.874		3.311

DALE for the pooled sample, the OECD countries, and the non-OECD countries, respectively. Superficially, there do not appear to be very large differences across the two subgroups. We first tested the joint significance of the additional variables, income distribution (GINI), per capita GDP, and so on. For each group, the F statistic is  $[(\mathbf{e}^* \mathbf{e}^* - \mathbf{e}' \mathbf{e}) / 7] / [\mathbf{e}' \mathbf{e} / (n - 11)]$ . These F statistics are shown in the last row of the table. The critical values for F[7,180] (all), F[7,19] (OECD), and F[7,150] (non-OECD) are 2.061, 2.543, and 2.071, respectively. We conclude that the additional explanatory variables are significant contributors to the fit for the non-OECD countries (and for all countries), but not for the OECD countries. Finally, to conduct the structural change test of OECD vs. non-OECD, we compute

$$F[11, 169] = \frac{[7757.007 - (69.74428 + 7378.598)] / 11}{(69.74428 + 7378.598) / (191 - 11 - 11)} = 0.637.$$

The 95 percent critical value for F[11,169] is 1.846. So, we do not reject the hypothesis that the regression model is the same for the two groups of countries. The Wald statistic in (6-18) tells a different story. The statistic is 35.221. The 95 percent critical value from the chi-squared table with 11 degrees of freedom is 19.675. On this basis, we would reject the hypothesis that the two coefficient vectors are the same.

#### 6.4.5 PREDICTIVE TEST OF MODEL STABILITY

The hypothesis test defined in (6-16) in Section 6.4.2 is equivalent to  $H_0: \beta_2 = \beta_1$  in the “model”

$$y_t = \mathbf{x}'_t \boldsymbol{\beta}_1 + \varepsilon_t, \quad t = 1, \dots, T_1$$

$$y_t = \mathbf{x}'_t \boldsymbol{\beta}_2 + \varepsilon_t, \quad t = T_1 + 1, \dots, T_1 + T_2.$$

(Note that the disturbance variance is assumed to be the same in both subperiods.) An alternative formulation of the model (the one used in the example) is

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{I} \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \gamma \end{pmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}.$$

## CHAPTER 6 ♦ Functional Form and Structural Change 175

This formulation states that

$$y_t = \mathbf{x}'_t \boldsymbol{\beta}_1 + \varepsilon_t, \quad t = 1, \dots, T_1$$

$$y_t = \mathbf{x}'_t \boldsymbol{\beta}_2 + \gamma_t + \varepsilon_t, \quad t = T_1 + 1, \dots, T_1 + T_2.$$

Because each  $\gamma_t$  is unrestricted, this alternative formulation states that the regression model of the first  $T_1$  periods ceases to operate in the second subperiod (and, in fact, no systematic model operates in the second subperiod). A test of the hypothesis  $\boldsymbol{\gamma} = \mathbf{0}$  in this framework would thus be a test of model stability. The least squares coefficients for this regression can be found by using the formula for the partitioned inverse matrix

$$\begin{aligned} \begin{pmatrix} \mathbf{b} \\ \mathbf{c} \end{pmatrix} &= \begin{bmatrix} \mathbf{X}'_1 \mathbf{X}_1 + \mathbf{X}'_2 \mathbf{X}_2 & \mathbf{X}'_2 \\ \mathbf{X}_2 & \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'_1 \mathbf{y}_1 + \mathbf{X}'_2 \mathbf{y}_2 \\ \mathbf{y}_2 \end{bmatrix} \\ &= \begin{bmatrix} (\mathbf{X}'_1 \mathbf{X}_1)^{-1} & -(\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_2 \\ -\mathbf{X}_2 (\mathbf{X}'_1 \mathbf{X}_1)^{-1} & \mathbf{I} + \mathbf{X}_2 (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_2 \end{bmatrix} \begin{bmatrix} \mathbf{X}'_1 \mathbf{y}_1 + \mathbf{X}'_2 \mathbf{y}_2 \\ \mathbf{y}_2 \end{bmatrix} \\ &= \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{c}_2 \end{pmatrix} \end{aligned}$$

where  $\mathbf{b}_1$  is the least squares slopes based on the first  $T_1$  observations and  $\mathbf{c}_2$  is  $\mathbf{y}_2 - \mathbf{X}_2 \mathbf{b}_1$ . The covariance matrix for the full set of estimates is  $s^2$  times the bracketed matrix. The two subvectors of residuals in this regression are  $\mathbf{e}_1 = \mathbf{y}_1 - \mathbf{X}_1 \mathbf{b}_1$  and  $\mathbf{e}_2 = \mathbf{y}_2 - (\mathbf{X}_2 \mathbf{b}_1 + \mathbf{I} \mathbf{c}_2) = \mathbf{0}$ , so the sum of squared residuals in this least squares regression is just  $\mathbf{e}'_1 \mathbf{e}_1$ . This is the same sum of squares as appears in (6-16). The degrees of freedom for the denominator is  $[T_1 + T_2 - (K + T_2)] = T_1 - K$  as well, and the degrees of freedom for the numerator is the number of elements in  $\boldsymbol{\gamma}$  which is  $T_2$ . The restricted regression with  $\boldsymbol{\gamma} = \mathbf{0}$  is the pooled model, which is likewise the same as appears in (6-16). This implies that the  $F$  statistic for testing the null hypothesis in this model is precisely that which appeared earlier in (6-16), which suggests why the test is labeled the “predictive test.”

## 6.5 SUMMARY AND CONCLUSIONS

This chapter has discussed the functional form of the regression model. We examined the use of dummy variables and other transformations to build nonlinearity into the model. We then considered other nonlinear models in which the parameters of the nonlinear model could be recovered from estimates obtained for a linear regression. The final sections of the chapter described hypothesis tests designed to reveal whether the assumed model had changed during the sample period, or was different for different groups of observations.

### Key Terms and Concepts

- Binary variable
- Dummy variable
- Intrinsicly linear
- Chow test
- Dummy variable trap
- Knots
- Control group
- Exactly identified
- Loglinear model
- Control observations
- Identification condition
- Marginal effect
- Difference in differences
- Interaction terms
- Natural experiment

## 176 PART I ♦ The Linear Regression Model

- Nonlinear restriction
- Qualification indices
- Threshold effects
- Overidentified
- Response
- Time profile
- Piecewise continuous
- Semilog equation
- Treatment
- Placebo effect
- Spline
- Treatment group
- Predictive test
- Structural change
- Wald test

### Exercises

1. A regression model with  $K = 16$  independent variables is fit using a panel of seven years of data. The sums of squares for the seven separate regressions and the pooled regression are shown below. The model with the pooled data allows a separate constant for each year. Test the hypothesis that the same coefficients apply in every year.

	<b>1954</b>	<b>1955</b>	<b>1956</b>	<b>1957</b>	<b>1958</b>	<b>1959</b>	<b>1960</b>	<b>All</b>
Observations	65	55	87	95	103	87	78	570
$\mathbf{e}'\mathbf{e}$	104	88	206	144	199	308	211	1425

2. *Reverse regression.* A common method of analyzing statistical data to detect discrimination in the workplace is to fit the regression

$$y = \alpha + \mathbf{x}'\boldsymbol{\beta} + \gamma d + \varepsilon, \quad (1)$$

where  $y$  is the wage rate and  $d$  is a dummy variable indicating either membership ( $d = 1$ ) or nonmembership ( $d = 0$ ) in the class toward which it is suggested the discrimination is directed. The regressors  $\mathbf{x}$  include factors specific to the particular type of job as well as indicators of the qualifications of the individual. The hypothesis of interest is  $H_0: \gamma \geq 0$  versus  $H_1: \gamma < 0$ . The regression seeks to answer the question, “In a given job, are individuals in the class ( $d = 1$ ) paid less than equally qualified individuals not in the class ( $d = 0$ )?” Consider an alternative approach. Do individuals in the class in the same job as others, and receiving the same wage, uniformly have higher qualifications? If so, this might also be viewed as a form of discrimination. To analyze this question, Conway and Roberts (1983) suggested the following procedure:

1. Fit (1) by ordinary least squares. Denote the estimates  $a$ ,  $\mathbf{b}$ , and  $c$ .
2. Compute the set of **qualification indices**,

$$\mathbf{q} = a\mathbf{i} + \mathbf{X}\mathbf{b}. \quad (2)$$

Note the omission of  $cd$  from the fitted value.

3. Regress  $\mathbf{q}$  on a constant,  $\mathbf{y}$  and  $\mathbf{d}$ . The equation is

$$\mathbf{q} = \alpha_* + \beta_*\mathbf{y} + \gamma_*\mathbf{d} + \varepsilon_*. \quad (3)$$

The analysis suggests that if  $\gamma_* < 0$ ,  $\gamma_* > 0$ .

- a. Prove that the theory notwithstanding, the least squares estimates  $c$  and  $c_*$  are related by

$$c_* = \frac{(\bar{y}_1 - \bar{y})(1 - R^2)}{(1 - P)(1 - r_{yd}^2)} - c, \quad (4)$$

## CHAPTER 6 ♦ Functional Form and Structural Change 177

where

$\bar{y}_1$  = mean of  $y$  for observations with  $d = 1$ ,

$\bar{y}$  = mean of  $y$  for all observations,

$P$  = mean of  $d$ ,

$R^2$  = coefficient of determination for (1),

$r_{yd}^2$  = squared correlation between  $y$  and  $d$ .

[Hint: The model contains a constant term. Thus, to simplify the algebra, assume that all variables are measured as deviations from the overall sample means and use a partitioned regression to compute the coefficients in (3). Second, in (2), use the result that based on the least squares results  $\mathbf{y} = \mathbf{ai} + \mathbf{Xb} + \mathbf{cd} + \mathbf{e}$ , so  $\mathbf{q} = \mathbf{y} - \mathbf{cd} - \mathbf{e}$ . From here on, we drop the constant term. Thus, in the regression in (3) you are regressing  $[\mathbf{y} - \mathbf{cd} - \mathbf{e}]$  on  $\mathbf{y}$  and  $\mathbf{d}$ .]

- b. Will the sample evidence necessarily be consistent with the theory? [Hint: Suppose that  $c = 0$ .]

A symposium on the Conway and Roberts paper appeared in the *Journal of Business and Economic Statistics* in April 1983.

3. *Reverse regression continued.* This and the next exercise continue the analysis of Exercise 2. In Exercise 2, interest centered on a particular dummy variable in which the regressors were accurately measured. Here we consider the case in which the crucial regressor in the model is measured with error. The paper by Kamlich and Polacheck (1982) is directed toward this issue.

Consider the simple errors in the variables model,

$$y = \alpha + \beta x^* + \varepsilon, \quad x = x^* + u,$$

where  $u$  and  $\varepsilon$  are uncorrelated and  $x$  is the erroneously measured, observed counterpart to  $x^*$ .

- a. Assume that  $x^*$ ,  $u$ , and  $\varepsilon$  are all normally distributed with means  $\mu^*$ , 0, and 0, variances  $\sigma_*^2$ ,  $\sigma_u^2$ , and  $\sigma_\varepsilon^2$ , and zero covariances. Obtain the probability limits of the least squares estimators of  $\alpha$  and  $\beta$ .
  - b. As an alternative, consider regressing  $x$  on a constant and  $y$ , and then computing the reciprocal of the estimate. Obtain the probability limit of this estimator.
  - c. Do the “direct” and “reverse” estimators bound the true coefficient?
4. *Reverse regression continued.* Suppose that the model in Exercise 3 is extended to  $y = \beta x^* + \gamma d + \varepsilon$ ,  $x = x^* + u$ . For convenience, we drop the constant term. Assume that  $x^*$ ,  $\varepsilon$ , and  $u$  are independent normally distributed with zero means. Suppose that  $d$  is a random variable that takes the values one and zero with probabilities  $\pi$  and  $1 - \pi$  in the population and is independent of all other variables in the model. To put this formulation in context, the preceding model (and variants of it) have appeared in the literature on discrimination. We view  $y$  as a “wage” variable,  $x^*$  as “qualifications,” and  $x$  as some imperfect measure such as education. The dummy variable  $d$  is membership ( $d = 1$ ) or nonmembership ( $d = 0$ ) in some protected class. The hypothesis of discrimination turns on  $\gamma < 0$  versus  $\gamma \geq 0$ .
- a. What is the probability limit of  $c$ , the least squares estimator of  $\gamma$ , in the least squares regression of  $y$  on  $x$  and  $d$ ? [Hints: The independence of  $x^*$  and  $d$  is important. Also,  $\text{plim } \mathbf{d}'\mathbf{d}/n = \text{Var}[d] + E^2[d] = \pi(1 - \pi) + \pi^2 = \pi$ . This minor modification does not affect the model substantively, but it greatly simplifies the

## 178 PART I ♦ The Linear Regression Model

algebra.] Now suppose that  $x^*$  and  $d$  are not independent. In particular, suppose that  $E[x^* | d = 1] = \mu^1$  and  $E[x^* | d = 0] = \mu^0$ . Repeat the derivation with this assumption.

- b. Consider, instead, a regression of  $x$  on  $y$  and  $d$ . What is the probability limit of the coefficient on  $d$  in this regression? Assume that  $x^*$  and  $d$  are independent.
- c. Suppose that  $x^*$  and  $d$  are not independent, but  $\gamma$  is, in fact, less than zero. Assuming that both preceding equations still hold, what is estimated by  $(\bar{y} | d = 1) - (\bar{y} | d = 0)$ ? What does this quantity estimate if  $\gamma$  does equal zero?

### Applications

1. In Application 1 in Chapter 3 and Application 1 in Chapter 5, we examined Koop and Tobias's data on wages, education, ability, and so on. We continue the analysis here. (The source, location and configuration of the data are given in the earlier application.) We consider the model

$$\begin{aligned}\ln \text{Wage} = & \beta_1 + \beta_2 \text{Educ} + \beta_3 \text{Ability} + \beta_4 \text{Experience} \\ & + \beta_5 \text{Mother's education} + \beta_6 \text{Father's education} + \beta_7 \text{Broken home} \\ & + \beta_8 \text{Siblings} + \varepsilon.\end{aligned}$$

- a. Compute the full regression by least squares and report your results. Based on your results, what is the estimate of the marginal value, in \$/hour, of an additional year of education, for someone who has 12 years of education when all other variables are at their means and  $\text{Broken home} = 0$ ?
- b. We are interested in possible nonlinearities in the effect of education on  $\ln \text{Wage}$ . (Koop and Tobias focused on experience. As before, we are not attempting to replicate their results.) A histogram of the education variable shows values from 9 to 20, a huge spike at 12 years (high school graduation) and, perhaps surprisingly, a second at 15—intuition would have anticipated it at 16. Consider aggregating the education variable into a set of dummy variables:

$$HS = 1 \text{ if } \text{Educ} \leq 12, 0 \text{ otherwise}$$

$$Col = 1 \text{ if } \text{Educ} > 12 \text{ and } \text{Educ} \leq 16, 0 \text{ otherwise}$$

$$Grad = 1 \text{ if } \text{Educ} > 16, 0 \text{ otherwise.}$$

Replace  $\text{Educ}$  in the model with  $(Col, Grad)$ , making high school ( $HS$ ) the base category, and recompute the model. Report all results. How do the results change? Based on your results, what is the marginal value of a college degree? (This is actually the marginal value of having 16 years of education—in recent years, college graduation has tended to require somewhat more than four years on average.) What is the marginal impact on  $\ln \text{Wage}$  of a graduate degree?

- c. The aggregation in part b actually loses quite a bit of information. Another way to introduce nonlinearity in education is through the function itself. Add  $\text{Educ}^2$  to the equation in part a and recompute the model. Again, report all results. What changes are suggested? Test the hypothesis that the quadratic term in the

CHAPTER 6 ♦ Functional Form and Structural Change **179**

equation is not needed—that is, that its coefficient is zero. Based on your results, sketch a profile of log wages as a function of education.

- d. One might suspect that the value of education is enhanced by greater ability. We could examine this effect by introducing an interaction of the two variables in the equation. Add the variable

$$\text{Educ\_Ability} = \text{Educ} \times \text{Ability}$$

to the base model in part a. Now, what is the marginal value of an additional year of education? The sample mean value of ability is 0.052374. Compute a confidence interval for the marginal impact on  $\ln \text{Wage}$  of an additional year of education for a person of average ability.

- e. Combine the models in c and d. Add both  $\text{Educ}^2$  and  $\text{Educ\_Ability}$  to the base model in part a and reestimate. As before, report all results and describe your findings. If we define “low ability” as less than the mean and “high ability” as greater than the mean, the sample averages are  $-0.798563$  for the 7,864 low-ability individuals in the sample and  $+0.717891$  for the 10,055 high-ability individuals in the sample. Using the formulation in part c, with this new functional form, sketch, describe, and compare the log wage profiles for low- and high-ability individuals.
2. (An extension of Application 1.) Here we consider whether different models as specified in Application 1 would apply for individuals who reside in “Broken homes.” Using the results in Sections 6.4.1 and 6.4.4, test the hypothesis that the same model (not including the *Broken home* dummy variable) applies to both groups of individuals, those with  $\text{Broken home} = 0$  and with  $\text{Broken home} = 1$ .
3. In Solow’s classic (1957) study of technical change in the U.S. economy, he suggests the following aggregate production function:  $q(t) = A(t)f[k(t)]$ , where  $q(t)$  is aggregate output per work hour,  $k(t)$  is the aggregate capital labor ratio, and  $A(t)$  is the technology index. Solow considered four static models,  $q/A = \alpha + \beta \ln k$ ,  $q/A = \alpha - \beta/k$ ,  $\ln(q/A) = \alpha + \beta \ln k$ , and  $\ln(q/A) = \alpha + \beta/k$ . Solow’s data for the years 1909 to 1949 are listed in Appendix Table F6.4.
- a. Use these data to estimate the  $\alpha$  and  $\beta$  of the four functions listed above. (*Note:* Your results will not quite match Solow’s. See the next exercise for resolution of the discrepancy.)
- b. In the aforementioned study, Solow states:

A scatter of  $q/A$  against  $k$  is shown in Chart 4. Considering the amount of a priori doctoring which the raw figures have undergone, the fit is remarkably tight. Except, that is, for the layer of points which are obviously too high. These maverick observations relate to the seven last years of the period, 1943–1949. From the way they lie almost exactly parallel to the main scatter, one is tempted to conclude that in 1943 the aggregate production function simply shifted.

Compute a scatter diagram of  $q/A$  against  $k$  and verify the result he notes above.

- c. Estimate the four models you estimated in the previous problem including a dummy variable for the years 1943 to 1949. How do your results change? (*Note:* These results match those reported by Solow, although he did not report the coefficient on the dummy variable.)

**180 PART I ♦ The Linear Regression Model**

- d. Solow went on to surmise that, in fact, the data were fundamentally different in the years before 1943 than during and after. Use a Chow test to examine the difference in the two subperiods using your four functional forms. Note that with the dummy variable, you can do the test by introducing an interaction term between the dummy and whichever function of  $k$  appears in the regression. Use an  $F$  test to test the hypothesis.
4. Data on the number of incidents of wave damage to a sample of ships, with the type of ship and the period when it was constructed, are given in Table 6.9. There are five types of ships and four different periods of construction. Use  $F$  tests and dummy variable regressions to test the hypothesis that there is no significant “ship type effect” in the expected number of incidents. Now, use the same procedure to test whether there is a significant “period effect.”

**TABLE 6.9 Ship Damage Incidents**

<i>Ship Type</i>	<i>Period Constructed</i>			
	<i>1960–1964</i>	<i>1965–1969</i>	<i>1970–1974</i>	<i>1975–1979</i>
A	0	4	18	11
B	29	53	44	18
C	1	1	2	1
D	0	0	11	4
E	0	7	12	1

*Source:* Data from McCullagh and Nelder (1983, p. 137).

## 7

# NONLINEAR, SEMIPARAMETRIC AND NONPARAMETRIC REGRESSION MODELS<sup>1</sup>

---

## 7.1 INTRODUCTION

Up to this point, the focus has been on a **linear regression model**

$$y = x_1\beta_1 + x_2\beta_2 + \cdots + \varepsilon. \quad (7-1)$$

Chapters 2 to 5 developed the least squares method of estimating the parameters and obtained the statistical properties of the estimator that provided the tools we used for point and interval estimation, hypothesis testing, and prediction. The modifications suggested in Chapter 6 provided a somewhat more general form of the linear regression model,

$$y = f_1(\mathbf{x})\beta_1 + f_2(\mathbf{x})\beta_2 + \cdots + \varepsilon. \quad (7-2)$$

By the definition we want to use in this chapter, this model is still “linear,” because the parameters appear in a linear form. Section 7.2 of this chapter will examine the **nonlinear regression model** (which includes (7-1) and (7-2) as special cases),

$$y = h(x_1, x_2, \dots, x_P; \beta_1, \beta_2, \dots, \beta_K) + \varepsilon, \quad (7-3)$$

where the conditional mean function involves  $P$  variables and  $K$  parameters. This form of the model changes the conditional mean function from  $E[y|\mathbf{x}, \boldsymbol{\beta}] = \mathbf{x}'\boldsymbol{\beta}$  to  $E[y|\mathbf{x}] = h(\mathbf{x}, \boldsymbol{\beta})$  for more general functions. This allows a much wider range of functional forms than the linear model can accommodate.<sup>2</sup> This change in the model form will require us to develop an alternative method of estimation, **nonlinear least squares**. We will also examine more closely the interpretation of parameters in nonlinear models. In particular, since  $\partial E[y|\mathbf{x}]/\partial \mathbf{x}$  is no longer equal to  $\boldsymbol{\beta}$ , we will want to examine how  $\boldsymbol{\beta}$  should be interpreted.

Linear and nonlinear least squares are used to estimate the parameters of the **conditional mean function**,  $E[y|\mathbf{x}]$ . As we saw in Example 4.5, other relationships between  $y$  and  $\mathbf{x}$ , such as the **conditional median**, might be of interest. Section 7.3 revisits this idea with an examination of the conditional median function and the least absolute

---

<sup>1</sup>This chapter covers some fairly advanced features of regression modeling and numerical analysis. It may be bypassed in a first course without loss of continuity.

<sup>2</sup>A complete discussion of this subject can be found in Amemiya (1985). Other important references are Jennrich (1969), Malinvaud (1970), and especially Goldfeld and Quandt (1971, 1972). A very lengthy authoritative treatment is the text by Davidson and MacKinnon (1993).

## 182 PART I ♦ The Linear Regression Model

deviations estimator. This section will also relax the restriction that the model coefficients are always the same in the different parts of the distribution of  $y$  (given  $\mathbf{x}$ ). The LAD estimator estimates the parameters of the conditional median, that is, 50<sup>th</sup> percentile function. The **quantile regression model** allows the parameters of the regression to change as we analyze different parts of the conditional distribution.

The model forms considered thus far are semiparametric in nature, and less parametric as we move from Section 7.2 to 7.3. The **partially linear regression** examined in Section 7.4 extends (7-1) such that  $y = f(x) + \mathbf{z}'\boldsymbol{\beta} + \varepsilon$ . The endpoint of this progression is a model in which the relationship between  $y$  and  $x$  is not forced to conform to a particular parameterized function. Using largely graphical and kernel density methods, we consider in Section 7.5 how to analyze a **nonparametric regression** relationship that essentially imposes little more than  $E[y|\mathbf{x}] = h(\mathbf{x})$ .

## 7.2 NONLINEAR REGRESSION MODELS

The general form of the nonlinear regression model is

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i. \quad (7-4)$$

The linear model is obviously a special case. Moreover, some models which appear to be nonlinear, such as

$$y = e^{\beta_1} x_1^{\beta_2} x_2^{\beta_3} e^\varepsilon,$$

become linear after a transformation, in this case after taking logarithms. In this chapter, we are interested in models for which there is no such transformation, such as the one in the following example.

### **Example 7.1 CES Production Function**

In Example 6.8, we examined a constant elasticity of substitution production function model:

$$\ln y = \ln \gamma - \frac{\nu}{\rho} \ln [\delta K^{-\rho} + (1-\delta)L^{-\rho}] + \varepsilon. \quad (7-5)$$

No transformation reduces this equation to one that is linear in the parameters. In Example 6.5, a linear Taylor series approximation to this function around the point  $\rho = 0$  is used to produce an intrinsically linear equation that can be fit by least squares. Nonetheless, the underlying model in (7.5) is nonlinear in the sense that interests us in this chapter.

This and the next section will extend the assumptions of the linear regression model to accommodate nonlinear functional forms such as the one in Example 7.1. We will then develop the nonlinear least squares estimator, establish its statistical properties, and then consider how to use the estimator for hypothesis testing and analysis of the model predictions.

### 7.2.1 ASSUMPTIONS OF THE NONLINEAR REGRESSION MODEL

We shall require a somewhat more formal definition of a nonlinear regression model. Sufficient for our purposes will be the following, which include the linear model as the special case noted earlier. We assume that there is an underlying probability distribution, or data generating process (DGP) for the observable  $y_i$  and a true parameter vector,  $\boldsymbol{\beta}$ ,

which is a characteristic of that DGP. The following are the assumptions of the nonlinear regression model:

- Functional form:** The conditional mean function for  $y_i$  given  $\mathbf{x}_i$  is

$$E[y_i | \mathbf{x}_i] = h(\mathbf{x}_i, \boldsymbol{\beta}), \quad i = 1, \dots, n,$$

where  $h(\mathbf{x}_i, \boldsymbol{\beta})$  is a continuously differentiable function of  $\boldsymbol{\beta}$ .

- Identifiability of the model parameters:** The parameter vector in the model is identified (estimable) if there is no nonzero parameter  $\boldsymbol{\beta}^0 \neq \boldsymbol{\beta}$  such that  $h(\mathbf{x}_i, \boldsymbol{\beta}^0) = h(\mathbf{x}_i, \boldsymbol{\beta})$  for all  $\mathbf{x}_i$ . In the linear model, this was the full rank assumption, but the simple absence of “multicollinearity” among the variables in  $\mathbf{x}$  is not sufficient to produce this condition in the nonlinear regression model. Example 7.2 illustrates the problem.
- Zero mean of the disturbance:** It follows from Assumption 1 that we may write

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i.$$

where  $E[\varepsilon_i | h(\mathbf{x}_i, \boldsymbol{\beta})] = 0$ . This states that the disturbance at observation  $i$  is uncorrelated with the conditional mean function for all observations in the sample. This is not quite the same as assuming that the disturbances and the exogenous variables are uncorrelated, which is the familiar assumption, however.

- Homoscedasticity and nonautocorrelation:** As in the linear model, we assume conditional homoscedasticity,

$$E[\varepsilon_i^2 | h(\mathbf{x}_j, \boldsymbol{\beta}), j = 1, \dots, n] = \sigma^2, \quad \text{a finite constant,} \quad (7-6)$$

and nonautocorrelation

$$E[\varepsilon_i \varepsilon_j | h(\mathbf{x}_i, \boldsymbol{\beta}), h(\mathbf{x}_j, \boldsymbol{\beta}), j = 1, \dots, n] = 0 \quad \text{for all } j \neq i.$$

- Data generating process:** The data-generating process for  $\mathbf{x}_i$  is assumed to be a well-behaved population such that first and second moments of the data can be assumed to converge to fixed, finite population counterparts. The crucial assumption is that the process generating  $\mathbf{x}_i$  is strictly exogenous to that generating  $\varepsilon_i$ . The data on  $\mathbf{x}_i$  are assumed to be “well behaved.”
- Underlying probability model:** There is a well-defined probability distribution generating  $\varepsilon_i$ . At this point, we assume only that this process produces a sample of uncorrelated, identically (marginally) distributed random variables  $\varepsilon_i$  with mean zero and variance  $\sigma^2$  conditioned on  $h(\mathbf{x}_i, \boldsymbol{\beta})$ . Thus, at this point, our statement of the model is **semiparametric**. (See Section 12.3.) We will not be assuming any particular distribution for  $\varepsilon_i$ . The conditional moment assumptions in 3 and 4 will be sufficient for the results in this chapter. In Chapter 14, we will fully parameterize the model by assuming that the disturbances are normally distributed. This will allow us to be more specific about certain test statistics and, in addition, allow some generalizations of the regression model. The assumption is not necessary here.

#### Example 7.2 Identification in a Translog Demand System

Christensen, Jorgenson, and Lau (1975), proposed the translog **indirect utility function** for a consumer allocating a budget among  $K$  commodities:

$$\ln V = \beta_0 + \sum_{k=1}^K \beta_k \ln(p_k/M) + \sum_{k=1}^K \sum_{j=1}^K \gamma_{kj} \ln(p_k/M) \ln(p_j/M),$$

## 184 PART I ♦ The Linear Regression Model

where  $V$  is indirect utility,  $p_k$  is the price for the  $k$ th commodity, and  $M$  is income. Utility, direct or indirect, is unobservable, so the utility function is not usable as an empirical model. **Roy's identity** applied to this logarithmic function produces a budget share equation for the  $k$ th commodity that is of the form

$$S_k = -\frac{\partial \ln V / \partial \ln p_k}{\partial \ln V / \partial \ln M} = \frac{\beta_k + \sum_{j=1}^K \gamma_{kj} \ln(p_j/M)}{\beta_M + \sum_{j=1}^K \gamma_{Mj} \ln(p_j/M)} + \varepsilon, k = 1, \dots, K,$$

where  $\beta_M = \sum_k \beta_k$  and  $\gamma_{Mj} = \sum_k \gamma_{kj}$ . No transformation of the budget share equation produces a linear model. This is an intrinsically nonlinear regression model. (It is also one among a system of equations, an aspect we will ignore for the present.) Although the share equation is stated in terms of observable variables, it remains unusable as an empirical model because of an **identification problem**. If every parameter in the budget share is multiplied by the same constant, then the constant appearing in both numerator and denominator cancels out, and the same value of the function in the equation remains. The indeterminacy is resolved by imposing the normalization  $\beta_M = 1$ . Note that this sort of identification problem does not arise in the linear model.

### 7.2.2 THE NONLINEAR LEAST SQUARES ESTIMATOR

The nonlinear least squares estimator is defined as the minimizer of the sum of squares,

$$S(\boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^n \varepsilon_i^2 = \frac{1}{2} \sum_{i=1}^n [y_i - h(\mathbf{x}_i, \boldsymbol{\beta})]^2. \quad (7-7)$$

The first order conditions for the minimization are

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n [y_i - h(\mathbf{x}_i, \boldsymbol{\beta})] \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}. \quad (7-8)$$

In the linear model, the vector of partial derivatives will equal the regressors,  $\mathbf{x}_i$ . In what follows, we will identify the derivatives of the conditional mean function with respect to the parameters as the “pseudoregressors,”  $\mathbf{x}_i^0(\boldsymbol{\beta}) = \mathbf{x}_i^0$ . We find that the nonlinear least squares estimator is found as the solutions to

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \mathbf{x}_i^0 \varepsilon_i = \mathbf{0}. \quad (7-9)$$

This is the nonlinear regression counterpart to the least squares normal equations in (3-5). Computation requires an iterative solution. (See Example 7.3.) The method is presented in Section 7.2.6.

Assumptions 1 and 3 imply that  $E[\varepsilon_i | h(\mathbf{x}_i, \boldsymbol{\beta})] = 0$ . In the linear model, it follows, *because of the linearity of the conditional mean*, that  $\varepsilon_i$  and  $\mathbf{x}_i$ , itself, are uncorrelated. However, *uncorrelatedness* of  $\varepsilon_i$  with a particular *nonlinear* function of  $\mathbf{x}_i$  (the regression function) does not necessarily imply uncorrelatedness with  $\mathbf{x}_i$ , itself, nor, for that matter, with other nonlinear functions of  $\mathbf{x}_i$ . On the other hand, the results we will obtain for the behavior of the estimator in this model are couched not in terms of  $\mathbf{x}_i$  but in terms of certain functions of  $\mathbf{x}_i$  (the derivatives of the regression function), so, in point of fact,  $E[\varepsilon | \mathbf{X}] = \mathbf{0}$  is not even the assumption we need.

The foregoing is not a theoretical fine point. Dynamic models, which are very common in the contemporary literature, would greatly complicate this analysis. If it can be assumed that  $\varepsilon_i$  is strictly uncorrelated with any *prior information* in the model, including

## CHAPTER 7 ♦ Nonlinear, Semiparametric 185

previous disturbances, then perhaps a treatment analogous to that for the linear model would apply. But the convergence results needed to obtain the asymptotic properties of the estimator still have to be strengthened. The dynamic nonlinear regression model is beyond the reach of our treatment here. Strict independence of  $\varepsilon_i$  and  $\mathbf{x}_i$  would be sufficient for uncorrelatedness of  $\varepsilon_i$  and every function of  $\mathbf{x}_i$ , but, again, in a dynamic model, this assumption might be questionable. Some commentary on this aspect of the nonlinear regression model may be found in Davidson and MacKinnon (1993, 2004).

If the disturbances in the nonlinear model are normally distributed, then the log of the normal density for the  $i$ th observation will be

$$\ln f(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = -(1/2)\{\ln 2\pi + \ln \sigma^2 + [y_i - h(\mathbf{x}_i, \boldsymbol{\beta})]^2/\sigma^2\}. \quad (7-10)$$

For this special case, we have from item D.2 in Theorem 14.2 (on maximum likelihood estimation), that the derivatives of the log density with respect to the parameters have mean zero. That is,

$$E\left[\frac{\partial \ln f(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}}\right] = E\left[\frac{1}{\sigma^2} \left(\frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right) \varepsilon_i\right] = \mathbf{0}, \quad (7-11)$$

so, in the normal case, the derivatives and the disturbances are uncorrelated. Whether this can be assumed to hold in other cases is going to be model specific, but under reasonable conditions, we would assume so. [See Ruud (2000, p. 540).]

In the context of the linear model, the **orthogonality condition**  $E[\mathbf{x}_i \varepsilon_i] = 0$  produces least squares as a **GMM estimator** for the model. (See Chapter 13.) The orthogonality condition is that the regressors and the disturbance in the model are uncorrelated. In this setting, the same condition applies to the first derivatives of the conditional mean function. The result in (7-11) produces a moment condition which will define the nonlinear least squares estimator as a GMM estimator.

**Example 7.3 First-Order Conditions for a Nonlinear Model**

The first-order conditions for estimating the parameters of the nonlinear regression model,

$$y_i = \beta_1 + \beta_2 e^{\beta_3 x_i} + \varepsilon_i,$$

by nonlinear least squares [see (7-13)] are

$$\begin{aligned} \frac{\partial S(\mathbf{b})}{\partial b_1} &= -\sum_{i=1}^n [y_i - b_1 - b_2 e^{b_3 x_i}] = 0, \\ \frac{\partial S(\mathbf{b})}{\partial b_2} &= -\sum_{i=1}^n [y_i - b_1 - b_2 e^{b_3 x_i}] e^{b_3 x_i} = 0, \\ \frac{\partial S(\mathbf{b})}{\partial b_3} &= -\sum_{i=1}^n [y_i - b_1 - b_2 e^{b_3 x_i}] b_2 x_i e^{b_3 x_i} = 0. \end{aligned}$$

These equations do not have an explicit solution.

Conceding the potential for ambiguity, we define a nonlinear regression model at this point as follows.

**186 PART I ♦ The Linear Regression Model**
**DEFINITION 7.1 Nonlinear Regression Model**

A **nonlinear regression model** is one for which the first-order conditions for least squares estimation of the parameters are nonlinear functions of the parameters.

Thus, nonlinearity is defined in terms of the techniques needed to estimate the parameters, not the shape of the regression function. Later we shall broaden our definition to include other techniques besides least squares.

### 7.2.3 LARGE SAMPLE PROPERTIES OF THE NONLINEAR LEAST SQUARES ESTIMATOR

Numerous analytical results have been obtained for the nonlinear least squares estimator, such as consistency and asymptotic normality. We cannot be sure that nonlinear least squares is the most efficient estimator, except in the case of normally distributed disturbances. (This conclusion is the same one we drew for the linear model.) But, in the semiparametric setting of this chapter, we can ask whether this estimator is optimal in some sense given the information that we do have; the answer turns out to be yes. Some examples that follow will illustrate the points.

It is necessary to make some assumptions about the regressors. The precise requirements are discussed in some detail in Judge et al. (1985), Amemiya (1985), and Davidson and MacKinnon (2004). In the linear regression model, to obtain our asymptotic results, we assume that the sample moment matrix  $(1/n)\mathbf{X}'\mathbf{X}$  converges to a positive definite matrix  $\mathbf{Q}$ . By analogy, we impose the same condition on the derivatives of the regression function, which are called the **pseudoregressors** in the linearized model *when they are computed at the parameter values*. Therefore, for the nonlinear regression model, the analog to (4-21) is

$$\text{plim } \frac{1}{n} \mathbf{X}^0 \mathbf{X}^0 = \text{plim } \frac{1}{n} \sum_{i=1}^n \left( \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0} \right) \left( \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}'_0} \right) = \mathbf{Q}^0, \quad (7-12)$$

where  $\mathbf{Q}^0$  is a positive definite matrix. To establish consistency of  $\mathbf{b}$  in the linear model, we required  $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\epsilon} = \mathbf{0}$ . We will use the counterpart to this for the pseudoregressors:

$$\text{plim } \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^0 \boldsymbol{\varepsilon}_i = \mathbf{0}.$$

This is the orthogonality condition noted earlier in (4-24). In particular, note that orthogonality of the disturbances and the data is not the same condition. Finally, asymptotic normality can be established under general conditions if

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i^0 \boldsymbol{\varepsilon}_i \xrightarrow{d} N[\mathbf{0}, \sigma^2 \mathbf{Q}^0].$$

With these in hand, the asymptotic properties of the nonlinear least squares estimator have been derived. They are, in fact, essentially those we have already seen for the

## CHAPTER 7 ♦ Nonlinear, Semiparametric 187

linear model, except that in this case we place the derivatives of the linearized function evaluated at  $\beta^0$ ,  $\mathbf{X}^0$  in the role of the regressors. [See Amemiya (1985).]

The nonlinear least squares criterion function is

$$S(\mathbf{b}) = \frac{1}{2} \sum_{i=1}^n [y_i - h(\mathbf{x}_i, \mathbf{b})]^2 = \frac{1}{2} \sum_{i=1}^n e_i^2, \quad (7-13)$$

where we have inserted what will be the solution value,  $\mathbf{b}$ . The values of the parameters that minimize (one half of) the sum of squared deviations are the nonlinear least squares estimators. The first-order conditions for a minimum are

$$\mathbf{g}(\mathbf{b}) = - \sum_{i=1}^n [y_i - h(\mathbf{x}_i, \mathbf{b})] \frac{\partial h(\mathbf{x}_i, \mathbf{b})}{\partial \mathbf{b}} = \mathbf{0}. \quad (7-14)$$

In the linear model of Chapter 3, this produces a set of linear equations, the normal equations (3-4). But in this more general case, (7-14) is a set of nonlinear equations that do not have an explicit solution. Note that  $\sigma^2$  is not relevant to the solution [nor was it in (3-4)]. At the solution,

$$\mathbf{g}(\mathbf{b}) = -\mathbf{X}^{0'} \mathbf{e} = \mathbf{0},$$

which is the same as (3-12) for the linear model.

Given our assumptions, we have the following general results:

**THEOREM 7.1 Consistency of the Nonlinear Least Squares Estimator**

If the following assumptions hold;

- a. The parameter space containing  $\beta$  is compact (has no gaps or nonconcave regions),
- b. For any vector  $\beta^0$  in that parameter space,  $\text{plim}(1/n)S(\beta^0) = q(\beta^0)$ , a continuous and differentiable function,
- c.  $q(\beta^0)$  has a unique minimum at the true parameter vector,  $\beta$ ,

then, the nonlinear least squares estimator defined by (7-13) and (7-14) is consistent. We will sketch the proof, then consider why the theorem and the proof differ as they do from the apparently simpler counterpart for the linear model. The proof, notwithstanding the underlying subtleties of the assumptions, is straightforward. The estimator, say,  $\mathbf{b}^0$ , minimizes  $(1/n)S(\beta^0)$ . If  $(1/n)S(\beta^0)$  is minimized for every  $n$ , then it is minimized by  $\mathbf{b}^0$  as  $n$  increases without bound. We also assumed that the minimizer of  $q(\beta^0)$  is uniquely  $\beta$ . If the minimum value of  $\text{plim}(1/n)S(\beta^0)$  equals the probability limit of the minimized value of the sum of squares, the theorem is proved. This equality is produced by the continuity in assumption b.

In the linear model, consistency of the least squares estimator could be established based on  $\text{plim}(1/n)\mathbf{X}'\mathbf{X} = \mathbf{Q}$  and  $\text{plim}(1/n)\mathbf{X}'\mathbf{e} = \mathbf{0}$ . To follow that approach here, we would use the linearized model and take essentially the same result. The loose

## 188 PART I ♦ The Linear Regression Model

end in that argument would be that the linearized model is not the true model, and there remains an approximation. For this line of reasoning to be valid, it must also be either assumed or shown that  $\text{plim}(1/n)\mathbf{X}^0\boldsymbol{\delta} = \mathbf{0}$  where  $\delta_i = h(\mathbf{x}_i, \boldsymbol{\beta})$  minus the Taylor series approximation. An argument to this effect appears in Mittelhammer et al. (2000, pp. 190–191).

Note that no mention has been made of unbiasedness. The linear least squares estimator in the linear regression model is essentially alone in the estimators considered in this book. It is generally not possible to establish unbiasedness for any other estimator. As we saw earlier, unbiasedness is of fairly limited virtue in any event—we found, for example, that the property would not differentiate an estimator based on a sample of 10 observations from one based on 10,000. Outside the linear case, consistency is the primary requirement of an estimator. Once this is established, we consider questions of efficiency and, in most cases, whether we can rely on asymptotic normality as a basis for statistical inference.

### THEOREM 7.2 Asymptotic Normality of the Nonlinear Least Squares Estimator

If the pseudoregressors defined in (7-12) are “well behaved,” then

$$\mathbf{b} \xrightarrow{a} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}(\mathbf{Q}^0)^{-1}\right],$$

where

$$\mathbf{Q}^0 = \text{plim} \frac{1}{n} \mathbf{X}^{0'} \mathbf{X}^0.$$

The sample estimator of the asymptotic covariance matrix is

$$\text{Est. Asy. Var}[\mathbf{b}] = \hat{\sigma}^2 (\mathbf{X}^{0'} \mathbf{X}^0)^{-1}. \quad (7-15)$$

Asymptotic efficiency of the nonlinear least squares estimator is difficult to establish without a distributional assumption. There is an indirect approach that is one possibility. The assumption of the orthogonality of the pseudoregressors and the true disturbances implies that the nonlinear least squares estimator is a GMM estimator in this context. With the assumptions of homoscedasticity and nonautocorrelation, the optimal weighting matrix is the one that we used, which is to say that in the class of GMM estimators for this model, nonlinear least squares uses the optimal weighting matrix. As such, it is asymptotically efficient in the class of GMM estimators.

The requirement that the matrix in (7-12) converges to a positive definite matrix implies that the columns of the regressor matrix  $\mathbf{X}^0$  must be linearly independent. This **identification condition** is analogous to the requirement that the independent variables in the linear model be linearly independent. Nonlinear regression models usually involve several independent variables, and at first blush, it might seem sufficient to examine the data directly if one is concerned with multicollinearity. However, this situation is not the case. Example 7.4 gives an application.

## CHAPTER 7 ♦ Nonlinear, Semiparametric 189

A consistent estimator of  $\sigma^2$  is based on the residuals:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n [y_i - h(\mathbf{x}_i, \boldsymbol{\beta})]^2. \quad (7-16)$$

A degrees of freedom correction,  $1/(n - K)$ , where  $K$  is the number of elements in  $\boldsymbol{\beta}$ , is not strictly necessary here, because all results are asymptotic in any event. Davidson and MacKinnon (2004) argue that on average, (7-16) will underestimate  $\sigma^2$ , and one should use the degrees of freedom correction. Most software in current use for this model does, but analysts will want to verify which is the case for the program they are using. With this in hand, the estimator of the asymptotic covariance matrix for the nonlinear least squares estimator is given in (7-15).

Once the nonlinear least squares estimates are in hand, inference and hypothesis tests can proceed in the same fashion as prescribed in Chapter 5. A minor problem can arise in evaluating the fit of the regression in that the familiar measure,

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (7-17)$$

is no longer guaranteed to be in the range of 0 to 1. It does, however, provide a useful descriptive measure.

#### 7.2.4 HYPOTHESIS TESTING AND PARAMETRIC RESTRICTIONS

In most cases, the sorts of hypotheses one would test in this context will involve fairly simple linear restrictions. The tests can be carried out using the familiar formulas discussed in Chapter 5 and the asymptotic covariance matrix presented earlier. For more involved hypotheses and for nonlinear restrictions, the procedures are a bit less clear-cut. Two principal testing procedures were discussed in Section 5.4: the Wald test, which relies on the consistency and asymptotic normality of the estimator, and the  $F$  test, which is appropriate in finite (all) samples, that relies on normally distributed disturbances. In the nonlinear case, we rely on large-sample results, so the Wald statistic will be the primary inference tool. An analog to the  $F$  statistic based on the fit of the regression will also be developed later. Finally, **Lagrange multiplier tests** for the general case can be constructed. Since we have not assumed normality of the disturbances (yet), we will postpone treatment of the likelihood ratio statistic until we revisit this model in Chapter 14.

The hypothesis to be tested is

$$H_0: \mathbf{r}(\boldsymbol{\beta}) = \mathbf{q}, \quad (7-18)$$

where  $\mathbf{r}(\boldsymbol{\beta})$  is a column vector of  $J$  continuous functions of the elements of  $\boldsymbol{\beta}$ . These restrictions may be linear or nonlinear. It is necessary, however, that they be **overidentifying restrictions**. Thus, in formal terms, if the original parameter vector has  $K$  free elements, then the hypothesis  $\mathbf{r}(\boldsymbol{\beta}) - \mathbf{q}$  must impose at least one functional relationship on the parameters. If there is more than one restriction, then they must be functionally independent. These two conditions imply that the  $J \times K$  **Jacobian**,

$$\mathbf{R}(\boldsymbol{\beta}) = \frac{\partial \mathbf{r}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \quad (7-19)$$

## 190 PART I ♦ The Linear Regression Model

must have full row rank and that  $J$ , the number of restrictions, must be strictly less than  $K$ . This situation is analogous to the linear model, in which  $\mathbf{R}(\boldsymbol{\beta})$  would be the matrix of coefficients in the restrictions. (See, as well, Section 5.4, where the methods examined here are applied to the linear model.)

Let  $\mathbf{b}$  be the unrestricted, nonlinear least squares estimator, and let  $\mathbf{b}_*$  be the estimator obtained when the constraints of the hypothesis are imposed.<sup>3</sup> Which test statistic one uses depends on how difficult the computations are. Unlike the linear model, the various testing procedures vary in complexity. For instance, in our example, the Lagrange multiplier is by far the simplest to compute. Of the four methods we will consider, only this test does not require us to compute a nonlinear regression.

The nonlinear analog to the familiar  $F$  statistic based on the fit of the regression (i.e., the sum of squared residuals) would be

$$F[J, n - K] = \frac{[S(\mathbf{b}_*) - S(\mathbf{b})]/J}{S(\mathbf{b})/(n - K)}. \quad (7-20)$$

This equation has the appearance of our earlier  $F$  ratio in (5-29). In the nonlinear setting, however, neither the numerator nor the denominator has exactly the necessary chi-squared distribution, so the  $F$  distribution is only approximate. Note that this  $F$  statistic requires that both the restricted and unrestricted models be estimated.

The Wald test is based on the distance between  $\mathbf{r}(\mathbf{b})$  and  $\mathbf{q}$ . If the unrestricted estimates fail to satisfy the restrictions, then doubt is cast on the validity of the restrictions. The statistic is

$$\begin{aligned} W &= [\mathbf{r}(\mathbf{b}) - \mathbf{q}]' \{ \text{Est. Asy. Var}[\mathbf{r}(\mathbf{b}) - \mathbf{q}] \}^{-1} [\mathbf{r}(\mathbf{b}) - \mathbf{q}] \\ &= [\mathbf{r}(\mathbf{b}) - \mathbf{q}]' \{ \mathbf{R}(\mathbf{b}) \hat{\mathbf{V}} \mathbf{R}'(\mathbf{b}) \}^{-1} [\mathbf{r}(\mathbf{b}) - \mathbf{q}], \end{aligned} \quad (7-21)$$

where

$$\hat{\mathbf{V}} = \text{Est. Asy. Var}[\mathbf{b}],$$

and  $\mathbf{R}(\mathbf{b})$  is evaluated at  $\mathbf{b}$ , the estimate of  $\boldsymbol{\beta}$ .

Under the null hypothesis, this statistic has a limiting chi-squared distribution with  $J$  degrees of freedom. If the restrictions are correct, the Wald statistic and  $J$  times the  $F$  statistic are asymptotically equivalent. The Wald statistic can be based on the estimated covariance matrix obtained earlier using the unrestricted estimates, which may provide a large savings in computing effort if the restrictions are nonlinear. It should be noted that the small-sample behavior of  $W$  can be erratic, and the more conservative  $F$  statistic may be preferable if the sample is not large.

The caveat about Wald statistics that applied in the linear case applies here as well. Because it is a pure significance test that does not involve the alternative hypothesis, the Wald statistic is not invariant to how the hypothesis is framed. In cases in which there are more than one equivalent ways to specify  $\mathbf{r}(\boldsymbol{\beta}) = \mathbf{q}$ ,  $W$  can give different answers depending on which is chosen.

The Lagrange multiplier test is based on the decrease in the sum of squared residuals that would result if the restrictions in the restricted model were released. The formalities of the test are given in Section 14.6.3. For the nonlinear regression model,

<sup>3</sup>This computational problem may be extremely difficult in its own right, especially if the constraints are nonlinear. We assume that the estimator has been obtained by whatever means are necessary.

CHAPTER 7 ♦ Nonlinear, Semiparametric **191**

the test has a particularly appealing form.<sup>4</sup> Let  $\mathbf{e}_*$  be the vector of residuals  $y_i - h(\mathbf{x}_i, \mathbf{b}_*)$  computed using the restricted estimates. Recall that we defined  $\mathbf{X}^0$  as an  $n \times K$  matrix of derivatives computed at a particular parameter vector in (7-29). Let  $\mathbf{X}_*^0$  be this matrix *computed at the restricted estimates*. Then the Lagrange multiplier statistic for the nonlinear regression model is

$$\text{LM} = \frac{\mathbf{e}'_* \mathbf{X}_*^0 [\mathbf{X}_*^0 \mathbf{X}_*^0]^{-1} \mathbf{X}_*^0 \mathbf{e}_*}{\mathbf{e}'_* \mathbf{e}_*/n}. \quad (7-22)$$

Under  $H_0$ , this statistic has a limiting chi-squared distribution with  $J$  degrees of freedom. What is especially appealing about this approach is that it requires only the restricted estimates. This method may provide some savings in computing effort if, as in our example, the restrictions result in a linear model. Note, also, that the Lagrange multiplier statistic is  $n$  times the uncentered  $R^2$  in the regression of  $\mathbf{e}_*$  on  $\mathbf{X}_*^0$ . Many Lagrange multiplier statistics are computed in this fashion.

### 7.2.5 APPLICATIONS

This section will present three applications of estimation and inference for nonlinear regression models. Example 7.4 illustrates a nonlinear consumption function that extends Examples 1.2 and 2.1. The model provides a simple demonstration of estimation and hypothesis testing for a nonlinear model. Example 7.5 analyzes the Box–Cox transformation. This specification is used to provide a more general functional form than the linear regression—it has the linear and loglinear models as special cases. Finally, Example 7.6 is a lengthy examination of an exponential regression model. In this application, we will explore some of the implications of nonlinear modeling, specifically “interaction effects.” We examined interaction effects in Section 6.3.3 in a model of the form

$$y = \beta_1 + \beta_2 x + \beta_3 z + \beta_4 xz + \varepsilon.$$

In this case, the interaction effect is  $\partial^2 E[y|x, z]/\partial x \partial z = \beta_4$ . There is no interaction effect if  $\beta_4$  equals zero. Example 7.6 considers the (perhaps unintended) implication of the nonlinear model that when  $E[y|x, z] = h(x, z, \boldsymbol{\beta})$ , there is an interaction effect even if the model is

$$h(x, z, \boldsymbol{\beta}) = h(\beta_1 + \beta_2 x + \beta_3 z).$$

#### **Example 7.4 Analysis of a Nonlinear Consumption Function**

The linear consumption function analyzed at the beginning of Chapter 2 is a restricted version of the more general consumption function

$$C = \alpha + \beta Y^\gamma + \varepsilon,$$

in which  $\gamma$  equals 1. With this restriction, the model is linear. If  $\gamma$  is free to vary, however, then this version becomes a nonlinear regression. Quarterly data on consumption, real disposable income, and several other variables for the U.S. economy for 1950 to 2000 are listed in Appendix Table F5.2. We will use these to fit the nonlinear consumption function. (Details of the computation of the estimates are given in Section 7.2.6 in Example 7.8.) The restricted linear and unrestricted nonlinear least squares regression results are shown in Table 7.1.

The procedures outlined earlier are used to obtain the asymptotic standard errors and an estimate of  $\sigma^2$ . (To make this comparable to  $s^2$  in the linear model, the value includes the degrees of freedom correction.)

---

<sup>4</sup>This test is derived in Judge et al. (1985). A lengthy discussion appears in Mittelhammer et al. (2000).

**192 PART I ♦ The Linear Regression Model**
**TABLE 7.1** Estimated Consumption Functions

<i>Parameter</i>	<i>Linear Model</i>		<i>Nonlinear Model</i>	
	<i>Estimate</i>	<i>Standard Error</i>	<i>Estimate</i>	<i>Standard Error</i>
$\alpha$	-80.3547	14.3059	458.7990	22.5014
$\beta$	0.9217	0.003872	0.10085	0.01091
$\gamma$	1.0000	—	1.24483	0.01205
$\mathbf{e}'\mathbf{e}$	1,536,321.881		504,403.1725	
$\sigma$	87.20983		50.0946	
$R^2$	0.996448		0.998834	
$\text{Var}[b]$	—		0.000119037	
$\text{Var}[c]$	—		0.00014532	
$\text{Cov}[b, c]$	—		-0.000131491	

In the preceding example, there is no question of collinearity in the data matrix  $\mathbf{X} = [\mathbf{i}, \mathbf{y}]$ ; the variation in  $Y$  is obvious on inspection. But, at the final parameter estimates, the  $R^2$  in the regression is 0.998834 and the correlation between the two pseudoregressors  $x_2^0 = Y^\gamma$  and  $x_3^0 = \beta Y^\gamma \ln Y$  is 0.999752. The condition number for the normalized matrix of sums of squares and cross products is 208.306. (The condition number is computed by computing the square root of the ratio of the largest to smallest characteristic root of  $\mathbf{L}_0' \mathbf{X}_0 \mathbf{D}^{-1}$  where  $x_1^0 = 1$  and  $\mathbf{D}$  is the diagonal matrix containing the square roots of  $\mathbf{x}_k^0 \mathbf{x}_k^0$  on the diagonal.) Recall that 20 was the benchmark for a problematic data set. By the standards discussed in Section 4.7.1 and A.6.6, the collinearity problem in this “data set” is severe. In fact, it appears not to be a problem at all.

For hypothesis testing and confidence intervals, the familiar procedures can be used, with the proviso that all results are only asymptotic. As such, for testing a restriction, the chi-squared statistic rather than the  $F$  ratio is likely to be more appropriate. For example, for testing the hypothesis that  $\gamma$  is different from 1, an asymptotic  $t$  test, based on the standard normal distribution, is carried out, using

$$z = \frac{1.24483 - 1}{0.01205} = 20.3178.$$

This result is larger than the critical values of 1.96 for the 5 percent significance level, and we thus reject the linear model in favor of the nonlinear regression. The three procedures for hypotheses produce the same conclusion.

- The  $F$  statistic is

$$F[1.204 - 3] = \frac{(1,536,321.881 - 504,403.17)/1}{504,403.17/(204 - 3)} = 411.29.$$

The critical value from the tables is 3.84, so the hypothesis is rejected.

- The Wald statistic is based on the distance of  $\hat{\gamma}$  from 1 and is simply the square of the asymptotic  $t$  ratio we computed earlier:

$$W = \frac{(1.24483 - 1)^2}{0.01205^2} = 412.805.$$

The critical value from the chi-squared table is 3.84.

- For the Lagrange multiplier statistic, the elements in  $\mathbf{x}_i^*$  are

$$\mathbf{x}_i^* = [1, Y^\gamma, \beta Y^\gamma \ln Y].$$

To compute this at the restricted estimates, we use the ordinary least squares estimates for  $\alpha$  and  $\beta$  and 1 for  $\gamma$  so that

$$\mathbf{x}_i^* = [1, Y, \beta Y \ln Y].$$

## CHAPTER 7 ♦ Nonlinear, Semiparametric 193

The residuals are the least squares residuals computed from the linear regression. Inserting the values given earlier, we have

$$LM = \frac{996,103.9}{(1,536,321.881/204)} = 132.267.$$

As expected, this statistic is also larger than the critical value from the chi-squared table.

 We are also interested in the marginal propensity to consume. In this expanded model,  $H_0 : \gamma = 1$  is a test—that the marginal propensity to consume is constant, not that it is 1. (That would be a joint test of both  $\gamma = 1$  and  $\beta = 1$ .) In this model, the marginal propensity to consume is

$$MPC = dC/dY = \beta\gamma Y^{\gamma-1},$$

which varies with  $Y$ . To  test the hypothesis that this value is 1, we require a particular value of  $Y$ . Because it is the most recent value, we choose  $DPI/2000.4 = 6634.9$ . At this value, the MPC is estimated as 0.86971. We estimate its standard error using the delta method, with the square root of

$$\begin{aligned} & [\partial MPC/\partial b \quad \partial MPC/\partial c] \begin{bmatrix} \text{Var}[b] & \text{Cov}[b, c] \\ \text{Cov}[b, c] & \text{Var}[c] \end{bmatrix} \begin{bmatrix} \partial MPC/\partial b \\ \partial MPC/\partial c \end{bmatrix} \\ &= [cY^{c-1} \quad bY^{c-1}(1 + c\ln Y)] \begin{bmatrix} 0.000119037 & -0.000131491 \\ -0.000131491 & 0.00014532 \end{bmatrix} \begin{bmatrix} cY^{c-1} \\ bY^{c-1}(1 + c\ln Y) \end{bmatrix} \\ &= 0.00007469, \end{aligned}$$

which gives a standard error of 0.0086423. For testing the hypothesis that the MPC is equal to 1.0 in 2000.4 we would refer  $z = (1.08264 - 1)/0.0086423 = -9.56299$  to the standard normal table. This difference is certainly statistically significant, so we would reject the hypothesis.

#### Example 7.5 The Box–Cox Transformation

The **Box–Cox transformation** [Box and Cox (1964), Zarembka (1974)] is used as a device for generalizing the linear model. The transformation is

$$x^{(\lambda)} = (x^\lambda - 1)/\lambda.$$

Special cases of interest are  $\lambda = 1$ , which produces a linear transformation,  $x^{(1)} = x - 1$ , and  $\lambda = 0$ . When  $\lambda$  equals zero, the transformation is, by L'Hôpital's rule,

$$\lim_{\lambda \rightarrow 0} \frac{x^\lambda - 1}{\lambda} = \lim_{\lambda \rightarrow 0} \frac{d(x^\lambda - 1)/d\lambda}{1} = \lim_{\lambda \rightarrow 0} x^\lambda \times \ln x = \ln x.$$

The regression analysis can be done *conditionally* on  $\lambda$ . For a given value of  $\lambda$ , the model,

$$y = \alpha + \sum_{k=2}^K \beta_k x_k^{(\lambda)} + \varepsilon, \tag{7-23}$$

is a linear regression that can be estimated by least squares. However, if  $\lambda$  in (7-23) is taken to be an unknown parameter, then the regression becomes nonlinear in the parameters.

In principle, each regressor could be transformed by a different value of  $\lambda$ , but, in most applications, this level of generality becomes excessively cumbersome, and  $\lambda$  is assumed to be the same for all the variables in the model.<sup>5</sup> To be defined for all values of  $\lambda$ ,  $x$  must be strictly positive. In most applications, some of the regressors—for example, a dummy variable—will not be transformed. For such a variable, say  $v_k$ ,  $v_k^{(\lambda)} = v_k$ , and the relevant derivatives in (7-24) will be zero. It is also possible to transform  $y$ , say, by  $y^{(\theta)}$ . Transformation of the dependent variable, however, amounts to a specification of the whole model, not just

<sup>5</sup>See, for example, Seaks and Layson (1983).

## 194 PART I ♦ The Linear Regression Model

the functional form of the conditional mean. For example,  $\theta = 1$  implies a linear equation while  $\theta = 0$  implies a logarithmic equation.

In some applications, the motivation for the transformation is to program around zero values in a loglinear model. Caves, Christensen, and Tretheway (1980) analyzed the costs of production for railroads providing freight and passenger service. Continuing a long line of literature on the costs of production in regulated industries, a translog cost function (see Section 10.4.2) would be a natural choice for modeling this multiple-output technology. Several of the firms in the study, however, produced no passenger service, which would preclude the use of the translog model. (This model would require the log of zero.) An alternative is the Box–Cox transformation, which is computable for zero output levels. A question does arise in this context (and other similar ones) as to whether zero outputs should be treated the same as nonzero outputs or whether an output of zero represents a discrete corporate decision distinct from other variations in the output levels. In addition, as can be seen in (7-24), this solution is only partial. The zero values of the regressors preclude computation of appropriate standard errors.

Nonlinear least squares is straightforward. In most instances, we can expect to find the least squares value of  $\lambda$  between  $-2$  and  $2$ . Typically, then,  $\lambda$  is estimated by scanning this range for the value that minimizes the sum of squares. Note what happens if there are zeros for  $x$  in the sample. Then, a constraint must still be placed on  $\lambda$  in their model, as  $0^{(\lambda)}$  is defined only if  $\lambda$  is strictly positive. A positive value of  $\lambda$  is not assured. Once the optimal value of  $\lambda$  is located, the least squares estimates, the mean squared residual, and this value of  $\lambda$  constitute the nonlinear least squares estimates of the parameters.

After determining the optimal value of  $\lambda$ , it is sometimes treated as if it were a known value in the least squares results. But  $\hat{\lambda}$  is an estimate of an unknown parameter. It is not hard to show that the least squares standard errors will always underestimate the correct asymptotic standard errors.<sup>6</sup> To get the appropriate values, we need the derivatives of the right-hand side of (7-23) with respect to  $\alpha$ ,  $\beta$ , and  $\lambda$ . The pseudoregressors are

$$\begin{aligned}\frac{\partial h(\cdot)}{\partial \alpha} &= 1, \\ \frac{\partial h(\cdot)}{\partial \beta_k} &= x_k^{(\lambda)}, \\ \frac{\partial h(\cdot)}{\partial \lambda} &= \sum_{k=1}^K \beta_k \frac{\partial x_k^{(\lambda)}}{\partial \lambda} = \sum_{k=1}^K \beta_k \left[ \frac{1}{\lambda} (x_k^\lambda \ln x_k - x_k^{(\lambda)}) \right].\end{aligned}\tag{7-24}$$

We can now use (7-15) and (7-16) to estimate the asymptotic covariance matrix of the parameter estimates. Note that  $\ln x_k$  appears in  $\partial h(\cdot)/\partial \lambda$ . If  $x_k = 0$ , then this matrix cannot be computed. This was the point noted earlier.

It is important to remember that the coefficients in a nonlinear model are not equal to the slope (or the elasticities) with respect to the variables. For the particular Box–Cox model in (7-23),

$$\frac{\partial E[\ln y|\mathbf{x}]}{\partial \ln x_k} = x_k \frac{\partial E[\ln y|\mathbf{x}]}{\partial x_k} = \beta_k x_k^\lambda = \eta_k.$$

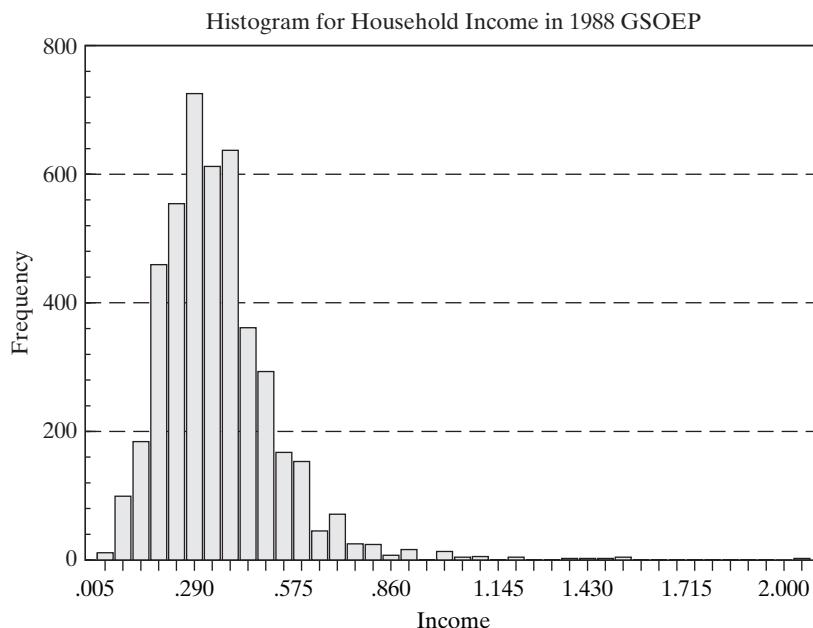
Standard errors for these estimates can be obtained using the **delta method**. The derivatives are  $\partial \eta / \partial \beta_k = x_k^\lambda = \eta_k / \beta_k$  and  $\partial \eta / \partial \lambda = \eta \ln x_k$ . Collecting terms, we obtain

$$\text{Asy.Var}[\hat{\eta}_k] = (\eta_k / \beta_k)^2 \{ \text{Asy.Var}[\hat{\beta}_k] + (\beta \ln x_k)^2 \text{Asy.Var}[\hat{\lambda}] + (2\beta \ln x_k) \text{Asy.Cov}[\hat{\beta}_k, \hat{\lambda}] \}$$

The application in Example 7.4 is a Box–Cox model of the sort discussed here. We can rewrite (7-23) as

$$\begin{aligned}y &= (\alpha - 1/\lambda) + (\beta/\lambda) X^\lambda + \varepsilon \\ &= \alpha^* + \beta^* x^\gamma + \varepsilon.\end{aligned}$$

<sup>6</sup>See Fomby, Hill, and Johnson (1984, pp. 426–431).



**FIGURE 7.1** Histogram for Income.

This shows that an alternative way to handle the Box–Cox regression model is to transform the model into a nonlinear regression and then use the Gauss–Newton regression (see Section 7.2.6) to estimate the parameters. The original parameters of the model can be recovered by  $\lambda = \gamma$ ,  $\alpha = \alpha^* + 1/\gamma$  and  $\beta = \gamma\beta^*$ .

#### **Example 7.6 Interaction Effects in a Loglinear Model for Income**

A recent study in health economics is “Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation” by Riphahn, Wambach, and Million (2003). The authors were interested in counts of physician visits and hospital visits and in the impact that the presence of private insurance had on the utilization counts of interest, that is, whether the data contain evidence of moral hazard. The sample used is an unbalanced panel of 7,293 households, the German Socioeconomic Panel (GSOEP) data set.<sup>7</sup> Among the variables reported in the panel are household income, with numerous other sociodemographic variables such as age, gender, and education. For this example, we will model the distribution of income using the last wave of the data set (1988), a cross section with 4,483 observations. Two of the individuals in this sample reported zero income, which is incompatible with the underlying models suggested in the development below. Deleting these two observations leaves a sample of 4,481 observations. Figures 7.1 and 7.2 display a histogram and a kernel density estimator for the household income variable for these observations.

We will fit an exponential regression model to the income variable, with

$$\begin{aligned} \text{Income} = & \exp(\beta_1 + \beta_2 \text{Age} + \beta_3 \text{Age}^2 + \beta_4 \text{Education} + \beta_5 \text{Female} \\ & + \beta_6 \text{Female} \times \text{Education} + \beta_7 \text{Age} \times \text{Education}) + \varepsilon. \end{aligned}$$

<sup>7</sup>The data are published on the *Journal of Applied Econometrics* data archive web site, at <http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/>. The variables in the data file are listed in Appendix Table F7.1. The number of observations in each year varies from one to seven with a total number of 27,326 observations. We will use these data in several examples here and later in the book.

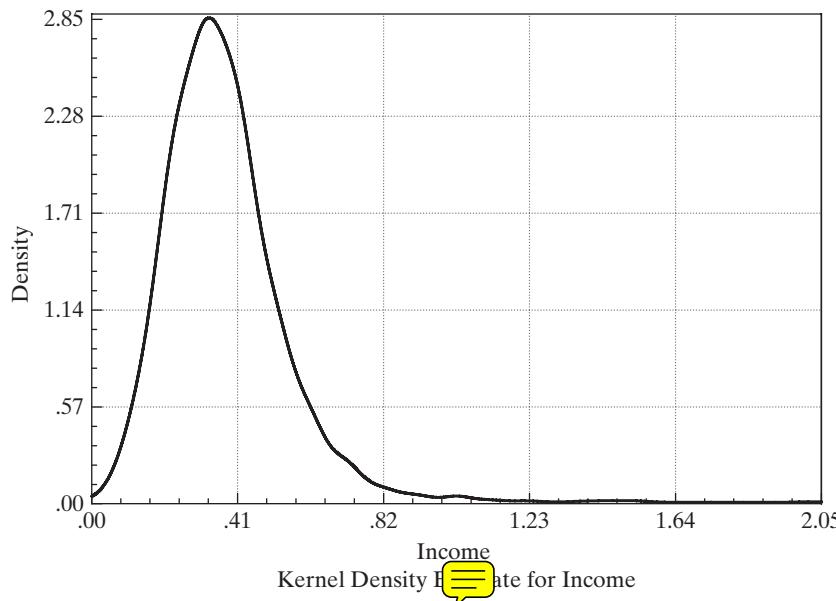
**196 PART I ♦ The Linear Regression Model**

**FIGURE 7.2** Kernel Density Estimator for Income.

Table 7.2 provides descriptive statistics for the variables used in this application.

**Loglinear models** play a prominent role in statistics. Many derive from a density function of the form  $f(y|\mathbf{x}) = p[y|\alpha^0 + \mathbf{x}'\beta, \theta]$ , where  $\alpha^0$  is a constant term and  $\theta$  is an additional parameter, and

$$E[y|\mathbf{x}] = g(\theta) \exp(\alpha^0 + \mathbf{x}'\beta),$$

(hence the name “loglinear models”). Examples include the Weibull, gamma, lognormal, and exponential models for continuous variables and the Poisson and negative binomial models for counts. We can write  $E[y|\mathbf{x}]$  as  $\exp[\ln g(\theta) + \alpha^0 + \mathbf{x}'\beta]$ , and then absorb  $\ln g(\theta)$  in the constant term in  $\ln E[y|\mathbf{x}] = \alpha + \mathbf{x}'\beta$ . The lognormal distribution (see Section B.4.4) is often used to model incomes. For the lognormal random variable,

$$p[y|\alpha^0 + \mathbf{x}'\beta, \theta] = \frac{\exp[-\frac{1}{2}(\ln y - \alpha^0 - \mathbf{x}'\beta)^2/\theta^2]}{\theta y \sqrt{2\pi}}, y > 0,$$

$$E[y|\mathbf{x}] = \exp(\alpha^0 + \mathbf{x}'\beta + \theta^2/2) = \exp(\alpha + \mathbf{x}'\beta).$$

**TABLE 7.2** Descriptive Statistics for Variables Used in Nonlinear Regression

Variable	Mean	Std.Dev.	Min	Maximum
INCOME	.348896	.164054	.0050	2
AGE	43.4452	11.2879	25.00	64
EDUC	11.4167	2.36615	7.000	18
FEMALE	.484267	.499808	.0000	1

## CHAPTER 7 ♦ Nonlinear, Semiparametric 197

The exponential regression model is also consistent with a gamma distribution. The density of a gamma distributed random variable is

$$p[y|\alpha^0 + \mathbf{x}'\beta, \theta] = \frac{\lambda^\theta \exp(-\lambda)y^{\theta-1}}{\Gamma(\theta)}, y > 0, \theta > 0, \lambda = \exp(-\alpha^0 - \mathbf{x}'\beta),$$

$$E[y|\mathbf{x}] = \theta/\lambda = \theta \exp(\alpha^0 + \mathbf{x}'\beta) = \exp(\ln \theta + \alpha^0 + \mathbf{x}'\beta) = \exp(\alpha + \mathbf{x}'\beta).$$

The parameter  $\theta$  determines the shape of the distribution. When  $\theta > 2$ , the gamma density has the shape of a chi-squared variable (which is a special case). Finally, the Weibull model has a similar form,

$$p[y|\alpha^0 + \mathbf{x}'\beta, \theta] = \theta \lambda^\theta \exp[-(\lambda y)^\theta] y^{\theta-1}, y \geq 0, \theta > 0, \lambda = \exp(-\alpha^0 - \mathbf{x}'\beta),$$

$$E[y|\mathbf{x}] = \Gamma(1 + 1/\theta) \exp(\alpha^0 + \mathbf{x}'\beta) = \exp[\ln \Gamma(1 + 1/\theta) + \alpha^0 + \mathbf{x}'\beta] = \exp(\alpha + \mathbf{x}'\beta).$$

In all cases, the maximum likelihood estimator is the most efficient estimator of the parameters. (Maximum likelihood estimation of the parameters of this model is considered in Chapter 14.) However, nonlinear least squares estimation of the model

$$E[y|\mathbf{x}] = \exp(\alpha + \mathbf{x}'\beta) + \varepsilon$$

has a virtue in that the nonlinear least squares estimator will be consistent even if the distributional assumption is incorrect—it is *robust* to this type of misspecification since it does not make explicit use of a distributional assumption.

Table 7.3 presents the nonlinear least squares regression results. Superficially, the pattern of signs and significance might be expected—with the exception of the dummy variable for female. However, two issues complicate the interpretation of the coefficients in this model. First, the model is nonlinear, so the coefficients do not give the magnitudes of the interesting effects in the equation. In particular, for this model,

$$\partial E[y|\mathbf{x}]/\partial x_k = \exp(\alpha + \mathbf{x}'\beta) \times \partial(\alpha + \mathbf{x}'\beta)/\partial x_k.$$

Second, as we have constructed our model, the second part of the derivative is not equal to the coefficient, because the variables appear either in a quadratic term or as a product with some other variable. Moreover, for the dummy variable, *Female*, we would want to compute the partial effect using

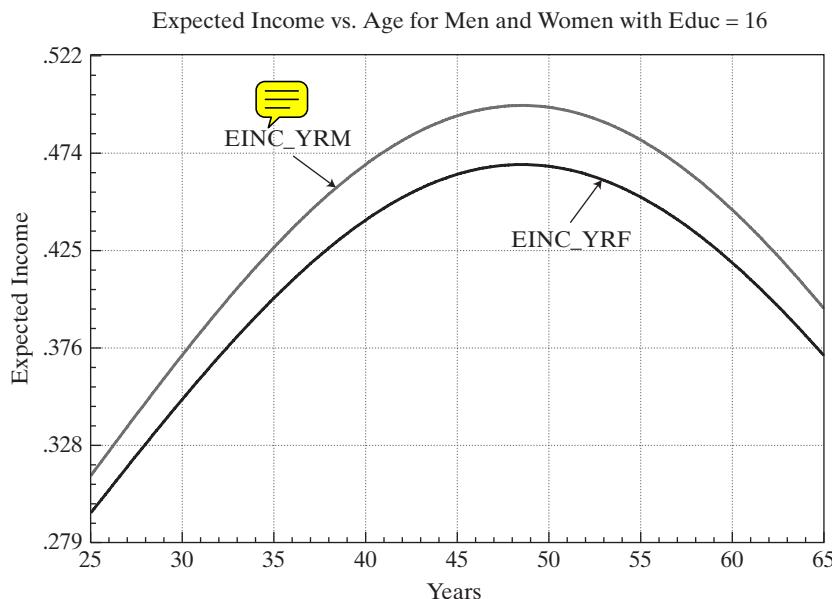
$$\Delta E[y|\mathbf{x}]/\Delta \text{Female} = E[y|\mathbf{x}, \text{Female} = 1] - E[y|\mathbf{x}, \text{Female} = 0]$$

A third consideration is how to compute the partial effects, as sample averages or at the means of the variables. For example,

$$\partial E[y|\mathbf{x}]/\partial \text{Age} = E[y|\mathbf{x}] \times (\beta_2 + 2\beta_3 \text{Age} + \beta_7 \text{Educ}).$$

**TABLE 7.3** Estimated Regression Equations

<i>Variable</i>	<i>Nonlinear Least Squares</i>			<i>Linear Least Squares</i>		
	<i>Estimate</i>	<i>Std. Error</i>	<i>t</i>	<i>Estimate</i>	<i>Std. Error</i>	<i>t</i>
<b>Constant</b>	-2.58070	.17455	14.78	-.13050	.06261	-2.08
<b>Age</b>	.06020	.00615	9.79	.01791	.00214	8.37
<b>Age<sup>2</sup></b>	-.00084	.00006082	-13.83	-.00027	.00001985	-13.51
<b>Education</b>	-.00616	.01095	-.56	-.00281	.00418	-.67
<b>Female</b>	.17497	.05986	2.92	.07955	.02339	3.40
<b>Female × Educ</b>	-.01476	.00493	-2.99	-.00685	.00202	-3.39
<b>Age × Educ</b>	.00134	.00024	5.59	.00055	.00009394	5.88
e'e		106.09825			106.24323	
<b>s</b>		.15387			.15410	
<b>R</b> <sup>2</sup>		.12005			.11880	

**198 PART I ♦ The Linear Regression Model**


**FIGURE 7.3** Expected Incomes.

The average value of *Age* in the sample is 43.4452 and the average *Education* is 11.4167. The partial effect of a year of education is estimated to be 0.000948 if it is computed by computing the partial effect for each individual and averaging the result. It is 0.000925 if it is computed by computing the conditional mean and the linear term at the averages of the three variables. The partial effect is difficult to interpret without information about the scale of the income variable. Since the average income in the data is about 0.35, these partial effects suggest that an additional year of education is associated with a change in expected income of about 2.6 percent (i.e., 0.009/0.35).

The rough calculation of partial effects with respect to *Age* does not reveal the model implications about the relationship between age and expected income. Note, for example, that the coefficient on *Age* is positive while the coefficient on *Age*<sup>2</sup> is negative. This implies (neglecting the interaction term at the end), that the *Age – Income* relationship implied by the model is parabolic. The partial effect is positive at some low values and negative at higher values. To explore this, we have computed the expected *Income* using the model separately for men and women, both with assumed college education (*Educ* = 16) and for the range of ages in the sample, 25 to 64. Figure 7.3 shows the result of this calculation. The upper curve is for men (*Female* = 0) and the lower one is for women. The parabolic shape is as expected; what the figure reveals is the relatively strong effect—ceteris paribus, incomes are predicted to rise by about 80 percent between ages 25 and 48. (There is an important aspect of this computation that the model builder would want to develop in the analysis. It remains to be argued whether this parabolic relationship describes the trajectory of expected income for an individual as they age, or the average incomes of different cohorts at a particular moment in time (1988). The latter would seem to be the more appropriate conclusion at this point, though one might be tempted to infer the former.)

The figure reveals a second implication of the estimated model that would not be obvious from the regression results. The coefficient on the dummy variable for *Female* is positive, highly significant, and, in isolation, by far the largest effect in the model. This might lead the analyst to conclude that on average, expected incomes in these data are higher for women than men. But, Figure 7.3 shows precisely the opposite. The difference is accounted

## CHAPTER 7 ♦ Nonlinear, Semiparametric 199

for by the interaction term, *Female* × *Education*. The negative sign on the latter coefficient is suggestive. But, the total effect would remain ambiguous without the sort of secondary analysis suggested by the figure.

Finally, in addition to the quadratic term in age, the model contains an interaction term, *Age* × *Education*. The coefficient is positive and highly significant. But, it is far from obvious how this should be interpreted. In a linear model,

$$\begin{aligned} \text{Income} = & \beta_1 + \beta_2 \text{Age} + \beta_3 \text{Age}^2 + \beta_4 \text{Education} + \beta_5 \text{Female} \\ & + \beta_6 \text{Female} \times \text{Education} + \beta_7 \text{Age} \times \text{Education} + \varepsilon, \end{aligned}$$

we would find that  $\beta_7 = \partial^2 E[\text{Income}|x]/\partial \text{Age} \partial \text{Education}$ . That is, the “interaction effect” is the change in the partial effect of *Age* associated with a change in *Education* (or vice versa). Of course, if  $\beta_7$  equals zero, that is, if there is no product term in the model, then there is no interaction effect—the second derivative equals zero. However, this simple interpretation usually does not apply in nonlinear models (i.e., in any nonlinear model). Consider our exponential regression, and suppose that in fact,  $\beta_7$  is indeed zero. For convenience, let  $\mu(x)$  equal the conditional mean function. Then, the partial effect with respect to *Age* is

$$\partial \mu(x) / \partial \text{Age} = \mu(x) \times (\beta_2 + 2\beta_3 \text{Age})$$

and

$$\partial^2 \mu(x) / \partial \text{Age} \partial \text{Educ} = \mu(x) \times (\beta_2 + 2\beta_3 \text{Age})(\beta_4 + \beta_6 \text{Female}), \quad (7-25)$$

which is nonzero even if there is no “**interaction term**” in the model. The interaction effect in the model that we estimated, which includes the product term, is

$$\partial^2 E[y|x] / \partial \text{Age} \partial \text{Educ} = \mu(x) \times [\beta_7 + (\beta_2 + 2\beta_3 \text{Age} + \beta_7 \text{Educ})(\beta_4 + \beta_6 \text{Female} + \beta_7 \text{Age})]. \quad (7-26)$$

At least some of what is being called the interaction effect in this model is attributable entirely to the fact the model is nonlinear. To isolate the “functional form effect” from the true “interaction effect,” we might subtract (7-25) from (7-26) and then reassemble the components:

$$\begin{aligned} \partial^2 \mu(x) / \partial \text{Age} \partial \text{Educ} = & \mu(x)[(\beta_2 + 2\beta_3 \text{Age})(\beta_4 + \beta_6 \text{Female})] \\ & + \mu(x) \beta_7 [1 + \text{Age}(\beta_2 + 2\beta_3) + \text{Educ}(\beta_4 + \beta_6 \text{Female}) + \text{Educ} \times \text{Age}(\beta_7)]. \quad (7-27) \end{aligned}$$

It is clear that the coefficient on the product term bears essentially no relationship to the quantity of interest (assuming it is the change in the partial effects that is of interest). On the other hand, the second term is nonzero if and only if  $\beta_7$  is nonzero. One might, therefore, identify the second part with the “interaction effect” in the model. Whether a behavioral interpretation could be attached to this is questionable, however. Moreover, that would leave unexplained the functional form effect. The point of this exercise is to suggest that one should proceed with some caution in interpreting interaction effects in nonlinear models. This sort of analysis has a focal point in the literature in Ai and Norton (2004). A number of comments and extensions of the result are to be found, including Greene (2010).

We make one final observation about the nonlinear regression. In a loglinear, single-index function model such as the one analyzed here, one might, “for comparison purposes,” compute simple linear least squares results. The coefficients in the right-hand side of Table 7.3 suggest superficially that nonlinear least squares and least squares are computing completely different relationships. To uncover the similarity (if there is one), it is useful to consider the partial effects rather than the coefficients. We found, for example, the partial effect of education in the nonlinear model, using the means of the variables, is 0.000925. Although the linear least squares coefficients are very different, if the partial effect for education is computed for the linear equation, we find  $-0.00281 - 0.00685(0.5) + 0.00055(43.4452) = 0.01766$ , where we have used 0.5 for *Female*. Dividing by 0.35, we obtain 0.0504, which is at least close to its counterpart in the nonlinear model. As a general result, at least approximately, the linear least squares coefficients are making this approximation.

## 200 PART I ♦ The Linear Regression Model

### 7.2.6 COMPUTING THE NONLINEAR LEAST SQUARES ESTIMATOR

Minimizing the sum of squared residuals for a nonlinear regression is a standard problem in nonlinear optimization that can be solved by a number of methods. (See Section E.3.) The method of Gauss–Newton is often used. This algorithm (and most of the sampling theory results for the asymptotic properties of the estimator) is based on a linear Taylor series approximation to the nonlinear regression function. The iterative estimator is computed by transforming the optimization to a series of linear least squares regressions.

The nonlinear regression model is  $y = h(\mathbf{x}, \boldsymbol{\beta}) + \varepsilon$ . (To save some notation, we have dropped the observation subscript). The procedure is based on a linear Taylor series approximation to  $h(\mathbf{x}, \boldsymbol{\beta})$  at a particular value for the parameter vector,  $\boldsymbol{\beta}^0$ :

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx h(\mathbf{x}, \boldsymbol{\beta}^0) + \sum_{k=1}^K \frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0} (\beta_k - \beta_k^0). \quad (7-28)$$

This form of the equation is called the **linearized regression model**. By collecting terms, we obtain

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx \left[ h(\mathbf{x}, \boldsymbol{\beta}^0) - \sum_{k=1}^K \beta_k^0 \left( \frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0} \right) \right] + \sum_{k=1}^K \beta_k \left( \frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0} \right). \quad (7-29)$$

Let  $x_k^0$  equal the  $k$ th partial derivative,<sup>8</sup>  $\partial h(\mathbf{x}, \boldsymbol{\beta}^0)/\partial \beta_k^0$ . For a given value of  $\boldsymbol{\beta}^0$ ,  $x_k^0$  is a function only of the data, not of the unknown parameters. We now have

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx \left[ h^0 - \sum_{k=1}^K x_k^0 \beta_k^0 \right] + \sum_{k=1}^K x_k^0 \beta_k,$$

which may be written

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx h^0 - \mathbf{x}' \boldsymbol{\beta}^0 + \mathbf{x}' \boldsymbol{\beta},$$

which implies that

$$y \approx h^0 - \mathbf{x}' \boldsymbol{\beta}^0 + \mathbf{x}' \boldsymbol{\beta} + \varepsilon.$$

By placing the known terms on the left-hand side of the equation, we obtain a linear equation:

$$y^0 = y - h^0 + \mathbf{x}' \boldsymbol{\beta}^0 = \mathbf{x}' \boldsymbol{\beta} + \varepsilon^0. \quad (7-30)$$

Note that  $\varepsilon^0$  contains both the true disturbance,  $\varepsilon$ , and the error in the first-order Taylor series approximation to the true regression, shown in (7-29). That is,

$$\varepsilon^0 = \varepsilon + \left[ h(\mathbf{x}, \boldsymbol{\beta}) - \left\{ h^0 - \sum_{k=1}^K x_k^0 \beta_k^0 + \sum_{k=1}^K x_k^0 \beta_k \right\} \right]. \quad (7-31)$$

Because all the errors are accounted for, (7-30) is an equality, not an approximation. With a value of  $\boldsymbol{\beta}^0$  in hand, we can  compute  $y^0$  and  $\mathbf{x}^0$  and then estimate the parameters of (7-30) by linear least squares. (Whether this estimator is consistent or not remains to be seen.)

<sup>8</sup>You should verify that for the linear regression model, these derivatives are the independent variables.

## CHAPTER 7 ♦ Nonlinear, Semiparametric 201

**Example 7.7 Linearized Regression**

For the model in Example 7.3, the regressors in the linearized equation would be

$$\begin{aligned}x_1^0 &= \frac{\partial h(\cdot)}{\partial \beta_1^0} = 1, \\x_2^0 &= \frac{\partial h(\cdot)}{\partial \beta_2^0} = e^{\beta_3^0 x}, \\x_3^0 &= \frac{\partial h(\cdot)}{\partial \beta_3^0} = \beta_2^0 x e^{\beta_3^0 x}.\end{aligned}$$

With a set of values of the parameters  $\beta^0$ ,

$$y^0 = y - h(x, \beta_1^0, \beta_2^0, \beta_3^0) + \beta_1^0 x_1^0 + \beta_2^0 x_2^0 + \beta_3^0 x_3^0$$

can be linearly regressed on the three variables previously defined to estimate  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$ .

The linearized regression model shown in (7-30) can be estimated by linear least squares. Once a parameter vector is obtained, it can play the role of a new  $\beta^0$ , and the computation can be done again. The **iteration** can continue until the difference between successive parameter vectors is small enough to assume convergence. One of the main virtues of this method is that at the last iteration the estimate of  $(\mathbf{Q}^0)^{-1}$  will, apart from the scale factor  $\hat{\sigma}^2/n$ , provide the correct estimate of the asymptotic covariance matrix for the parameter estimator.

This iterative solution to the minimization problem is

$$\begin{aligned}\mathbf{b}_{t+1} &= \left[ \sum_{i=1}^n \mathbf{x}_i^0 \mathbf{x}_i^{0'} \right]^{-1} \left[ \sum_{i=1}^n \mathbf{x}_i^0 (y_i - h_i^0 + \mathbf{x}_i^{0'} \mathbf{b}_t) \right] \\&= \mathbf{b}_t + \left[ \sum_{i=1}^n \mathbf{x}_i^0 \mathbf{x}_i^{0'} \right]^{-1} \left[ \sum_{i=1}^n \mathbf{x}_i^0 (y_i - h_i^0) \right] \\&= \mathbf{b}_t + (\mathbf{X}^0 \mathbf{X}^0)^{-1} \mathbf{X}^0 \mathbf{e}^0 \\&= \mathbf{b}_t + \Delta_t,\end{aligned}\tag{7-32}$$

where all terms on the right-hand side are evaluated at  $\mathbf{b}_t$  and  $\mathbf{e}^0$  is the vector of nonlinear least squares residuals. This algorithm has some intuitive appeal as well. For each iteration, we update the previous parameter estimates by regressing the nonlinear least squares residuals on the derivatives of the regression functions. The process will have converged (i.e., the update will be  $\mathbf{0}$ ) when  $\mathbf{X}^0 \mathbf{e}^0$  is close enough to  $\mathbf{0}$ . This derivative has a direct counterpart in the normal equations for the linear model,  $\mathbf{X}' \mathbf{e} = \mathbf{0}$ .

As usual, when using a digital computer, we will not achieve exact convergence with  $\mathbf{X}^0 \mathbf{e}^0$  exactly equal to zero. A useful, scale-free counterpart to the convergence criterion discussed in Section E.3.6 is  $\delta = \mathbf{e}^0 \mathbf{X}^0 (\mathbf{X}^0 \mathbf{X}^0)^{-1} \mathbf{X}^0 \mathbf{e}^0$ . [See (7-22).] We note, finally, that iteration of the linearized regression, although a very effective algorithm for many problems, does not always work. As does Newton's method, this algorithm sometimes "jumps off" to a wildly errant second iterate, after which it may be impossible to compute the residuals for the next iteration. The choice of starting values for the iterations can be crucial. There is art as well as science in the computation of nonlinear least squares estimates. [See McCullough and Vinod (1999).] In the absence of information about starting values, a workable strategy is to try the Gauss–Newton iteration first. If it

## 202 PART I ♦ The Linear Regression Model

fails, go back to the initial starting values and try one of the more general algorithms, such as BFGS, treating minimization of the sum of squares as an otherwise ordinary optimization problem.

### Example 7.8 Nonlinear Least Squares

Example 7.4 considered analysis of a nonlinear consumption function



$$C = \alpha + \beta Y^\gamma + \varepsilon.$$

The linearized regression model is

$$C - (\alpha^0 + \beta^0 Y^{\gamma^0}) + (\alpha^0 1 + \beta^0 Y^{\gamma^0} + \gamma^0 \beta^0 Y^{\gamma^0} \ln Y) = \alpha + \beta(Y^{\gamma^0}) + \gamma(\beta^0 Y^{\gamma^0} \ln Y) + \varepsilon^0.$$

Combining terms, we find that the nonlinear least squares procedure reduces to iterated regression of

$$C^0 = C + \gamma^0 \beta^0 Y^{\gamma^0} \ln Y$$

on

$$\mathbf{x}^0 = \left[ \frac{\partial h(\cdot)}{\partial \alpha} \frac{\partial h(\cdot)}{\partial \beta} \frac{\partial h(\cdot)}{\partial \gamma} \right]' = \begin{bmatrix} 1 \\ Y^{\gamma^0} \\ \beta^0 Y^{\gamma^0} \ln Y \end{bmatrix}.$$

Finding the **starting values** for a nonlinear procedure can be difficult. Simply trying a convenient set of values can be unproductive. Unfortunately, there are no good rules for starting values, except that they should be as close to the final values as possible (not particularly helpful). When it is possible, an initial consistent estimator of  $\beta$  will be a good starting value. In many cases, however, the only consistent estimator available is the one we are trying to compute by least squares. For better or worse, trial and error is the most frequently used procedure. For the present model, a natural set of values can be obtained because a simple linear model is a special case. Thus, we can start  $\alpha$  and  $\beta$  at the linear least squares values that would result in the special case of  $\gamma = 1$  and use 1 for the starting value for  $\gamma$ . The **iterations** are begun at the least squares estimates for  $\alpha$  and  $\beta$  and 1 for  $\gamma$ .

The solution is reached in eight iterations, after which any further iteration is merely “fine tuning” the hidden digits (i.e., those that the analyst would not be reporting to their reader. “Gradient” is the scale-free convergence measure,  $\delta$ , noted earlier.) Note that the coefficient vector takes a very errant step after the first iteration—the sum of squares becomes huge—but the iterations settle down after that and converge routinely.

Begin NLSQ iterations. Linearized regression.

```
Iteration = 1; Sum of squares = 1536321.88; Gradient = 996103.930
Iteration = 2; Sum of squares = 0.184780956E+12; Gradient = 0.184780452E+12 ( $\times 10^{12}$ )
Iteration = 3; Sum of squares = 20406917.6; Gradient = 19902415.7
Iteration = 4; Sum of squares = 581703.598; Gradient = 77299.6342
Iteration = 5; Sum of squares = 504403.969; Gradient = 0.752189847
Iteration = 6; Sum of squares = 504403.216; Gradient = 0.526642396E-04
Iteration = 7; Sum of squares = 504403.216; Gradient = 0.511324981E-07
Iteration = 8; Sum of squares = 504403.216; Gradient = 0.606793426E-10
```

## 7.3 MEDIAN AND QUANTILE REGRESSION

We maintain the essential assumptions of the linear regression model,

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$$

where  $E[\varepsilon|\mathbf{x}] = 0$  and  $E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$ . If  $\varepsilon|\mathbf{x}$  is normally distributed, so that the distribution of  $\varepsilon|\mathbf{x}$  is also symmetric, then the median,  $\text{Med}[\varepsilon|\mathbf{x}]$ , is also zero and  $\text{Med}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$ .

CHAPTER 7 ♦ Nonlinear, Semiparametric **203**

Under these assumptions, least squares remains a natural choice for estimation of  $\beta$ . But, as we explored in Example 4.5, **least absolute deviations** (LAD) is a possible alternative that might even be preferable in a small sample. Suppose, however, that we depart from the second assumption directly. That is, the statement of the model is

$$\text{Med}[y|\mathbf{x}] = \mathbf{x}'\beta.$$

This result suggests a motivation for LAD in its own right, rather than as a robust (to outliers) alternative to least squares.<sup>9</sup> The conditional median of  $y_i|\mathbf{x}_i$  might be an interesting function. More generally, other quantiles of the distribution of  $y_i|\mathbf{x}_i$  might also be of interest. For example, we might be interested in examining the various quantiles of the distribution of income or spending. Quantile regression (rather than least squares) is used for this purpose. The (linear) quantile regression model can be defined as

$$Q[y|\mathbf{x}, q] = \mathbf{x}'\beta_q \text{ such that } \text{Prob}[y \leq \mathbf{x}'\beta_q|\mathbf{x}] = q, 0 < q < 1. \quad (7-33)$$

The **median regression** would be defined for  $q = \frac{1}{2}$ . Other focal points are the lower and upper quartiles,  $q = \frac{1}{4}$  and  $q = \frac{3}{4}$ , respectively. We will develop the median regression in detail in Section 7.3.1, once again largely as an alternative estimator in the linear regression setting.

The quantile regression model is a richer specification than the linear model that we have studied thus far, because the coefficients in (7-33) are indexed by  $q$ . The model is nonparametric—it requires a much less detailed specification of the distribution of  $y|\mathbf{x}$ . In the simplest linear model with fixed coefficient vector,  $\beta$ , the quantiles of  $y|\mathbf{x}$  would be defined by variation of the constant term. The implication of the model is shown in Figure 7.4. For a fixed  $\beta$  and conditioned on  $x$ , the value of  $\alpha_q + \beta x$  such that  $\text{Prob}(y < \alpha_q + \beta x) = q$  is shown for  $q = 0.15, 0.5$ , and  $0.9$  in Figure 7.4. There is a value of  $\alpha_q$  for each quantile. In Section 7.3.2, we will examine the more general specification of the quantile regression model in which the entire coefficient vector plays the role of  $\alpha_q$  in Figure 7.4.

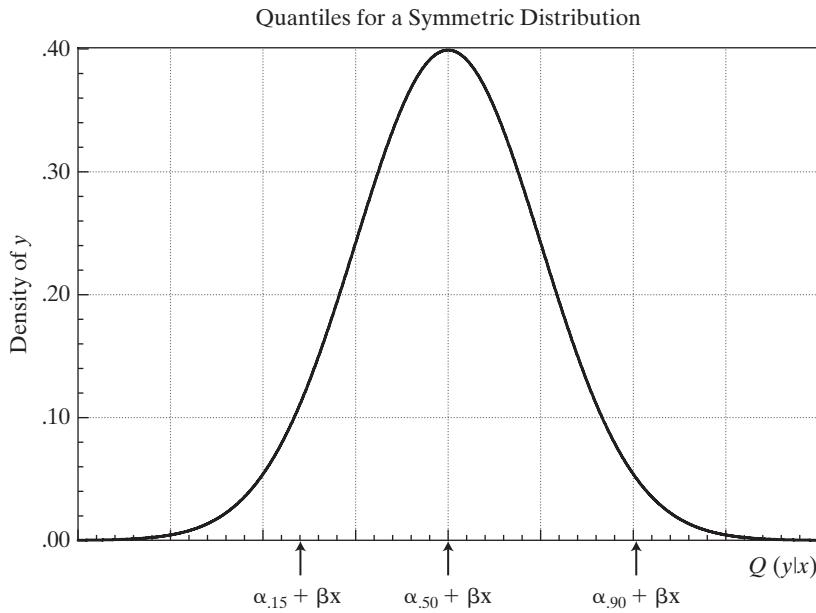
### 7.3.1 LEAST ABSOLUTE DEVIATIONS ESTIMATION

Least squares can be severely distorted by outlying observations. Recent applications in microeconomics and financial economics involving thick-tailed disturbance distributions, for example, are particularly likely to be affected by precisely these sorts of observations. (Of course, in those applications in finance involving hundreds of thousands of observations, which are becoming commonplace, this discussion is moot.) These applications have led to the proposal of “robust” estimators that are unaffected by outlying observations.<sup>10</sup> In this section, we will examine one of these, the least absolute deviations, or LAD estimator.

That least squares gives such large weight to large deviations from the regression causes the results to be particularly sensitive to small numbers of atypical data points when the sample size is small or moderate. The least absolute deviations (LAD) estimator has been suggested as an alternative that remedies (at least to some degree) the

<sup>9</sup>In Example 4.5, we considered the possibility that in small samples with possibly thick-tailed disturbance distributions, the LAD estimator might have smaller variance than least squares.

<sup>10</sup>For some applications, see Taylor (1974), Amemiya (1985, pp. 70–80), Andrews (1974), Koenker and Bassett (1978), and a survey written at a very accessible level by Birkes and Dodge (1993). A somewhat more rigorous treatment is given by Hardle (1990).

**204 PART I ♦ The Linear Regression Model**


**FIGURE 7.4** Quantile Regression Model.

problem. The LAD estimator is the solution to the optimization problem,

$$\text{Min}_{\mathbf{b}_0} \sum_{i=1}^n |y_i - \mathbf{x}'_i \mathbf{b}_0|.$$

The LAD estimator's history predates least squares (which itself was proposed over 200 years ago). It has seen little use in econometrics, primarily for the same reason that Gauss's method (LS) supplanted LAD at its origination; LS is vastly easier to compute. Moreover, in a more modern vein, its statistical properties are more firmly established than LAD's and samples are usually large enough that the small sample advantage of LAD is not needed.

The LAD estimator is a special case of the quantile regression:

$$\text{Prob}[y_i \leq \mathbf{x}'_i \boldsymbol{\beta}_g] = q.$$

The LAD estimator estimates the *median regression*. That is, it is the solution to the quantile regression when  $q = 0.5$ . Koenker and Bassett (1978, 1982), Huber (1967), and Rogers (1993) have analyzed this regression.<sup>11</sup> Their results suggest an estimator for the asymptotic covariance matrix of the quantile regression estimator,

$$\text{Est. Asy. Var}[\mathbf{b}_q] = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{D} \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1},$$

<sup>11</sup> Powell (1984) has extended the LAD estimator to produce a robust estimator for the case in which some observations on the dependent variable are censored, that is, when negative values of  $y_i$  are recorded as zero. See Example 14.7 for discussion and Melenberg and van Soest (1996) for an application. For some related results on other semiparametric approaches to regression, see Butler et al. (1990) and McDonald and White (1993).

## CHAPTER 7 ♦ Nonlinear, Semiparametric 205

where  $\mathbf{D}$  is a diagonal matrix containing weights

$$d_i = \left[ \frac{q}{f(0)} \right]^2 \text{ if } y_i - \mathbf{x}'_i \boldsymbol{\beta} \text{ is positive and } \left[ \frac{1-q}{f(0)} \right]^2 \text{ otherwise,}$$

and  $f(0)$  is the true density of the disturbances evaluated at 0.<sup>12</sup> [It remains to obtain an estimate of  $f(0)$ .] There is a useful symmetry in this result. Suppose that the true density were normal with variance  $\sigma^2$ . Then the preceding would reduce to  $\sigma^2(\pi/2)(\mathbf{X}'\mathbf{X})^{-1}$ , which is the result we used in Example 4.5. For more general cases, some other empirical estimate of  $f(0)$  is going to be required. Nonparametric methods of density estimation are available [see Section 12.4 and, e.g., Johnston and DiNardo (1997, pp. 370–375)]. But for the small sample situations in which techniques such as this are most desirable (our application below involves 25 observations), nonparametric kernel density estimation of a single ordinate is optimistic; these are, after all, asymptotic results. But asymptotically, as suggested by Example 4.5, the results begin overwhelmingly to favor least squares. For better or worse, a convenient estimator would be a **kernel density estimator** as described in Section 12.4.1. Looking ahead, the computation would be

$$\hat{f}(0) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left[\frac{e_i}{h}\right]$$

where  $h$  is the **bandwidth** (to be discussed shortly),  $K[.]$  is a weighting, or kernel function and  $e_i, i = 1, \dots, n$  is the set of residuals. There are no hard and fast rules for choosing  $h$ ; one popular choice is that used by Stata (2006),  $h = .9s/n^{1/5}$ . The kernel function is likewise discretionary, though it rarely matters much which one chooses; the logit kernel (see Table 12.2) is a common choice.

The **bootstrap** method of inferring statistical properties is well suited for this application. Since the efficacy of the bootstrap has been established for this purpose, the search for a formula for standard errors of the LAD estimator is not really necessary. The bootstrap estimator for the asymptotic covariance matrix can be computed as follows:

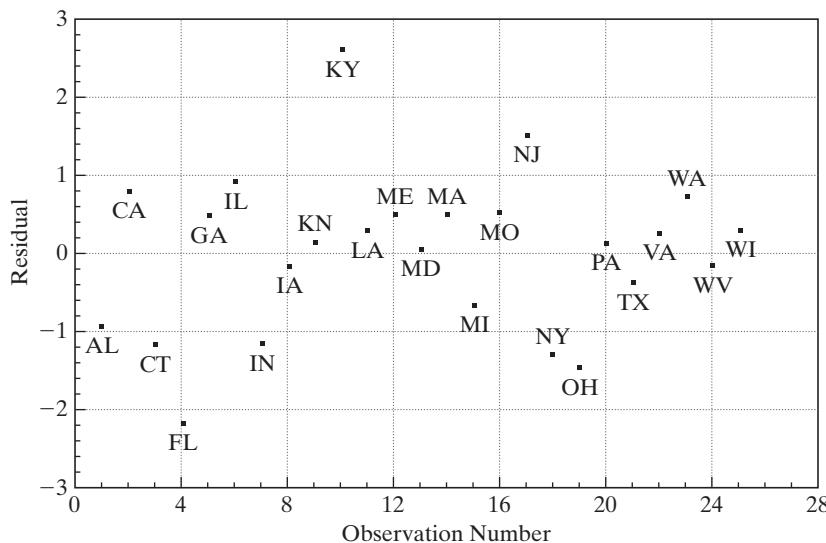
$$\text{Est. Var}[\mathbf{b}_{LAD}] = \frac{1}{R} \sum_{r=1}^R (\mathbf{b}_{LAD}(r) - \mathbf{b}_{LAD})(\mathbf{b}_{LAD}(r) - \mathbf{b}_{LAD})'$$

where  $\mathbf{b}_{LAD}$  is the LAD estimator and  $\mathbf{b}_{LAD}(r)$  is the  $r$ th LAD estimate of  $\boldsymbol{\beta}$  based on a sample of  $n$  observations, drawn with replacement, from the original data set.

**Example 7.9 LAD Estimation of a Cobb-Douglas Production Function**

Zellner and Revankar (1970) proposed a generalization of the Cobb-Douglas production function that allows economies of scale to vary with output. Their statewide data on  $Y$  = value added (output),  $K$  = capital,  $L$  = labor, and  $N$  = the number of establishments in the transportation industry are given in Appendix Table F7.2. For this application, estimates of the

<sup>12</sup>Koenker suggests that for independent and identically distributed observations, one should replace  $d_i$  with the constant  $a = q(1-q)/[f(F^{-1}(q))]^2 = [.50/f(0)]^2$  for the median (LAD) estimator. This reduces the expression to the true asymptotic covariance matrix,  $a(\mathbf{X}'\mathbf{X})^{-1}$ . The one given is a sample estimator which will behave the same in large samples. (Personal communication to the author.)

**206 PART I ♦ The Linear Regression Model**

**FIGURE 7.5** Standardized Residuals for Production Function.

**TABLE 7.4** LS and LAD Estimates of a Production Function

Least Squares				LAD				
Coefficient	Estimate	Standard Error	t Ratio	Estimate	Std. Error	t Ratio	Std. Error	t Ratio
Constant	2.293	0.107	21.396	2.275	0.202	11.246	0.183	12.374
$\beta_k$	0.279	0.081	3.458	0.261	0.124	2.099	0.138	1.881
$\beta_l$	0.927	0.098	9.431	0.927	0.121	7.637	0.169	5.498
$\Sigma e^2$	0.7814			0.7984				
$\Sigma  e $	3.3652			3.2541				

Cobb–Douglas production function,

$$\ln(Y_i/N_i) = \beta_1 + \beta_2 \ln(K_i/N_i) + \beta_3 \ln(L_i/N_i) + \varepsilon_i,$$

are obtained by least squares and LAD. The standardized least squares residuals shown in Figure 7.5 suggest that two observations (Florida and Kentucky) are outliers by the usual construction. The least squares coefficient vectors with and without these two observations are  $(2.293, 0.279, 0.927)$  and  $(2.205, 0.261, 0.879)$ , respectively, which bears out the suggestion that these two points do exert considerable influence. Table 7.4 presents the LAD estimates of the same parameters, with standard errors based on 500 bootstrap replications. The LAD estimates with and without these two observations are identical, so only the former are presented. Using the simple approximation of multiplying the corresponding OLS standard error by  $(\pi/2)^{1/2} = 1.2533$  produces a surprisingly close estimate of the bootstrap estimated standard errors for the two slope parameters  $(0.102, 0.123)$  compared with the bootstrap estimates of  $(0.124, 0.121)$ . The second set of estimated standard errors are based on Koenker's suggested estimator,  $.25/\hat{f}^2(0) = 0.25/1.5467^2 = 0.104502$ . The bandwidth and kernel function are those suggested earlier. The results are surprisingly consistent given the small sample size.

### 7.3.2 QUANTILE REGRESSION MODELS

The quantile regression model is

$$Q[y|\mathbf{x}, q] = \mathbf{x}'\boldsymbol{\beta}_q \text{ such that } \text{Prob}[y \leq \mathbf{x}'\boldsymbol{\beta}_q | \mathbf{x}] = q, 0 < q < 1.$$

This is essentially a nonparametric specification. No assumption is made about the distribution of  $y|\mathbf{x}$  or about its conditional variance. The fact that  $q$  can vary continuously (strictly) between zero and one means that there are an infinite number of possible “parameter vectors.” It seems reasonable to view the coefficients, which we might write  $\boldsymbol{\beta}(q)$  less as fixed “parameters,” as we do in the linear regression model, than loosely as *features* of the distribution of  $y|\mathbf{x}$ . For example, it is not likely to be meaningful to view  $\boldsymbol{\beta}(.49)$  to be discretely different from  $\boldsymbol{\beta}(.50)$  or to compute precisely a particular difference such as  $\boldsymbol{\beta}(.5) - \boldsymbol{\beta}(.3)$ . On the other hand, the qualitative difference, or possibly the lack of a difference, between  $\boldsymbol{\beta}(.3)$  and  $\boldsymbol{\beta}(.5)$  as displayed in our following example, may well be an interesting characteristic of the sample.

The estimator,  $\mathbf{b}_q$  of  $\boldsymbol{\beta}_q$  for a specific quantile is computed by minimizing the function

$$\begin{aligned} F_n(\boldsymbol{\beta}_q | \mathbf{y}, \mathbf{X}) &= \sum_{i:y_i \geq \mathbf{x}'_i \boldsymbol{\beta}_q}^n q|y_i - \mathbf{x}'_i \boldsymbol{\beta}_q| + \sum_{i:y_i < \mathbf{x}'_i \boldsymbol{\beta}_q}^n (1-q)|y_i - \mathbf{x}'_i \boldsymbol{\beta}_q| \\ &= \sum_{i=1}^n g(y_i - \mathbf{x}'_i \boldsymbol{\beta}_q | q) \end{aligned}$$

where

$$g(e_{i,q} | q) = \begin{cases} qe_{i,q} & \text{if } e_{i,q} \geq 0 \\ (1-q)e_{i,q} & \text{if } e_{i,q} < 0 \end{cases}, e_{i,q} = y_i - \mathbf{x}'_i \boldsymbol{\beta}_q.$$

When  $q = 0.5$ , the estimator is the least absolute deviations estimator we examined in Example 4.5 and Section 7.3.1. Solving the minimization problem requires an iterative estimator. It can be set up as a linear programming problem.<sup>13</sup> [See Keonker and D’Oray (1987).]

We cannot use the methods of Chapter 4 to determine the asymptotic covariance matrix of the estimator. But, the fact that the estimator is obtained by minimizing a sum does lead to a set of results similar to those we obtained in Section 4.4 for least squares. [See Buchinsky (1998).] Assuming that the regressors are “well behaved,” the quantile regression estimator of  $\boldsymbol{\beta}_q$  is consistent and asymptotically normally distributed with asymptotic covariance matrix

$$\text{Asy.Var.}[\mathbf{b}_q] = \frac{1}{n} \mathbf{H}^{-1} \mathbf{G} \mathbf{H}^{-1}$$

where

$$\mathbf{H} = \text{plim} \frac{1}{n} \sum_{i=1}^n f_q(0 | \mathbf{x}_i) \mathbf{x}_i \mathbf{x}'_i$$

---

<sup>13</sup>Quantile regression is supported as a built in procedure in contemporary software such as Stata, SAS, and NLOGIT.

## 208 PART I ♦ The Linear Regression Model

and

$$\mathbf{G} = \text{plim} \frac{q(1-q)}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i.$$

This is the result we had earlier for the LAD estimator, now with quantile  $q$  instead of 0.5. As before, computation is complicated by the need to compute the density of  $\varepsilon_q$  at zero. This will require either an approximation of uncertain quality or a specification of the particular density, which we have hoped to avoid. The usual approach, as before, is to use bootstrapping.

### **Example 7.10 Income Elasticity of Credit Card Expenditure**

Greene (1992, 2007) analyzed the default behavior and monthly expenditure behavior of a large sample (13,444 observations) of credit card users. Among the results of interest in the study was an estimate of the income elasticity of the monthly expenditure. A conventional regression approach might be based on

$$Q[\ln \text{Spending} | \mathbf{x}, q] = \beta_{1,q} + \beta_{2,q} \ln \text{Income} + \beta_{3,q} \text{Age} + \beta_{4,q} \text{Dependents}$$



The data in Appendix Table F7.3 contain these and numerous other covariates that might explain spending; we have chosen these three for this example only. The 13,444 observations in the data set are based on credit card applications. Of the full sample, 10,499 applications were approved and the next 12 months of spending and default behavior were observed.<sup>14</sup> Spending is the average monthly expenditure in the 12 months after the account was initiated. Average monthly income and number of household dependents are among the demographic data in the application. Table 7.5 presents least squares estimates of the coefficients of the conditional mean function as well as full results for several quantiles.<sup>15</sup> Standard errors are shown for the least squares and median ( $1 = 0.5$ ) results. The results for the other quantiles are essentially the same. The least squares estimate of 1.08344 is slightly and significantly greater than one—the estimated standard error is 0.03212 so the  $t$  statistic is  $(1.08344 - 1)/0.03212 = 2.60$ . This suggests an aspect of consumption behavior that might not be surprising. However, the very large amount of variation over the range of quantiles might not have been expected. We might guess that at the highest levels of spending for any income level, there is (comparably so) some saturation in the response of spending to changes in income.

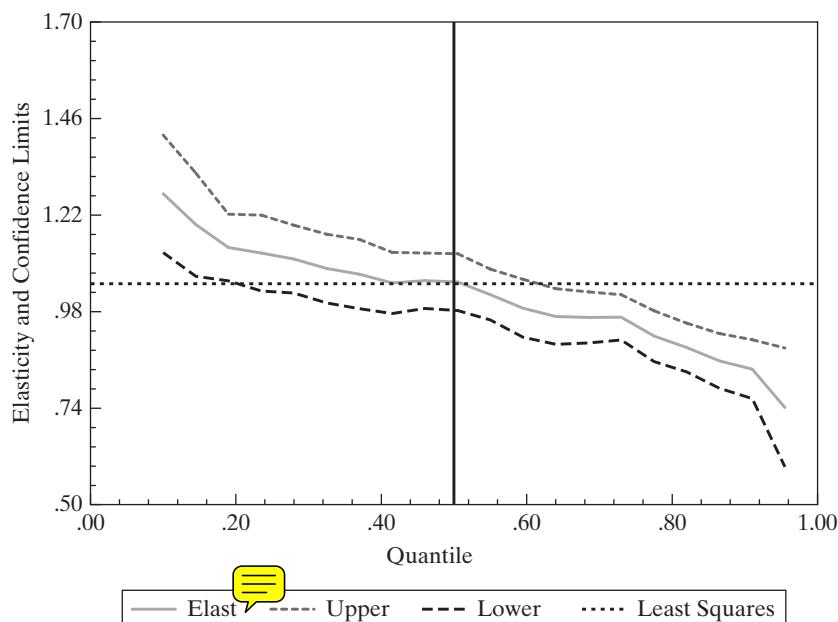
Figure 7.6 displays the estimates of the income elasticity of expenditure for the range of quantiles from 0.1 to 0.9, with the least squares estimate which would correspond to the fixed value at all quantiles shown in the center of the figure. Confidence limits shown in the figure are based on the asymptotic normality of the estimator. They are computed as the estimated income elasticity plus and minus 1.96 times the estimated standard error. Figure 7.7 shows the implied quantile regressions for  $q = .1, .3, .5, .7$ , and  $.9$ . The relatively large increase from the .1 quantile to the .3 suggests some skewness in the spending distribution. In broad

<sup>14</sup>The expenditure data are taken from the credit card records while the income and demographic data are taken from the applications. While it might be tempting to use, for example, Powell's (1986a,b) censored quantile regression estimator to accommodate this large cluster of zeros for the dependent variable, this approach would misspecify the model—the “zeros” represent nonexistent observations, not missing ones. A more detailed approach—the one used in the 1992 study—would model separately the presence or absence of the observation on spending, then model spending conditionally on acceptance of the application. We will revisit this issue in Chapter 17 in the context of the sample selection model. The income data are censored at 100,000 and 220 of the observations have expenditures that are filled with \$1 or less. We have not “cleaned” the data set for these aspects. The full 10,499 observations have been used as they are in the original data set.

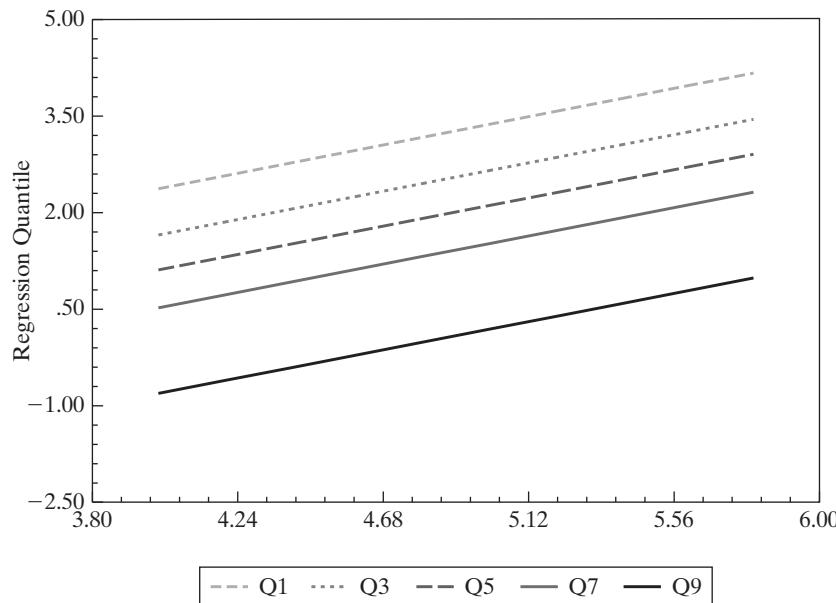
<sup>15</sup>We would note, if (7-33) is the statement of the model, then it does not follow that the conditional mean function is a linear regression. That would be an additional assumption.

**TABLE 7.5** Estimated Quantile Regression Models

<i>Quantile</i>	<i>Estimated Parameters</i>			
	<i>Constant</i>	<i>In Income</i>	<i>Age</i>	<i>Dependents</i>
0.1	-6.73560	1.40306	-.03081	-.04297
0.2	-4.31504	1.16919	-.02460	-.04630
0.3	-3.62455	1.12240	-.02133	-.04788
0.4	-2.98830	1.07109	-.01859	-.04731
(Median) 0.5	-2.80376	1.07493	-.01699	-.04995
Std.Error	(.24564)	(.03223)	(.00157)	(.01080)
<i>t</i>	-11.41	33.35	-10.79	-4.63
Least Squares	-3.05581	1.08344	-.01736	-.04461
Std.Error	(.23970)	(.03212)	(.00135)	(.01092)
<i>t</i>	-12.75	33.73	-12.88	-4.08
0.6	-2.05467	1.00302	-.01478	-.04609
0.7	-1.63875	.97101	-.01190	-.03803
0.8	-.94031	.91377	-.01126	-.02245
0.9	-.05218	.83936	-.00891	-.02009

**FIGURE 7.6** Estimates of Income Elasticity of Expenditure.

terms, the results do seem to be largely consistent with our earlier result of the quantiles largely being differentiated by shifts in the constant term, in spite of the seemingly large change in the coefficient on *In Income* in the results.

**210 PART I ♦ The Linear Regression Model**

**FIGURE 7.7** Quantile Regressions for Ln Spending.

**7.4 PARTIALLY LINEAR REGRESSION**

The proper functional form in the linear regression is an important specification issue. We examined this in detail in Chapter 6. Some approaches, including the use of dummy variables, logs, quadratics, and so on, were considered as means of capturing nonlinearity. The translog model in particular (Example 2.4) is a well-known approach to approximating an unknown nonlinear function. Even with these approaches, the researcher might still be interested in relaxing the assumption of functional form in the model. The partially linear model [analyzed in detail by Yatchew (1998, 2000) and Härdle, Liang, and Gao (2000)] is another approach. Consider a regression model in which one variable,  $x$ , is of particular interest, and the functional form with respect to  $x$  is problematic. Write the model as

$$y_i = f(x_i) + \mathbf{z}'_i \boldsymbol{\beta} + \varepsilon_i,$$

where the data are assumed to be well behaved and, save for the functional form, the assumptions of the classical model are met. The function  $f(x_i)$  remains unspecified. As stated, estimation by least squares is not feasible until  $f(x_i)$  is specified. Suppose the data were such that they consisted of pairs of observations  $(y_{j1}, y_{j2})$ ,  $j = 1, \dots, n/2$ , in which  $x_{j1} = x_{j2}$  within every pair. If so, then estimation of  $\boldsymbol{\beta}$  could be based on the simple transformed model

$$y_{j2} - y_{j1} = (\mathbf{z}_{j2} - \mathbf{z}_{j1})' \boldsymbol{\beta} + (\varepsilon_{j2} - \varepsilon_{j1}), \quad j = 1, \dots, n/2.$$

As long as observations are independent, the constructed disturbances,  $v_i$ , still have zero mean, variance now  $2\sigma^2$ , and remain uncorrelated across pairs, so a classical model applies and least squares is actually optimal. Indeed, with the estimate of  $\boldsymbol{\beta}$ , say,  $\hat{\boldsymbol{\beta}}_d$  in

## CHAPTER 7 ♦ Nonlinear, Semiparametric 211

hand, a noisy estimate of  $f(x_i)$  could be estimated with  $y_i - \mathbf{z}'_i \hat{\beta}_d$  (the estimate contains the estimation error as well as  $\varepsilon_i$ ).<sup>16</sup>

The problem, of course, is that the enabling assumption is heroic. Data would not behave in that fashion unless they were generated experimentally. The logic of the partially linear regression estimator is based on this observation nonetheless. Suppose that the observations are sorted so that  $x_1 < x_2 < \dots < x_n$ . Suppose, as well, that this variable is well behaved in the sense that as the sample size increases, this sorted data vector more tightly and uniformly fills the space within which  $x_i$  is assumed to vary. Then, intuitively, the difference is “almost” right, and becomes better as the sample size grows. [Yatchew (1997, 1998) goes more deeply into the underlying theory.] A theory is also developed for a better differencing of groups of two or more observations. The transformed observation is  $y_{d,i} = \sum_{m=0}^M d_m y_{i-m}$  where  $\sum_{m=0}^M d_m = 0$  and  $\sum_{m=0}^M d_m^2 = 1$ . (The data are not separated into nonoverlapping groups for this transformation—we merely used that device to motivate the technique.) The pair of weights for  $M = 1$  is obviously  $\pm\sqrt{0.5}$ —this is just a scaling of the simple difference, 1, −1. Yatchew [1998, p. 697] tabulates “optimal” differencing weights for  $M = 1, \dots, 10$ . The values for  $M = 2$  are (0.8090, −0.500, −0.3090) and for  $M = 3$  are (0.8582, −0.3832, −0.2809, −0.1942). This estimator is shown to be consistent, asymptotically normally distributed, and have asymptotic covariance matrix<sup>17</sup>

$$\text{Asy. Var}[\hat{\beta}_d] = \left(1 + \frac{1}{2M}\right) \frac{\sigma_v^2}{n} E_x[\text{Var}[\mathbf{z} | x]].$$

The matrix can be estimated using the sums of squares and cross products of the differenced data. The residual variance is likewise computed with

$$\hat{\sigma}_v^2 = \frac{\sum_{i=M+1}^n (y_{d,i} - \mathbf{z}'_{d,i} \hat{\beta}_d)^2}{n - M}.$$

Yatchew suggests that the partial residuals,  $y_{d,i} - \mathbf{z}'_{d,i} \hat{\beta}_d$  be smoothed with a kernel density estimator to provide an improved estimator of  $f(x_i)$ . Manzan and Zeron (2010) present an application of this model to the U.S. gasoline market.

#### **Example 7.11 Partially Linear Translog Cost Function**

Yatchew (1998, 2000) applied this technique to an analysis of scale effects in the costs of electricity supply. The cost function, following Nerlove (1963) and Christensen and Greene (1976), was specified to be a translog model (see Example 2.4 and Section 10.5.2) involving labor and capital input prices, other characteristics of the utility, and the variable of interest, the number of customers in the system,  $C$ . We will carry out a similar analysis using Christensen and Greene's 1970 electricity supply data. The data are given in Appendix Table F4.4. (See Section 10.5.1 for description of the data.) There are 158 observations in the data set, but the last 35 are holding companies which are comprised of combinations of the others. In addition, there are several extremely small New England utilities whose costs are clearly unrepresentative of the best practice in the industry. We have done the analysis using firms 6–123 in the data set. Variables in the data set include  $Q$  = output,  $C$  = total cost, and  $PK$ ,  $PL$ , and  $PF$  = unit cost measures for capital, labor, and fuel, respectively. The parametric model

<sup>16</sup>See Estes and Honoré (1995) who suggest this approach (with simple differencing of the data).

<sup>17</sup>Yatchew (2000, p. 191) denotes this covariance matrix  $E[\text{Cov}[\mathbf{z} | x]]$ .

## 212 PART I ♦ The Linear Regression Model

specified is a restricted version of the Christensen and Greene model,

$$\ln c = \beta_1 k + \beta_2 l + \beta_3 q + \beta_4 (q^2/2) + \beta_5 + \varepsilon,$$

where  $c = \ln[C/(Q \times P)]$ ,  $k = \ln(PK/PF)$ ,  $l = \ln(PL/PF)$ , and  $q = \ln Q$ . The partially linear model substitutes  $f(Q)$  for the last three terms. The division by  $PF$  ensures that average cost is homogeneous of degree one in the prices, a theoretical necessity. The estimated equations, with estimated standard errors, are shown here.

$$(\text{parametric}) \quad c = -7.32 + 0.069k + 0.241 - 0.569q + 0.057q^2/2 + \varepsilon, \quad s = 0.13949$$

$$(\text{partially linear}) \quad c_d = 0.108k_d + 0.163l_d + f(q) + \varepsilon \quad s = 0.16529$$

## 7.5 NONPARAMETRIC REGRESSION

The regression function of a variable  $y$  on a single variable  $x$  is specified as

$$y = \mu(x) + \varepsilon.$$

No assumptions about distribution, homoscedasticity, serial correlation or, most importantly, functional form are made at the outset;  $\mu(x)$  may be quite nonlinear. Because this is the conditional mean, the only substantive restriction would be that deviations from the conditional mean function are not a function of (correlated with)  $x$ . We have already considered several possible strategies for allowing the conditional mean to be nonlinear, including spline functions, polynomials, logs, dummy variables, and so on. But, each of these is a “global” specification. The functional form is still the same for all values of  $x$ . Here, we are interested in methods that do not assume any particular functional form.

The simplest case to analyze would be one in which several (different) observations on  $y_i$  were made with each specific value of  $x_i$ . Then, the conditional mean function could be estimated naturally using the simple group means. The approach has two shortcomings, however. Simply connecting the points of means,  $(x_i, \bar{y} | x_i)$  does not produce a smooth function. The method would still be assuming something specific about the function between the points, which we seek to avoid. Second, this sort of data arrangement is unlikely to arise except in an experimental situation. Given that data are not likely to be grouped, another possibility is a piecewise regression in which we define “neighborhoods” of points around each  $x$  of interest and fit a separate linear or quadratic regression in each neighborhood. This returns us to the problem of continuity that we noted earlier, but the method of splines, discussed in Section 6.3.1, is actually designed specifically for this purpose. Still, unless the number of neighborhoods is quite large, such a function is still likely to be crude.

Smoothing techniques are designed to allow construction of an estimator of the conditional mean function without making strong assumptions about the behavior of the function between the points. They retain the usefulness of the **nearest neighbor** concept but use more elaborate schemes to produce smooth, well-behaved functions. The general class may be defined by a conditional mean estimating function

$$\hat{\mu}(x^*) = \sum_{i=1}^n w_i(x^* | x_1, x_2, \dots, x_n) y_i = \sum_{i=1}^n w_i(x^* | \mathbf{x}) y_i,$$

## CHAPTER 7 ♦ Nonlinear, Semiparametric 213

where the weights sum to 1. The linear least squares regression line is such an estimator. The predictor is

$$\hat{\mu}(x^*) = a + bx^*.$$

where  $a$  and  $b$  are the least squares constant and slope. For this function, you can show that

$$w_i(x^*|\mathbf{x}) = \frac{1}{n} + \frac{x^*(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

The problem with this particular weighting function, which we seek to avoid here, is that it allows every  $x_i$  to be in the neighborhood of  $x^*$ , but it does not reduce the weight of any  $x_i$  when it is far from  $x^*$ . A number of **smoothing functions** have been suggested that are designed to produce a better behaved regression function. [See Cleveland (1979) and Schimek (2000).] We will consider two.

The locally weighted smoothed regression estimator (“loess” or “lowess” depending on your source) is based on explicitly defining a neighborhood of points that is close to  $x^*$ . This requires the choice of a bandwidth,  $h$ . The **neighborhood** is the set of points for which  $|x^* - x_i|$  is small. For example, the set of points that are within the range  $x^* \pm h/2$  might constitute the neighborhood. The choice of bandwidth is crucial, as we will explore in the following example, and is also a challenge. There is no single best choice. A common choice is **Silverman’s (1986) rule of thumb**,

$$h_{Silverman} = \frac{.9[\min(s, IQR)]}{1.349 n^{0.2}}$$

where  $s$  is the sample standard deviation and  $IQR$  is the interquartile range (.75 quantile minus .25 quantile). A suitable weight is then required. Cleveland (1979) recommends the tricube weight,

$$T_i(x^*|\mathbf{x}, h) = \left[ 1 - \left( \frac{|x_i - x^*|}{h} \right)^3 \right]^3.$$

Combining terms, then the weight for the loess smoother is

$$w_i(x^*|\mathbf{x}, h) = 1(x_i \text{ in the neighborhood}) \times T_i(x^*|\mathbf{x}, h).$$

The bandwidth is essential in the results. A wider neighborhood will produce a smoother function, but the wider neighborhood will track the data less closely than a narrower one. A second possibility, similar to the least squares approach, is to allow the neighborhood to be all points but make the weighting function decline smoothly with the distance between  $x^*$  and any  $x_i$ . A variety of **kernel functions** are used for this purpose. Two common choices are the **logistic kernel**,

$$K(x^*|x_i, h) = \Lambda(v_i)[1 - \Lambda(v_i)] \text{ where } \Lambda(v_i) = \exp(v_i)/[1 + \exp(v_i)], v_i = (x_i - x^*)/h,$$

**214 PART I ♦ The Linear Regression Model**


and the Epanechnikov kernel,

$$K(x^*|x_i, h) = 0.75(1 - 0.2v_i^2)/\sqrt{5} \text{ if } |v_i| \leq 5 \text{ and } 0 \text{ otherwise.}$$

This produces the kernel weighted regression estimator,

$$\hat{\mu}(x^*|\mathbf{x}, h) = \frac{\sum_{i=1}^n \frac{1}{k} K\left[\frac{x_i - x^*}{h}\right] y_i}{\sum_{i=1}^n \frac{1}{k} K\left[\frac{x_i - x^*}{h}\right]},$$

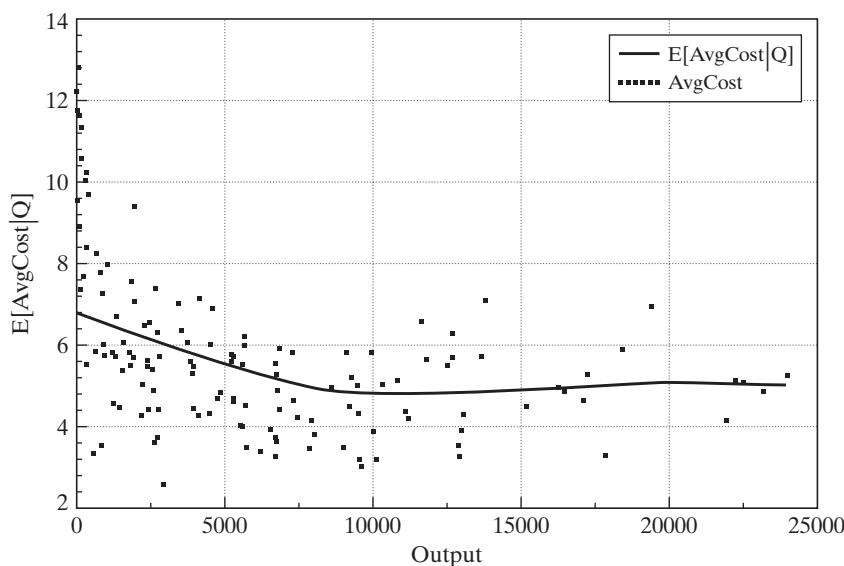
which has become a standard tool in nonparametric analysis.

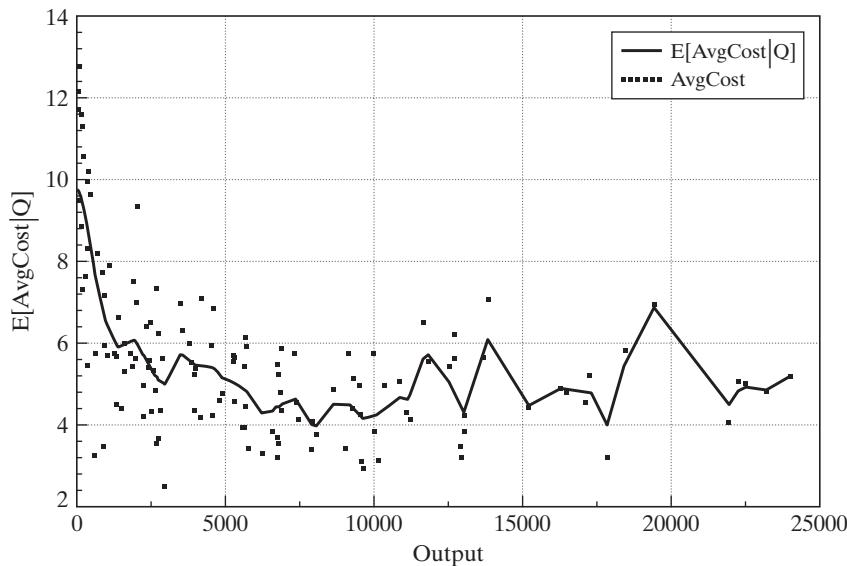
**Example 7.12 A Nonparametric Average Cost Function**

In Example 7.11, we fit a partially linear regression for the relationship between average cost and output for electricity supply. Figures 7.8 and 7.9 show the less ambitious nonparametric regressions of average cost on output. The overall picture is the same as in the earlier example. The kernel function is the logistic density in both cases. The function in Figure 7.8 uses a bandwidth of 2,000. Because this is a fairly large proportion of the range of variation of output, the function is quite smooth. The regression in Figure 7.9 uses a bandwidth of only 200. The function tracks the data better, but at an obvious cost. The example demonstrates what we and others have noted often. The choice of bandwidth in this exercise is crucial.

Data smoothing is essentially data driven. As with most nonparametric techniques, inference is not part of the analysis—this body of results is largely descriptive. As can be seen in the example, nonparametric regression can reveal interesting characteristics of the data set. For the econometrician, however, there are a few drawbacks. There is no danger of misspecifying the conditional mean function, for example. But, the great

**FIGURE 7.8** Nonparametric Cost Function.





**FIGURE 7.9** Nonparametric Cost Function.

generality of the approach limits the ability to test one's specification or the underlying theory. [See, for example, Blundell, Browning, and Crawford's (2003) extensive study of British expenditure patterns.] Most relationships are more complicated than a simple conditional mean of one variable. In the Example 7.12, some of the variation in average cost relates to differences in factor prices (particularly fuel) and in load factors. Extensions of the fully nonparametric regression to more than one variable is feasible, but very cumbersome. [See Härdle (1990) and Li and Racine (2007).] A promising approach is the partially linear model considered earlier.

## 7.6 SUMMARY AND CONCLUSIONS

In this chapter, we extended the regression model to a form that allows nonlinearity in the parameters in the regression function. The results for interpretation, estimation, and hypothesis testing are quite similar to those for the linear model. The two crucial differences between the two models are, first, the more involved estimation procedures needed for the nonlinear model and, second, the ambiguity of the interpretation of the coefficients in the nonlinear model (because the derivatives of the regression are often nonconstant, in contrast to those in the linear model).

### Key Terms and Concepts

- |                             |                            |                          |
|-----------------------------|----------------------------|--------------------------|
| • Bandwidth                 | • Conditional median       | • Identification problem |
| • Bootstrap                 | • Delta method             | • Incidental parameters  |
| • Box-Cox transformation    | • GMM estimator            | problem                  |
| • Conditional mean function | • Identification condition | • Index function model   |



## 216 PART I ♦ The Linear Regression Model

- Indirect utility function
- Interaction term
- Iteration
- Jacobian
- Kernel density estimator
- Kernel functions
- Least absolute deviations (LAD)
- Linear regression model
- Linearized regression model
- Lagrange multiplier test
- Logistic kernel
- Logit model
- Loglinear model
- Median regression
- Nearest neighbor
- Neighborhood
- Nonlinear least squares
- Nonlinear regression model
- Nonparametric estimators
- Nonparametric regression
- Normalization
- Orthogonality condition
- Overidentifying restrictions
- Partially linear model
- Pseudoregressors
- Quantile regression
- Roy's identity
- Semiparametric
- Semiparametric estimation
- Silverman's rule of thumb
- Smoothing function
- Starting values
- Two-step estimation
- Wald test

### Exercises

1. Describe how to obtain nonlinear least squares estimates of the parameters of the model  $y = \alpha x^\beta + \varepsilon$ .
2. Verify the following differential equation, which applies to the Box–Cox transformation:

$$\frac{d^i x^{(\lambda)}}{d\lambda^i} = \left( \frac{1}{\lambda} \right) \left[ x^\lambda (\ln x)^i - \frac{i d^{i-1} x^{(\lambda)}}{d\lambda^{i-1}} \right]. \quad (7-34)$$

Show that the limiting sequence for  $\lambda = 0$  is

$$\lim_{\lambda \rightarrow 0} \frac{d^i x^{(\lambda)}}{d\lambda^i} = \frac{(\ln x)^{i+1}}{i+1}. \quad (7-35)$$

These results can be used to great advantage in deriving the actual second derivatives of the log-likelihood function for the Box–Cox model.

### Applications



1. The data in Appendix table F5.3 present 27 statewide observations on value added (output), labor input (labor), and capital stock (capital) for SIC 33 (primary metals). We are interested in determining whether a linear or loglinear production model is more appropriate for these data. Use MacKinnon, White, and Davidson's (1983)  $P_F$  test to determine whether a linear or loglinear production model is preferred.
2. Using the Box–Cox transformation, we may specify an alternative to the Cobb–Douglas model as

$$\ln Y = \alpha + \beta_k \frac{(K^\lambda - 1)}{\lambda} + \beta_l \frac{(L^\lambda - 1)}{\lambda} + \varepsilon.$$

Using Zellner and Revankar's data in Appendix Table , estimate  $\alpha$ ,  $\beta_k$ ,  $\beta_l$ , and  $\lambda$  by using the scanning method suggested in Section 11.5.2. (Do not forget to scale  $Y$ ,  $K$ , and  $L$  by the number of establishments.) Use (7-24), (7-15), and (7-16) to compute the appropriate asymptotic standard errors for your estimates. Compute the two output elasticities,  $\partial \ln Y / \partial \ln K$  and  $\partial \ln Y / \partial \ln L$ , at the sample means of  $K$  and  $L$ . (Hint:  $\partial \ln Y / \partial \ln K = K \partial \ln Y / \partial K$ .)

## CHAPTER 7 ♦ Nonlinear, Semiparametric 217

3. For the model in Application 2, test the hypothesis that  $\lambda = 0$  using a Wald test and a Lagrange multiplier test. Note that the restricted model is the Cobb–Douglas loglinear model. The LM test statistic is shown in (7-22). To carry out the test, you will need to compute the elements of the fourth column of  $\mathbf{X}^0$ , the pseudoregressor corresponding to  $\lambda$  is  $\partial E[y|x]/\partial\lambda | \lambda = 0$ . Result (7-35) will be useful.
4. The National Institute of Standards and Technology (NIST) has created a web site that contains a variety of estimation problems, with data sets, designed to test the accuracy of computer programs. (The URL is <http://www.itl.nist.gov/div898/strd/>.) One of the five suites of test problems is a set of 27 nonlinear least squares problems, divided into three groups: easy, moderate, and difficult. We have chosen one of them for this application. You might wish to try the others (perhaps see if the software you are using can solve the problems). This is the Misralc problem (<http://www.itl.nist.gov/div898/strd/nls/data/misralc.shtml>). The nonlinear regression model is

$$\begin{aligned}y_i &= h(x, \beta) + \varepsilon \\&= \beta_1 \left( 1 - \frac{1}{\sqrt{1 + 2\beta_2 x_i}} \right) + \varepsilon_i.\end{aligned}$$

The data are as follows:

<b>Y</b>	<b>X</b>
10.07	77.6
14.73	114.9
17.94	141.1
23.93	190.8
29.61	239.9
35.18	289.0
40.02	332.8
44.82	378.4
50.76	434.8
55.05	477.3
61.01	536.8
66.40	593.1
75.47	689.1
81.78	760.0

For each problem posed, NIST also provides the “certified solution,” (i.e., the right answer). For the Misralc problem, the solutions are as follows:

	<b>Estimate</b>	<b>Estimated Standard Error</b>
$\beta_1$	6.3642725809E + 02	4.6638326572E + 00
$\beta_2$	2.0813627256E – 04	1.7728423155E – 06
$\mathbf{e}'\mathbf{e}$		4.0966836971E – 02
$s^2 = \mathbf{e}'\mathbf{e}/(n - K)$		5.8428615257E – 02

Finally, NIST provides two sets of starting values for the iterations, generally one set that is “far” from the solution and a second that is “close” from the solution. For this problem, the starting values provided are  $\beta^1 = (500, 0.0001)$  and  $\beta^2 = (600, 0.0002)$ . The exercise here is to reproduce the NIST results with your software. [For a detailed

## 218 PART I ♦ The Linear Regression Model

analysis of the NIST nonlinear least squares benchmarks with several well-known computer programs, see McCullough (1999).]

5. In Example 7.1, the CES function is suggested as a model for production;

$$\ln y = \ln \gamma - \frac{\nu}{\rho} \ln [\delta K^{-\rho} + (1-\delta)L^{-\rho}] + \varepsilon. \quad (7-36)$$

Example 6.8 suggested an indirect method of estimating the parameters of this model. The function is linearized around  $\rho = 0$ , which produces an intrinsically linear approximation to the function,

$$\ln y = \beta_1 + \beta_2 \ln K + \beta_3 \ln L + \beta_4 [1/2(\ln K - \ln L)^2] + \varepsilon \quad \text{[Talk icon]}$$

where  $\beta_1 = \ln \gamma$ ,  $\beta_2 = \nu \delta$ ,  $\beta_3 = \nu(1-\delta)$  and  $\beta_4 = \rho \nu \delta(1-\delta)$ . The approximation can be estimated by linear least squares. Estimates of the structural parameters are found by inverting the preceding four equations. An estimator of the asymptotic covariance matrix is suggested using the delta method. The parameters of (7-36) can also be estimated directly using nonlinear least squares and the results given earlier in this chapter.

Christensen and Greene's (1976) data on U.S. electricity generation are given in Appendix Table F4.4. The data file contains 158 observations. Using the first 123, fit the CES production function, using capital and fuel as the two factors of production rather than capital and labor. Compare the results obtained by the two approaches, and comment on why the differences (which are substantial) arise.

The following exercises require specialized software. The relevant techniques are available in several packages that might be in use, such as SAS, Stata, or LIMDEP. The exercises are suggested as departure points for explorations using a few of the many estimation techniques listed in this chapter.

6. Using the gasoline market data in Appendix Table F2.2, use the partially linear regression method in Section 7.4 to fit an equation of the form

$$\ln(G/Pop) = \beta_1 \ln(Income) + \beta_2 \ln P_{new\ cars} + \beta_3 \ln P_{used\ cars} + g(\ln P_{gasoline}) + \varepsilon. \quad \text{[Talk icon]}$$

7. To continue the analysis in Question 6, consider a nonparametric regression of  $G/Pop$  on the price. Using the nonparametric estimation method in Section 7.5, fit the nonparametric estimator using a range of bandwidth values to explore the effect of bandwidth.

## 8

# ENDOGENEITY AND INSTRUMENTAL VARIABLE ESTIMATION

---

## 8.1 INTRODUCTION

The assumption that  $\mathbf{x}_i$  and  $\varepsilon_i$  are uncorrelated in the linear regression model,

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i, \quad (8-1)$$

has been crucial in the development thus far. But, there are many applications in which this assumption is untenable. Examples include models of treatment effects such as that in Example 6.5, models that contain variables that are measured with error, dynamic models involving expectations, and a large variety of common situations that involve variables that are unobserved, or for other reasons are omitted from the equation. Without the assumption that the disturbances and the regressors are uncorrelated, none of the proofs of consistency or unbiasedness of the least squares estimator that were obtained in Chapter 4 will remain valid, so the least squares estimator loses its appeal. This chapter will develop an estimation method that arises in situations such as these.

It is convenient to partition  $\mathbf{x}$  in (8-1) into two sets of variables,  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , with the assumption that  $\mathbf{x}_1$  is not correlated with  $\varepsilon$  and  $\mathbf{x}_2$  is, or may be, (part of the empirical investigation). We are assuming that  $\mathbf{x}_1$  is **exogenous** in the model—see assumption A.3 in the statement of the linear regression model in Section 2.3. It will follow that  $\mathbf{x}_2$  is, by this definition, **endogenous** in the model. How does endogeneity arise? Example 8.1 suggests some common settings.

### **Example 8.1 Models with Endogenous Right Hand Side Variables**

The following models and settings will appear at various points in this book.

**Omitted Variables:** In Example 4.2, we examined an equation for gasoline consumption of the form

$$\ln G = \beta_1 + \beta_2 \ln Price + \beta_3 \ln Income + \varepsilon.$$

When income is improperly omitted from this (any) demand equation, the resulting “model” is

$$\ln G = \beta_1 + \beta_2 \ln Price + w,$$

where  $w = \beta_3 \ln Income + \varepsilon$ . Linear regression of  $\ln G$  on a constant and  $\ln Price$  does not consistently estimate  $(\beta_1, \beta_2)$  if  $\ln Price$  is correlated with  $w$ . It surely will be in aggregate time-series data. The omitted variable reappears in the equation in the disturbance, causing **omitted variable bias** in the least squares estimator of the misspecified equation.

**Endogenous Treatment Effects:** Kreuger and Dale (1999) examined the effect of attendance at an elite college on lifetime earnings. The regression model with a “treatment effect” dummy variable,  $T$ , which equals one for those who attended an elite college and

## 220 PART I ♦ The Linear Regression Model

zero otherwise, appears as

$$\ln y = \mathbf{x}'\boldsymbol{\beta} + \delta T + \varepsilon.$$

Least squares regression of a measure of earnings,  $\ln y$ , on  $\mathbf{x}$  and  $T$  attempts to produce an estimate of  $\delta$ , the impact of the treatment. It seems inevitable, however, that some unobserved determinants of lifetime earnings, such as ambition, inherent abilities, persistence, and so on would also determine whether the individual had an opportunity to attend an elite college. If so, then the least squares estimate of  $\delta$  will inappropriately attribute the effect to the treatment, rather than to these underlying factors. Least squares will not consistently estimate  $\delta$ , ultimately because of the correlation between  $T$  and  $\varepsilon$ .

In order to quantify definitively the impact of attendance at an elite college on the individuals who did so, the researcher would have to conduct an impossible experiment. Individuals in the sample would have to be observed twice, once having attended the elite college and a second time (in a second lifetime) without having done so. Whether comparing individuals who attended elite colleges to other individuals who did not adequately measures the **effect of the treatment on the treated** individuals is the subject of a vast current literature. See, for example, Imbens and Wooldridge (2009) for a survey.

**Simultaneous Equations:** In an equilibrium model of price and output determination in a market, there would be equations for both supply and demand. For example, a model of output and price determination in a product market might appear

$$(Demand) \quad \text{Quantity}_D = \alpha_0 + \alpha_1 \text{Price} + \alpha_2 \text{Income} + \varepsilon_D,$$

$$(Supply) \quad \text{Quantity}_S = \beta_0 + \beta_1 \text{Price} + \beta_2 \text{InputPrice} + \varepsilon_S,$$

$$(Equilibrium) \quad \text{Quantity}_D = \text{Quantity}_S.$$

Consider attempting to estimate the parameters of the demand equation by regression of a time series of equilibrium quantities on equilibrium prices and incomes. The equilibrium price is determined by the equation of the two quantities. By imposing the equilibrium condition, we can solve for  $\text{Price} = (\alpha_0 - \beta_0 + \alpha_2 \text{Income} - \beta_2 \text{InputPrice} + \varepsilon_D - \varepsilon_S)/(\beta_1 - \alpha_1)$ . The implication is that Price is correlated with  $\varepsilon_D$ —if an external shock causes  $\varepsilon_D$  to change, that induces a shift in the demand curve and ultimately causes a new equilibrium price. Least squares regression of quantity on price and income does not estimate the parameters of the demand equation consistently. This “feedback” between  $\varepsilon_D$  and Price in this model produces **simultaneous equations bias** in the least squares estimator.

**Dynamic Panel Data Models:** In Chapter 11, we will examine a **random effects** dynamic model of the form  $y_{it} = x_{it}\boldsymbol{\beta} + \gamma y_{i,t-1} + \varepsilon_{it} + u_i$ , where  $u_i$  contains the time-invariant unobserved features of individual  $i$ . Clearly, in this case, the regressor  $y_{i,t-1}$  is correlated with the disturbance,  $(\varepsilon_{it} + u_i)$ —the unobserved heterogeneity is present in  $y_{it}$  in every period. In Chapter 13, we will examine a model for municipal expenditure of the form  $S_{it} = f(S_{i,t-1}, \dots) + \varepsilon_{it}$ . The disturbances are assumed to be freely correlated across periods, so both  $S_{i,t-1}$  and  $\varepsilon_{it}$  are correlated with  $\varepsilon_{i,t-1}$ . It follows that they are correlated with each other, which means that this model, even without time persistent effects, does not satisfy the assumptions of the linear regression model. The regressors and disturbances are correlated.

**Omitted Parameter Heterogeneity:** Many cross-country studies of economic growth have the following structure (greatly simplified for purposes of this example),

$$\Delta \ln Y_{it} = \alpha_i + \theta_i t + \beta_i \Delta \ln Y_{i,t-1} + \varepsilon_{it},$$

where  $\Delta \ln Y_{it}$  is the growth rate of country  $i$  in year  $t$ . [See, for example, Lee, Pesaran and Smith (1997).] Note that the coefficients in the model are country specific. What does least squares regression of growth rates of income on a time trend and lagged growth rates estimate? Rewrite the growth equation as

$$\begin{aligned} \Delta \ln Y_{it} &= \alpha + \theta t + \beta \Delta \ln Y_{i,t-1} + (\alpha_i - \alpha) + (\theta_i - \theta)t + (\beta_i - \beta)\Delta \ln Y_{i,t-1} + \varepsilon_{it} \\ &= \alpha + \theta t + \beta \Delta \ln Y_{i,t-1} + w_{it}. \end{aligned}$$

## CHAPTER 8 ♦ Endogeneity and Instrumental Variable 221

We assume that the “average” parameters,  $\alpha$ ,  $\theta$ , and  $\beta$ , are meaningful fixed parameters to be estimated. Does the least squares regression of  $\Delta \ln Y_{it}$  on a constant,  $t$ , and  $\Delta \ln Y_{i,t-1}$  estimate these parameters consistently? We might assume that the cross-country variation in the constant terms is purely random, and the time trends,  $\theta_i$ , are driven by purely exogenous factors. But, the differences across countries of the convergence parameters,  $\beta_i$ , are likely at least to be correlated with the growth in incomes in those countries, which will induce a correlation between the lagged income growth and the term  $(\beta_i - \beta)$  embedded in  $w_{it}$ . If  $(\beta_i - \beta)$  is random noise that is uncorrelated with  $\Delta \ln Y_{i,t-1}$ , then  $(\beta_i - \beta) \Delta \ln Y_{i,t-1}$  will be also.

**Measurement Error:** Ashenfelter and Krueger (1994), Ashenfelter and Zimmerman (1997) and Bonjour et al. (2003) examined applications in which an earnings equation

$$y_{i,t} = f(Education_{i,t}, \dots) + \varepsilon_{i,t}$$

is specified for sibling pairs (twins)  $t = 1, 2$  for  $n$  families. Education is a variable that is inherently unmeasurable; years of schooling is typically the best **proxy variable** available. Consider, in a very simple model, attempting to estimate the parameters of

$$y_{it} = \beta_1 + \beta_2 Education_{it} + \varepsilon_{it},$$

by a regression of *Earnings<sub>it</sub>* on a constant and *Schooling<sub>it</sub>* with

$$Schooling_{it} = Education_{it} + u_{it},$$

where  $u_{it}$  is the measurement error. By a simple substitution, we find

$$y_{it} = \beta_1 + \beta_2 Schooling_{it} + w_{it},$$

where  $w_{it} = \varepsilon_{it} - \beta_2 u_{it}$ . *Schooling* is clearly correlated with  $w_{it} = (\varepsilon_{it} - \beta_2 u_{it})$ . The interpretation is that at least some of the variation in *Schooling* is due to variation in the measurement error,  $u_{it}$ . Since *Schooling* is correlated with  $w_{it}$ , it is endogenous, and least squares is not a suitable estimator of the earnings equation. As we will show later, in cases such as this one, the mismeasurement of a relevant variable causes a particular form of inconsistency, **attenuation bias**, in the estimator of  $\beta_2$ .

**Nonrandom Sampling:** In a model of the effect of a training program, an employment program, or the labor supply behavior of a particular segment of the labor force, the sample of observations may have voluntarily selected themselves into the observed sample. The Job Training Partnership Act (JTPA) was a job training program intended to provide employment assistance to disadvantaged youth. Anderson et al. (1991) found that for a sample that they examined, the program appeared to be administered most often to the best qualified applicants. In an earnings equation estimated for such a nonrandom sample, the implication is that the disturbances are not truly random. For the application just described, for example, on average, the disturbances are unusually high compared to the full population. Merely unusually high would not be a problem save for the general finding that the explanation for the nonrandomness is found at least in part in the variables that appear elsewhere in the model. This nonrandomness of the sample of the  sample translates to a form of omitted variable bias known as **sample selection bias**.

**Attrition:** We can observe two closely related important cases of nonrandom sampling. In panel data studies of firm performance, the firms still in the sample at the end of the observation period are likely to be a subset of those present at the beginning—those firms that perform badly, “fail” or drop out of the sample. Those that remain are unusual in the same fashion as the previous sample of JTPA participants. In these cases, least squares regression of the performance variable on the covariates (whatever they are), suffers from a form of selection bias known as **survivorship bias**. In this case, the distribution of outcomes, firm performances, for the survivors is systematically higher than that for the population of firms as a whole. This produces a phenomenon known as **truncation bias**. In clinical trials and other statistical analysis of health interventions, subjects often drop out of the study for reasons related to the intervention, itself—for a quality of life intervention such as a drug treatment for cancer, subjects may leave because they recover and feel uninterested in returning for the exit interview, or they may pass away or become incapacitated and be

**222 PART I ♦ The Linear Regression Model**

unable to return. In either case, the statistical analysis is subject to **attrition bias**. The same phenomenon may impact the analysis of panel data in health econometrics studies. For example, Contoyannis, Jones, and Rice (2004) examined self-assessed health outcomes in a long panel data set extracted from the British Household Panel Data survey. In each year of the study, a significant number of the observations were absent from the next year's data set, with the result that the sample was winnowed significantly from the beginning to the end of the study.

In all the cases listed in Example 8.1, the term “bias” refers to the result that least squares (or other conventional modifications of least squares) is an inconsistent (persistently biased) estimator of the coefficients of the model of interest. Though the source of the result differs considerably from setting to setting, all ultimately trace back to endogeneity of some or all of the right-hand-side variables and this, in turn, translates to correlation between the regressors and the disturbances. These can be broadly viewed in terms of some specific effects:

- Omitted variables, either observed or unobserved
- Feedback effects
- Dynamic effects
- Endogenous sample design



and so on. There are two general solutions to the problem of constructing a consistent estimator. In some cases, a more detailed, “**structural**” specification of the model can be developed. These usually involve specifying additional equations that explain the correlation between  $\mathbf{x}_i$  and  $\varepsilon_i$  in a way that enables estimation of the full set of parameters of interest. We will develop a few of these models in later chapters, including, for example, Chapter 16, where we consider Heckman’s (1979) model of sample selection. The second approach, which is becoming increasingly common in contemporary research, is the method of **instrumental variables**. The method of instrumental variables is developed around the following estimation strategy: Suppose that in the model of (8-1), the  $K$  variables  $\mathbf{x}_i$  may be correlated with  $\varepsilon_i$ . Suppose as well that there exists a set of  $L$  variables  $\mathbf{z}_i$ , such that  $\mathbf{z}_i$  is correlated with  $\mathbf{x}_i$ , but not with  $\varepsilon_i$ . We cannot estimate  $\beta$  consistently by using the familiar least squares estimator. But, the assumed lack of correlation between  $\mathbf{z}_i$  and  $\varepsilon_i$  implies a set of relationships that may allow us construct a consistent estimator of  $\beta$  by using the assumed relationships among  $\mathbf{z}_i$ ,  $\mathbf{x}_i$ , and  $\varepsilon_i$ .

This chapter will develop the method of instrumental variables as an extension of the models and estimators that have been considered in Chapters 2–7. Section 8.2 will formalize the model in a way that provides an estimation framework. The method of instrumental variables (IV) estimation and two-stage least squares (2SLS) is developed in detail in Section 8.3. Two tests of the model specification are considered in Section 8.4. A particular application of the estimation, measurement error, is developed in detail in Section 8.5. Section 8.6 will consider nonlinear models and begin the development of the generalized method of moments (GMM) estimator. The IV estimator is a powerful tool that underlies a great deal of contemporary empirical research. A shortcoming, the problem of weak instruments is considered in Section 8.7. Finally, some observations about instrumental variables and the search for causal effects are presented in Section 8.8.

This chapter will develop the fundamental results for IV estimation. The use of instrumental variables will appear in many applications in the chapters to follow,

## CHAPTER 8 ♦ Endogeneity and Instrumental Variable 223

including multiple equations models in Chapter 10, the panel data methods in Chapter 11, and in the development of the generalized method of moments in Chapter 13.

## 8.2 ASSUMPTIONS OF THE EXTENDED MODEL

The assumptions of the linear regression model, laid out in Chapters 2 and 4 are

- A.1. Linearity:**  $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$ .
- A.2. Full rank:** The  $n \times K$  sample data matrix,  $\mathbf{X}$  has full column rank 
- A.3. Exogeneity of the independent variables:**  $E[\varepsilon_i | x_{j1}, x_{j2}, \dots, x_{jk}] = 0, i, j = 1, \dots, n$ . There is no correlation between the disturbances and the independent variables.
- A.4. Homoscedasticity and nonautocorrelation:** Each disturbance,  $\varepsilon_i$ , has the same finite variance,  $\sigma^2$  and is uncorrelated with every other disturbance,  $\varepsilon_j$ , conditioned on  $\mathbf{X}$ .
- A.5. Stochastic or nonstochastic data:**  $(x_{i1}, x_{i2}, \dots, x_{iK}) i = 1, \dots, n$ .
- A.6. Normal distribution:** The disturbances are normally distributed.

We will maintain the important result that  $\text{plim}(\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}_{xx}$ . The basic assumptions of the regression model have changed, however. First, A.3 (no correlation between  $\mathbf{x}$  and  $\varepsilon$ ), under our new assumptions,

$$\mathbf{A.I3.} \quad E[\varepsilon_i | \mathbf{x}_i] = \eta_i.$$

We interpret Assumption A.I3 to mean that the regressors now provide information about the expectations of the disturbances. The important implication of A.I3 is that the disturbances and the regressors are now correlated. Assumption A.I3 implies that

$$E[\mathbf{x}_i \varepsilon_i] = \boldsymbol{\gamma} \tag{8-2}$$

for some nonzero  $\boldsymbol{\gamma}$ . If the data are “well behaved,” then we can apply Theorem D.5 (Khinchine’s theorem) to assert that

$$\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \boldsymbol{\gamma}. \tag{8-3}$$

Notice that the original model results if  $\eta_i = 0$ . The implication of (8-3) is that the regressors,  $\mathbf{X}$ , are no longer exogenous.

We now assume that there is an additional set of variables,  $\mathbf{Z}$ , that have two properties:

- 1. Exogeneity:** They are uncorrelated with the disturbance.
- 2. Relevance:** They are correlated with the independent variables,  $\mathbf{X}$ .

We will formalize these notions as we proceed. In the context of our model, variables that have these two properties are instrumental variables. We assume the following:

- A.I7.**  $[\mathbf{x}_i, \mathbf{z}_i, \varepsilon_i], i = 1, \dots, n$ , are an i.i.d. sequence of random variables.
- A.I8a.**  $E[x_{ik}^2] = \mathbf{Q}_{xx,kk} < \infty$ , a finite constant,  $k = 1, \dots, K$ .
- A.I8b.**  $E[z_{il}^2] = \mathbf{Q}_{zz,ll} < \infty$ , a finite constant,  $l = 1, \dots, L$ .
- A.I8c.**  $E[z_{il}x_{ik}] = \mathbf{Q}_{zx,lk} < \infty$ , a finite constant,  $l = 1, \dots, L, k = 1, \dots, K$ .
- A.I9.**  $E[\varepsilon_i | \mathbf{z}_i] = 0$ .

## 224 PART I ♦ The Linear Regression Model

In later work in time series models, it will be important to relax assumption A.I7. Finite means of  $\mathbf{z}_l$  follows from A.I8b. Using the same analysis as in Section 4.4, we have

$$\text{plim}(1/n)\mathbf{Z}'\mathbf{Z} = \mathbf{Q}_{zz}, \text{ a finite, positive definite matrix (well-behaved data),}$$

$$\text{plim}(1/n)\mathbf{Z}'\mathbf{X} = \mathbf{Q}_{zx}, \text{ a finite, } L \times K \text{ matrix with rank } K \text{ (relevance),}$$

$$\text{plim}(1/n)\mathbf{Z}'\boldsymbol{\epsilon} = \mathbf{0} \text{ (exogeneity).}$$

In our statement of the classical regression model, we have assumed thus far the special case of  $\eta_i = 0$ ;  $\boldsymbol{\gamma} = \mathbf{0}$  follows. There is no need to dispense with Assumption A.I7—it may well continue to be true—but in this special case, it becomes irrelevant.

For the present, we will assume that  $L = K$ —there are the same number of instrumental variables as there are right-hand-side variables in the equation. Recall in the introduction and in Example 8.1, we partitioned  $\mathbf{x}$  into  $\mathbf{x}_1$  a set of  $K_1$  exogenous variables and  $\mathbf{x}_2$ , a set of  $K_2$  endogenous variables on the right hand side of (8-1). In nearly all cases in practice, the “problem of endogeneity” is attributable to one or a small number of variables in  $\mathbf{x}$ . In the Kreuger and Dale (1999) study of endogenous treatment effects in Example 8.1, we have a single endogenous variable in the equation, the treatment dummy variable,  $T$ . The implication for our formulation here is that in such a case, the  $K_1$  variables  $\mathbf{x}_1$  will be among the instrumental variables in  $\mathbf{Z}$  and the  $K_2$  remaining variables will be other exogenous variables that are not the same as  $\mathbf{x}_2$ . The usual interpretation will be that these  $K_2$  variables,  $\mathbf{z}_2$ , are the “instruments for  $\mathbf{x}_2$ ” while the  $\mathbf{x}_1$  variables are instruments for themselves. To continue the example, the matrix  $\mathbf{Z}$  for the endogenous treatment effects model would contain the  $K_1$  columns of  $\mathbf{X}$  and an additional instrumental variable,  $\mathbf{z}$ , for the treatment dummy variable. In the simultaneous equations model of supply and demand, the endogenous right-hand-side variable is the  $x_2 = \text{price}$  while the exogenous variables are  $(1, \text{Income})$ . One might suspect (correctly), that in this model, a set of instrumental variables would be  $\mathbf{z} = (1, \text{Income}, \text{InputPrice})$ . In terms of the underlying relationships among the variables, this intuitive understanding will provide a reliable guide. For reasons that will be clear shortly, however, it is necessary statistically to treat  $\mathbf{Z}$  as the instruments for  $\mathbf{X}$  in its entirety.

There is a second subtle point about the use of instrumental variables that will likewise be more evident below. The “relevance condition” must actually be a statement of conditional correlation. Consider, once again, the treatment effects example, and suppose that  $z$  is the instrumental variable in question for the treatment dummy variable  $T$ . The relevance condition as stated implies that the correlation between  $z$  and  $(\mathbf{x}, T)$  is nonzero. Formally, what will be required is that the conditional correlation of  $z$  with  $T|\mathbf{x}$  be nonzero. One way to view this is in terms of a projection; the instrumental variable  $z$  is relevant if the coefficient on  $z$  in the regression of  $T$  on  $(\mathbf{x}, z)$  is nonzero. Intuitively,  $z$  must provide information about the movement of  $T$  that is not provided by the  $\mathbf{x}$  variables that are already in the model.

### 8.3 ESTIMATION

For the general model of Section 8.2, we lose most of the useful results we had for least squares. We will consider the implications for least squares and then construct an alternative estimator for  $\boldsymbol{\beta}$  in this extended model.

### 8.3.1 LEAST SQUARES

The least squares estimator,  $\mathbf{b}$ , is no longer unbiased;

$$E[\mathbf{b}|\mathbf{X}] = \boldsymbol{\beta} + \mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\boldsymbol{\eta} \neq \boldsymbol{\beta},$$

so the Gauss–Markov theorem no longer holds. The estimator is also inconsistent;

$$\text{plim } \mathbf{b} = \boldsymbol{\beta} + \text{plim} \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \text{plim} \left( \frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n} \right) = \boldsymbol{\beta} + \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\gamma} \neq \boldsymbol{\beta}. \quad (8-4)$$

(The asymptotic distribution is considered in the exercises). The inconsistency of least squares is not confined to the coefficients on the endogenous variables. To see this, apply (8-4) to the treatment effects example discussed earlier. In that case, all but the last variable in  $\mathbf{X}$  are uncorrelated with  $\boldsymbol{\varepsilon}$ . This means that

$$\text{plim} \left( \frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n} \right) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \gamma_K \end{pmatrix} = \gamma_K \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

It follows that for this special case, the result in (8-4) is

$$\text{plim } \mathbf{b} = \boldsymbol{\beta} + \gamma_K \times \text{the last column of } \mathbf{Q}_{\mathbf{XX}}^{-1}.$$

There is no reason to expect that any of the elements of the last column of  $\mathbf{Q}_{\mathbf{XX}}^{-1}$  will equal zero. The implication is that even though only one of the variables in  $\mathbf{X}$  is correlated with  $\boldsymbol{\varepsilon}$ , all of the elements of  $\mathbf{b}$  are inconsistent, not just the estimator of the coefficient on the endogenous variable. This effect is called **smearing**; the inconsistency due to the endogeneity of the one variable is smeared across all of the least squares estimators.

### 8.3.2 THE INSTRUMENTAL VARIABLES ESTIMATOR

Because  $E[\mathbf{z}_i \boldsymbol{\varepsilon}_i] = 0$  and all terms have finite variances, it follows that

$$\text{plim} \left( \frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n} \right) = \text{plim} \left( \frac{\mathbf{Z}'\mathbf{y}}{n} \right) - \text{plim} \left( \frac{\mathbf{Z}'\mathbf{X}\boldsymbol{\beta}}{n} \right) = \mathbf{0}.$$

Therefore,

$$\text{plim} \left( \frac{\mathbf{Z}'\mathbf{y}}{n} \right) = \left[ \text{plim} \left( \frac{\mathbf{Z}'\mathbf{X}}{n} \right) \right] \boldsymbol{\beta} + \text{plim} \left( \frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n} \right) = \left[ \text{plim} \left( \frac{\mathbf{Z}'\mathbf{X}}{n} \right) \right] \boldsymbol{\beta}. \quad (8-5)$$

We have assumed that  $\mathbf{Z}$  has the same number of variables as  $\mathbf{X}$ . For example, suppose in our consumption function that  $\mathbf{x}_t = [1, Y_t]$  when  $\mathbf{z}_t = [1, Y_{t-1}]$ . We have assumed that the rank of  $\mathbf{Z}'\mathbf{X}$  is  $K$ , so now  $\mathbf{Z}'\mathbf{X}$  is a square matrix. It follows that

$$\left[ \text{plim} \left( \frac{\mathbf{Z}'\mathbf{X}}{n} \right) \right]^{-1} \text{plim} \left( \frac{\mathbf{Z}'\mathbf{y}}{n} \right) = \boldsymbol{\beta}, \quad (8-6)$$

which leads us to the **instrumental variable estimator**,

$$\mathbf{b}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}.$$

## 226 PART I ♦ The Linear Regression Model

We have already proved that  $\mathbf{b}_{IV}$  is consistent. We now turn to the **asymptotic distribution**. We will use the same method as in Section 4.4.2. First,

$$\sqrt{n}(\mathbf{b}_{IV} - \boldsymbol{\beta}) = \left( \frac{\mathbf{Z}'\mathbf{X}}{n} \right)^{-1} \frac{1}{\sqrt{n}} \mathbf{Z}'\boldsymbol{\varepsilon},$$

which has the same **limiting distribution** as  $\mathbf{Q}_{zz}^{-1}[(1/\sqrt{n})\mathbf{Z}'\boldsymbol{\varepsilon}]$ . Our analysis of  $(1/\sqrt{n})\mathbf{Z}'\boldsymbol{\varepsilon}$  can be the same as that of  $(1/\sqrt{n})\mathbf{X}'\boldsymbol{\varepsilon}$  in Section 4.4.2, so it follows that

$$\left( \frac{1}{\sqrt{n}} \mathbf{Z}'\boldsymbol{\varepsilon} \right) \xrightarrow{d} N[\mathbf{0}, \sigma^2 \mathbf{Q}_{zz}],$$

and

$$\left( \frac{\mathbf{Z}'\mathbf{X}}{n} \right)^{-1} \left( \frac{1}{\sqrt{n}} \mathbf{Z}'\boldsymbol{\varepsilon} \right) \xrightarrow{d} N[\mathbf{0}, \sigma^2 \mathbf{Q}_{zx}^{-1} \mathbf{Q}_{zz} \mathbf{Q}_{zx}^{-1}].$$

This step completes the derivation for the next theorem.

### THEOREM 8.1 Asymptotic Distribution of the Instrumental Variables Estimator

If Assumptions A.1, A.2, A.I3, A.4, A.5, A.I7, A.I8a–c, and A.I9 all hold for  $[y_i, \mathbf{x}_i, \mathbf{z}_i, \varepsilon_i]$ , where  $\mathbf{z}$  is a valid set of  $L = K$  instrumental variables, then the asymptotic distribution of the instrumental variables estimator  $\mathbf{b}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$  is

$$\mathbf{b}_{IV} \xrightarrow{a} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n} \mathbf{Q}_{zx}^{-1} \mathbf{Q}_{zz} \mathbf{Q}_{zx}^{-1}\right]. \quad (8-7)$$

where  $\mathbf{Q}_{zx} = \text{plim}(\mathbf{Z}'\mathbf{X}/n)$  and  $\mathbf{Q}_{zz} = \text{plim}(\mathbf{Z}'\mathbf{Z}/n)$ .

To estimate the **asymptotic covariance matrix**, we will require an estimator of  $\sigma^2$ . The natural estimator is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}'_i \mathbf{b}_{IV})^2.$$

A correction for degrees of freedom is superfluous, as all results here are asymptotic, and  $\hat{\sigma}^2$  would not be unbiased in any event. (Nonetheless, it is standard practice in most software to make the degrees of freedom correction.) Write the vector of residuals as

$$\mathbf{y} - \mathbf{X}\mathbf{b}_{IV} = \mathbf{y} - \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}.$$

Substitute  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$  and collect terms to obtain  $\hat{\boldsymbol{\varepsilon}} = [\mathbf{I} - \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}']\boldsymbol{\varepsilon}$ . Now,

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}$$

$$= \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n} + \left( \frac{\boldsymbol{\varepsilon}'\mathbf{Z}}{n} \right) \left( \frac{\mathbf{X}'\mathbf{Z}}{n} \right)^{-1} \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right) \left( \frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n} \right)^{-1} \left( \frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n} \right) - 2 \left( \frac{\boldsymbol{\varepsilon}'\mathbf{X}}{n} \right) \left( \frac{\mathbf{Z}'\mathbf{X}}{n} \right)^{-1} \left( \frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n} \right).$$

## CHAPTER 8 ♦ Endogeneity and Instrumental Variable 227

We found earlier that we could (after a bit of manipulation) apply the product result for probability limits to obtain the probability limit of an expression such as this. Without repeating the derivation, we find that  $\hat{\sigma}^2$  is a **consistent estimator** of  $\sigma^2$ , by virtue of the first term. The second and third product terms converge to zero. To complete the derivation, then, we will estimate  $\text{Asy. Var}[\mathbf{b}_{\text{IV}}]$  with

$$\begin{aligned}\text{Est. Asy. Var}[\mathbf{b}_{\text{IV}}] &= \frac{1}{n} \left\{ \left( \frac{\hat{\boldsymbol{\epsilon}}' \hat{\boldsymbol{\epsilon}}}{n} \right) \left( \frac{\mathbf{Z}' \mathbf{X}}{n} \right)^{-1} \left( \frac{\mathbf{Z}' \mathbf{Z}}{n} \right) \left( \frac{\mathbf{X}' \mathbf{Z}}{n} \right)^{-1} \right\} \\ &= \hat{\sigma}^2 (\mathbf{Z}' \mathbf{X})^{-1} (\mathbf{Z}' \mathbf{Z}) (\mathbf{X}' \mathbf{Z})^{-1}.\end{aligned}\quad (8-8)$$

### 8.3.3 MOTIVATING THE INSTRUMENTAL VARIABLES ESTIMATOR

In obtaining the IV estimator, we relied on the solutions to the equations in (8-5),

$$\text{plim}(\mathbf{Z}' \mathbf{y}) = \text{plim}(\mathbf{Z}' \mathbf{X}/n) \boldsymbol{\beta}$$

or

$$\mathbf{Q}_{\mathbf{Zy}} = \mathbf{Q}_{\mathbf{ZX}} \boldsymbol{\beta}.$$

The IV estimator is obtained by solving this set of  $K$  **moment equations**. Since this is a set of  $K$  equations in  $K$  unknowns, if  $\mathbf{Q}_{\mathbf{ZX}}^{-1}$  exists, then there is an exact solution for  $\boldsymbol{\beta}$ , given in (8-6). The corresponding moment equations if only  $\mathbf{X}$  is used would be

$$\text{plim}(\mathbf{X}' \mathbf{y}/n) = \text{plim}(\mathbf{X}' \mathbf{X}/n) \boldsymbol{\beta} + \text{plim}(\mathbf{X}' \boldsymbol{\epsilon}/n) = \text{plim}(\mathbf{X}' \mathbf{X}/n) \boldsymbol{\beta} + \boldsymbol{\gamma}$$

or

$$\mathbf{Q}_{\mathbf{xy}} = \mathbf{Q}_{\mathbf{xx}} \boldsymbol{\beta} + \boldsymbol{\gamma},$$

which is, without further restrictions,  $K$  equations in  $2K$  unknowns. There are insufficient equations to solve this system for either  $\boldsymbol{\beta}$  or  $\boldsymbol{\gamma}$ . The further restrictions that would allow estimation of  $\boldsymbol{\beta}$  would be  $\boldsymbol{\gamma} = \mathbf{0}$ ; this is precisely the exogeneity assumption A.3. The implication is that the parameter vector  not **identified** in terms of the moments of  $\mathbf{X}$  and  $\mathbf{y}$  alone—there does not exist a solution. But, it is identified in terms of the moments of  $\mathbf{Z}$ ,  $\mathbf{X}$  and  $\mathbf{y}$ , plus the  $K$  restrictions imposed by the exogeneity assumption, and the relevance assumption that allows computation of  $\mathbf{b}_{\text{IV}}$ .

Consider these results in the context of a simplified model 

$$y = \beta x + \delta T + \varepsilon.$$

In order for least squares consistently to estimate  $\delta$  (and  $\beta$ ), it is assumed that movements in  $T$  are exogenous to the model, so that covariation of  $y$  and  $T$  is explainable by the movement of  $T$  and not by the movement of  $\varepsilon$ . When  $T$  and  $\varepsilon$  are correlated and  $\varepsilon$  varies through some factor not in the equation, the movement of  $y$  will appear to be induced by variation in  $T$  when it is actually induced by variation in  $\varepsilon$  which is transmitted through  $T$ . If  $T$  is exogenous, that is, not correlated with  $\varepsilon$ , then movements in  $\varepsilon$  will not “cause” movements in  $T$  (we use the term “cause” very loosely here) and will thus not be mistaken for exogenous variation in  $T$ . The exogeneity assumption plays precisely this role. To summarize, then, in order for a regression model to identify  $\delta$  correctly, it must be assumed that variation in  $T$  is not associated with variation in  $\varepsilon$ . If it is, then as seen in (8-4), variation in  $y$  comes about through an additional source, variation in

## 228 PART I ♦ The Linear Regression Model

$\varepsilon$  that is transmitted through variation in  $T$ . That is the influence of  $y$  in (8-4). What is needed, then, to identify  $\delta$  is movement in  $T$  that is definitely not induced by movement in  $\varepsilon$ . Enter the instrumental variable,  $z$ . If  $z$  is an instrumental variable with  $\text{cov}(z, T) \neq 0$  and  $\text{cov}(z, \varepsilon) = 0$ , then movement in  $z$  provides the variation that we need. If we can consider doing this exercise experimentally, in order to measure the “causal effect” of movement in  $T$ , we would change  $z$  and then measure the per unit change in  $y$  associated with the change in  $T$ , knowing that the change in  $T$  was induced only by the change in  $z$ , not  $\varepsilon$ , that is,  $(\Delta y / \Delta z) / (\Delta T / \Delta z)$ .

### Example 8.2 Instrumental Variable Analysis

Grootendorst (2007) and Deaton (1997) recount what appears to be the earliest application of the method of instrumental variables:

Although IV theory has been developed primarily by economists, the method originated in epidemiology. IV was used to investigate the route of cholera transmission during the London cholera epidemic of 1853–54. A scientist from that era, John Snow, hypothesized that cholera was waterborne. To test this, he could have tested whether those who drank purer water had lower risk of contracting cholera. In other words, he could have assessed the correlation between water purity ( $x$ ) and cholera incidence ( $y$ ). Yet, as Deaton (1997) notes, this would not have been convincing: “The people who drank impure water were also more likely to be poor, and to live in an environment contaminated in many ways, not least by the ‘poison miasmas’ that were then thought to be the cause of cholera.” Snow instead identified an instrument that was strongly correlated with water purity yet uncorrelated with other determinants of cholera incidence, both observed and unobserved. This instrument was the identity of the company supplying households with drinking water. At the time, Londoners received drinking water directly from the Thames River. One company, the Lambeth Water Company, drew water at a point in the Thames above the main sewage discharge; another, the Southwark and Vauxhall Company, took water below the discharge. Hence the instrument  $z$  was strongly correlated with water purity  $x$ . The instrument was also uncorrelated with the unobserved determinants of cholera incidence ( $y$ ). According to Snow (1844, pp. 74–75), the households served by the two companies were quite similar; indeed: “the mixing of the supply is of the most intimate kind. The pipes of each Company go down all the streets, and into nearly all the courts and alleys. . . . The experiment, too, is on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice, and in most cases, without their knowledge; one group supplied with water containing the sewage of London, and amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity.”

### Example 8.3 Streams as Instruments

In Hoxby (2000), the author was interested in the effect of the amount of school “choice” in a school “market” on educational achievement in the market. The equations of interest were of the form

$$\frac{A_{ikm}}{\ln E_{km}} = \beta_1 C_m + \mathbf{x}'_{ikm} \beta_2 + \bar{\mathbf{x}}'_{km} \beta_3 + \bar{\mathbf{x}}'_{.m} \beta_4 + \varepsilon_{ikm} + \varepsilon_{km} + \varepsilon_m$$

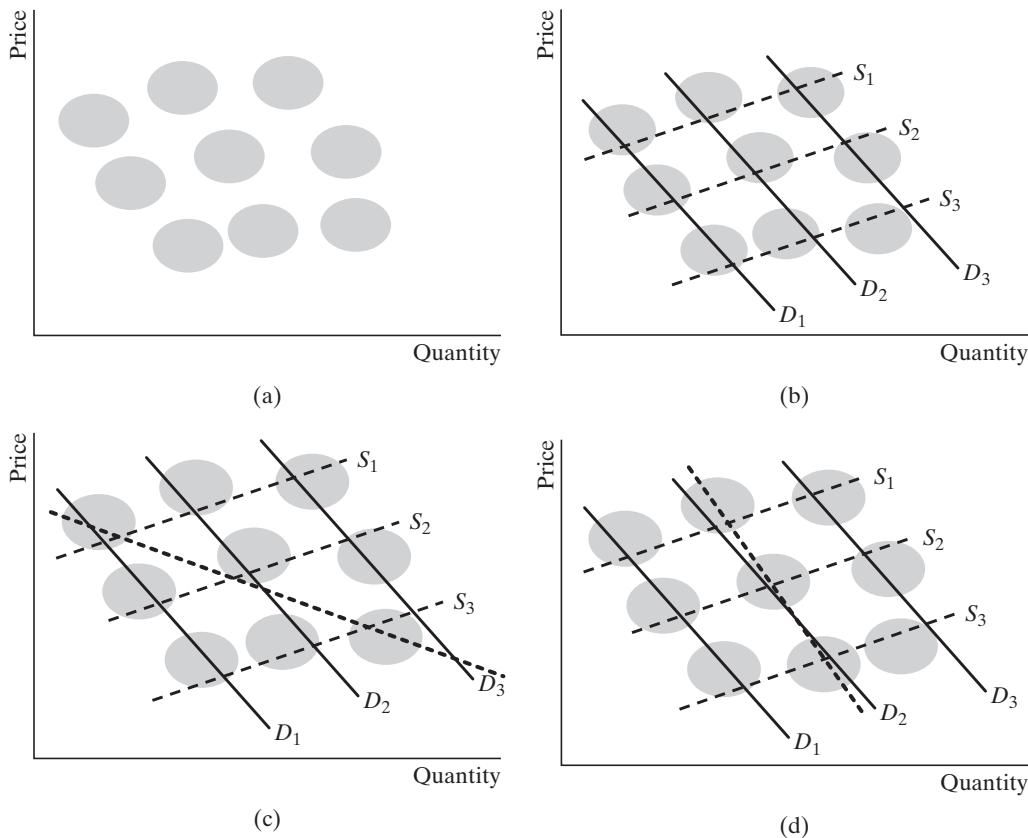
where “ $ikm$ ” denotes household  $i$  in district  $k$  in market  $m$ ,  $A_{ikm}$  is a measure of achievement and  $E_{ikm}$  is per capita expenditures. The equation contains individual level data, district means, and market means. The exogenous variables are intended to capture the different sources of heterogeneity at all three levels of aggregation. (The compound disturbance, which we will revisit when we examine panel data specifications in Chapter 10, is intended to allow for random effects at all three levels as well.) Reasoning that the amount of choice available to students,  $C_m$ , would be endogenous in this equation, the author sought a valid instrumental variable that would “explain” (be correlated with)  $C_m$  but uncorrelated with the disturbances in the equation. In the U.S. market, to a large degree, school district boundaries were set

## CHAPTER 8 ♦ Endogeneity and Instrumental Variable 229

in the late 18th and through the 19th centuries and handed down to present-day administrators by historical precedent. In the formative years, the author noted, district boundaries were set in response to natural travel barriers, such as rivers and streams. It follows, as she notes, that “the number of districts in a given land area is an increasing function of the number of natural barriers”; hence, the number of streams in the physical market area provides the needed instrumental variable. [The controversial topic of the study and the unconventional choice of instruments caught the attention of the popular press, for example, <http://gsppi.berkeley.edu/faculty/jrothstein/hoxby/wsj.pdf>, and academic observers including Rothstein (2004).] This study is an example of a “natural experiment” as described in Angrist and Pischke (2009).

### **Example 8.4 Instrumental Variable in Regression**

The role of an instrumental variable in identifying parameters in regression models was developed in Working's (1926) classic application, adapted here for our market equilibrium example in Example 8.1. Figure 8.1a displays the “observed data” for the market equilibria in a market in which there are random disturbances ( $\varepsilon_S$ ,  $\varepsilon_D$ ) and variation in demanders' incomes and input prices faced by suppliers. The market equilibria in Figure 8.1a are scattered about as the aggregates of all these effects. Figure 8.1b suggests the underlying conditions of supply and demand that give rise to these equilibria. Different outcomes in the supply equation



**FIGURE 8.1** Identifying a Demand Curve with an Instrumental Variable.

## 230 PART I ♦ The Linear Regression Model

corresponding to different values of the input price and different income values on the demand side produce nine regimes, punctuated by the random variation induced by the disturbances. Given the ambiguous mass of points, linear regression of quantity on price (and income) is likely to produce a result such as that shown by the heavy dotted line in Figure 8.1c. The slope of this regression barely resembles the slope of the demand equations. Faced with this prospect, how is it possible to learn about the slope of the demand curve? The experiment needed, shown in Figure 8.1d, would involve two elements: (1) Hold Income constant, so we can focus on the demand curve in a particular demand setting. That is the function of multiple regression—Income is included as a conditioning variable in the equation. (2) Now that we have focused on a particular set of demand outcomes, move the supply curve so that the equilibria now trace out the demand function. That is the function of the changing *InputPrice*, which is the instrumental variable that we need for identification of the demand function(s) for this experiment.

### 8.3.4 TWO-STAGE LEAST SQUARES

Thus far, we have assumed that the number of instrumental variables in  $\mathbf{Z}$  is the same as the number of variables (exogenous plus endogenous) in  $\mathbf{X}$ . (In the typical application, the researcher provides the necessary instrumental variable for the single endogenous variable in their equation.) However, it is possible that the data contain additional instruments. Recall the market equilibrium application considered in Examples 8.1 and 8.4. Suppose this were an agricultural market in which there are two exogenous conditions of supply, *InputPrice* and *Rainfall*. Then, the equations of the model are

$$(Demand) \quad Quantity_D = \alpha_0 + \alpha_1 Price + \alpha_2 Income + \varepsilon_D,$$

$$(Supply) \quad Quantity_S = \beta_0 + \beta_1 Price + \beta_2 InputPrice + \beta_3 Rainfall + \varepsilon_S,$$

$$(Equilibrium) \quad Quantity_D = Quantity_S.$$

Given the approach taken in Example 8.4, it would appear that the researcher could simply choose either of the two exogenous variables (instruments) in the supply equation for purpose of identifying the demand equation. (We will turn to the now apparent problem of how to identify the supply equation in Section 8.4.2.) Intuition should suggest that simply choosing a subset of the available instrumental variables would waste sample information—it seems inevitable that it will be preferable to use the full matrix  $\mathbf{Z}$ , even when  $L > K$ . The method of two-stage least squares solves the problem of how to use all the information in the sample when  $\mathbf{Z}$  contains more variables than are necessary to construct an instrumental variable estimator.

If  $\mathbf{Z}$  contains more variables than  $\mathbf{X}$ , then much of the preceding derivation is unusable, because  $\mathbf{Z}'\mathbf{X}$  will be  $L \times K$  with rank  $K < L$  and will thus not have an inverse. The crucial result in all the preceding is  $\text{plim}(\mathbf{Z}'\boldsymbol{\varepsilon}/n) = \mathbf{0}$ . That is, every column of  $\mathbf{Z}$  is asymptotically uncorrelated with  $\boldsymbol{\varepsilon}$ . That also means that every linear combination of the columns of  $\mathbf{Z}$  is also uncorrelated with  $\boldsymbol{\varepsilon}$ , which suggests that one approach would be to choose  $K$  linear combinations of the columns of  $\mathbf{Z}$ . Which to choose? One obvious possibility, discarded in the preceding paragraph, is simply to choose  $K$  variables among the  $L$  in  $\mathbf{Z}$ . Discarding the information contained in the “extra”  $L-K$  columns will turn out to be inefficient. A better choice is the projection of the columns of  $\mathbf{X}$  in the column space of  $\mathbf{Z}$ :

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}.$$

CHAPTER 8 ♦ Endogeneity and Instrumental Variable **231**

We will return shortly to the virtues of this choice. With this choice of instrumental variables,  $\hat{\mathbf{X}}$  for  $\mathbf{Z}$ , we have

$$\mathbf{b}_{IV} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}. \quad (8-9)$$

The estimator of the asymptotic covariance matrix will be  $\hat{\sigma}^2$  times the bracketed matrix in (8-9). The proofs of consistency and asymptotic normality for this estimator are exactly the same as before, because our proof was generic for any valid set of instruments, and  $\hat{\mathbf{X}}$  qualifies.

There are two reasons for using this estimator—one practical, one theoretical. If any column of  $\mathbf{X}$  also appears in  $\mathbf{Z}$ , then that column of  $\mathbf{X}$  is reproduced exactly in  $\hat{\mathbf{X}}$ . This is easy to show. In the expression for  $\hat{\mathbf{X}}$ , if the  $k$ th column in  $\mathbf{X}$  is one of the columns in  $\mathbf{Z}$ , say the  $l$ th, then the  $k$ th column in  $(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$  will be the  $l$ th column of an  $L \times L$  identity matrix. This result means that the  $k$ th column in  $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$  will be the  $l$ th column in  $\mathbf{Z}$ , which is the  $k$ th column in  $\mathbf{X}$ . This result is important and useful. Consider what is probably the typical application. Suppose that the regression contains  $K$  variables, only one of which, say, the  $k$ th, is correlated with the disturbances. We have one or more instrumental variables in hand, as well as the other  $K - 1$  variables that certainly qualify as instrumental variables in their own right. Then what we would use is  $\mathbf{Z} = [\mathbf{X}_{(k)}, \mathbf{z}_1, \mathbf{z}_2, \dots]$ , where we indicate omission of the  $k$ th variable by  $(k)$  in the subscript. Another useful interpretation of  $\hat{\mathbf{X}}$  is that each column is the set of fitted values when the corresponding column of  $\mathbf{X}$  is regressed on all the columns of  $\mathbf{Z}$ , which is obvious from the definition. It also makes clear why each  $\mathbf{x}_k$  that appears in  $\mathbf{Z}$  is perfectly replicated. Every  $\mathbf{x}_k$  provides a perfect predictor for itself, without any help from the remaining variables in  $\mathbf{Z}$ . In the example, then, every column of  $\mathbf{X}$  except the one that is omitted from  $\mathbf{X}_{(k)}$  is replicated exactly, whereas the one that is omitted is replaced in  $\hat{\mathbf{X}}$  by the predicted values in the regression of this variable on all the  $\mathbf{z}$ 's.

Of all the different linear combinations of  $\mathbf{Z}$  that we might choose,  $\hat{\mathbf{X}}$  is the most efficient in the sense that the asymptotic covariance matrix of an IV estimator based on a linear combination  $\mathbf{ZF}$  is smaller when  $\mathbf{F} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$  than with any other  $\mathbf{F}$  that uses all  $L$  columns of  $\mathbf{Z}$ ; a fortiori, this result eliminates linear combinations obtained by dropping any columns of  $\mathbf{Z}$ . This important result was proved in a seminal paper by Brundy and Jorgenson (1971). [See, also, Wooldridge (2002a, pp. 96–97).]

We close this section with some practical considerations in the use of the instrumental variables estimator. By just multiplying out the matrices in the expression, you can show that

$$\begin{aligned} \mathbf{b}_{IV} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\ &= (\mathbf{X}'(\mathbf{I} - \mathbf{M}_{\mathbf{Z}})\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{M}_{\mathbf{Z}})\mathbf{y} \\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} \end{aligned} \quad (8-10)$$

because  $\mathbf{I} - \mathbf{M}_{\mathbf{Z}}$  is idempotent. Thus, when (*and only when*)  $\hat{\mathbf{X}}$  is the set of instruments, the IV estimator is computed by least squares regression of  $\mathbf{y}$  on  $\hat{\mathbf{X}}$ . This conclusion suggests (only logically; one need not actually do this in two steps), that  $\mathbf{b}_{IV}$  can be computed in two steps, first by computing  $\hat{\mathbf{X}}$ , then by the least squares regression. For this reason, this is called the **two-stage least squares** (2SLS) estimator. We will revisit this form of estimator at great length at several points later, particularly in our discussion of simultaneous equations models in Section 10.5. One should be careful of this approach,

## 232 PART I ♦ The Linear Regression Model

however, in the computation of the asymptotic covariance matrix;  $\hat{\sigma}^2$  should not be based on  $\hat{\mathbf{X}}$ . The estimator

$$s_{IV}^2 = \frac{(\mathbf{y} - \hat{\mathbf{X}}\mathbf{b}_{IV})'(\mathbf{y} - \hat{\mathbf{X}}\mathbf{b}_{IV})}{n}$$

is inconsistent for  $\sigma^2$ , with or without a correction for degrees of freedom.

An obvious question is where one is likely to find a suitable set of instrumental variables. The recent literature on “natural experiments” focuses on policy changes such as the Mariel Boatlift (Example 6.5) or natural outcomes such as occurrences of streams (Example 8.3) or birthdays [Angrist (1992, 1994)]. In many time-series settings, lagged values of the variables in the model provide natural candidates. In other cases, the answer is less than obvious. The asymptotic covariance matrix of the IV estimator can be rather large if  $\mathbf{Z}$  is not highly correlated with  $\mathbf{X}$ ; the elements of  $(\mathbf{Z}'\mathbf{X})^{-1}$  grow large. (See Section 12.1 on “weak” instruments.) Unfortunately, there usually is not much choice in the selection of instrumental variables. The choice of  $\mathbf{Z}$  is often ad hoc.<sup>1</sup> There is a bit of a dilemma in this result. It would seem to suggest that the best choices of instruments are variables that are highly correlated with  $\mathbf{X}$ . But the more highly correlated a variable is with the problematic columns of  $\mathbf{X}$ , the less defensible the claim that these same variables are *uncorrelated* with the disturbances.

### Example 8.5 Instrumental Variable Estimation of a Labor Supply Equation

A leading example of a model in which correlation between a regressor and the disturbance is likely to arise is in market equilibrium models. Cornwell and Rupert (1988) analyzed the returns to schooling in a panel data set of 595 observations on heads of households. The sample data are drawn from years 1976 to 1982 from the “Non-Survey of Economic Opportunity” from the Panel Study of Income Dynamics. The estimating equation is

$$\begin{aligned} \ln \text{Wage}_{it} = & \beta_1 + \beta_2 \text{Exp}_{it} + \beta_3 \text{Exp}_{it}^2 + \beta_4 \text{Wks}_{it} + \beta_5 \text{Occ}_{it} + \beta_6 \text{Ind}_{it} + \beta_7 \text{South}_{it} + \\ & \beta_8 \text{SMSA}_{it} + \beta_9 \text{MS}_{it} + \beta_{10} \text{Union}_{it} + \beta_{11} \text{Ed}_i + \beta_{12} \text{Fem}_i + \beta_{13} \text{Blk}_i + \varepsilon_{it} \end{aligned}$$

where the variables are

- $\text{Exp}$  = years of full time work experience, 0 if 
- $\text{Wks}$  = weeks worked, 0 if not,
- $\text{Occ}$  = 1 if blue-collar occupation, 0 if not,
- $\text{Ind}$  = 1 if the individual works in a manufacturing industry, 0 if not,
- $\text{South}$  = 1 if the individual resides in the south, 0 if not,
- $\text{SMSA}$  = 1 if the individual resides in an SMSA, 0 if not,
- $\text{MS}$  = 1 if the individual is married, 0 if not,
- $\text{Union}$  = 1 if the individual wage is set by a union contract, 0 if not,
- $\text{Ed}$  = years of education,
- $\text{Fem}$  = 1 if the individual is female, 0 if not,
- $\text{Blk}$  = 1 if the individual is black, 0 if not.

See Appendix Table F8.1 for the data source. The main interest of the study, beyond comparing various estimation methods, is  $\beta_{11}$ , the return to education. The equation suggested is a **reduced form equation**; it contains all the variables in the model but does not specify the underlying structural relationships. In contrast, the three-equation, model specified in Section 8.3.4 is a **structural equation system**. The reduced form for this model would

<sup>1</sup>Results on “optimal instruments” appear in White (2001) and Hansen (1982). In the other direction, there is a contrary literature on “weak” instruments, such as Staiger and Stock (1997), which we will explore in Section 12.1.

**TABLE 8.1** Estimated Labor Supply Equation

Variable	OLS		IV with Z <sub>1</sub>		IV with Z <sub>2</sub>	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Constant	44.7665	1.2153	18.8987	13.0590	30.7044	4.9997
In Wage	0.7326	0.1972	5.1828	2.2454	3.1518	0.8572
Education	-0.1532	0.03206	-0.4600	0.1578	-0.3200	0.06607
Union	-1.9960	0.1701	-2.3602	0.2567	-2.1940	0.1860
Female	-1.3498	0.2642	0.6957	1.0650	-0.2378	0.4679

consist of separate regressions of *Price* and *Quantity* on (1, *Income*, *InputPrice*, *Rainfall*). We will return to the idea of reduced forms in the setting of simultaneous equations models in Chapter 10. For the present, the implication for the suggested model is that this market equilibrium equation represents the outcome of the interplay of supply and demand in a labor market. Arguably, the supply side of this market might consist of a household labor supply equation such as

$$Wks_{it} = \gamma_1 + \gamma_2 \ln Wage_{it} + \gamma_3 Ed_i + \gamma_4 Union_{it} + \gamma_5 Fem_i + u_{it}.$$

(One might prefer a different set of right-hand-side variables in this structural equation. Structural equations are more difficult to specify than reduced forms. If the number of weeks worked and the accepted wage offer are determined jointly, then  $\ln Wage_{it}$  and  $u_{it}$  in this equation are correlated. We consider two instrumental variable estimators based on

$$\mathbf{Z}_1 = [1, Ind_{it}, Ed_i, Union_{it}, Fem_i]$$

and

$$\mathbf{Z}_2 = [1, Ind_{it}, Ed_i, Union_{it}, Fem_i, SMSA_{it}].$$

Table 8.1 presents the three sets of estimates. The least squares estimates are computed using the standard results in Chapters 3 and 4. One noteworthy result is the very small coefficient on the log wage variable. The second set of results is the instrumental variable estimate developed in Section 8.3.2. Note that here, the single instrument is  $Ind_{it}$ . As might be expected, the log wage coefficient becomes considerably larger. The other coefficients are, perhaps, contradictory. One might have different expectations about all three coefficients. The third set of coefficients are the two-stage least squares estimates based on the larger set of instrumental variables. In this case,  $SMSA$  and  $Ind$  are both used as instrumental variables.

## 8.4 TWO SPECIFICATION TESTS

There are two aspects of the model that we would be interested in verifying if possible, rather than assuming them at the outset. First, it will emerge in the derivation in Section 8.4.1 that of the two estimators considered here, least squares and instrumental variables, the first is unambiguously more efficient. The IV estimator is robust; it is consistent whether or not  $\text{plim}(\mathbf{X}'\boldsymbol{\varepsilon}/n) = \mathbf{0}$ . However, if not needed, that is if  $\boldsymbol{\gamma} = \mathbf{0}$ , then least squares would be a better estimator by virtue of its smaller variance.<sup>2</sup> For this reason, and possibly in the interest of a test of the theoretical specification of the model,

<sup>2</sup>It is possible, of course, that even if least squares is inconsistent, it might still be more precise. If LS is only slightly biased but has a much smaller variance than IV, then by the expected squared error criterion, variance plus squared bias, least squares might still prove the preferred estimator. This turns out to be nearly impossible to verify empirically. We will revisit the issue in passing at a few points later in the text.

## 234 PART I ♦ The Linear Regression Model

a test that reveals information about the bias of least squares will be useful. Second, the use of two-stage least squares with  $L > K$ , that is, with “additional” instruments, entails  $L - K$  restrictions on the relationships among the variables in the model. As might be apparent from the derivation so far, when there are  $K$  variables in  $\mathbf{X}$ , some of which may be endogenous, then there must be at least  $K$  variables in  $\mathbf{Z}$  in order to identify the parameters of the model, that is, to obtain consistent estimators of the parameters using the information in the sample. When there is an excess of instruments, one is actually imposing additional, arguably superfluous restrictions on the process generating the data. Consider, once again, the agricultural market example at the end of Section 8.3.3. In that structure, it is certainly safe to assume that *Rainfall* is an exogenous event that is uncorrelated with the disturbances in the demand equation. But, it is conceivable that the interplay of the markets involved might be such that the *InputPrice* is correlated with the shocks in the demand equation. In the market for biofuels, corn is both an input in the market supply and an output in other markets. In treating *InputPrice* as exogenous in that example, we would be imposing the assumption that *InputPrice* is uncorrelated with  $\epsilon_D$ , at least by some measure unnecessarily since the parameters of the demand equation can be estimated without this assumption. This section will describe two specification tests that consider these aspects of the IV estimator.

### 8.4.1 THE HAUSMAN AND WU SPECIFICATION TESTS

It might not be obvious that the regressors in the model are correlated with the disturbances or that the regressors are measured with error. If not, there would be some benefit to using the least squares (LS) estimator rather than the IV estimator. Consider a comparison of the two covariance matrices *under the hypothesis that both estimators are consistent, that is, assuming  $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\epsilon} = \mathbf{0}$* . The difference between the asymptotic covariance matrices of the two estimators is

$$\begin{aligned}\text{Asy. Var}[\mathbf{b}_{\text{IV}}] - \text{Asy. Var}[\mathbf{b}_{\text{LS}}] &= \frac{\sigma^2}{n} \text{plim} \left( \frac{\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}}{n} \right)^{-1} - \frac{\sigma^2}{n} \text{plim} \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \\ &= \frac{\sigma^2}{n} \text{plim } n[(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}].\end{aligned}$$

To compare the two matrices in the brackets, we can compare their inverses. The inverse of the first is  $\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{X}'(\mathbf{I} - \mathbf{M}_Z)\mathbf{X} = \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{M}_Z\mathbf{X}$ . Because  $\mathbf{M}_Z$  is a non-negative definite matrix, it follows that  $\mathbf{X}'\mathbf{M}_Z\mathbf{X}$  is also. So,  $\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$  equals  $\mathbf{X}'\mathbf{X}$  minus a nonnegative definite matrix. Because  $\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$  is smaller, in the matrix sense, than  $\mathbf{X}'\mathbf{X}$ , its inverse is larger. Under the hypothesis, the asymptotic covariance matrix of the LS estimator is never larger than that of the IV estimator, and it will actually be smaller unless all the columns of  $\mathbf{X}$  are perfectly predicted by regressions on  $\mathbf{Z}$ . Thus, we have established that if  $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\epsilon} = \mathbf{0}$ —that is, if LS is consistent—then it is a preferred estimator. (Of course, we knew that from all our earlier results on the virtues of least squares.)

Our interest in the difference between these two estimators goes beyond the question of efficiency. The null hypothesis of interest will usually be specifically whether  $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\epsilon} = \mathbf{0}$ . Seeking the covariance between  $\mathbf{X}$  and  $\boldsymbol{\epsilon}$  through  $(1/n)\mathbf{X}'\boldsymbol{\epsilon}$  is fruitless, of course, because the normal equations produce  $(1/n)\mathbf{X}'\boldsymbol{\epsilon} = \mathbf{0}$ . In a seminal paper, Hausman (1978) suggested an alternative testing strategy. [Earlier work by Wu (1973) and Durbin (1954) produced what turns out to be the same test.] The logic of Hausman’s

## CHAPTER 8 ♦ Endogeneity and Instrumental Variable 235

approach is as follows. Under the null hypothesis, we have two consistent estimators of  $\beta$ ,  $\mathbf{b}_{LS}$  and  $\mathbf{b}_{IV}$ . Under the alternative hypothesis, only one of these,  $\mathbf{b}_{IV}$ , is consistent. The suggestion, then, is to examine  $\mathbf{d} = \mathbf{b}_{IV} - \mathbf{b}_{LS}$ . Under the null hypothesis,  $\text{plim } \mathbf{d} = \mathbf{0}$ , whereas under the alternative,  $\text{plim } \mathbf{d} \neq \mathbf{0}$ . Using a strategy we have used at various points before, we might test this hypothesis with a Wald statistic,

$$H = \mathbf{d}' \{ \text{Est. Asy. Var}[\mathbf{d}] \}^{-1} \mathbf{d}.$$

The asymptotic covariance matrix we need for the test is

$$\begin{aligned} \text{Asy. Var}[\mathbf{b}_{IV} - \mathbf{b}_{LS}] &= \text{Asy. Var}[\mathbf{b}_{IV}] + \text{Asy. Var}[\mathbf{b}_{LS}] \\ &\quad - \text{Asy. Cov}[\mathbf{b}_{IV}, \mathbf{b}_{LS}] - \text{Asy. Cov}[\mathbf{b}_{LS}, \mathbf{b}_{IV}]. \end{aligned}$$

At this point, the test is straightforward, save for the considerable complication that we do not have an expression for the covariance term. Hausman gives a fundamental result that allows us to proceed. Paraphrased slightly,

*the covariance between an efficient estimator,  $\mathbf{b}_E$ , of a parameter vector,  $\beta$ , and its difference from an inefficient estimator,  $\mathbf{b}_I$ , of the same parameter vector,  $\mathbf{b}_E - \mathbf{b}_I$ , is zero.*

For our case,  $\mathbf{b}_E$  is  $\mathbf{b}_{LS}$  and  $\mathbf{b}_I$  is  $\mathbf{b}_{IV}$ . By Hausman's result we have

$$\text{Cov}[\mathbf{b}_E, \mathbf{b}_E - \mathbf{b}_I] = \text{Var}[\mathbf{b}_E] - \text{Cov}[\mathbf{b}_E, \mathbf{b}_I] = \mathbf{0}$$

or

$$\text{Cov}[\mathbf{b}_E, \mathbf{b}_I] = \text{Var}[\mathbf{b}_E],$$

so

$$\text{Asy. Var}[\mathbf{b}_{IV} - \mathbf{b}_{LS}] = \text{Asy. Var}[\mathbf{b}_{IV}] - \text{Asy. Var}[\mathbf{b}_{LS}].$$

Inserting this useful result into our Wald statistic and reverting to our empirical estimates of these quantities, we have

$$H = (\mathbf{b}_{IV} - \mathbf{b}_{LS})' \{ \text{Est. Asy. Var}[\mathbf{b}_{IV}] - \text{Est. Asy. Var}[\mathbf{b}_{LS}] \}^{-1} (\mathbf{b}_{IV} - \mathbf{b}_{LS}).$$

Under the null hypothesis, we are using two different, but consistent, estimators of  $\sigma^2$ . If we use  $s^2$  as the common estimator, then the statistic will be

$$H = \frac{\mathbf{d}' [(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]^{-1} \mathbf{d}}{s^2}.$$

It is tempting to invoke our results for the full rank quadratic form in a normal vector and conclude the degrees of freedom for this chi-squared statistic is  $K$ . But that method will usually be incorrect, and worse yet, unless  $\mathbf{X}$  and  $\mathbf{Z}$  have no variables in common, the rank of the matrix in this statistic is less than  $K$ , and the ordinary inverse will not even exist. In most cases, at least some of the variables in  $\mathbf{X}$  will also appear in  $\mathbf{Z}$ . (In almost any application,  $\mathbf{X}$  and  $\mathbf{Z}$  will both contain the constant term.) That is, some of the variables in  $\mathbf{X}$  are known to be uncorrelated with the disturbances. For example, the usual case will involve a single variable that is thought to be problematic or that is measured with error. In this case, our hypothesis,  $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\epsilon} = \mathbf{0}$ , does not really involve all  $K$  variables, because a subset of the elements in this vector, say,  $K_0$ , are known to be zero. As such, the quadratic form in the Wald test is being used to test

## 236 PART I ♦ The Linear Regression Model

only  $K^* = K - K_0$  hypotheses. It is easy (and useful) to show that, in fact,  $H$  is a rank  $K^*$  quadratic form. Since  $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$  is an idempotent matrix,  $(\hat{\mathbf{X}}'\hat{\mathbf{X}}) = \hat{\mathbf{X}}'\mathbf{X}$ . Using this result and expanding  $\mathbf{d}$ , we find

$$\begin{aligned}\mathbf{d} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}[\hat{\mathbf{X}}'\mathbf{y} - (\hat{\mathbf{X}}'\hat{\mathbf{X}})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'(\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) \\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{e},\end{aligned}$$

where  $\mathbf{e}$  is the vector of least squares residuals. Recall that  $K_0$  of the columns in  $\hat{\mathbf{X}}$  are the original variables in  $\mathbf{X}$ . Suppose that these variables are the first  $K_0$ . Thus, the first  $K_0$  rows of  $\hat{\mathbf{X}}'\mathbf{e}$  are the same as the first  $K_0$  rows of  $\mathbf{X}'\mathbf{e}$ , which are, of course  $\mathbf{0}$ . (This statement does not mean that the first  $K_0$  elements of  $\mathbf{d}$  are zero.) So, we can write  $\mathbf{d}$  as

$$\mathbf{d} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} \begin{bmatrix} \mathbf{0} \\ \hat{\mathbf{X}}'^*\mathbf{e} \end{bmatrix} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{q}^* \end{bmatrix},$$

where  $\mathbf{X}^*$  is the  $K^*$  variables in  $\mathbf{x}$  that are not in  $\mathbf{z}$ .

Finally, denote the entire matrix in  $H$  by  $\mathbf{W}$ . (Because that ordinary inverse may not exist, this matrix will have to be a generalized inverse; see Section A.6.12.) Then, denoting the whole matrix product by  $\mathbf{P}$ , we obtain

$$H = [\mathbf{0}' \mathbf{q}^*'](\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\mathbf{W}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{q}^* \end{bmatrix} = [\mathbf{0}' \mathbf{q}^*']\mathbf{P} \begin{bmatrix} \mathbf{0} \\ \mathbf{q}^* \end{bmatrix} = \mathbf{q}^{*'}\mathbf{P}_{**}\mathbf{q}^*,$$

where  $\mathbf{P}_{**}$  is the lower right  $K^* \times K^*$  submatrix of  $\mathbf{P}$ . We now have the end result. Algebraically,  $H$  is actually a quadratic form in a  $K^*$  vector, so  $K^*$  is the degrees of freedom for the test.

The preceding Wald test requires a generalized inverse [see Hausman and Taylor (1981)], so it is going to be a bit cumbersome. In fact, one need not actually approach the test in this form, and it can be carried out with any regression program. The alternative **variable addition test** approach devised by Wu (1973) is simpler. An  $F$  statistic with  $K^*$  and  $n - K - K^*$  degrees of freedom can be used to test the joint significance of the elements of  $\gamma$  in the augmented regression

$$\mathbf{y} = \mathbf{X}\beta + \hat{\mathbf{X}}^*\gamma + \varepsilon^*, \quad (8-11)$$

where  $\hat{\mathbf{X}}^*$  are the fitted values in regressions of the variables in  $\mathbf{X}^*$  on  $\mathbf{Z}$ . This result is equivalent to the Hausman test for this model. [Algebraic derivations of this result can be found in the articles and in Davidson and MacKinnon (2004, Section 8.7).]

### Example 8.6 (Continued) Labor Supply Model

For the labor supply equation estimated in Example 8.5, we used the Wu (variable addition) test to examine the endogeneity of the  $\ln \text{Wage}_{it}$  variable. For the first step,  $\ln \text{Wage}_{it}$  is regressed on  $\mathbf{z}_{1,it}$ . The predicted value from this equation is then added to the least squares regression of  $\text{Wks}_{it}$  on  $\mathbf{x}_{it}$ . The results of this regression are

$$\begin{aligned}\widehat{\text{Wks}}_{it} &= 18.8987 + 0.6938 \ln \text{Wage}_{it} - 0.4600 \text{Ed}_i - 2.3602 \text{Union}_{it} \\ &\quad (12.3284) \quad (0.1980) \quad (0.1490) \quad (0.2423) \\ &\quad + 0.6958 \text{Fem}_i + 4.4891 \ln \widehat{\text{Wage}}_{it} + u_{it}, \\ &\quad (1.0054) \quad (2.1290)\end{aligned}$$

CHAPTER 8 ♦ Endogeneity and Instrumental Variable **237**

where the estimated standard errors are in parentheses. The  $t$  ratio on the fitted log wage coefficient is 2.108, which is larger than the critical value from the standard normal table of 1.96. Therefore, the hypothesis of exogeneity of the log  $Wage$  variable is rejected.

Although most of the preceding results are specific to this test of correlation between some of the columns of  $\mathbf{X}$  and the disturbances,  $\boldsymbol{\varepsilon}$ , the Hausman test is general. To reiterate, when we have a situation in which we have a pair of estimators,  $\hat{\theta}_E$  and  $\hat{\theta}_I$ , such that under  $H_0$ :  $\hat{\theta}_E$  and  $\hat{\theta}_I$  are both consistent and  $\hat{\theta}_E$  is efficient relative to  $\hat{\theta}_I$ , while under  $H_I$ :  $\hat{\theta}_I$  remains consistent while  $\hat{\theta}_E$  is inconsistent, then we can form a test of the hypothesis by referring the **Hausman statistic**,

$$H = (\hat{\theta}_I - \hat{\theta}_E)' \{ \text{Est. Asy. Var}[\hat{\theta}_I] - \text{Est. Asy. Var}[\hat{\theta}_E] \}^{-1} (\hat{\theta}_I - \hat{\theta}_E) \xrightarrow{d} \chi^2[J],$$

to the appropriate critical value for the chi-squared distribution. The appropriate degrees of freedom for the test,  $J$ , will depend on the context. Moreover, some sort of generalized inverse matrix may be needed for the matrix, although in at least one common case, the random effects regression model (see Chapter 11), the appropriate approach is to extract some rows and columns from the matrix instead. The short rank issue is not general. Many applications can be handled directly in this form with a full rank quadratic form. Moreover, the Wu approach is specific to this application. Another applications that we will consider, the independence from irrelevant alternatives test for the multinomial logit model, does not lend itself to the regression approach and is typically handled using the Wald statistic and the full rank quadratic form. As a final note, observe that the short rank of the matrix in the Wald statistic is an algebraic result. The failure of the matrix in the Wald statistic to be positive definite, however, is sometimes a finite-sample problem that is not part of the model structure. In such a case, forcing a solution by using a generalized inverse may be misleading. Hausman suggests that in this instance, the appropriate conclusion might be simply to take the result as zero and, by implication, not reject the null hypothesis.

**Example 8.7 Hausman Test for a Consumption Function**

Quarterly data for 1950.1 to 2000.4 on a number of macroeconomic variables appear in Appendix Table F5.2. A consumption function of the form  $C_t = \alpha + \beta Y_t + \varepsilon_t$  is estimated using the 203 observations on aggregate U.S. real consumption and real disposable personal income, omitting the first. This model is a candidate for the possibility of bias due to correlation between  $Y_t$  and  $\varepsilon_t$ . Consider instrumental variables estimation using  $Y_{t-1}$  and  $C_{t-1}$  as the instruments for  $Y_t$ , and, of course, the constant term is its own instrument. One observation is lost because of the lagged values, so the results are based on 203 quarterly observations. The Hausman statistic can be computed in two ways:

1. Use the Wald statistic for  $H$  with the Moore–Penrose generalized inverse. The common  $s^2$  is the one computed by least squares under the null hypothesis of no correlation. With this computation,  $H = 8.481$ . There is  $K^* = 1$  degree of freedom. The 95 percent critical value from the chi-squared table is 3.84. Therefore, we reject the null hypothesis of no correlation between  $Y_t$  and  $\varepsilon_t$ .
2. Using the Wu statistic based on (8-11), we regress  $C_t$  on a constant,  $Y_t$ , and the predicted value in a regression of  $Y_t$  on a constant,  $Y_{t-1}$  and  $C_{t-1}$ . The  $t$  ratio on the prediction is 2.968, so the  $F$  statistic with 1 and 200 degrees of freedom is 8.809. The critical value for this  $F$  distribution is 3.888, so, again, the null hypothesis is rejected.

## 238 PART I ♦ The Linear Regression Model

### 8.4.2 A TEST FOR OVERIDENTIFICATION

The motivation for choosing the IV estimator is not efficiency. The estimator is constructed to be consistent; efficiency is not a consideration. In Chapter 13, we will revisit the issue of efficient method of moments estimation. The observation that 2SLS represents the most efficient use of all  $L$  instruments establishes only the efficiency of the estimator in the class of estimators that use  $K$  linear combinations of the columns of  $\mathbf{Z}$ . The IV estimator is developed around the **orthogonality conditions**

$$E[\text{yellow speech bubble}] = 0. \quad (8-12)$$

The sample counterpart to this is the **moment equation**,

$$\frac{1}{n} \sum_{i=1}^n \text{yellow speech bubble} = 0. \quad (8-13)$$

The solution, when  $L = K$ , is  $\mathbf{b}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$ , as we have seen. If  $L > K$ , then there is no single solution, and we arrived at 2SLS as a strategy. Estimation is still based on (8-13). However, the sample counterpart is now  $L$  equations in  $K$  unknowns and (8-13) has no solution. Nonetheless, under the hypothesis of the model, (8-12) remains true. We can consider the additional restrictions as a hypothesis that might or might not be supported by the sample evidence. The excess of moment equations provides a way to test the **overidentification** of the model. The test will be based on (8-13), which, when evaluated at  $\mathbf{b}_{IV}$ , will not equal zero when  $L > K$ , though the hypothesis in (8-12) might still be true.

The test statistic will be a Wald statistic. (See Section 5.4.) The sample statistic, based on (8-13) and the IV estimator, is

$$\bar{\mathbf{m}} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i e_{IV,i} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i (y_i - \mathbf{x}'_i \mathbf{b}_{IV}).$$

The Wald statistic is

$$\chi^2[L - K] = \bar{\mathbf{m}}' [\text{Var}(\bar{\mathbf{m}})]^{-1} \bar{\mathbf{m}}.$$

To complete the construction, we require an estimator of the variance. There are two ways to proceed. Under the assumption of the model,

$$\text{Var}[\bar{\mathbf{m}}] = \frac{\sigma^2}{n^2} \mathbf{Z}' \mathbf{Z},$$

which can be estimated easily using the sample estimator of  $\sigma^2$ . Alternatively, we might base the estimator on (8-12), which would imply that an appropriate estimator would be

$$\text{Est.Var}[\bar{\mathbf{m}}] = \frac{1}{n^2} \sum_{i=1}^n (\mathbf{z}_i e_{IV,i})(\mathbf{z}_i e_{IV,i})' = \frac{1}{n^2} \sum_{i=1}^n e_{IV,i}^2 \mathbf{z}_i \mathbf{z}'_i.$$

These two estimators will be numerically different in a finite sample, but under the assumptions that we have made so far, both (multiplied by  $n$ ) will converge to the same matrix, so the choice is immaterial. Current practice favors the second. The Wald

statistic is, then

$$\left( \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i e_{IV,i} \right)' \left[ \frac{1}{n^2} \sum_{i=1}^n e_{IV,i}^2 \mathbf{z}_i \mathbf{z}'_i \right]^{-1} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i e_{IV,i} \right).$$

A remaining detail is the number of degrees of freedom. The test can only detect the failure of  $L - K$  moment equations, so that is the rank of the quadratic form; the limiting distribution of the statistic is chi squared with  $L - K$  degrees of freedom.

#### **Example 8.8 Overidentification of the Labor Supply Equation**

In Example 8.5, we computed 2SLS estimates of the parameters of an equation for weeks worked. The estimator is based on

$$x = [1, \ln \text{Wage}, \text{Education}, \text{Union}, \text{Female}]$$

and

$$z = [1, \text{Ind}, \text{Education}, \text{Union}, \text{Female}, \text{SMSA}].$$

There is one overidentifying restriction. The sample moment based on the 2SLS results in Table 8.1 is

(1/4165)   $\mathbf{Z} \mathbf{e}_{2\text{SLS}} = [0, .03476, 0, 0, 0, -.01543]'.$

The chi-squared statistic is 1.09399 with one degree of freedom. If the first suggested variance estimator is used, the statistic is 1.05241. Both are well under the 95 percent critical value of 3.84, so the hypothesis of overidentification is not rejected.

We note a final implication of the test. One might conclude, based on the underlying theory of the model, that the overidentification test relates to one particular instrumental variable and not another. For example, in our market equilibrium example with two instruments for the demand equation, *Rainfall* and *InputPrice*, rainfall is obviously exogenous, so a rejection of the overidentification restriction would eliminate *InputPrice* as a valid instrument. However, this conclusion would be inappropriate; the test suggests only that one or more of the elements in (8-12) are nonzero. It does not suggest which elements in particular these are.

## 8.5 MEASUREMENT ERROR

Thus far, it has been assumed (at least implicitly) that the data used to estimate the parameters of our models are true measurements on their theoretical counterparts. In practice, this situation happens only in the best of circumstances. All sorts of measurement problems creep into the data that must be used in our analyses. Even carefully constructed survey data do not always conform exactly to the variables the analysts have in mind for their regressions. Aggregate statistics such as GDP are only estimates of their theoretical counterparts, and some variables, such as depreciation, the services of capital, and “the interest rate,” do not even exist in an agreed-upon theory. At worst, there may be no physical measure corresponding to the variable in our model; intelligence, education, and permanent income are but a few examples. Nonetheless, they all have appeared in very precisely defined regression models.

## 240 PART I ♦ The Linear Regression Model

### 8.5.1 LEAST SQUARES ATTENUATION

In this section, we examine some of the received results on regression analysis with badly measured data. The general assessment of the problem is not particularly optimistic. The biases introduced by measurement error can be rather severe. There are almost no known finite-sample results for the models of measurement error; nearly all the results that have been developed are asymptotic.<sup>3</sup> The following presentation will use a few simple asymptotic results for the classical regression model.

The simplest case to analyze is that of a regression model with a single regressor and no constant term. Although this case is admittedly unrealistic, it illustrates the essential concepts, and we shall generalize it presently. Assume that the model,

$$y^* = \beta x^* + \varepsilon, \quad (8-14)$$

conforms to all the assumptions of the classical normal regression model. If data on  $y^*$  and  $x^*$  were available, then  $\beta$  would be estimable by least squares. Suppose, however, that the observed data are only imperfectly measured versions of  $y^*$  and  $x^*$ . In the context of an example, suppose that  $y^*$  is  $\ln(\text{output/labor})$  and  $x^*$  is  $\ln(\text{capital/labor})$ . Neither factor input can be measured with precision, so the observed  $y$  and  $x$  contain errors of measurement. We assume that

$$y = y^* + v \quad \text{with } v \sim N[0, \sigma_v^2], \quad (8-15a)$$

$$x = x^* + u \quad \text{with } u \sim N[0, \sigma_u^2]. \quad (8-15b)$$

Assume, as well, that  $u$  and  $v$  are independent of each other and of  $y^*$  and  $x^*$ . (As we shall see, adding these restrictions is not sufficient to rescue a bad situation.)

As a first step, insert (8-15a) into (8-14), assuming for the moment that only  $y^*$  is measured with error:

$$y = \beta x^* + \varepsilon + v = \beta x^* + \varepsilon'.$$

This result conforms to the assumptions of the classical regression model. As long as the regressor is measured properly, measurement error on the dependent variable can be absorbed in the disturbance of the regression and ignored. To save some cumbersome notation, therefore, we shall henceforth assume that the measurement error problems concern only the independent variables in the model.

Consider, then, the regression of  $y$  on the observed  $x$ . By substituting (8-15b) into (8-14), we obtain

$$y = \beta x + [\varepsilon - \beta u] = \beta x + w. \quad (8-16)$$

Because  $x$  equals  $x^* + u$ , the regressor in (8-16) is correlated with the disturbance:

$$\text{Cov}[x, w] = \text{Cov}[x^* + u, \varepsilon - \beta u] = -\beta \sigma_u^2. \quad (8-17)$$

This result violates one of the central assumptions of the classical model, so we can expect the least squares estimator,

$$b = \frac{(1/n) \sum_{i=1}^n x_i y_i}{(1/n) \sum_{i=1}^n x_i^2},$$

---

<sup>3</sup>See, for example, Imbens and Hyslop (2001).

CHAPTER 8 ♦ Endogeneity and Instrumental Variable **241**

to be inconsistent. To find the probability limits, insert (8-14) and (8-15b) and use the Slutsky theorem:

$$\text{plim } b = \frac{\text{plim}(1/n) \sum_{i=1}^n (x_i^* + u_i)(\beta x_i^* + \varepsilon_i)}{\text{plim}(1/n) \sum_{i=1}^n (x_i^* + u_i)^2}.$$

Because  $x^*$ ,  $\varepsilon$ , and  $u$  are mutually independent, this equation reduces to

$$\text{plim } b = \frac{\beta Q^*}{Q^* + \sigma_u^2} = \frac{\beta}{1 + \sigma_u^2/Q^*}, \quad (8-18)$$

where  $Q^* = \text{plim}(1/n) \sum_i x_i^{*2}$ . As long as  $\sigma_u^2$  is positive,  $b$  is inconsistent, with a persistent bias toward zero. Clearly, the greater the variability in the measurement error, the worse the bias. The effect of biasing the coefficient toward zero is called **attenuation**.

In a multiple regression model, matters only get worse. Suppose, to begin, we assume that  $\mathbf{y} = \mathbf{X}^* \boldsymbol{\beta} + \boldsymbol{\varepsilon}$  and  $\mathbf{X} = \mathbf{X}^* + \mathbf{U}$ , allowing every observation on every variable to be measured with error. The extension of the earlier result is

$$\text{plim} \left( \frac{\mathbf{X}' \mathbf{X}}{n} \right) = \mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}, \quad \text{and} \quad \text{plim} \left( \frac{\mathbf{X}' \mathbf{y}}{n} \right) = \mathbf{Q}^* \boldsymbol{\beta}.$$

Hence,

$$\text{plim } \mathbf{b} = [\mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}]^{-1} \mathbf{Q}^* \boldsymbol{\beta} = \boldsymbol{\beta} - [\mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}]^{-1} \boldsymbol{\Sigma}_{uu} \boldsymbol{\beta}. \quad (8-19)$$

This probability limit is a mixture of all the parameters in the model. In the same fashion as before, bringing in outside information could lead to **identification**. The amount of information necessary is extremely large, however, and this approach is not particularly promising.

It is common for only a single variable to be measured with error. One might speculate that the problems would be isolated to the single coefficient. Unfortunately, this situation is not the case. For a single bad variable—assume that it is the first—the matrix  $\boldsymbol{\Sigma}_{uu}$  is of the form

$$\boldsymbol{\Sigma}_{uu} = \begin{bmatrix} \sigma_u^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

It can be shown that for this special case,

$$\text{plim } b_1 = \frac{\beta_1}{1 + \sigma_u^2 q^{*11}} \quad (8-20a)$$

[note the similarity of this result to (8-18)], and, for  $k \neq 1$ ,

$$\text{plim } b_k = \beta_k - \beta_1 \left[ \frac{\sigma_u^2 q^{*k1}}{1 + \sigma_u^2 q^{*11}} \right], \quad (8-20b)$$

where  $q^{*k1}$  is the  $(k, 1)$ th element in  $(\mathbf{Q}^*)^{-1}$ .<sup>4</sup> This result depends on several unknowns and cannot be estimated. The coefficient on the badly measured variable is still biased

<sup>4</sup>Use (A-66) to invert  $[\mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}] = [\mathbf{Q}^* + (\sigma_u \mathbf{e}_1)(\sigma_u \mathbf{e}_1)']$ , where  $\mathbf{e}_1$  is the first column of a  $K \times K$  identity matrix. The remaining results are then straightforward.

## 242 PART I ♦ The Linear Regression Model

toward zero. The other coefficients are all biased as well, although in unknown directions. A badly measured variable contaminates all the least squares estimates.<sup>5</sup> If more than one variable is measured with error, there is very little that can be said.<sup>6</sup> Although expressions can be derived for the biases in a few of these cases, they generally depend on numerous parameters whose signs and magnitudes are unknown and, presumably, unknowable.

### 8.5.2 INSTRUMENTAL VARIABLES ESTIMATION

An alternative set of results for estimation in this model (and numerous others) is built around the method of instrumental variables. Consider once again the errors in variables model in (8-14) and (8-15a,b). The parameters,  $\beta$ ,  $\sigma_\varepsilon^2$ ,  $q^*$ , and  $\sigma_u^2$  are not identified in terms of the moments of  $x$  and  $y$ . Suppose, however, that there exists a variable  $z$  such that  $z$  is correlated with  $x^*$  but not with  $u$ . For example, in surveys of families, income is notoriously badly reported, partly deliberately and partly because respondents often neglect some minor sources. Suppose, however, that one could determine the total amount of checks written by the head(s) of the household. It is quite likely that this  $z$  would be highly correlated with income, but perhaps not significantly correlated with the errors of measurement. If  $\text{Cov}[x^*, z]$  is not zero, then the parameters of the model become estimable, as

$$\text{plim} \frac{(1/n) \sum_i y_i z_i}{(1/n) \sum_i x_i z_i} = \frac{\beta \text{Cov}[x^*, z]}{\text{Cov}[x^*, z]} = \beta. \quad (8-21)$$

For the general case,  $\mathbf{y} = \mathbf{X}^* \boldsymbol{\beta} + \boldsymbol{\epsilon}$ ,  $\mathbf{X} = \mathbf{X}^* + \mathbf{U}$ , suppose that there exists a matrix of variables  $\mathbf{Z}$  that is not correlated with the disturbances or the measurement error but is correlated with regressors,  $\mathbf{X}$ . Then the instrumental variables estimator based on  $\mathbf{Z}$ ,  $\mathbf{b}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$ , is consistent and asymptotically normally distributed with asymptotic covariance matrix that is estimated with

$$\text{Est. Asy. Var}[\mathbf{b}_{\text{IV}}] = \hat{\sigma}^2 [\mathbf{Z}'\mathbf{X}]^{-1} [\mathbf{Z}'\mathbf{Z}] [\mathbf{X}'\mathbf{Z}]^{-1}. \quad (8-22)$$

For more general cases, Theorem 8.1 and the results in Section 8.3 apply.

### 8.5.3 PROXY VARIABLES

In some situations, a variable in a model simply has no observable counterpart. Education, intelligence, ability, and like factors are perhaps the most common examples. In this instance, unless there is some observable indicator for the variable, the model will have to be treated in the framework of missing variables. Usually, however, such an indicator can be obtained; for the factors just given, years of schooling and test scores of various sorts are familiar examples. The usual treatment of such variables is in the measurement error framework. If, for example,

$$\text{income} = \beta_1 + \beta_2 \text{education} + \varepsilon$$

<sup>5</sup>This point is important to remember when the presence of measurement error is suspected.

<sup>6</sup>Some firm analytic results have been obtained by Levi (1973), Theil (1961), Klepper and Leamer (1983), Garber and Klepper (1980), Griliches (1986), and Cragg (1997).

## CHAPTER 8 ♦ Endogeneity and Instrumental Variable 243

and

$$\text{years of schooling} = \text{education} + u,$$

then the model of Section 8.5.1 applies. The only difference here is that the true variable in the model is “latent.” No amount of improvement in reporting or measurement would bring the proxy closer to the variable for which it is proxying.

The preceding is a pessimistic assessment, perhaps more so than necessary. Consider a **structural model**,

$$\text{Earnings} = \beta_1 + \beta_2 \text{Experience} + \beta_3 \text{Industry} + \beta_4 \text{Ability} + \varepsilon.$$

*Ability* is unobserved, but suppose that an indicator, say, *IQ*, is. If we suppose that *IQ* is related to *Ability* through a relationship such as

$$IQ = \alpha_1 + \alpha_2 \text{Ability} + v,$$

then we may solve the second equation for *Ability* and insert it in the first to obtain the **reduced form equation**

$$\text{Earnings} = (\beta_1 - \beta_4 \alpha_1 / \alpha_2) + \beta_2 \text{Experience} + \beta_3 \text{Industry} + (\beta_4 / \alpha_2) IQ + (\varepsilon - v \beta_4 / \alpha_2).$$

This equation is intrinsically linear and can be estimated by least squares. We do not have consistent estimators of  $\beta_1$  and  $\beta_4$ , but we do have them for the coefficients of interest,  $\beta_2$  and  $\beta_3$ . This would appear to “solve” the problem. We should note the essential ingredients; we require that the **indicator**, *IQ*, not be related to the other variables in the model, and we also require that *v* not be correlated with any of the variables. In this instance, some of the parameters of the structural model are identified in terms of observable data. Note, though, that *IQ* is not a proxy variable, it is an indicator of the latent variable, *Ability*. This form of modeling has figured prominently in the education and educational psychology literature. Consider, in the preceding small model how one might proceed with not just a single indicator, but say with a battery of test scores, all of which are indicators of the same latent ability variable.

It is to be emphasized that a proxy variable is not an instrument (or the reverse). Thus, in the instrumental variables framework, it is implied that we do not regress *y* on *Z* to obtain the estimates. To take an extreme example, suppose that the full model was

$$\mathbf{y} = \mathbf{X}^* \boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\mathbf{X} = \mathbf{X}^* + \mathbf{U},$$

$$\mathbf{Z} = \mathbf{X}^* + \mathbf{W}.$$

That is, we happen to have two badly measured estimates of  $\mathbf{X}^*$ . The parameters of this model can be estimated without difficulty if  $\mathbf{W}$  is uncorrelated with  $\mathbf{U}$  and  $\mathbf{X}^*$ , *but not by regressing y on Z*. The instrumental variables technique is called for.

When the model contains a variable such as education or ability, the question that naturally arises is; If interest centers on the other coefficients in the model, why not just discard the problem variable?<sup>7</sup> This method produces the familiar problem of an omitted variable, compounded by the least squares estimator in the full model being inconsistent anyway. Which estimator is worse? McCallum (1972) and Wickens (1972)

---

<sup>7</sup>This discussion applies to the measurement error and latent variable problems equally.

## 244 PART I ♦ The Linear Regression Model

show that the asymptotic bias (actually, degree of inconsistency) is worse if the proxy is omitted, even if it is a bad one (has a high proportion of measurement error). This proposition neglects, however, the precision of the estimates. Aigner (1974) analyzed this aspect of the problem and found, as might be expected, that it could go either way. He concluded, however, that “there is evidence to broadly support use of the proxy.”

### **Example 8.9 Income and Education in a Study of Twins**

The traditional model used in labor economics to study the effect of education on income is an equation of the form

$$y_i = \beta_1 + \beta_2 \text{age}_i + \beta_3 \text{age}_i^2 + \beta_4 \text{education}_i + \mathbf{x}'\boldsymbol{\beta}_5 + \varepsilon_i,$$

where  $y_i$  is typically a wage or yearly income (perhaps in log form) and  $\mathbf{x}_i$  contains other variables, such as an indicator for sex, region of the country, and industry. The literature contains discussion of many possible problems in estimation of such an equation by least squares using measured data. Two of them are of interest here:

1. Although “education” is the variable that appears in the equation, the data available to researchers usually include only “years of schooling.” This variable is a proxy for education, so an equation fit in this form will be tainted by this problem of measurement error. Perhaps surprisingly so, researchers also find that reported data on years of schooling are themselves subject to error, so there is a second source of measurement error. For the present, we will not consider the first (much more difficult) problem.
2. Other variables, such as “ability”—we denote these  $\mu_i$ —will also affect income and are surely correlated with education. If the earnings equation is estimated in the form shown above, then the estimates will be further biased by the absence of this “omitted variable.” For reasons we will explore in Chapter 24, this bias has been called the **selectivity effect** in recent studies.

Simple cross-section studies will be considerably hampered by these problems. But, in a study of twins, Ashenfelter and Kreuger (1994) analyzed a data set that allowed them, with a few simple assumptions, to ameliorate these problems.<sup>8</sup>

Annual “twins festivals” are held at many places in the United States. The largest is held in Twinsburg, Ohio. The authors interviewed about 500 individuals over the age of 18 at the August 1991 festival. Using pairs of twins as their observations enabled them to modify their model as follows: Let  $(y_{ij}, A_{ij})$  denote the earnings and age for twin  $j$ ,  $j = 1, 2$ , for pair  $i$ . For the education variable, only self-reported “schooling” data,  $S_{ij}$ , are available. The authors approached the measurement problem in the schooling variable,  $S_{ij}$ , by asking each twin how much schooling they had and how much schooling their sibling had. Denote reported schooling by sibling  $m$  of sibling  $j$  by  $S_{ij}(m)$ . So, the self-reported years of schooling of twin 1 is  $S_{i1}(1)$ . When asked how much schooling twin 1 has, twin 2 reports  $S_{i1}(2)$ . The measurement error model for the schooling variable is

$$S_{ij}(m) = S_{ij} + u_{ij}(m), \quad j, m = 1, 2, \quad \text{where } S_{ij} = \text{“true” schooling for twin } j \text{ of pair } i.$$

We assume that the two sources of measurement error,  $u_{ij}(m)$ , are uncorrelated and  $S_{ij}$  have zero means. Now, consider a simple bivariate model such as the one in (12-17).

$$y_{ij} = \beta S_{ij} + \varepsilon_{ij}.$$

As we saw earlier, a least squares estimate of  $\beta$  using the reported data will be attenuated:

$$\text{plim } b = \frac{\beta \times \text{Var}[S_{ij}]}{\text{Var}[S_{ij}] + \text{Var}[u_{ij}(j)]} = \beta q.$$

<sup>8</sup>Other studies of twins and siblings include Bound, Chorkas, Haskel, Hawkes, and Spector (2003). Ashenfelter and Rouse (1998), Ashenfelter and Zimmerman (1997), Behrman and Rosengweig (1999), Isacsson (1999), Miller, Mulvey, and Martin (1995), Rouse (1999), and Taubman (1976).

## CHAPTER 8 ♦ Endogeneity and Instrumental Variable 245

(Because there is no natural distinction between twin 1 and twin 2, the assumption that the variances of the two measurement errors are equal is innocuous.) The factor  $q$  is sometimes called the reliability ratio. In this simple model, if the reliability ratio were known, then  $\beta$  could be consistently estimated. In fact, the construction of this model allows just that. Since the two measurement errors are uncorrelated,

$$\text{Corr}[S_{i1}(1), S_{i1}(2)] = \text{Corr}[S_{i2}(1), S_{i2}(2)]$$

$$= \frac{\text{Var}[S_{i1}]}{\{\{\text{Var}[S_{i1}] + \text{Var}[u_{i1}(1)]\} \times \{\text{Var}[S_{i1}] + \text{Var}[u_{i1}(2)]\}\}^{1/2}} = q.$$

In words, the correlation between the two reported education attainments measures the reliability ratio. The authors obtained values of 0.920 and 0.877 for 298 pairs of identical twins and 0.869 and 0.951 for 92 pairs of fraternal twins, thus providing a quick assessment of the extent of measurement error in their schooling data.

The earnings equation is a multiple regression, so this result is useful for an overall assessment of the problem, but the numerical values are not sufficient to undo the overall biases in the least squares regression coefficients. An instrumental variables estimator was used for that purpose. The estimating equation for  $y_{ij} = \ln \text{Wage}_{ij}$ , with the least squares (LS) and instrumental variable (IV) estimates is as follows:

$$\begin{array}{l} y_{ij} = \beta_1 + \beta_2 \text{age}_i + \beta_3 \text{age}_i^2 + \beta_4 S_{ij}(j) + \beta_5 S_{im}(m) + \beta_6 \text{sex}_i + \beta_7 \text{race}_i + \varepsilon_{ij} \\ \text{LS} \quad (0.088) \quad (-0.087) \quad (0.084) \quad \quad \quad (0.204) \quad (-0.410) \\ \text{IV} \quad (0.088) \quad (-0.087) \quad (0.116) \quad (-0.037) \quad (0.206) \quad (-0.428). \end{array}$$

In the equation,  $S_{ij}(j)$  is the person's report of his or her own years of schooling and  $S_{im}(m)$  is the sibling's report of the sibling's own years of schooling. The problem variable is schooling. To obtain a consistent estimator, the method of instrumental variables was used, using each sibling's report of the other sibling's years of schooling as a pair of instrumental variables. The estimates reported by the authors are shown below the equation. (The constant term was not reported, and for reasons not given, the second schooling variable was not included in the equation when estimated by LS.) This preliminary set of results is presented to give a comparison to other results in the literature. The age, schooling, and gender effects are comparable with other received results, whereas the effect of race is vastly different, -40 percent here compared with a typical value of +9 percent in other studies. The effect of using the instrumental variable estimator on the estimates of  $\beta_4$  is of particular interest. Recall that the reliability ratio was estimated at about 0.9, which suggests that the IV estimate would be roughly 11 percent higher ( $1/0.9$ ). Because this result is a multiple regression, that estimate is only a crude guide. The estimated effect shown above is closer to 38 percent.

The authors also used a different estimation approach. Recall the issue of selection bias caused by unmeasured effects. The authors reformulated their model as

$$y_{ij} = \beta_1 + \beta_2 \text{age}_i + \beta_3 \text{age}_i^2 + \beta_4 S_{ij}(j) + \beta_6 \text{sex}_i + \beta_7 \text{race}_i + \mu_i + \varepsilon_{ij}.$$

Unmeasured latent effects, such as "ability," are contained in  $\mu_i$ . Because  $\mu_i$  is not observable but is, it is assumed, correlated with other variables in the equation, the least squares regression of  $y_{ij}$  on the other variables produces a biased set of coefficient estimates. [This is a "fixed effects model—see Section 9." The assumption that the latent effect, "ability," is common between the twins and fully accounted for is a controversial assumption that ability is accounted for by "nature" rather than "nurture." See, e.g., Behrman and Taubman (1989). A search of the Internet on the subject of the "nature versus nurture debate" will turn up millions of citations. We will not visit the subject here.] The difference between the two earnings equations is

$$y_{i1} - y_{i2} = \beta_4[S_{i1}(1) - S_{i2}(2)] + \varepsilon_{i1} - \varepsilon_{i2}.$$

This equation removes the latent effect but, it turns out, worsens the measurement error problem. As before,  $\beta_4$  can be estimated by instrumental variables. There are two instrumental variables available,  $S_{i2}(1)$  and  $S_{i1}(2)$ . (It is not clear in the paper whether the authors used

## 246 PART I ♦ The Linear Regression Model

the two separately or the difference of the two.) The least squares estimate is 0.092, which is comparable to the earlier estimate. The instrumental variable estimate is 0.167, which is nearly 82 percent higher. The two reported standard errors are 0.024 and 0.043, respectively. With these figures, it is possible to carry out Hausman's test;

$$H = \frac{(0.167 - 0.092)^2}{0.043^2 - 0.024^2} = 4.418.$$

The 95 percent critical value from the chi-squared distribution with one degree of freedom is 3.84, so the hypothesis that the LS estimator is consistent would be rejected. (The square root of  $H$ , 2.102, would be treated as a value from the standard normal distribution, from which the critical value would be 1.96. The authors reported a  $t$  statistic for this regression of 1.97. The source of the difference is unclear.)

## 8.6 NONLINEAR INSTRUMENTAL VARIABLES ESTIMATION

In Section 8.2, we extended the linear regression model to allow for the possibility that the regressors might be correlated with the disturbances. The same problem can arise in nonlinear models. The consumption function estimated in Section 7.2.5 is almost surely a case in point, and we reestimated it using the instrumental variables technique for linear models in Example 8.7. In this section, we will extend the method of instrumental variables to nonlinear regression models.

In the nonlinear model,

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i,$$

the covariates  $\mathbf{x}_i$  may be correlated with the disturbances. We would expect this effect to be transmitted to the pseudoregressors,  $\mathbf{x}_i^0 = \partial h(\mathbf{x}_i, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ . If so, then the results that we derived for the linearized regression would no longer hold. Suppose that there is a set of variables  $[\mathbf{z}_1, \dots, \mathbf{z}_L]$  such that

$$\text{plim}(1/n)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0} \quad (8-23)$$

and

$$\text{plim}(1/n)\mathbf{Z}'\mathbf{X}^0 = \mathbf{Q}_{\mathbf{zx}}^0 \neq \mathbf{0},$$

where  $\mathbf{X}^0$  is the matrix of pseudoregressors in the linearized regression, evaluated at the true parameter values. If the analysis that we used for the linear model in Section 8.3 can be applied to this set of variables, then we will be able to construct a consistent estimator for  $\boldsymbol{\beta}$  using the instrumental variables. As a first step, we will attempt to replicate the approach that we used for the linear model. The linearized regression model is given in (7-30),

$$\mathbf{y} = \mathbf{h}(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon} \approx \mathbf{h}^0 + \mathbf{X}^0(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \boldsymbol{\varepsilon}$$

or

$$\mathbf{y}^0 \approx \mathbf{X}^0\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{y}^0 = \mathbf{y} - \mathbf{h}^0 + \mathbf{X}^0\boldsymbol{\beta}^0.$$

## CHAPTER 8 ♦ Endogeneity and Instrumental Variable 247

For the moment, we neglect the approximation error in linearizing the model. In (8-23), we have assumed that

$$\text{plim}(1/n)\mathbf{Z}'\mathbf{y}^0 = \text{plim}(1/n)\mathbf{Z}'\mathbf{X}^0\boldsymbol{\beta}. \quad (8-24)$$

Suppose, as we assumed before, that there are the same number of instrumental variables as there are parameters, that is, columns in  $\mathbf{X}^0$ . (Note: This number need not be the number of variables.) Then the “estimator” used before is suggested:

$$\mathbf{b}_{\text{IV}} = (\mathbf{Z}'\mathbf{X}^0)^{-1}\mathbf{Z}'\mathbf{y}^0. \quad (8-25)$$

The logic is sound, but there is a problem with this estimator. The unknown parameter vector  $\boldsymbol{\beta}$  appears on both sides of (8-24). We might consider the approach we used for our first solution to the nonlinear regression model. That is, with some initial estimator in hand, iterate back and forth between the instrumental variables regression and recomputing the pseudoregressors until the process converges to the fixed point that we seek. Once again, the logic is sound, and in principle, this method does produce the estimator we seek.

If we add to our preceding assumptions

$$\frac{1}{\sqrt{n}}\mathbf{Z}'\boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{0}, \sigma^2 \mathbf{Q}_{zz}],$$

then we will be able to use the same form of the asymptotic distribution for this estimator that we did for the linear case. Before doing so, we must fill in some gaps in the preceding. First, despite its intuitive appeal, the suggested procedure for finding the estimator is very unlikely to be a good algorithm for locating the estimates. Second, we do not wish to limit ourselves to the case in which we have the same number of instrumental variables as parameters. So, we will consider the problem in general terms. The estimation criterion for nonlinear instrumental variables is a quadratic form,

$$\begin{aligned} \text{Min}_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) &= \frac{1}{2}\{[\mathbf{y} - \mathbf{h}(\mathbf{X}, \boldsymbol{\beta})]'\mathbf{Z}\}(\mathbf{Z}'\mathbf{Z})^{-1}\{\mathbf{Z}'[\mathbf{y} - \mathbf{h}(\mathbf{X}, \boldsymbol{\beta})]\} \\ &= \frac{1}{2}\boldsymbol{\varepsilon}(\boldsymbol{\beta})'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}(\boldsymbol{\beta}).^9 \end{aligned} \quad (8-26)$$

The first-order conditions for minimization of this weighted sum of squares are

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\mathbf{X}^{0r}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}(\boldsymbol{\beta}) = \mathbf{0}. \quad (8-27)$$

This result is the same one we had for the linear model with  $\mathbf{X}^0$  in the role of  $\mathbf{X}$ . This problem, however, is highly nonlinear in most cases, and the repeated least squares approach is unlikely to be effective. But it is a straightforward minimization problem in the frameworks of Appendix E, and instead, we can just treat estimation here as a problem in nonlinear optimization.

We have approached the formulation of this instrumental variables estimator more or less strategically. However, there is a more structured approach. The

---

<sup>9</sup>Perhaps the more natural point to begin the minimization would be  $S^0(\boldsymbol{\beta}) = [\boldsymbol{\varepsilon}(\boldsymbol{\beta})'\mathbf{Z}] [\mathbf{Z}'\boldsymbol{\varepsilon}(\boldsymbol{\beta})]$ . We have bypassed this step because the criterion in (8-26) and the estimator in (8-27) will turn out (following and in Chapter 13) to be a simple yet more efficient GMM estimator.

## 248 PART I ♦ The Linear Regression Model

orthogonality condition

$$\text{plim}(1/n)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0}$$

defines a GMM estimator. With the homoscedasticity and nonautocorrelation assumption, the resultant **minimum distance estimator** produces precisely the criterion function suggested above. We will revisit this estimator in this context, in Chapter 13.

With well-behaved *pseudoregressors* and instrumental variables, we have the general result for the nonlinear instrumental variables estimator; this result is discussed at length in Davidson and MacKinnon (2004).

### THEOREM 8.2 Asymptotic Distribution of the Nonlinear Instrumental Variables Estimator

*With well-behaved instrumental variables and pseudoregressors,*

$$\mathbf{b}_{\text{IV}} \xrightarrow{a} N[\boldsymbol{\beta}, (\sigma^2/n)(\mathbf{Q}_{xz}^0(\mathbf{Q}_{zz})^{-1}\mathbf{Q}_{zx}^0)^{-1}].$$

We estimate the asymptotic covariance matrix with

$$\text{Est. Asy. Var}[\mathbf{b}_{\text{IV}}] = \hat{\sigma}^2 [\hat{\mathbf{X}}^0 \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' \hat{\mathbf{X}}^0]^{-1},$$

where  $\hat{\mathbf{X}}^0$  is  $\mathbf{X}^0$  computed using  $\mathbf{b}_{\text{IV}}$ .

As a final observation, note that the “two-stage least squares” interpretation of the instrumental variables estimator for the linear model still applies here, with respect to the IV estimator. That is, at the final estimates, the first-order conditions (normal equations) imply that

$$\mathbf{X}^0 \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y} = \mathbf{X}^0 \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X}^0 \boldsymbol{\beta},$$

which says that the estimates satisfy the normal equations for a linear regression of  $\mathbf{y}$  (not  $\mathbf{y}^0$ ) on the predictions obtained by regressing the columns of  $\mathbf{X}^0$  on  $\mathbf{Z}$ . The interpretation is not quite the same here, because to compute the predictions of  $\mathbf{X}^0$ , we must have the estimate of  $\boldsymbol{\beta}$  in hand. Thus, this two-stage least squares approach does not show *how to compute*  $\mathbf{b}_{\text{IV}}$ ; it shows a characteristic of  $\mathbf{b}_{\text{IV}}$ .

#### Example 8.10 Instrumental Variables Estimates of the Consumption Function

The consumption function in Section 7.2.5 was estimated by nonlinear least squares without accounting for the nature of the data that would certainly induce correlation between  $\mathbf{X}^0$  and  $\boldsymbol{\varepsilon}$ . As we did earlier, we will reestimate this model using the technique of instrumental variables. For this application, we will use the one-period lagged value of consumption and one- and two-period lagged values of income as instrumental variables . Table 8.2 reports the nonlinear least squares and instrumental variables estimates. Because we are using two periods of lagged values, two observations are lost. Thus, the least squares estimates are not the same as those reported earlier.

The instrumental variable estimates differ considerably from the least squares estimates. The differences can be deceiving, however. Recall that the MPC in the model is  $\beta\gamma Y^{\gamma-1}$ . The 2000.4 value for *DPI* that we examined earlier was 6634.9. At this value, the instrumental variables and least squares estimates of the MPC are 1.1543 with an estimated standard

CHAPTER 8 ♦ Endogeneity and Instrumental Variable **249****TABLE 8.2** Nonlinear Least Squares and Instrumental Variable Estimates

<b>Parameter</b>	<b>Instrumental Variables</b>		<b>Least Squares</b>	
	<b>Estimate</b>	<b>Standard Error</b>	<b>Estimate</b>	<b>Standard Error</b>
$\alpha$	627.031	26.6063	468.215	22.788
$\beta$	0.040291	0.006050	0.0971598	0.01064
$\gamma$	1.34738	0.016816	1.24892	0.1220
$\sigma$	57.1681	—	49.87998	—
$\mathbf{e}'\mathbf{e}$	650,369.805	—	495,114.490	—

error of 0.01234 and 1.08406 with an estimated standard error of 0.008694, respectively. These values do differ a bit, but less than the quite large differences in the parameters might have led one to expect. We do note that the IV estimate is considerably greater than the estimate in the linear model, 0.9217 (and greater than one, which seems a bit implausible).

## 8.7 WEAK INSTRUMENTS

Our analysis thus far has focused on the “identification” condition for IV estimation, that is, the “exogeneity assumption,” A.I9, which produces

$$\text{plim} (1/n) \mathbf{Z}' \boldsymbol{\epsilon} = \mathbf{0}. \quad (8-28)$$

Taking the “relevance” assumption,

$$\text{plim} (1/n) \mathbf{Z}' \mathbf{X} = \mathbf{Q}_{\mathbf{Z}\mathbf{X}}, \text{ a finite, nonzero, } L \times K \text{ matrix with rank } K, \quad (8-29)$$

as given produces a consistent IV estimator. In absolute terms, with (8-28) in place, (8-29) is sufficient to assert consistency. As such, researchers have focused on *exogeneity* as the defining problem to be solved in constructing the IV estimator. A growing literature has argued that greater attention needs to be given to the relevance condition. While strictly speaking, (8-29) is indeed sufficient for the asymptotic results we have claimed, the common case of “weak instruments,” in which (8-29) is only barely true has attracted considerable scrutiny. In practical terms, instruments are “weak” when they are only slightly correlated with the right-hand-side variables,  $\mathbf{X}$ ; that is,  $(1/n) \mathbf{Z}' \mathbf{X}$  is close to zero. (We will quantify this theoretically when we revisit the issue in Chapter 10.) Researchers have begun to examine these cases, finding in some an explanation for perverse and contradictory empirical results.<sup>10</sup>

Superficially, the problem of weak instruments shows up in the asymptotic covariance matrix of the IV estimator,

$$\text{Asy. Var}[\mathbf{b}_{\text{IV}}] = \frac{\sigma_{\boldsymbol{\epsilon}}^2}{n} \left[ \left( \frac{\mathbf{Z}' \mathbf{Z}}{n} \right) \left( \frac{\mathbf{Z}' \mathbf{Z}}{n} \right)^{-1} \left( \frac{\mathbf{Z}' \mathbf{X}}{n} \right) \right]^{-1},$$

which will be “large” when the instruments are weak, and, other things equal, larger the weaker they are. However, the problems run deeper than that. Nelson and Startz

<sup>10</sup>Important references are Nelson and Startz (1990a,b), Staiger and Stock (1997), Stock, Wright, and Yogo (2002), Hahn and Hausman (2002, 2003), Kleibergen (2002), Stock and Yogo (2005), and Hausman, Stock, and Yogo (2005).

## 250 PART I ♦ The Linear Regression Model

(1990a,b) and Hahn and Hausman (2003) list two implications: (i) The two-stage least squares estimator is badly biased toward the ordinary least squares estimator, which is known to be inconsistent, and (ii) the standard first-order asymptotics (such as those we have used in the preceding) will not give an accurate framework for statistical inference. Thus, the problem is worse than simply lack of precision. There is also at least some evidence that the issue goes well beyond “small sample problems.” [See Bound, Jaeger, and Baker (1995).]

Current research offers several prescriptions for detecting weakness in instrumental variables. For a single endogenous variable ( $\mathbf{x}$  that is correlated with  $\boldsymbol{\epsilon}$ ), the standard approach is based on the first-step least squares regression of two-stage least squares. The conventional  $F$  statistic for testing the hypothesis that all the coefficients in the regression

$$x_i = \mathbf{Z}'_i \boldsymbol{\pi} + v_i$$

are zero is used to test the “hypothesis” that the instruments are weak. An  $F$  statistic less than 10 signals the problem. [See Nelson and Startz (1990b), Staiger and Stock (1997), and Stock and Watson (2007, Chapter 12) for motivation of this specific test.] When there are more than one endogenous variable in the model, testing each one separately using this test is not sufficient, since collinearity among the variables could impact the result but would not show up in either test. Shea (1997) proposes a four-step multivariate procedure that can be used. Godfrey (1999) derived a surprisingly simple alternative method of doing the computation. For endogenous variable  $k$ , the Godfrey statistic is the ratio of the estimated variances of the two estimators, OLS and 2SLS,

$$R_k^2 = \frac{v_k(\text{OLS})/\mathbf{e}'\mathbf{e}(\text{OLS})}{v_k(\text{2SLS})/\mathbf{e}'\mathbf{e}(\text{2SLS})}$$

where  $v_k(\text{OLS})$  is the  $k$ th diagonal element of  $[\mathbf{e}'\mathbf{e}(\text{OLS})/(n-K)](\mathbf{X}'\mathbf{X})^{-1}$  and  $v_k(\text{2SLS})$  is defined likewise. With the scalings, the statistic reduces to

$$R_k^2 = \frac{(\mathbf{X}'\mathbf{X})^{kk}}{(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{kk}}$$

where the superscript indicates the element of the inverse matrix. The  $F$  statistic can then be based on this measure;  $F = [R_k^2/(L-1)]/[(1-R_k^2)/(n-L)]$  assuming that  $\mathbf{Z}$  contains a constant term.

It is worth noting that the test for weak instruments is not a specification test, nor is it a constructive test for building the model. Rather, it is a strategy for helping the researcher avoid basing inference on unreliable statistics whose properties are not well represented by the familiar asymptotic results, for example, distributions under assumed null model specifications. Several extensions are of interest. Other statistical procedures are proposed in Hahn and Hausman (2002) and Kleibergen (2002). We are also interested in cases of more than a single endogenous variable. We will take another look at this issue in Chap. 10, where we can cast the modeling framework as a simultaneous equations model.

The stark results of this section call the IV estimator into question. In a fairly narrow circumstance, an alternative estimator is the “moment”-free LIML estimator discussed in the next chapter. Another, perhaps somewhat unappealing, approach is to revert to least squares. The OLS estimator is not without virtue. The asymptotic variance of the

CHAPTER 8 ♦ Endogeneity and Instrumental Variable **251**

OLS estimator

$$\text{Asy. Var}[\mathbf{b}_{\text{LS}}] = (\sigma^2/n) \mathbf{Q}_{\mathbf{XX}}^{-1}$$

is unambiguously smaller than the asymptotic variance of the IV estimator

$$\text{Asy. Var}[\mathbf{b}_{\text{IV}}] = (\sigma^2/n) (\mathbf{Q}_{\mathbf{XZ}} \mathbf{Q}_{\mathbf{ZZ}}^{-1} \mathbf{Q}_{\mathbf{ZX}})^{-1}.$$

(The proof is considered in the exercises.) Given the preceding results, it could be far smaller. The OLS estimator is inconsistent, however,

$$\text{plim } \mathbf{b}_{\text{LS}} - \boldsymbol{\beta} = \mathbf{Q}_{\mathbf{XX}}^{-1} \boldsymbol{\gamma}$$

[see (8-4)]. By a mean squared error comparison, it is unclear whether the OLS estimator with

$$M(\mathbf{b}_{\text{LS}} | \boldsymbol{\beta}) = (\sigma^2/n) \mathbf{Q}_{\mathbf{XX}}^{-1} + \mathbf{Q}_{\mathbf{XX}}^{-1} \boldsymbol{\gamma} \boldsymbol{\gamma}' \mathbf{Q}_{\mathbf{XX}}^{-1},$$

or the IV estimator, with

$$M(\mathbf{b}_{\text{IV}} | \boldsymbol{\beta}) = (\sigma^2/n) (\mathbf{Q}_{\mathbf{XZ}} \mathbf{Q}_{\mathbf{ZZ}}^{-1} \mathbf{Q}_{\mathbf{ZX}})^{-1},$$

is more precise. The natural recourse in the face of weak instruments is to drop the endogenous variable from the model or improve the instrument set. Each of these is a specification issue. Strictly in terms of estimation strategy within the framework of the data and specification in hand, there is scope for OLS to be the preferred strategy.

## 8.8 NATURAL EXPERIMENTS AND THE SEARCH FOR CAUSAL EFFECTS

Econometrics and statistics have historically been taught, understood, and operated under the credo that “correlation is not causation.” But, much of the still-growing field of microeconomics and some of what we have done in this chapter have been advanced as “causal modeling.”<sup>11</sup> In the contemporary literature on treatment effects and program evaluation, the point of the econometric exercise really is to establish more than mere statistical association—in short, the answer to the question “does the program *work*?” requires an econometric response more committed than “the data seem to be consistent with that hypothesis.” A cautious approach to econometric modeling has nonetheless continued to base its view of “causality” essentially on statistical grounds.<sup>12</sup>

An example of the sort of causal model considered here is a structural equation such as Krueger and Dale’s (1999) model for earnings attainment and elite college attendance,

$$\ln Earnings = \mathbf{x}' \boldsymbol{\beta} + \delta T + \varepsilon,$$

<sup>11</sup>See, for example, Chapter 2 of Cameron and Trivedi (2005), which is entitled “Causal and Noncausal Models” and, especially, Angrist, Imbens, and Rubin (1996), Angrist and Krueger (2001), and Angrist and Pischke (2009, 2010).

<sup>12</sup>See, among many recent commentaries on this line of inquiry, Heckman and Vytlacil (2007).

**252 PART I ♦ The Linear Regression Model**

in which  $\delta$  is the “causal effect” of attendance at an elite college. In this model,  $T$  cannot vary autonomously, outside the model. Variation in  $T$  is determined partly by the same hidden influences that determine lifetime earnings. Though a causal effect can be attributed to  $T$ , measurement of that effect,  $\delta$ , cannot be done with multiple linear regression. The technique of linear instrumental variables estimation has evolved as a mechanism for disentangling causal influences. As does least squares regression, the method of instrumental variables must be defended against the possibility that the underlying statistical relationships uncovered could be due to “something else.” But, when the instrument is the outcome of a “natural experiment,” true exogeneity is claimed. It is this purity of the result that has fueled the enthusiasm of the most strident advocates of this style of investigation. The power of the method lends an inevitability and stability to the findings. This has produced a willingness of contemporary researchers to step beyond their cautious roots.<sup>13</sup> Example 8.11 describes a recent controversial contribution to this literature. On the basis of a natural experiment, the authors identify a cause-and-effect relationship that would have been viewed as beyond the reach of regression modeling under earlier paradigms.<sup>14</sup>

**Example 8.11 Does Television Cause Autism?**

The following is the abstract of economists Waldman, Nicholson and Adilov's (2008) study of autism.<sup>15</sup>

Autism is currently estimated to affect approximately one in every 166 children, yet the cause or causes of the condition are not well understood. One of the current theories concerning the condition is that among a set of children vulnerable to developing the condition because of their underlying genetics, the condition manifests itself when such a child is exposed to a (currently unknown) environmental trigger. In this paper we empirically investigate the hypothesis that early childhood television viewing serves as such a trigger. Using the Bureau of Labor Statistics' American Time Use Survey, we first establish that the amount of television a young child watches is positively related to the amount of precipitation in the child's community. This suggests that, if television is a trigger for autism, then autism should be more prevalent in communities that receive substantial precipitation. We then look at county-level autism data for three states—California, Oregon, and Washington—characterized by high precipitation variability. Employing a variety of tests, we show that in each of the three states (and across all three states when pooled) there is substantial evidence that county autism rates are indeed positively related to county-wide levels of precipitation. In our final set of tests we use California and Pennsylvania data on children born between 1972 and 1989 to show, again consistent with the television as trigger hypothesis, that county autism rates are also positively related to the percentage of households that subscribe to cable television. *Our precipitation tests indicate that just under forty percent of autism diagnoses in the three states studied is the result of television watching due to precipitation, while our cable tests indicate that approximately seventeen percent of the growth in autism in California and Pennsylvania during the 1970s and 1980s is due to the growth of cable television. These findings are consistent with early childhood television viewing being an important trigger for autism.* (Emphasis added.) We also discuss further tests that can be conducted to explore the hypothesis more directly.

<sup>13</sup>See, e.g., Angrist and Pischke (2009, 2010). In reply, Keane (2010, p. 48) opines “What has always bothered me about the ‘experimentalist’ school is the false sense of certainty it conveys. The basic idea is that if we [have] a ‘really good instrument,’ we can come up with ‘convincing’ estimates of ‘causal effects’ that are not too sensitive to assumptions.”

<sup>14</sup>See the symposium in the Spring 2010 *Journal of Economic Perspectives*, Angrist and Pischke (2010), Leamer (2010), Sims (2010), Keane (2010), Stock (2010), and Nevo and Whinston (2010).

<sup>15</sup>Extracts from <http://www.johnson.cornell.edu/faculty/profiles/waldman/autism-waldman-nicholson-adilov.pdf>.

CHAPTER 8 ♦ Endogeneity and Instrumental Variable **253**

The authors add (at page 3), “Although consistent with the hypothesis that early childhood television watching is an important trigger for autism, our first main finding is also consistent with another possibility. Specifically, since precipitation is likely correlated with young children spending more time indoors generally, not just young children watching more television, our first main finding could be due to any indoor toxin. *Therefore, we also employ a second instrumental variable or natural experiment, that is correlated with early childhood television watching but unlikely to be substantially correlated with time spent indoors.*” (Emphasis added.) They conclude (on pages 39-40): “Using the results found in Table 3’s pooled cross-sectional analysis of California, Oregon, and Washington’s county-level autism rates, we find that if early childhood television watching is the sole trigger driving the positive correlation between autism and precipitation then thirty-eight percent of autism diagnoses are due to the incremental television watching due to precipitation.”

Waldman, Nicholson and Adilov’s (2008)<sup>16</sup> study provoked an intense and widespread response among academics, autism researchers, and the public. Whitehouse (2007) surveyed some of the discussion, which touches upon the methodological implications of the search for “causal effects” in econometric research:

Prof. Waldman’s willingness to hazard an opinion on a delicate matter of science reflects the growing ambition of economists—and also their growing hubris, in the view of critics. Academic economists are increasingly venturing beyond their traditional stomping ground, a wanderlust that has produced some powerful results but also has raised concerns about whether they’re sometimes going too far.

Such debates are likely to grow as economists delve into issues in education, politics, history and even epidemiology. Prof. Waldman’s use of precipitation illustrates one of the tools that has emboldened them: the instrumental variable, a statistical method that, by introducing some random or natural influence, helps economists sort out questions of cause and effect. Using the technique, they can create “natural experiments” that seek to approximate the rigor of randomized trials—the traditional gold standard of medical research.

Instrumental variables have helped prominent researchers shed light on sensitive topics. Joshua Angrist of the Massachusetts Institute of Technology has studied the cost of war, the University of Chicago’s Steven Levitt has examined the effect of adding police on crime, and Harvard’s Caroline Hoxby has studied school performance. Their work has played an important role in public-policy debates. But as enthusiasm for the approach has grown, so too have questions. One concern: When economists use one variable as a proxy for another—rainfall patterns instead of TV viewing, for example—it’s not always clear what the results actually measure. Also, the experiments on their own offer little insight into why one thing affects another. “There’s a saying that ignorance is bliss,” says James Heckman, an economics professor at the University of Chicago who won a Nobel Prize in 2000 for his work on statistical methods. “I think that characterizes a lot of the enthusiasm for these instruments.” Says MIT economist Jerry Hausman, “If your instruments aren’t perfect, you could go seriously wrong.

<sup>16</sup>Published as NBER working paper 12632 in 2006.

**254 PART I ♦ The Linear Regression Model****Example 8.12 Is Season of Birth a Valid Instrument?**

Buckles and Hungerman (BH, 2008) list more than 20 studies of long-term economic outcomes that use season of birth as an instrumental variable, beginning with one of the earliest and best known papers in the “natural experiments” literature, Angrist and Krueger (1991). The assertion of the validity of season of birth as a proper instrument is that family background is unrelated to season of birth, but it is demonstrably related to long-term outcomes such as income and education. The assertion justifies using dummy variables for season of birth as instrumental variables in outcome equations. If, on the other hand, season of birth is correlated with family background, then it will “fail the exclusion restriction in most IV settings where it has been used” (BH, page 2). According to the authors, the randomness of quarter of birth over the population [see, e.g., Kleibergen (2002)] has been taken as a given, without scientific investigation of the claim. Using data from live birth certificates and census data, BH found a numerically modest, but statistically significant relationship between birth dates and family background. They found “women giving birth in the winter look different from other women; they are younger, less educated, and less likely to be married.... The fraction of children born to women without a high school degree is about 10 percent higher (2 percentage points) in January than in May ... We also document a 10 percent decline in the fraction of children born to teenagers from January to May.” Precisely why there should be such a relationship remains uncertain. Researchers differ (of course) on the numerical implications of BH’s finding. [See Lahart (2009).] But, the methodological implication of their finding is consistent with Hausman’s observation.

**8.9 SUMMARY AND CONCLUSIONS**

The instrumental variable (IV) estimator, in various forms, is among the most fundamental tools in econometrics. Broadly interpreted, it encompasses most of the estimation methods that we will examine in this book. This chapter has developed the basic results for IV estimation of linear models. The essential departure point is the exogeneity and relevance assumptions that define an instrumental variable. We then analyzed linear IV estimation in the form of the two-stage least squares estimator. With only a few special exceptions related to simultaneous equations models with two variables, almost no finite-sample properties have been established for the IV estimator. (We temper that, however, with the results in Section 8.7 on weak instruments, where we saw evidence that whatever the finite-sample properties of the IV estimator might be, under some well-discernible circumstances, these properties are not attractive.) We then examined the asymptotic properties of the IV estimator for linear and nonlinear regression models. Finally, some cautionary notes about using IV estimators when the instruments are only weakly relevant in the model are examined in Section 8.7.

**Key Terms and Concepts**

- Asymptotic covariance matrix
- Asymptotic distribution
- Attenuation
- Attenuation bias
- Attrition
- Attrition bias
- Consistent estimator
- Dynamic panel data model
- Effect of the treatment on the treated
- Endogenous
- Endogenous treatment effect
- Exogenous
- Hausman statistic
- Identification
- Indicator
- Instrumental variables
- Instrumental variable estimator

## CHAPTER 8 ♦ Endogeneity and Instrumental Variable 255

Limiting distribution  
Measurement error

- Minimum distance estimator
- Moment equations
- Natural experiment
- Nonrandom sampling
- Omitted parameter heterogeneity
- Omitted variables
- Omitted variable bias
- Orthogonality conditions
- Overidentification

- Panel data
- Proxy variable
- Random effects
- Reduced form equation
- Relevance
- Reliability ratio
- Sample selection bias
- Selectivity effect
- Simultaneous equations
- Simultaneous equations bias
- Smearing
- Specification test
- Strongly exogenous
- Structural equation system
- Structural Model
- Structural specification
- Survivorship bias
- Truncation bias
- Two-stage least squares (2SLS)
- Variable addition test
- Weak instruments
- Weakly exogenous
- Wu test

**Exercises**

1. In the discussion of the instrumental variable estimator, we showed that the least squares estimator  $\mathbf{b}_{LS}$  is biased and inconsistent. Nonetheless,  $\mathbf{b}_{LS}$ 's estimate something—see derive the asymptotic covariance matrix of  $\mathbf{b}_{LS}$  and show that  $\mathbf{b}_{LS}$  is asymptotically normally distributed.
2. For the measurement error model in (8-14) and (8-15), prove that when only  $x$  is measured with error, the squared correlation between  $y$  and  $x$  is less than that between  $y^*$  and  $x^*$ . (Note the assumption that  $y^* = y$ .) Does the same hold true if  $y^*$  is also measured with error?
3. Derive the results in (8-20a) and (8-20b) for the measurement error model. Note the hint in footnote 4 in Section 8.5.1 that suggests you use result (A-66) when you need to invert

$$[\mathbf{Q}^* + \Sigma_{uu}] = [\mathbf{Q}^* + (\sigma_u \mathbf{e}_L \mathbf{e}_L' \sigma_u \mathbf{e}_1)'].$$

4. At the end of Section 8.7, it is suggested that the OLS estimator could have a smaller mean squared error than the 2SLS estimator. Using (8-4), the results of Exercise 1, and Theorem 8.1, show that the result will be true if

$$\mathbf{Q}_{XX} - \mathbf{Q}_{XZ} \mathbf{Q}_{ZZ}^{-1} \mathbf{Q}_{ZX} >> \frac{1}{(\sigma^2/n) + \gamma' \mathbf{Q}_{XX}^{-1} \gamma} \gamma \gamma'.$$

How can you verify that this is at least possible? The right-hand side is a rank one, nonnegative definite matrix. What can be said about the left-hand side?

5. Consider the linear model  $y_i = \alpha + \beta x_i + \varepsilon_i$  in which  $\text{Cov}[x_i, \varepsilon_i] = \gamma \neq 0$ . Let  $z$  be an exogenous, relevant instrumental variable for this model. Assume, as well, that  $z$  is binary—it takes only values 1 and 0. Show the algebraic forms of the LS estimator and the IV estimator for both  $\alpha$  and  $\beta$ .
6. In the discussion of the instrumental variables estimator, we showed that the least squares estimator  $\mathbf{b}$  is biased and inconsistent. Nonetheless,  $\mathbf{b}$  does estimate something:  $\text{plim } \mathbf{b} = \theta = \beta + \mathbf{Q}^{-1} \gamma$ . Derive the asymptotic covariance matrix of  $\mathbf{b}$ , and show that  $\mathbf{b}$  is asymptotically normally distributed.

**256 PART I ♦ The Linear Regression Model****Application**

1. In Example 8.5, we have suggested a model of a labor market. From the “reduced form” equation given first, you can see the full set of variables that appears in the model—that is the “endogenous variables,”  $\ln Wage_{it}$ , and  $Wks_{it}$ , and all other exogenous variables. The labor supply equation suggested next contains these two variables and three of the exogenous variables. From these facts, you can deduce what variables would appear in a labor “demand” equation for  $\ln Wage_{it}$ . Assume (for purpose of our example) that  $\ln Wage_{it}$  is determined by  $Wks_{it}$  and the remaining appropriate exogenous variables. (We should emphasize that this exercise is purely to illustrate the computations—the structure here would not provide a theoretically sound model for labor market equilibrium.)
  - a. What is the labor demand equation implied?
  - b. Estimate the parameters of this equation by OLS and by 2SLS and compare the results.
  - c. Are the instruments used in this equation relevant? How do you know?



Explore the panel nature of the data set. Just pool the data.)

## 9

# THE GENERALIZED REGRESSION MODEL AND HETEROSCEDASTICITY

---

## 9.1 INTRODUCTION

In this and the next several chapters, we will extend the multiple regression model to disturbances that violate Assumption A.4 of the classical regression model. The **generalized linear regression model** is

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\ E[\boldsymbol{\varepsilon} | \mathbf{X}] &= \mathbf{0}, \\ E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}] &= \sigma^2\boldsymbol{\Omega} = \Sigma, \end{aligned} \tag{9-1}$$

where  $\boldsymbol{\Omega}$  is a positive definite matrix. (The covariance matrix is written in the form  $\sigma^2\boldsymbol{\Omega}$  at several points so that we can obtain the classical model,  $\sigma^2\mathbf{I}$ , as a convenient special case.)

The two leading cases we will consider in detail are **heteroscedasticity** and **autocorrelation**. Disturbances are heteroscedastic when they have different variances. Heteroscedasticity arises in volatile high-frequency time-series data such as daily observations in financial markets and in cross-section data where the scale of the dependent variable and the explanatory power of the model tend to vary across observations. Microeconomic data such as expenditure surveys are typical. The disturbances are still assumed to be uncorrelated across observations, so  $\sigma^2\boldsymbol{\Omega}$  would be

$$\sigma^2\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ & \vdots & & \\ 0 & 0 & \cdots & \omega_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}.$$

(The first mentioned situation involving financial data is more complex than this and is examined in detail in Chapter 20.)

Autocorrelation is usually found in time-series data. Economic time series often display a “memory” in that variation around the regression function is not independent from one period to the next. The seasonally adjusted price and quantity series published by government agencies are examples. Time-series data are usually homoscedastic, so  $\sigma^2\boldsymbol{\Omega}$  might be

$$\sigma^2\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ & & \ddots & \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{bmatrix}.$$

## 258 PART II ♦ Generalized Regression Model and Equation Systems

The values that appear off the diagonal depend on the model used for the disturbance. In most cases, consistent with the notion of a fading memory, the values decline as we move away from the diagonal.

**Panel data** sets, consisting of cross sections observed at several points in time, may exhibit both characteristics. We shall consider them in Chapter 11. This chapter presents some general results for this extended model. We will examine the model of heteroscedasticity in this chapter and in Chapter 14. A general model of autocorrelation appears in Chapter 20. Chapters 10 and 11 examine in detail specific types of generalized regression models.

Our earlier results for the classical model will have to be modified. We will take the following approach on general results and in the specific cases of heteroscedasticity and serial correlation:

1. We first consider the consequences for the least squares estimator of the more general form of the regression model. This will include assessing the effect of ignoring the complication of the generalized model and of devising an appropriate estimation strategy, still based on least squares.
2. We will examine alternative estimation approaches that can make better use of the characteristics of the model. Minimal assumptions about  $\Omega$  are made at this point.
3. We then narrow the assumptions and begin to look for methods of detecting the failure of the classical model—that is, we formulate procedures for testing the specification of the classical model against the generalized regression.
4. The final step in the analysis is to formulate **parametric models** that make specific assumptions about  $\Omega$ . Estimators in this setting are some form of generalized least squares or maximum likelihood which is developed in Chapter 14.

The model is examined in general terms in this chapter. Major applications to panel data and multiple equation systems are considered in Chapters 11 and 10, respectively.

### 9.2 INEFFICIENT ESTIMATION BY LEAST SQUARES AND INSTRUMENTAL VARIABLES

The essential results for the classical model with **spherical disturbances**

$$E[\epsilon | \mathbf{X}] = \mathbf{0}$$

and

$$E[\epsilon\epsilon' | \mathbf{X}] = \sigma^2 \mathbf{I} \quad (9-2)$$

are presented in Chapters 2 through 6. To reiterate, we found that the **ordinary least squares (OLS) estimator**

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon \quad (9-3)$$

is best linear unbiased (BLU), consistent and asymptotically normally distributed (CAN), and if the disturbances are normally distributed, like other maximum likelihood estimators considered in Chapter 14, asymptotically efficient among all CAN estimators. We now consider which of these properties continue to hold in the model of (9-1).

To summarize, the least squares estimators retain only some of their desirable properties in this model. Least squares remains unbiased, consistent, and asymptotically



## CHAPTER 9 ♦ The Generalized Regression Model 259

normally distributed. It will, however, no longer be efficient—this claim remains to be verified—and the usual inference procedures are no longer appropriate.

### 9.2.1 FINITE-SAMPLE PROPERTIES OF ORDINARY LEAST SQUARES

By taking expectations on both sides of (9-3), we find that if  $E[\boldsymbol{\epsilon} | \mathbf{X}] = \mathbf{0}$ , then

$$E[\mathbf{b}] = E_{\mathbf{X}}[E[\mathbf{b} | \mathbf{X}]] = \boldsymbol{\beta}. \quad (9-4)$$

Therefore, we have the following theorem.

#### **THEOREM 9.1** Finite-Sample Properties of $\mathbf{b}$ in the Generalized Regression Model

*If the regressors and disturbances are uncorrelated, then the unbiasedness of least squares is unaffected by violations of assumption (9-2). The least squares estimator is unbiased in the generalized regression model. With nonstochastic regressors, or conditional on  $\mathbf{X}$ , the sampling variance of the least squares estimator is*

$$\begin{aligned} \text{Var}[\mathbf{b} | \mathbf{X}] &= E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})' | \mathbf{X}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} | \mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\Omega)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \frac{\sigma^2}{n} \left( \frac{1}{n}\mathbf{X}'\mathbf{X} \right)^{-1} \left( \frac{1}{n}\mathbf{X}'\Omega\mathbf{X} \right) \left( \frac{1}{n}\mathbf{X}'\mathbf{X} \right)^{-1}. \end{aligned} \quad (9-5)$$

*If the regressors are stochastic, then the unconditional variance is  $E_{\mathbf{X}}[\text{Var}[\mathbf{b} | \mathbf{X}]]$ . In (9-3),  $\mathbf{b}$  is a linear function of  $\boldsymbol{\epsilon}$ . Therefore, if  $\boldsymbol{\epsilon}$  is normally distributed, then*

$$\mathbf{b} | \mathbf{X} \sim N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Omega\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}].$$

The end result is that  $\mathbf{b}$  has properties that are similar to those in the classical regression case. Because the variance of the least squares estimator is not  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ , however, statistical inference based on  $s^2(\mathbf{X}'\mathbf{X})^{-1}$  may be misleading. Not only is this the wrong matrix to be used, but  $s^2$  may be a biased estimator of  $\sigma^2$ . There is usually no way to know whether  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$  is larger or smaller than the true variance of  $\mathbf{b}$ , so even with a good estimator of  $\sigma^2$ , the conventional estimator of  $\text{Var}[\mathbf{b} | \mathbf{X}]$  may not be particularly useful. Finally, because we have dispensed with the fundamental underlying assumption, the familiar inference procedures based on the  $F$  and  $t$  distributions will no longer be appropriate. One issue we will explore at several points following is how badly one is likely to go awry if the result in (9-5) is ignored and if the use of the familiar procedures based on  $s^2(\mathbf{X}'\mathbf{X})^{-1}$  is continued.

### 9.2.2 ASYMPTOTIC PROPERTIES OF ORDINARY LEAST SQUARES

If  $\text{Var}[\mathbf{b} | \mathbf{X}]$  converges to zero, then  $\mathbf{b}$  is mean square consistent. With well-behaved regressors,  $(\mathbf{X}'\mathbf{X}/n)^{-1}$  will converge to a constant matrix. But  $(\sigma^2/n)(\mathbf{X}'\Omega\mathbf{X}/n)$  need

## 260 PART II ♦ Generalized Regression Model and Equation Systems

not converge at all. By writing this product as

$$\frac{\sigma^2}{n} \left( \frac{\mathbf{X}' \boldsymbol{\Omega} \mathbf{X}}{n} \right) = \left( \frac{\sigma^2}{n} \right) \left( \frac{\sum_{i=1}^n \sum_{j=1}^n \omega_{ij} \mathbf{x}_i \mathbf{x}'_j}{n} \right) \quad (9-6)$$

we see that though the leading constant will, by itself, converge to zero, the matrix is a sum of  $n^2$  terms, divided by  $n$ . Thus, the product is a scalar that is  $O(1/n)$  times a matrix that is, at least at this juncture,  $O(n)$ , which is  $O(1)$ . So, it does appear at first blush that if the product in (9-6) does converge, it might converge to a matrix of nonzero constants. In this case, the covariance matrix of the least squares estimator would not converge to zero, and consistency would be difficult to establish. We will examine in some detail, the conditions under which the matrix in (9-6) converges to a constant matrix.<sup>1</sup> If it does, then because  $\sigma^2/n$  does vanish, ordinary least squares is consistent as well as unbiased.

### THEOREM 9.2 Consistency of OLS in the Generalized Regression Model

If  $\mathbf{Q} = \text{plim}(\mathbf{X}'\mathbf{X}/n)$  and  $\text{plim}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}/n)$  are both finite positive definite matrices, then  $\mathbf{b}$  is consistent for  $\boldsymbol{\beta}$ . Under the assumed conditions,

$$\text{plim } \mathbf{b} = \boldsymbol{\beta}.$$

The conditions in Theorem 9.2 depend on both  $\mathbf{X}$  and  $\boldsymbol{\Omega}$ . An alternative formula<sup>2</sup> that separates the two components is as follows. Ordinary least squares is consistent in the generalized regression model if:

1. The smallest characteristic root of  $\mathbf{X}'\mathbf{X}$  increases without bound as  $n \rightarrow \infty$ , which implies that  $\text{plim}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0}$ . If the regressors satisfy the Grenander conditions **G1** through **G3** of Section 4.4.1, Table 4.2, then they will meet this requirement.
2. The largest characteristic root of  $\boldsymbol{\Omega}$  is finite for all  $n$ . For the heteroscedastic model, the variances are the characteristic roots, which requires them to be finite. For models with autocorrelation, the requirements are that the elements of  $\boldsymbol{\Omega}$  be finite and that the off-diagonal elements not be too large relative to the diagonal elements. We will examine this condition at several points below.

The least squares estimator is asymptotically normally distributed if the limiting distribution of

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) = \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \frac{1}{\sqrt{n}} \mathbf{X}' \boldsymbol{\varepsilon} \quad (9-7)$$

is normal. If  $\text{plim}(\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}$ , then the limiting distribution of the right-hand side is the same as that of

$$\mathbf{v}_{n,LS} = \mathbf{Q}^{-1} \frac{1}{\sqrt{n}} \mathbf{X}' \boldsymbol{\varepsilon} = \mathbf{Q}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i \boldsymbol{\varepsilon}_i, \quad (9-8)$$

<sup>1</sup>In order for the product in (9-6) to vanish, it would be sufficient for  $(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}/n)$  to be  $O(n^\delta)$  where  $\delta < 1$ .

<sup>2</sup>Amemiya (1985, p. 184).

CHAPTER 9 ♦ The Generalized Regression Model **261**

where  $\mathbf{x}'_i$  is a row of  $\mathbf{X}$  (assuming, of course, that the limiting distribution exists at all). The question now is whether a central limit theorem can be applied directly to  $\mathbf{v}$ . If the disturbances are merely heteroscedastic and still uncorrelated, then the answer is generally yes. In fact, we already showed this result in Section 4.4.2 when we invoked the Lindeberg–Feller central limit theorem (D.19) or the Lyapounov theorem (D.20). The theorems allow unequal variances in the sum. The exact variance of the sum is

$$E_{\mathbf{x}} \left[ \text{Var} \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i \boldsymbol{\varepsilon}_i \right] \middle| \mathbf{x}'_i \right] = \frac{\sigma^2}{n} \sum_{i=1}^n \omega_i \mathbf{Q}_i,$$

which, for our purposes, we would require to converge to a positive definite matrix. In our analysis of the classical model, the heterogeneity of the variances arose because of the regressors, but we still achieved the limiting normal distribution in (4-27) through (4-33). All that has changed here is that the variance of  $\boldsymbol{\varepsilon}$  varies across observations *as well*. Therefore, *the proof of asymptotic normality in Section 4.4.2 is general enough to include this model without modification*. As long as  $\mathbf{X}$  is well behaved and the diagonal elements of  $\Omega$  are finite and well behaved, the least squares estimator is asymptotically normally distributed, with the covariance matrix given in (9-5). That is;

*In the heteroscedastic case, if the variances of  $\varepsilon_i$  are finite and are not dominated by any single term, so that the conditions of the Lindeberg–Feller central limit theorem apply to  $\mathbf{v}_{n,LS}$  in (9-8), then the least squares estimator is asymptotically normally distributed with covariance matrix*

$$\text{Asy. Var}[\mathbf{b}] = \frac{\sigma^2}{n} \mathbf{Q}^{-1} \text{plim} \left( \frac{1}{n} \mathbf{X}' \Omega \mathbf{X} \right) \mathbf{Q}^{-1}. \quad (9-9)$$

For the most general case, asymptotic normality is much more difficult to establish because the sums in (9-8) are not necessarily sums of independent or even uncorrelated random variables. Nonetheless, Amemiya (1985, p. 187) and Anderson (1971) have established the asymptotic normality of  $\mathbf{b}$  in a model of autocorrelated disturbances general enough to include most of the settings we are likely to meet in practice. We will revisit this issue in Chapters 20 and 21 when we examine time-series modeling. We can conclude that, except in particularly unfavorable cases, we have the following theorem.

**THEOREM 9.3 Asymptotic Distribution of  $\mathbf{b}$  in the GR Model**

*If the regressors are sufficiently well behaved and the off-diagonal terms in  $\Omega$  diminish sufficiently rapidly, then the least squares estimator is asymptotically normally distributed with mean  $\beta$  and covariance matrix given in (9-9).*

### 9.2.3 ROBUST ESTIMATION OF ASYMPTOTIC COVARIANCE MATRICES

There is a remaining question regarding all the preceding results. In view of (9-5), is it necessary to discard ordinary least squares as an estimator? Certainly if  $\Omega$  is known, then, as shown in Section 9.6.1, there is a simple and efficient estimator available based

## 262 PART II ♦ Generalized Regression Model and Equation Systems

on it, and the answer is yes. If  $\Omega$  is unknown, but its structure is known and we can estimate  $\Omega$  using sample information, then the answer is less clear-cut. In many cases, basing estimation of  $\beta$  on some alternative procedure that uses an  $\hat{\Omega}$  will be preferable to ordinary least squares. This subject is covered in Chapters 10 and 11. The third possibility is that  $\Omega$  is completely unknown, both as to its structure and the specific values of its elements. In this situation, least squares or instrumental variables may be the only estimator available, and as such, the only available strategy is to try to devise an estimator for the appropriate asymptotic covariance matrix of  $\mathbf{b}$ .

If  $\sigma^2\Omega$  were known, then the *estimator* of the asymptotic covariance matrix of  $\mathbf{b}$  in (9-10) would be

$$\mathbf{V}_{OLS} = \frac{1}{n} \left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1} \left( \frac{1}{n} \mathbf{X}' [\sigma^2 \Omega] \mathbf{X} \right) \left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1}.$$

The matrix of sums of squares and cross products in the left and right matrices are sample data that are readily estimable. The problem is the center matrix that involves the unknown  $\sigma^2\Omega$ . For estimation purposes, note that  $\sigma^2$  is not a separate unknown parameter. Because  $\Omega$  is an unknown matrix, it can be scaled arbitrarily, say, by  $\kappa$ , and with  $\sigma^2$  scaled by  $1/\kappa$ , the same product remains. In our applications, we will remove the indeterminacy by assuming that  $\text{tr}(\Omega) = n$ , as it is when  $\sigma^2\Omega = \sigma^2\mathbf{I}$  in the classical model. For now, just let  $\Sigma = \sigma^2\Omega$ . It might seem that to estimate  $(1/n)\mathbf{X}'\Sigma\mathbf{X}$ , an estimator of  $\Sigma$ , which contains  $n(n+1)/2$  unknown parameters, is required. But fortunately (because with  $n$  observations, this method is going to be hopeless), this observation is not quite right. What is required is an estimator of the  $K(K+1)/2$  unknown elements in the matrix

$$\text{plim } \mathbf{Q}_* = \text{plim} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \sigma_{ij} \mathbf{x}_i \mathbf{x}_j'.$$

The point is that  $\mathbf{Q}_*$  is a matrix of sums of squares and cross products that involves  $\sigma_{ij}$  and the rows of  $\mathbf{X}$ . The least squares estimator  $\mathbf{b}$  is a consistent estimator of  $\beta$ , which implies that the least squares residuals  $e_i$  are “pointwise” consistent estimators of their population counterparts  $\varepsilon_i$ . The general approach, then, will be to use  $\mathbf{X}$  and  $\mathbf{e}$  to devise an estimator of  $\mathbf{Q}_*$ .

This (perhaps somewhat counterintuitive) principle is exceedingly useful in modern research. Most important applications, including general models of heteroscedasticity, autocorrelation, and a variety of panel data models, can be estimated in this fashion. The payoff is that the estimator frees the analyst from the necessity to assume a particular structure for  $\Omega$ . With tools such as the robust covariance estimator in hand, one of the distinct trends in current research is away from narrow assumptions and toward broad, robust models such as these. The heteroscedasticity and autocorrelation cases are considered in Section 9.4 and Chapter 20, respectively, while several models for panel data are detailed in Chapter 11.

### 9.2.4 INSTRUMENTAL VARIABLE ESTIMATION

Chapter 8 considered cases in which the regressors,  $\mathbf{X}$ , are correlated with the disturbances,  $\mathbf{e}$ . The instrumental variables (IV) estimator developed there enjoys a kind of robustness that least squares lacks in that it achieves consistency whether or not  $\mathbf{X}$  and  $\mathbf{e}$  are correlated, while  $\mathbf{b}$  is neither unbiased nor consistent. However, efficiency was not

CHAPTER 9 ♦ The Generalized Regression Model **263**

a consideration in constructing the IV estimator. We will reconsider the IV estimator here, but since it is inefficient to begin with, there is little to say about the implications of nonspherical disturbances for the efficiency of the estimator, as we examined for  $\mathbf{b}$  in the previous section. As such, the relevant question for us to consider here would be, essentially, does IV still “work” in the generalized regression model? Consistency and asymptotic normality will be the useful properties.

The IV estimator is

$$\begin{aligned}\mathbf{b}_{\text{IV}} &= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \\ &= \boldsymbol{\beta} + [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon},\end{aligned}\quad (9-10)$$

where  $\mathbf{X}$  is the set of  $K$  regressors and  $\mathbf{Z}$  is a set of  $L \geq K$  instrumental variables. We now consider the extension of Theorems 9.2 and 9.3 to the IV estimator when  $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\boldsymbol{\Omega}$ .

Suppose that  $\mathbf{X}$  and  $\mathbf{Z}$  are well behaved as assumed in Section 8.2. That is,

$$\text{plim}(1/n)\mathbf{Z}'\mathbf{Z} = \mathbf{Q}_{\mathbf{Z}\mathbf{Z}}, \text{ a positive definite matrix,}$$

$$\text{plim}(1/n)\mathbf{Z}'\mathbf{X} = \mathbf{Q}_{\mathbf{Z}\mathbf{X}} = \mathbf{Q}'_{\mathbf{X}\mathbf{Z}}, \text{ a nonzero matrix,}$$

$$\text{plim}(1/n)\mathbf{X}'\mathbf{X} = \mathbf{Q}_{\mathbf{X}\mathbf{X}}, \text{ a positive definite matrix.}$$

To avoid a string of matrix computations that may not fit on a single line, for convenience let

$$\begin{aligned}\mathbf{Q}_{\mathbf{X}\mathbf{X},\mathbf{Z}} &= [\mathbf{Q}_{\mathbf{X}\mathbf{Z}}\mathbf{Q}_{\mathbf{Z}\mathbf{Z}}^{-1}\mathbf{Q}_{\mathbf{Z}\mathbf{X}}]^{-1}\mathbf{Q}_{\mathbf{X}\mathbf{Z}}\mathbf{Q}_{\mathbf{Z}\mathbf{Z}}^{-1} \\ &= \text{plim}\left[\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)\right]^{-1}\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1}.\end{aligned}$$

If  $\mathbf{Z}$  is a valid set of instrumental variables, that is, if the second term in (9-10) vanishes asymptotically, then

$$\text{plim } \mathbf{b}_{\text{IV}} = \boldsymbol{\beta} + \mathbf{Q}_{\mathbf{X}\mathbf{X},\mathbf{Z}} \text{ plim}\left(\frac{1}{n}\mathbf{Z}'\boldsymbol{\varepsilon}\right) = \boldsymbol{\beta}.$$

This result is exactly the same one we had before. We might note that at the several points where we have established unbiasedness or consistency of the least squares or instrumental variables estimator, the covariance matrix of the disturbance vector has played no role; unbiasedness is a property of the means. As such, this result should come as no surprise. The large sample behavior of  $\mathbf{b}_{\text{IV}}$  depends on the behavior of

$$\mathbf{v}_{n,\text{IV}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{z}_i \boldsymbol{\varepsilon}_i.$$

This result is exactly the one we analyzed in Section 4.4.2. If the sampling distribution of  $\mathbf{v}_n$  converges to a normal distribution, then we will be able to construct the asymptotic distribution for  $\mathbf{b}_{\text{IV}}$ . This set of conditions is the same that was necessary for  $\mathbf{X}$  when we considered  $\mathbf{b}$  above, with  $\mathbf{Z}$  in place of  $\mathbf{X}$ . We will once again rely on the results of Anderson (1971) or Amemiya (1985) that under very general conditions,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{z}_i \boldsymbol{\varepsilon}_i \xrightarrow{d} \mathbf{N}\left[\mathbf{0}, \sigma^2 \text{plim}\left(\frac{1}{n}\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}\right)\right].$$

With the other results already in hand, we now have the following.

**264 PART II ♦ Generalized Regression Model and Equation Systems**
**THEOREM 9.4 Asymptotic Distribution of the IV Estimator in the Generalized Regression Model**

If the regressors and the instrumental variables are well behaved in the fashions discussed above, then

$$\mathbf{b}_{IV} \xrightarrow{a} N[\boldsymbol{\beta}, \mathbf{V}_{IV}],$$

where

$$\mathbf{V}_{IV} = \frac{\sigma^2}{n} (\mathbf{Q}_{XX,Z}) \text{plim} \left( \frac{1}{n} \mathbf{Z}' \Omega \mathbf{Z} \right) (\mathbf{Q}'_{XX,Z}).$$

 Theorem 9.4 is the instrumental variable estimation counterpart to Theorems 9.2 and 9.3 for least squares.

### 9.3 EFFICIENT ESTIMATION BY GENERALIZED LEAST SQUARES

Efficient estimation of  $\boldsymbol{\beta}$  in the generalized regression model requires knowledge of  $\Omega$ . To begin, it is useful to consider cases in which  $\Omega$  is a known, symmetric, positive definite matrix. This assumption will occasionally be true, though in most models,  $\Omega$  will contain unknown parameters that must also be estimated. We shall examine this case in Section 9.6.2.

#### 9.3.1 GENERALIZED LEAST SQUARES (GLS)

Because  $\Omega$  is a positive definite symmetric matrix, it can be factored into

$$\Omega = \mathbf{C}\Lambda\mathbf{C}',$$

where the columns of  $\mathbf{C}$  are the characteristic vectors of  $\Omega$  and the characteristic roots of  $\Omega$  are arrayed in the diagonal matrix  $\Lambda$ . Let  $\Lambda^{1/2}$  be the diagonal matrix with  $i$ th diagonal element  $\sqrt{\lambda_i}$ , and let  $\mathbf{T} = \mathbf{C}\Lambda^{1/2}$ . Then  $\Omega = \mathbf{T}\mathbf{T}'$ . Also, let  $\mathbf{P}' = \mathbf{C}\Lambda^{-1/2}$ , so  $\Omega^{-1} = \mathbf{P}'\mathbf{P}$ . Premultiply the model in (9-1) by  $\mathbf{P}$  to obtain

$$\mathbf{Py} = \mathbf{PX}\boldsymbol{\beta} + \mathbf{Pe}$$

or

$$\mathbf{y}_* = \mathbf{X}_*\boldsymbol{\beta} + \boldsymbol{\epsilon}_*. \quad (9-11)$$

The conditional variance of  $\boldsymbol{\epsilon}_*$  is

$$E[\boldsymbol{\epsilon}_* \boldsymbol{\epsilon}'_* | \mathbf{X}_*] = \mathbf{P}\sigma^2\Omega\mathbf{P}' = \sigma^2\mathbf{I},$$

so the classical regression model applies to this transformed model. Because  $\Omega$  is assumed to be known,  $\mathbf{y}_*$  and  $\mathbf{X}_*$  are observed data. In the classical model, ordinary least squares is efficient; hence,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'_* \mathbf{X}_*)^{-1} \mathbf{X}'_* \mathbf{y}_* \\ &= (\mathbf{X}' \mathbf{P}' \mathbf{P} \mathbf{X})^{-1} \mathbf{X}' \mathbf{P}' \mathbf{P} \mathbf{y} \\ &= (\mathbf{X}' \Omega^{-1} \mathbf{X})^{-1} \mathbf{X}' \Omega^{-1} \mathbf{y} \end{aligned}$$

CHAPTER 9 ♦ The Generalized Regression Model **265**

is the **efficient estimator** of  $\beta$ . This estimator is the **generalized least squares (GLS)** or Aitken (1935) estimator of  $\beta$ . This estimator is in contrast to the ordinary least squares (OLS) estimator, which uses a “weighting matrix,”  $\mathbf{I}$ , instead of  $\Omega^{-1}$ . By appealing to the classical regression model in (9-11), we have the following theorem, which includes the generalized regression model analogs to our results of Chapter 4:

**THEOREM 9.5 Properties of the Generalized Least Squares Estimator**

If  $E[\boldsymbol{\varepsilon}_* | \mathbf{X}_*] = \mathbf{0}$ , then

$$E[\hat{\beta} | \mathbf{X}_*] = E[(\mathbf{X}'_* \mathbf{X}_*)^{-1} \mathbf{X}'_* \mathbf{y}_* | \mathbf{X}_*] = \beta + E[(\mathbf{X}'_* \mathbf{X}_*)^{-1} \mathbf{X}'_* \boldsymbol{\varepsilon}_* | \mathbf{X}_*] = \beta.$$

The GLS estimator  $\hat{\beta}$  is unbiased. This result is equivalent to  $E[\mathbf{P}\boldsymbol{\varepsilon} | \mathbf{P}\mathbf{X}] = \mathbf{0}$ , but because  $\mathbf{P}$  is a matrix of known constants, we return to the familiar requirement  $E[\boldsymbol{\varepsilon} | \mathbf{X}] = \mathbf{0}$ . The requirement that the regressors and disturbances be uncorrelated is unchanged.

The GLS estimator is consistent if  $\text{plim}(1/n)\mathbf{X}'_* \mathbf{X}_* = \mathbf{Q}_*$ , where  $\mathbf{Q}_*$  is a finite positive definite matrix. Making the substitution, we see that this implies

$$\text{plim}[(1/n)\mathbf{X}' \Omega^{-1} \mathbf{X}]^{-1} = \mathbf{Q}_*^{-1}. \quad (9-12)$$

We require the transformed data  $\mathbf{X}_* = \mathbf{P}\mathbf{X}$ , not the original data  $\mathbf{X}$ , to be well behaved.<sup>3</sup> Under the assumption in (9-1), the following hold:

The GLS estimator is asymptotically normally distributed, with mean  $\beta$  and sampling variance

$$\text{Var}[\hat{\beta} | \mathbf{X}_*] = \sigma^2 (\mathbf{X}'_* \mathbf{X}_*)^{-1} = \sigma^2 (\mathbf{X}' \Omega^{-1} \mathbf{X})^{-1}. \quad (9-13)$$

The GLS estimator  $\hat{\beta}$  is the minimum variance linear unbiased estimator in the generalized regression model. This statement follows by applying the Gauss–Markov theorem to the model in (9-11). The result in Theorem 9.5 is Aitken’s (1935) **theorem**, and  $\hat{\beta}$  is sometimes called the Aitken estimator. This broad result includes the Gauss–Markov theorem as a special case when  $\Omega = \mathbf{I}$ .

For testing hypotheses, we can apply the full set of results in Chapter 5 to the transformed model in (9-11). For testing the  $J$  linear restrictions,  $\mathbf{R}\beta = \mathbf{q}$ , the appropriate statistic is

$$F[J, n - K] = \frac{(\mathbf{R}\hat{\beta} - \mathbf{q})' [\mathbf{R}\hat{\sigma}^2 (\mathbf{X}'_* \mathbf{X}_*)^{-1} \mathbf{R}']^{-1} (\mathbf{R}\hat{\beta} - \mathbf{q})}{J} = \frac{(\hat{\epsilon}'_c \hat{\epsilon}_c - \hat{\epsilon}' \hat{\epsilon})/J}{\hat{\sigma}^2},$$

where the residual vector is

$$\hat{\epsilon} = \mathbf{y}_* - \mathbf{X}_* \hat{\beta}$$

and

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}' \hat{\epsilon}}{n - K} = \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})' \Omega^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta})}{n - K}. \quad (9-14)$$

<sup>3</sup>Once again, to allow a time trend, we could weaken this assumption a bit.

## 266 PART II ♦ Generalized Regression Model and Equation Systems

The constrained GLS residuals,  $\hat{\epsilon}_c = \mathbf{y}_* - \mathbf{X}_* \hat{\beta}_c$ , are based on

$$\hat{\beta}_c = \hat{\beta} - [\mathbf{X}' \Omega^{-1} \mathbf{X}]^{-1} \mathbf{R}' [\mathbf{R} (\mathbf{X}' \Omega^{-1} \mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R} \hat{\beta} - \mathbf{q}).^4$$

To summarize, all the results for the classical model, including the usual inference procedures, apply to the transformed model in (9-11).

There is no precise counterpart to  $R^2$  in the generalized regression model. Alternatives have been proposed, but care must be taken when using them. For example, one choice is the  $R^2$  in the transformed regression, (9-11). But this regression need not have a constant term, so the  $R^2$  is not bounded by zero and one. Even if there is a constant term, the transformed regression is a computational device, not the model of interest. That a good (or bad) fit is obtained in the “model” in (9-11) may be of no interest; the dependent variable in that model,  $y_*$ , is different from the one in the model as originally specified. The usual  $R^2$  often suggests that the fit of the model is improved by a correction for heteroscedasticity and degraded by a correction for autocorrelation, but both changes can often be attributed to the computation of  $y_*$ . A more appealing fit measure might be based on the residuals from the original model once the GLS estimator is in hand, such as

$$R_G^2 = 1 - \frac{(\mathbf{y} - \mathbf{X} \hat{\beta})' (\mathbf{y} - \mathbf{X} \hat{\beta})}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Like the earlier contender, however, this measure is not bounded in the unit interval. In addition, this measure cannot be reliably used to compare models. The generalized least squares estimator minimizes the **generalized sum of squares**

$$\epsilon_*' \epsilon_* = (\mathbf{y} - \mathbf{X} \beta)' \Omega^{-1} (\mathbf{y} - \mathbf{X} \beta),$$

not  $\epsilon' \epsilon$ . As such, there is no assurance, for example, that dropping a variable from the model will result in a decrease in  $R_G^2$ , as it will in  $R^2$ . Other goodness-of-fit measures, designed primarily to be a function of the sum of squared residuals (raw or weighted by  $\Omega^{-1}$ ) and to be bounded by zero and one, have been proposed.<sup>5</sup> Unfortunately, they all suffer from at least one of the previously noted shortcomings. The  $R^2$ -like measures in this setting are purely descriptive. That being the case, the squared sample correlation between the actual and predicted values,  $r_{y,\hat{y}}^2 = \text{corr}^2(y, \hat{y}) = \text{corr}^2(y, \mathbf{x}' \hat{\beta})$ , would likely be a useful descriptor. Note, though, that this is not a proportion of variation explained, as is  $R^2$ ; it is a measure of the agreement of the model predictions with the actual data.

### 9.3.2 FEASIBLE GENERALIZED LEAST SQUARES (FGLS)

To use the results of Section 9.3.1,  $\Omega$  must be known. If  $\Omega$  contains unknown parameters that must be estimated, then generalized least squares is not feasible. But with an unrestricted  $\Omega$ , there are  $n(n + 1)/2$  additional parameters in  $\sigma^2 \Omega$ . This number is far too many to estimate with  $n$  observations. Obviously, some structure must be imposed on the model if we are to proceed.

---

<sup>4</sup>Note that this estimator is the constrained OLS estimator using the transformed data. [See (5-23).]

<sup>5</sup>See, example, Judge et al. (1985, p. 32) and Buse (1973).

## CHAPTER 9 ♦ The Generalized Regression Model 267

The typical problem involves a small set of parameters such that  $\Omega = \Omega(\theta)$ . For example, a commonly used formula in time-series settings is

$$\Omega(\rho) = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{n-2} \\ & & & & \vdots & \\ \rho^{n-1} & \rho^{n-2} & \cdots & & & 1 \end{bmatrix},$$

which involves only one additional unknown parameter. A model of heteroscedasticity that also has only one new parameter is

$$\sigma_i^2 = \sigma^2 z_i^\theta. \quad (9-15)$$

Suppose, then, that  $\hat{\theta}$  is a consistent estimator of  $\theta$ . (We consider later how such an estimator might be obtained.) To make GLS estimation feasible, we shall use  $\hat{\Omega} = \Omega(\hat{\theta})$  instead of the true  $\Omega$ . The issue we consider here is whether using  $\Omega(\hat{\theta})$  requires us to change any of the results of Section 9.3.1.

It would seem that if  $\text{plim } \hat{\theta} = \theta$ , then using  $\hat{\Omega}$  is asymptotically equivalent to using the true  $\Omega$ .<sup>6</sup> Let the **feasible generalized least squares (FGLS)** estimator be denoted

$$\hat{\beta} = (\mathbf{X}' \hat{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}' \hat{\Omega}^{-1} \mathbf{y}.$$

Conditions that imply that  $\hat{\beta}$  is asymptotically equivalent to  $\hat{\beta}$  are

$$\text{plim} \left[ \left( \frac{1}{n} \mathbf{X}' \hat{\Omega}^{-1} \mathbf{X} \right) - \left( \frac{1}{n} \mathbf{X}' \Omega^{-1} \mathbf{X} \right) \right] = \mathbf{0} \quad (9-16)$$

and

$$\text{plim} \left[ \left( \frac{1}{\sqrt{n}} \mathbf{X}' \hat{\Omega}^{-1} \boldsymbol{\varepsilon} \right) - \left( \frac{1}{\sqrt{n}} \mathbf{X}' \Omega^{-1} \boldsymbol{\varepsilon} \right) \right] = \mathbf{0}. \quad (9-17)$$

The first of these equations states that if the weighted sum of squares matrix based on the true  $\Omega$  converges to a positive definite matrix, then the one based on  $\hat{\Omega}$  converges to the same matrix. We are assuming that this is true. In the second condition, if the *transformed* regressors are well behaved, then the right-hand-side sum will have a limiting normal distribution. This condition is exactly the one we used in Chapter 4 to obtain the asymptotic distribution of the least squares estimator; here we are using the same results for  $\mathbf{X}_*$  and  $\boldsymbol{\varepsilon}_*$ . Therefore, (9-17) requires the same condition to hold when  $\Omega$  is replaced with  $\hat{\Omega}$ .<sup>7</sup>

These conditions, in principle, must be verified on a case-by-case basis. Fortunately, in most familiar settings, they are met. If we assume that they are, then the FGLS estimator based on  $\hat{\theta}$  has the same asymptotic properties as the GLS estimator. This result is extremely useful. Note, especially, the following theorem.

<sup>6</sup>This equation is sometimes denoted  $\text{plim } \hat{\Omega} = \Omega$ . Because  $\Omega$  is  $n \times n$ , it cannot have a probability limit. We use this term to indicate convergence element by element.

<sup>7</sup>The condition actually requires only that if the right-hand sum has *any* limiting distribution, then the left-hand one has the same one. Conceivably, this distribution might not be the normal distribution, but that seems unlikely except in a specially constructed, theoretical case.

**THEOREM 9.6 Efficiency of the FGLS Estimator**

An asymptotically efficient FGLS estimator does not require that we have an efficient estimator of  $\Omega$ ; only a consistent one is required to achieve full efficiency for the FGLS estimator.

Except for the simplest cases, the finite-sample properties and exact distributions of FGLS estimators are unknown. The asymptotic efficiency of FGLS estimators may not carry over to small samples because of the variability introduced by the estimated  $\Omega$ . Some analyses for the case of heteroscedasticity are given by Taylor (1977). A model of autocorrelation is analyzed by Griliches and Rao (1969). In both studies, the authors find that, over a broad range of parameters, FGLS is more efficient than least squares. But if the departure from the classical assumptions is not too severe, then least squares may be more efficient than FGLS in a small sample.

#### 9.4 HETROSCEDASTICITY AND WEIGHTED LEAST SQUARES

Regression disturbances whose variances are not constant across observations are heteroscedastic. **Heteroscedasticity** arises in numerous applications, in both cross-section and time-series data. For example, even after accounting for firm sizes, we expect to observe greater variation in the profits of large firms than in those of small ones. The variance of profits might also depend on product diversification, research and development expenditure, and industry characteristics and therefore might also vary across firms of similar sizes. When analyzing family spending patterns, we find that there is greater variation in expenditure on certain commodity groups among high-income families than low ones due to the greater discretion allowed by higher incomes.<sup>8</sup>

In the heteroscedastic regression model,

$$\text{Var}[\varepsilon_i | \mathbf{X}] = \sigma_i^2, \quad i = 1, \dots, n.$$

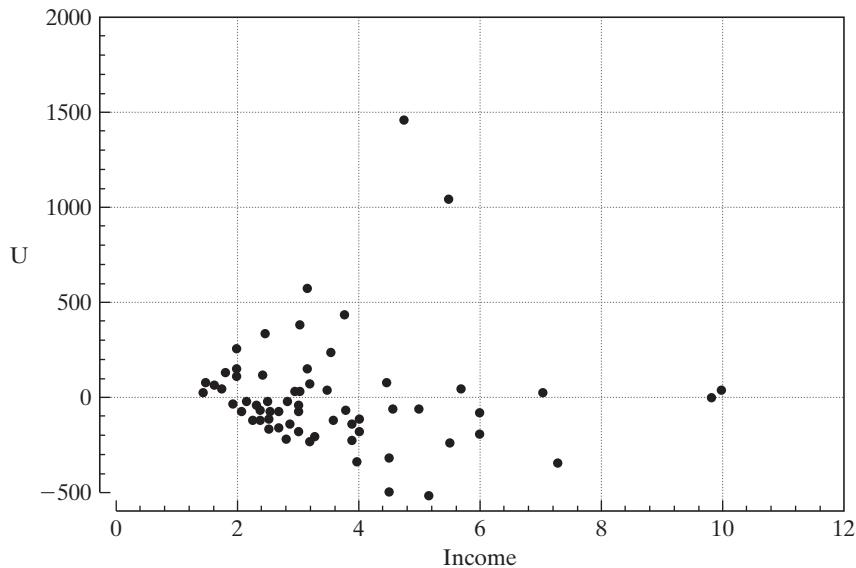
We continue to assume that the disturbances are pairwise uncorrelated. Thus,

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}] = \sigma^2 \boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} \omega_1 & 0 & 0 & \cdots & 0 \\ 0 & \omega_2 & 0 & \cdots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \cdots & \omega_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}.$$

It will sometimes prove useful to write  $\sigma_i^2 = \sigma^2 \omega_i$ . This form is an arbitrary scaling which allows us to use a normalization,

$$\text{tr}(\boldsymbol{\Omega}) = \sum_{i=1}^n \omega_i = n.$$

<sup>8</sup>Prais and Houthakker (1955).



**FIGURE 9.1** Plot of Residuals Against Income.

This makes the classical regression with homoscedastic disturbances a simple special case with  $\omega_i = 1, i = 1, \dots, n$ . Intuitively, one might then think of the  $\omega$ 's as weights that are scaled in such a way as to reflect only the variety in the disturbance variances. The scale factor  $\sigma^2$  then provides the overall scaling of the disturbance process.

**Example 9.1 Heteroscedastic Regression**

The data in Appendix Table F7.3 give monthly credit card expenditure for 13,444 individuals. Linear regression of monthly expenditure on a constant, age, income and its square, and a dummy variable for home ownership using the 72 of the observations for which expenditure was nonzero produces the residuals plotted in Figure 9.1. The pattern of the residuals is characteristic of a regression with heteroscedasticity. (The subsample of 72 observations is given in Appendix Table F9.1.)

We will examine the heteroscedastic regression model, first in general terms, then with some specific forms of the disturbance covariance matrix. We begin by examining the consequences of heteroscedasticity for least squares estimation. We then consider **robust estimation**. Section 9.4.4 presents appropriate estimators of the asymptotic covariance matrix of the least squares estimator. Specification tests for heteroscedasticity are considered in Section 9.5. Section 9.6 considers generalized (weighted) least squares, which requires knowledge at least of the form of  $\Omega$ . Finally, two common applications are examined in Section 9.7.

#### 9.4.1 ORDINARY LEAST SQUARES ESTIMATION

We showed in Section 9.2 that in the presence of heteroscedasticity, the least squares estimator  $\mathbf{b}$  is still unbiased, consistent, and asymptotically normally distributed. The asymptotic covariance matrix is

$$\text{Asy. Var}[\mathbf{b}] = \frac{\sigma^2}{n} \left( \text{plim} \frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1} \left( \text{plim} \frac{1}{n} \mathbf{X}' \Omega \mathbf{X} \right) \left( \text{plim} \frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1}.$$



## 270 PART II ♦ Generalized Regression Model and Equation Systems

Estimation of the asymptotic covariance matrix would be based on

$$\text{Var}[\mathbf{b} | \mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1} \left( \sigma^2 \sum_{i=1}^n \omega_i \mathbf{x}_i \mathbf{x}_i' \right) (\mathbf{X}'\mathbf{X})^{-1}. \quad (9-18)$$

[See (9-5).] Assuming, as usual, that the regressors are well behaved, so that  $(\mathbf{X}'\mathbf{X}/n)^{-1}$  converges to a positive definite matrix, we find that the mean square consistency of  $\mathbf{b}$  depends on the limiting behavior of the matrix:

$$\mathbf{Q}_n^* = \frac{\mathbf{X}'\Omega\mathbf{X}}{n} = \frac{1}{n} \sum_{i=1}^n \omega_i \mathbf{x}_i \mathbf{x}_i'.$$

If  $\mathbf{Q}_n^*$  converges to a positive definite matrix  $\mathbf{Q}^*$ , then as  $n \rightarrow \infty$ ,  $\mathbf{b}$  will converge to  $\boldsymbol{\beta}$  in mean square. Under most circumstances, if  $\omega_i$  is finite for all  $i$ , then we would expect this result to be true. Note that  $\mathbf{Q}_n^*$  is a weighted sum of the squares and cross products of  $\mathbf{x}$  with weights  $\omega_i/n$ , which sum to 1. We have already assumed that another weighted sum,  $\mathbf{X}'\mathbf{X}/n$ , in which the weights are  $1/n$ , converges to a positive definite matrix  $\mathbf{Q}$ , so it would be surprising if  $\mathbf{Q}_n^*$  did not converge as well. In general, then, we would expect that

$$\mathbf{b} \xrightarrow{a} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n} \mathbf{Q}^{-1} \mathbf{Q}^* \mathbf{Q}^{-1}\right], \quad \text{with } \mathbf{Q}^* = \text{plim } \mathbf{Q}_n^*.$$

A formal proof is based on Section 4.4 with  $\mathbf{Q}_i = \omega_i \mathbf{x}_i \mathbf{x}_i'$ .

### 9.4.2 INEFFICIENCY OF ORDINARY LEAST SQUARES

It follows from our earlier results that  $\mathbf{b}$  is inefficient relative to the GLS estimator. By how much will depend on the setting, but there is some generality to the pattern. As might be expected, the greater is the dispersion in  $\omega_i$  across observations, the greater the efficiency of GLS over OLS. The impact of this on the efficiency of estimation will depend crucially on the nature of the disturbance variances. In the usual cases, in which  $\omega_i$  depends on variables that appear elsewhere in the model, the greater is the dispersion in these variables, the greater will be the gain to using GLS. It is important to note, however, that both these comparisons are based on knowledge of  $\Omega$ . In practice, one of two cases is likely to be true. If we do have detailed knowledge of  $\Omega$ , the performance of the inefficient estimator is a moot point. We will use GLS or feasible GLS anyway. In the more common case, we will not have detailed knowledge of  $\Omega$ , so the comparison is not possible.

### 9.4.3 THE ESTIMATED COVARIANCE MATRIX OF $\mathbf{b}$

If the type of heteroscedasticity is known with certainty, then the ordinary least squares estimator is undesirable; we should use generalized least squares instead. The precise form of the heteroscedasticity is usually unknown, however. In that case, generalized least squares is not usable, and we may need to salvage what we can from the results of ordinary least squares.

The conventionally estimated covariance matrix for the least squares estimator  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$  is inappropriate; the appropriate matrix is  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Omega\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$ . It is unlikely that these two would coincide, so the usual estimators of the standard errors are likely to be erroneous. In this section, we consider how erroneous the conventional estimator is likely to be.

CHAPTER 9 ♦ The Generalized Regression Model **271**

As usual,

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K} = \frac{\mathbf{e}'\mathbf{M}\mathbf{e}}{n-K}, \quad (9-19)$$

where  $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . Expanding this equation, we obtain

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K} - \frac{\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}}{n-K}. \quad (9-20)$$

Taking the two parts separately yields

$$E\left[\frac{\mathbf{e}'\mathbf{e}}{n-K} \mid \mathbf{X}\right] = \frac{\text{tr}E[\mathbf{e}\mathbf{e}' \mid \mathbf{X}]}{n-K} = \frac{n\sigma^2}{n-K}. \quad (9-21)$$

[We have used the scaling  $\text{tr}(\Omega) = n$ .] In addition,

$$\begin{aligned} E\left[\frac{\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}}{n-K} \mid \mathbf{X}\right] &= \frac{\text{tr}\{E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X} \mid \mathbf{X}]\}}{n-K} \\ &= \frac{\text{tr}\left[\sigma^2\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\Omega\mathbf{X}}{n}\right)\right]}{n-K} = \frac{\sigma^2}{n-K} \text{tr}\left[\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\mathbf{Q}_n^*\right], \end{aligned} \quad (9-22)$$

where  $\mathbf{Q}_n^*$  is defined after (9-18). As  $n \rightarrow \infty$ , the term in (9-21) will converge to  $\sigma^2$ . The term in (9-22) will converge to zero if  $\mathbf{b}$  is consistent because both matrices in the product are finite. Therefore;

$$\text{If } \mathbf{b} \text{ is consistent, then } \lim_{n \rightarrow \infty} E[s^2] = \sigma^2.$$

It can also be shown—we leave it as an exercise—that if the fourth moment of every disturbance is finite and all our other assumptions are met, then

$$\lim_{n \rightarrow \infty} \text{Var}\left[\frac{\mathbf{e}'\mathbf{e}}{n-K}\right] = \lim_{n \rightarrow \infty} \text{Var}\left[\frac{\mathbf{e}'\mathbf{e}}{n-K}\right] = 0.$$

This result implies, therefore, that

$$\text{If } \text{plim } \mathbf{b} = \boldsymbol{\beta}, \text{ then } \text{plim } s^2 = \sigma^2.$$

Before proceeding, it is useful to pursue this result. The normalization  $\text{tr}(\Omega) = n$  implies that

$$\sigma^2 = \bar{\sigma}^2 = \frac{1}{n} \sum_i \sigma_i^2 \quad \text{and} \quad \omega_i = \frac{\sigma_i^2}{\bar{\sigma}^2}.$$

Therefore, our previous convergence result implies that the least squares estimator  $s^2$  converges to  $\text{plim } \bar{\sigma}^2$ , that is, the probability limit of the average variance of the disturbances, *assuming that this probability limit exists*. Thus, some further assumption about these variances is necessary to obtain the result.

The difference between the conventional estimator and the appropriate (true) covariance matrix for  $\mathbf{b}$  is

$$\text{Est. Var}[\mathbf{b} \mid \mathbf{X}] - \text{Var}[\mathbf{b} \mid \mathbf{X}] = s^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Omega\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}. \quad (9-23)$$

## 272 PART II ♦ Generalized Regression Model and Equation Systems

In a large sample (so that  $s^2 \approx \sigma^2$ ), this difference is approximately equal to

$$\mathbf{D} = \frac{\sigma^2}{n} \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \left[ \frac{\mathbf{X}'\mathbf{X}}{n} - \frac{\mathbf{X}'\Omega\mathbf{X}}{n} \right] \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1}. \quad (9-24)$$

The difference between the two matrices hinges on

$$\Delta = \frac{\mathbf{X}'\mathbf{X}}{n} - \frac{\mathbf{X}'\Omega\mathbf{X}}{n} = \sum_{i=1}^n \left( \frac{1}{n} \right) \mathbf{x}_i \mathbf{x}_i' - \sum_{i=1}^n \left( \frac{\omega_i}{n} \right) \mathbf{x}_i \mathbf{x}_i' = \frac{1}{n} \sum_{i=1}^n (1 - \omega_i) \mathbf{x}_i \mathbf{x}_i', \quad (9-25)$$

where  $\mathbf{x}_i'$  is the  $i$ th row of  $\mathbf{X}$ . These are two weighted averages of the matrices  $\mathbf{Q}_i = \mathbf{x}_i \mathbf{x}_i'$ , using weights 1 for the first term and  $\omega_i$  for the second. The scaling  $\text{tr}(\Omega) = n$  implies that  $\sum_i (\omega_i/n) = 1$ . Whether the weighted average based on  $\omega_i/n$  differs much from the one using  $1/n$  depends on the weights. If the weights are related to the values in  $\mathbf{x}_i$ , then the difference can be considerable. If the weights are uncorrelated with  $\mathbf{x}_i \mathbf{x}_i'$ , however, then the weighted average will tend to equal the unweighted average.<sup>9</sup>

Therefore, the comparison rests on whether the heteroscedasticity is related to any of  $x_k$  or  $x_j \times x_k$ . The conclusion is that, in general: *If the heteroscedasticity is not correlated with the variables in the model, then at least in large samples, the ordinary least squares computations, although not the optimal way to use the data, will not be misleading.* For example, in the groupwise heteroscedasticity model of Section 9.7.2, if the observations are grouped in the subsamples in a way that is unrelated to the variables in  $\mathbf{X}$ , then the usual OLS estimator of  $\text{Var}[\mathbf{b}]$  will, at least in large samples, provide a reliable estimate of the appropriate covariance matrix. It is worth remembering, however, that the least squares estimator will be inefficient, the more so the larger are the differences among the variances of the groups.<sup>10</sup>

The preceding is a useful result, but one should not be overly optimistic. First, it remains true that ordinary least squares is demonstrably inefficient. Second, if the primary assumption of the analysis—that the heteroscedasticity is unrelated to the variables in the model—is incorrect, then the conventional standard errors may be quite far from the appropriate values.

### 9.4.4 ESTIMATING THE APPROPRIATE COVARIANCE MATRIX FOR ORDINARY LEAST SQUARES

It is clear from the preceding that heteroscedasticity has some potentially serious implications for inferences based on the results of least squares. The application of more appropriate estimation techniques requires a detailed formulation of  $\Omega$ , however. It may well be that the form of the heteroscedasticity is unknown. White (1980a) has shown that it is still possible to obtain an appropriate estimator for the variance of the least squares estimator, even if the heteroscedasticity is related to the variables in  $\mathbf{X}$ .

<sup>9</sup>Suppose, for example, that  $\mathbf{X}$  contains a single column and that both  $\mathbf{x}_i$  and  $\omega_i$  are independent and identically distributed random variables. Then  $\mathbf{x}'\mathbf{x}/n$  converges to  $E[x_i^2]$ , whereas  $\mathbf{x}'\Omega\mathbf{x}/n$  converges to  $\text{Cov}[\omega_i, x_i^2] + E[\omega_i]E[x_i^2]$ .  $E[\omega_i] = 1$ , so if  $\omega$  and  $x^2$  are uncorrelated, then the sums have the same probability limit.

<sup>10</sup>Some general results, including analysis of the properties of the estimator based on estimated variances, are given in Taylor (1977).

## CHAPTER 9 ♦ The Generalized Regression Model 273

Referring to (9-18), we seek an estimator of

$$\mathbf{Q}_* = \frac{1}{n} \sum_{i=1}^n \sigma_i^2 \mathbf{x}_i \mathbf{x}'_i.$$

White (1980a) shows that under very general conditions, the estimator

$$\mathbf{S}_0 = \frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{x}_i \mathbf{x}'_i \quad (9-26)$$

has

$$\text{plim } \mathbf{S}_0 = \text{plim } \mathbf{Q}_*.^{11}$$

We can sketch a proof of this result using the results we obtained in Section 4.4.<sup>12</sup> Note first that  $\mathbf{Q}_*$  is not a parameter matrix in itself. It is a weighted sum of the outer products of the rows of  $\mathbf{X}$  (or  $\mathbf{Z}$  for the instrumental variables case). Thus, we seek not to “estimate”  $\mathbf{Q}_*$ , but to find a function of the sample data that will be arbitrarily close to this function of the population parameters as the sample size grows large. The distinction is important. We are not estimating the middle matrix in (9-9) or (9-18); we are attempting to construct a matrix from the sample data that will behave the same way that this matrix behaves. In essence, if  $\mathbf{Q}_*$  converges to a finite positive matrix, then we would be looking for a function of the sample data that converges to the same matrix. Suppose that the true disturbances  $\varepsilon_i$  could be observed. Then each term in  $\mathbf{Q}_*$  would equal  $E[\varepsilon_i^2 \mathbf{x}_i \mathbf{x}'_i | \mathbf{x}_i]$ . With some fairly mild assumptions about  $\mathbf{x}_i$ , then, we could invoke a law of large numbers (see Theorems D.4 through D.9) to state that if  $\mathbf{Q}_*$  has a probability limit, then

$$\text{plim} \frac{1}{n} \sum_{i=1}^n \sigma_i^2 \mathbf{x}_i \mathbf{x}'_i = \text{plim} \frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{x}_i \mathbf{x}'_i.$$

The final detail is to justify the replacement of  $\varepsilon_i$  with  $e_i$  in  $\mathbf{S}_0$ . The consistency of  $\mathbf{b}$  for  $\beta$  is sufficient for the argument. (Actually, residuals based on *any* consistent estimator of  $\beta$  would suffice for this estimator, but as of now,  $\mathbf{b}$  or  $\mathbf{b}_{IV}$  is the only one in hand.) The end result is that the **White heteroscedasticity consistent estimator**

$$\begin{aligned} \text{Est. Asy. Var}[\mathbf{b}] &= \frac{1}{n} \left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{x}_i \mathbf{x}'_i \right) \left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1} \\ &= n(\mathbf{X}' \mathbf{X})^{-1} \mathbf{S}_0 (\mathbf{X}' \mathbf{X})^{-1} \end{aligned} \quad (9-27)$$

can be used to estimate the asymptotic covariance matrix of  $\mathbf{b}$ .

This result is extremely important and useful.<sup>13</sup> It implies that without actually specifying the type of heteroscedasticity, we can still make appropriate inferences based on the results of least squares. This implication is especially useful if we are unsure of the precise nature of the heteroscedasticity (which is probably most of the time). We will pursue some examples in Section 8.7.

<sup>11</sup>See also Eicker (1967), Horn, Horn, and Duncan (1975), and MacKinnon and White (1985).

<sup>12</sup>We will give only a broad sketch of the proof. Formal results appear in White (1980) and (2001).

<sup>13</sup>Further discussion and some refinements may be found in Cragg (1982). Cragg shows how White's observation can be extended to devise an estimator that improves on the efficiency of ordinary least squares.

**274 PART II ♦ Generalized Regression Model and Equation Systems**
**TABLE 9.1** Least Squares Regression Results

	<i>Constant</i>	<i>Age</i>	<i>OwnRent</i>	<i>Income</i>	<i>Income</i> <sup>2</sup>
Sample mean		32.08	0.36	3.369	
Coefficient	-237.15	-3.0818	27.941	234.35	-14.997
Standard error	199.35	5.5147	82.922	80.366	7.008
<i>t</i> ratio	-1.19	-0.5590	0.337	2.916	-2.008
White S.E.	212.99	3.3017	92.188	88.866	6.9446
D. and M. (1)	220.79	3.4227	95.566	92.122	7.1991
D. and M. (2)	221.09	3.4477	95.672	92.083	7.1995
$R^2 = 0.243578, s = 284.75080$					
Mean expenditure = \$262.53. Income is $\times \$10,000$					
Tests for heteroscedasticity: White = 14.329,					
Breusch-Pagan = 41.920, Koenker-Bassett = 6.187.					
<del>(Two degrees of freedom. <math>\chi^2_*</math> = 5.99.)</del>					

A number of studies have sought to improve on the White estimator for OLS.<sup>14</sup> The asymptotic properties of the estimator are unambiguous, but its usefulness in small samples is open to question. The possible problems stem from the general result that the squared OLS residuals tend to underestimate the squares of the true disturbances. [That is why we use  $1/(n - K)$  rather than  $1/n$  in computing  $s^2$ .] The end result is that in small samples, at least as suggested by some Monte Carlo studies [e.g., MacKinnon and White (1985)], the White estimator is a bit too optimistic; the matrix is a bit too small, so asymptotic *t* ratios are a little too large. Davidson and MacKinnon (1993, p. 554) suggest a number of fixes, which include (1) scaling up the end result by a factor  $n/(n - K)$  and (2) using the squared residual scaled by its true variance,  $e_i^2/m_{ii}$ , instead of  $e_i^2$ , where  $m_{ii} = 1 - \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ .<sup>15</sup> [See Exercise 9.6.b.] On the basis of their study, Davidson and MacKinnon strongly advocate one or the other correction. Their admonition “One should *never* use [the White estimator] because [(2)] *always* performs better” seems a bit strong, but the point is well taken. The use of sharp asymptotic results in small samples can be problematic. The last two rows of Table 9.1 show the recomputed standard errors with these two modifications.

**Example 9.2 The White Estimator**

Using White's estimator for the regression in Example 9.1 produces the results in the row labeled “White S. E.” in Table 9.1. The two income coefficients are individually and jointly statistically significant based on the individual *t* ratios and  $F(2, 67) = [(0.244 - 0.064)/2]/[0.756/(72 - 5)] = 7.976$ . The 1 percent critical value is 4.94.

The differences in the estimated standard errors seem fairly minor given the extreme heteroscedasticity. One surprise is the decline in the standard error of the age coefficient. The *F* test is no longer available for testing the joint significance of the two income coefficients because it relies on homoscedasticity. A **Wald test**, however, may be used in any event. The chi-squared test is based on

$$W = (\mathbf{R}\mathbf{b})' [\mathbf{R}(\text{Est. Asy. Var}[\mathbf{b}])\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b}) \quad \text{where } \mathbf{R} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

<sup>14</sup>See, e.g., MacKinnon and White (1985) and Messer and White (1984).

<sup>15</sup>This is the standardized residual in (4-61). The authors also suggest a third correction,  $e_i^2/m_{ii}^2$ , as an approximation to an estimator based on the “jackknife” technique, but their advocacy of this estimator is much weaker than that of the other two.

## CHAPTER 9 ♦ The Generalized Regression Model 275

and the estimated asymptotic covariance matrix is the White estimator. The  $F$  statistic based on least squares is 7.976. The Wald statistic based on the White estimator is 20.604; the 95 percent critical value for the chi-squared distribution with two degrees of freedom is 5.99, so the conclusion is unchanged.

### 9.5 TESTING FOR HETEROSCEDASTICITY

Heteroscedasticity poses potentially severe problems for inferences based on least squares. One can rarely be certain that the disturbances are heteroscedastic, however, and unfortunately, what form the heteroscedasticity takes if they are. As , it is useful to be able to test for homoscedasticity and, if necessary, modify our estimation procedures accordingly.<sup>16</sup> Several types of tests have been suggested. They can be roughly grouped in descending order in terms of their generality and, as might be expected, in ascending order in terms of their power.<sup>17</sup> We will examine the two most commonly used tests.

Tests for heteroscedasticity are based on the following strategy. Ordinary least squares is a consistent estimator of  $\beta$  even in the presence of heteroscedasticity. As such, the ordinary least squares residuals will mimic, albeit imperfectly because of sampling variability, the heteroscedasticity of the true disturbances. Therefore, tests designed to detect heteroscedasticity will, in general, be applied to the ordinary least squares residuals.

#### 9.5.1 WHITE'S GENERAL TEST

To formulate most of the available tests, it is necessary to specify, at least in rough terms, the nature of the heteroscedasticity. It would be desirable to be able to test a general hypothesis of the form

$$H_0 : \sigma_i^2 = \sigma^2 \quad \text{for all } i,$$

$$H_1 : \text{Not } H_0.$$

In view of our earlier findings on the difficulty of estimation in a model with unknown parameters, this is rather ambitious. Nonetheless, such a test has been devised by White (1980b). The correct covariance matrix for the least squares estimator is 

$$\text{Var}[\mathbf{b} | \mathbf{X}] = \sigma^2 [\mathbf{X}' \mathbf{X}]^{-1} [\mathbf{X}' \Omega \mathbf{X}] [\mathbf{X}' \mathbf{X}]^{-1},$$

which, as we have seen, can be estimated using (9-27). The conventional estimator is  $\mathbf{V} = s^2 [\mathbf{X}' \mathbf{X}]^{-1}$ . If there is no heteroscedasticity, then  $\mathbf{V}$  will give a consistent estimator of  $\text{Var}[\mathbf{b} | \mathbf{X}]$ , whereas if there is, then it will not. White has devised a statistical test based on this observation. A simple operational version of his test is carried out by obtaining  $nR^2$  in the regression of  $e_i^2$  on a constant and all unique variables contained in  $\mathbf{x}$  and

---

<sup>16</sup>There is the possibility that a preliminary test for heteroscedasticity will incorrectly lead us to use weighted least squares or fail to alert us to heteroscedasticity and lead us improperly to use ordinary least squares. Some limited results on the properties of the resulting estimator are given by Ohtani and Toyoda (1980). Their results suggest that it is best to test first for heteroscedasticity rather than merely to assume that it is present.

<sup>17</sup>A study that examines the power of several tests for heteroscedasticity is Ali and Giaccotto (1984).

## 276 PART II ♦ Generalized Regression Model and Equation Systems

all the squares and cross products of the variables in  $\mathbf{x}$ . The statistic is asymptotically distributed as chi-squared with  $P - 1$  degrees of freedom, where  $P$  is the number of regressors in the equation, including the constant.

The **White test** is extremely general. To carry it out, we need not make any specific assumptions about the nature of the heteroscedasticity. Although this characteristic is a virtue, it is, at the same time, a potentially serious shortcoming. The test may reveal heteroscedasticity, but it may instead simply identify some other specification error (such as the omission of  $x^2$  from a simple regression).<sup>18</sup> Except in the context of a specific problem, little can be said about the power of White's test; it may be very low against some alternatives. In addition, unlike some of the other tests we shall discuss, the White test is **nonconstructive**. If we reject the null hypothesis, then the result of the test gives no indication of what to do next.

### 9.5.2 THE BREUSCH-PAGAN/GODFREY LM TEST

Breusch and Pagan<sup>19</sup> have devised a **Lagrange multiplier test** of the hypothesis that  $\sigma_i^2 = \sigma^2 f(\alpha_0 + \boldsymbol{\alpha}' \mathbf{z}_i)$ , where  $\mathbf{z}_i$  is a vector of independent variables.<sup>20</sup> The model is homoscedastic if  $\boldsymbol{\alpha} = \mathbf{0}$ . The test can be carried out with a simple regression:

$$\text{LM} = \frac{1}{2} \text{ explained sum of squares in the regression of } e_i^2 / (\mathbf{e}' \mathbf{e} / n) \text{ on } \mathbf{z}_i.$$

For computational purposes, let  $\mathbf{Z}$  be the  $n \times P$  matrix of observations on  $(1, \mathbf{z}_i)$ , and let  $\mathbf{g}$  be the vector of observations of  $g_i = e_i^2 / (\mathbf{e}' \mathbf{e} / n) - 1$ . Then

$$\text{LM} = \frac{1}{2} [\mathbf{g}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{g}]. \quad (9-28)$$

Under the null hypothesis of homoscedasticity, LM has a limiting chi-squared distribution with degrees of freedom equal to the number of variables in  $\mathbf{z}_i$ . This test can be applied to a variety of models, including, for example, those examined in Example 9.3 (2) and in Sections 9.7.1 and 9.7.2.<sup>21</sup>

It has been argued that the **Breusch-Pagan Lagrange multiplier test** is sensitive to the assumption of normality. Koenker (1981) and Koenker and Bassett (1982) suggest that the computation of LM be based on a more **robust estimator** of the variance of  $e_i^2$ ,

$$V = \frac{1}{n} \sum_{i=1}^n \left[ e_i^2 - \frac{\mathbf{e}' \mathbf{e}}{n} \right]^2.$$

The variance of  $e_i^2$  is not necessarily equal to  $2\sigma^4$  if  $e_i$  is not normally distributed. Let  $\mathbf{u}$  equal  $(e_1^2, e_2^2, \dots, e_n^2)$  and  $\mathbf{i}$  be an  $n \times 1$  column of 1s. Then  $\bar{u} = \mathbf{e}' \mathbf{e} / n$ . With this change, the computation becomes

$$\text{LM} = \left[ \frac{1}{V} \right] (\mathbf{u} - \bar{u} \mathbf{i})' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' (\mathbf{u} - \bar{u} \mathbf{i}).$$

<sup>18</sup>Thursby (1982) considers this issue in detail.

<sup>19</sup>Breusch and Pagan (1979).

<sup>20</sup>Lagrange multiplier tests are discussed in Section 14.6.3.

<sup>21</sup>The model  $\sigma_i^2 = \sigma^2 \exp(\boldsymbol{\alpha}' \mathbf{z}_i)$  is one of these cases. In analyzing this model specifically, Harvey (1976) derived the same test statistic.

## CHAPTER 9 ♦ The Generalized Regression Model 277

Under normality, this modified statistic will have the same asymptotic distribution as the Breusch-Pagan statistic, but absent normality, there is some evidence that it provides a more powerful test. Waldman (1983) has shown that if the variables in  $\mathbf{z}_i$  are the same as those used for the White test described earlier, then the two tests are algebraically the same.

**Example 9.3 Testing for Heteroscedasticity**

**1. White's Test:** For the data used in Example 9.1, there are 15 variables in  $\mathbf{x} \otimes \mathbf{x}$  including the constant term. But since  $\text{Ownrent}^2 = \text{OwnRent}$  and  $\text{Income} \times \text{Income} = \text{Income}^2$ , only 13 are unique. Regression of the squared least squares residuals on these 13 variables produces  $R^2 = 0.199013$ . The chi-squared statistic is therefore  $72(0.199013) = 14.329$ . The 95 percent critical value of chi-squared with 12 degrees of freedom is 21.03, so despite what might seem to be obvious in Figure 9.1, the hypothesis of homoscedasticity is not rejected by this test.

**2. Breusch-Pagan Test:** This test requires a specific alternative hypothesis. For this purpose, we specify the test based on  $\mathbf{z} = [1, \text{Income}, \text{Income}^2]$ . Using the least squares residuals, we compute  $g_i = e_i^2 / (\mathbf{e}'\mathbf{e}/72) - 1$ ; then  $\text{LM} = \frac{1}{2}\mathbf{g}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{g}$ . The sum of squares is 5,432,562.033. The computation produces  $\text{LM} = 41.920$ . The critical value for the chi-squared distribution with two degrees of freedom is 5.99, so the hypothesis of homoscedasticity is rejected. The Koenker and Bassett variant of this statistic is only 6.187, which is still significant but much smaller than the LM statistic. The wide difference between these two statistics suggests that the assumption of normality is erroneous. Absent any knowledge of the heteroscedasticity, we might use the Bera and Jarque (1981, 1982) and Kiefer and Salmon (1983) test for normality,

$$\chi^2[2] = n[1/6(m_3/s^3)^2 + 1/25((m_4 - 3)/s^4)^2]$$

where  $m_j = (1/n) \sum_i e_i^j$ . Under the null hypothesis of homoscedastic and normally distributed disturbances, this statistic has a limiting chi-squared distribution with two degrees of freedom. Based on the least squares residuals, the value is 497.35, which certainly does lead to rejection of the hypothesis. Some caution is warranted here, however. It is unclear what part of the hypothesis should be rejected. We have convincing evidence in Figure 9.1 that the disturbances are heteroscedastic, so the assumption of homoscedasticity underlying this test is questionable. This does suggest the need to examine the data before applying a **specification test** such as this one.

## 9.6 WEIGHTED LEAST SQUARES

Having tested for and found evidence of heteroscedasticity, the logical next step is to revise the estimation technique to account for it. The GLS estimator is

$$\hat{\beta} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y}. \quad (9-29)$$

Consider the most general case,  $\text{Var}[\varepsilon_i | \mathbf{X}] = \sigma_i^2 = \sigma^2\omega_i$ . Then  $\Omega^{-1}$  is a diagonal matrix whose  $i$ th diagonal element is  $1/\omega_i$ . The GLS estimator is obtained by regressing

$$\mathbf{P}\mathbf{y} = \begin{bmatrix} y_1/\sqrt{\omega_1} \\ y_2/\sqrt{\omega_2} \\ \vdots \\ y_n/\sqrt{\omega_n} \end{bmatrix} \text{ on } \mathbf{P}\mathbf{X} = \begin{bmatrix} \mathbf{x}'_1/\sqrt{\omega_1} \\ \mathbf{x}'_2/\sqrt{\omega_2} \\ \vdots \\ \mathbf{x}'_n/\sqrt{\omega_n} \end{bmatrix}.$$

## 278 PART II ♦ Generalized Regression Model and Equation Systems

Applying ordinary least squares to the transformed model, we obtain the **weighted least squares (WLS)** estimator.

$$\hat{\beta} = \left[ \sum_{i=1}^n w_i \mathbf{x}_i \mathbf{x}'_i \right]^{-1} \left[ \sum_{i=1}^n w_i \mathbf{x}_i y_i \right],$$

where  $w_i = 1/\sigma_i^2$ .<sup>22</sup> The logic of the computation is that observations with smaller variances receive a larger weight in the computations of the sums and therefore have greater influence in the estimates obtained.

### 9.6.1 WEIGHTED LEAST SQUARES WITH KNOWN $\Omega$

A common specification is that the variance is proportional to one of the regressors or its square. Our earlier example of family expenditures is one in which the relevant variable is usually income. Similarly, in studies of firm profits, the dominant variable is typically assumed to be firm size. If

$$\sigma_i^2 = \sigma^2 x_{ik}^2,$$

then the transformed regression model for GLS is

$$\frac{y}{x_k} = \beta_k + \beta_1 \left( \frac{x_1}{x_k} \right) + \beta_2 \left( \frac{x_2}{x_k} \right) + \cdots + \frac{\varepsilon}{x_k}. \quad (9-30)$$

If the variance is proportional to  $x_k$  instead of  $x_k^2$ , then the weight applied to each observation is  $1/\sqrt{x_k}$  instead of  $1/x_k$ .

In (9-30), the coefficient on  $x_k$  becomes the constant term. But if the variance is proportional to any power of  $x_k$  other than two, then the transformed model will no longer contain a constant, and we encounter the problem of interpreting  $R^2$  mentioned earlier. For example, no conclusion should be drawn if the  $R^2$  in the regression of  $y/z$  on  $1/z$  and  $x/z$  is higher than in the regression of  $y$  on a constant and  $x$  for any  $z$ , including  $x$ . The good fit of the weighted regression might be due to the presence of  $1/z$  on both sides of the equality.

It is rarely possible to be certain about the nature of the heteroscedasticity in a regression model. In one respect, this problem is only minor. The weighted least squares estimator

$$\hat{\beta} = \left[ \sum_{i=1}^n w_i \mathbf{x}_i \mathbf{x}'_i \right]^{-1} \left[ \sum_{i=1}^n w_i \mathbf{x}_i y_i \right]$$

is consistent regardless of the weights used, as long as the weights are uncorrelated with the disturbances.

But using the wrong set of weights has two other consequences that may be less benign. First, the improperly weighted least squares estimator is inefficient. This point might be moot if the correct weights are unknown, but the GLS standard errors will

---

<sup>22</sup>The weights are often denoted  $w_i = 1/\sigma_i^2$ . This expression is consistent with the equivalent  $\hat{\beta} = [\mathbf{X}'(\sigma^2 \Omega)^{-1} \mathbf{X}]^{-1} \mathbf{X}'(\sigma^2 \Omega)^{-1} \mathbf{y}$ . The  $\sigma^2$ 's cancel, leaving the expression given previously.

CHAPTER 9 ♦ The Generalized Regression Model **279**

also be incorrect. The asymptotic covariance matrix of the estimator

$$\hat{\beta} = [\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad (9-31)$$

is

$$\text{Asy. Var}[\hat{\beta}] = \sigma^2[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}\Omega\mathbf{V}^{-1}\mathbf{X}[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}. \quad (9-32)$$

This result may or may not resemble the usual estimator, which would be the matrix in brackets, and underscores the usefulness of the White estimator in (9-27).

The standard approach in the literature is to use OLS with the White estimator or some variant for the asymptotic covariance matrix. One could argue both flaws and virtues in this approach. In its favor, **robustness to unknown heteroscedasticity** is a compelling virtue. In the clear presence of heteroscedasticity, however, least squares can be extremely inefficient. The question becomes whether using the wrong weights is better than using no weights at all. There are several layers to the question. If we use one of the models mentioned earlier—Harvey’s, for example, is a versatile and flexible candidate—then we may use the wrong set of weights and, in addition, estimation of the variance parameters introduces a new source of variation into the slope estimators for the model. A heteroscedasticity robust estimator for weighted least squares can be formed by combining (9-32) with the White estimator. The weighted least squares estimator in (9-31) is consistent with any set of weights  $\mathbf{V} = \text{diag}[v_1, v_2, \dots, v_n]$ . Its asymptotic covariance matrix can be estimated with

$$\text{Est. Asy. Var}[\hat{\beta}] = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \left[ \sum_{i=1}^n \left( \frac{e_i^2}{v_i^2} \right) \mathbf{x}_i \mathbf{x}_i' \right] (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \quad (9-33)$$

Any consistent estimator can be used to form the residuals. The weighted least squares estimator is a natural candidate.

### 9.6.2 ESTIMATION WHEN $\Omega$ CONTAINS UNKNOWN PARAMETERS

The general form of the heteroscedastic regression model has too many parameters to estimate by ordinary methods. Typically, the model is restricted by formulating  $\sigma^2\Omega$  as a function of a few parameters, as in  $\sigma_i^2 = \sigma^2 x_i^\alpha$  or  $\sigma_i^2 = \sigma^2 [x_i \alpha]$ . Write this as  $\Omega(\alpha)$ . FGLS based on a consistent estimator of  $\Omega(\alpha)$  (meaning a consistent estimator of  $\alpha$ ) is asymptotically equivalent to full GLS. The new problem is that we must first find consistent estimators of the unknown parameters in  $\Omega(\alpha)$ . Two methods are typically used, two-step GLS and maximum likelihood. We consider the two-step estimator here and the maximum likelihood estimator in Chapter 14.

For the heteroscedastic model, the GLS estimator is

$$\hat{\beta} = \left[ \sum_{i=1}^n \left( \frac{1}{\sigma_i^2} \right) \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \left[ \sum_{i=1}^n \left( \frac{1}{\sigma_i^2} \right) \mathbf{x}_i y_i \right]. \quad (9-34)$$

The **two-step estimators** are computed by first obtaining estimates  $\hat{\sigma}_i^2$ , usually using some function of the ordinary least squares residuals. Then,  $\hat{\beta}$  uses (9-34) and  $\hat{\sigma}_i^2$ . The ordinary least squares estimator of  $\beta$ , although inefficient, is still consistent. As such, statistics computed using the ordinary least squares residuals,  $e_i = (y_i - \mathbf{x}_i' \mathbf{b})$ , will have the same asymptotic properties as those computed using the true disturbances,  $\varepsilon_i = (y_i - \mathbf{x}_i' \beta)$ .

## 280 PART II ♦ Generalized Regression Model and Equation Systems

This result suggests a regression approach for the true disturbances and variables  $\mathbf{z}_i$  that may or may not coincide with  $\mathbf{x}_i$ . Now  $E[\varepsilon_i^2 | \mathbf{z}_i] = \sigma_i^2$ , so

$$\varepsilon_i^2 = \sigma_i^2 + v_i,$$

where  $v_i$  is just the difference between  $\varepsilon_i^2$  and its conditional expectation. Because  $\varepsilon_i$  is unobservable, we would use the least squares residual, for which  $e_i = \varepsilon_i - \mathbf{x}'_i(\mathbf{b} - \boldsymbol{\beta}) = \varepsilon_i + u_i$ . Then,  $e_i^2 = \varepsilon_i^2 + u_i^2 + 2\varepsilon_i u_i$ . But, in large samples, as  $\mathbf{b} \xrightarrow{P} \boldsymbol{\beta}$ , terms in  $u_i$  will become negligible, so that at least approximately,<sup>23</sup>

$$e_i^2 = \sigma_i^2 + v_i^*.$$

The procedure suggested is to treat the variance function as a regression and use the squares or some other functions of the least squares residuals as the dependent variable.<sup>24</sup> For example, if  $\sigma_i^2 = \mathbf{z}'_i \boldsymbol{\alpha}$ , then a consistent estimator of  $\boldsymbol{\alpha}$  will be the least squares slopes,  $\mathbf{a}$ , in the “model,”

$$e_i^2 = \mathbf{z}'_i \boldsymbol{\alpha} + v_i^*.$$

In this model,  $v_i^*$  is both heteroscedastic and autocorrelated, so  $\mathbf{a}$  is consistent but inefficient. But, consistency is all that is required for asymptotically efficient estimation of  $\boldsymbol{\beta}$  using  $\Omega(\hat{\boldsymbol{\alpha}})$ . It remains to be settled whether improving the estimator of  $\boldsymbol{\alpha}$  in this and the other models we will consider would improve the small sample properties of the two-step estimator of  $\boldsymbol{\beta}$ .<sup>25</sup>

The two-step estimator may be iterated by recomputing the residuals after computing the FGLS estimates and then reentering the computation. The asymptotic properties of the iterated estimator are the same as those of the two-step estimator, however. In some cases, this sort of iteration will produce the maximum likelihood estimator at convergence. Yet none of the estimators based on regression of squared residuals on other variables satisfy the requirement. Thus, iteration in this context provides little additional benefit, if any.

## 9.7 APPLICATIONS

This section will present two common applications of the heteroscedastic regression model, Harvey's model of **multiplicative heteroscedasticity** and a model of **groupwise heteroscedasticity** that extends to the disturbance variance some concepts that are usually associated with variation in the regression function.

### 9.7.1 MULTIPLICATIVE HETEROSEDASTICITY

Harvey's (1976) model of multiplicative heteroscedasticity is a very flexible, general model that includes most of the useful formulations as special cases. The general formulation is

$$\sigma_i^2 = \sigma^2 \exp(\mathbf{z}'_i \boldsymbol{\alpha}).$$

<sup>23</sup>See Amemiya (1985) and Harvey (1976) for formal analyses.

<sup>24</sup>See, for example, Jobson and Fuller (1980).

<sup>25</sup>Fomby, Hill, and Johnson (1984, pp. 177–186) and Amemiya (1985, pp. 203–207; 1977a) examine this model.

CHAPTER 9 ♦ The Generalized Regression Model **281**

A model with heteroscedasticity of the form

$$\sigma_i^2 = \sigma^2 \prod_{m=1}^M z_{im}^{\alpha_m}$$

results if the logs of the variables are placed in  $\mathbf{z}_i$ . The groupwise heteroscedasticity model described in Example 9.4 is produced by making  $\mathbf{z}_i$  a set of group dummy variables (one must be omitted). In this case,  $\sigma^2$  is the disturbance variance for the base group whereas for the other groups,  $\sigma_g^2 = \sigma^2 \exp(\alpha_g)$ .

**Example 9.4 Multiplicative Heteroscedasticity**

In Example 6.4, we fit a cost function for the U.S. airline industry of the form

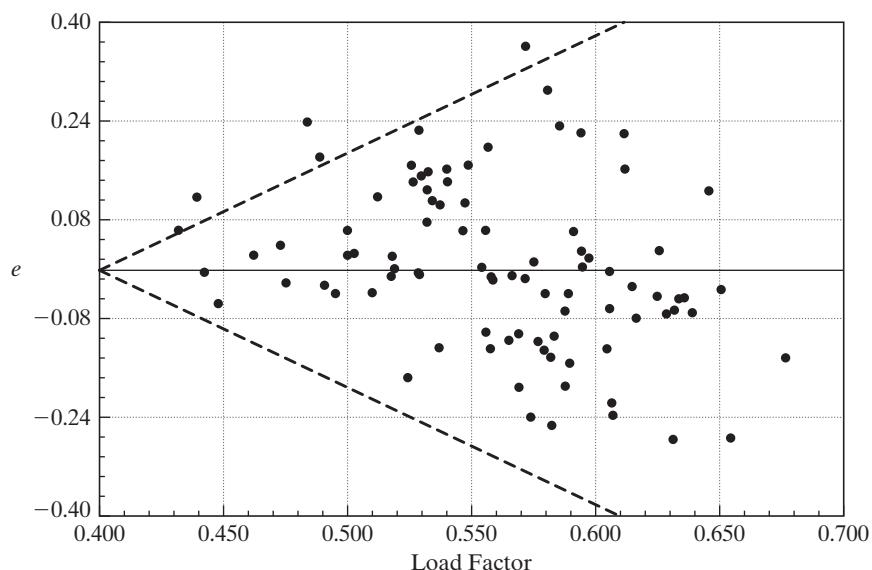
$$\ln C_{it} = \beta_1 + \beta_2 \ln Q_{it} + \beta_3 [\ln Q_{it}]^2 + \beta_4 \ln P_{fuel,i,t} + \beta_5 \text{Loadfactor}_{i,t} + \varepsilon_{i,t}$$

where  $C_{i,t}$  is total cost,  $Q_{i,t}$  is output, and  $P_{fuel,i,t}$  is the price of fuel and the 90 observations in the data set are for six firms observed for 15 years. (The model also included dummy variables for firm and year, which we will omit for simplicity.) We now consider a revised model in which the load factor appears in the variance of  $\varepsilon_{i,t}$  rather than in the regression function. The model is

$$\begin{aligned}\sigma_{i,t}^2 &= \sigma^2 \exp(\gamma \text{Loadfactor}_{i,t}) \\ &= \exp(\gamma_1 + \gamma_2 \text{Loadfactor}_{i,t}).\end{aligned}$$

The constant in the implied regression is  $\gamma_1 = \ln \sigma^2$ . Figure 9.2 shows a plot of the least squares residuals against *Load factor* for the 90 observations. The figure does suggest the presence of heteroscedasticity. (The dashed lines are placed to highlight the effect.) We computed the LM statistic using (9-28). The chi-squared statistic is 2.959. This is smaller than the critical value of 3.84 for one degree of freedom, so on this basis, the null hypothesis of homoscedasticity with respect to the load factor is not rejected.

**FIGURE 9.2** Plot of Residuals Against Load Factor.



**282 PART II ♦ Generalized Regression Model and Equation Systems**
**TABLE 9.2** Multiplicative Heteroscedasticity Model

	<i>Constant</i>	<i>Ln Q</i>	<i>Ln<sup>2</sup> Q</i>	<i>Ln P<sub>f</sub></i>	<i>R</i> <sup>2</sup>	<i>Sum of Squares</i>
OLS	9.1382	0.92615	0.029145	0.41006	0.9861674 <sup>c</sup>	1.577479 <sup>d</sup>
	0.24507 <sup>a</sup>	0.032306	0.012304	0.018807		
	0.22595 <sup>b</sup>	0.030128	0.011346	0.017524		
Two step	9.2463	0.92136	0.024450	0.40352	0.986119	1.612938
	0.21896	0.033028	0.011412	0.016974		
Iterated <sup>e</sup>	9.2774	0.91609	0.021643	0.40174	0.986071	1.645693
	0.20977	0.032993	0.011017	0.016332		

<sup>a</sup>Conventional OLS standard errors<sup>b</sup>White robust standard errors<sup>c</sup>Squared correlation between actual and fitted values<sup>d</sup>Sum of squared residuals<sup>e</sup>Values of  $c_2$  by iteration: 8.254344, 11.622473, 11.705029, 11.710618, 11.711012, 11.711040, 11.711042

To begin, we use OLS to estimate the parameters of the cost function and the set of residuals,  $e_{i,t}$ . Regression of  $\log(e_{i,t}^2)$  on a constant and the load factor provides estimates of  $\gamma_1$  and  $\gamma_2$ , denoted  $c_1$  and  $c_2$ . The results are shown in Table 9.2. As Harvey notes,  $\exp(c_1)$  does not necessarily estimate  $\sigma^2$  consistently—for normally distributed disturbances, it is low by a factor of 1.2704. However, as seen in (9-29), the estimate of  $\sigma^2$  (biased or otherwise) is not needed to compute the FGLS estimator. Weights  $w_{i,t} = \exp(-c_1 - c_2 \text{Loadfactor}_{i,t})$  are computed using these estimates, then weighted least squares using (9-30) is used to obtain the FGLS estimates of  $\beta$ . The results of the computations are shown in Table 9.2.

We might consider iterating the procedure. Using the results of FGLS at step 2, we can recompute the residuals, then recompute  $c_1$  and  $c_2$  and the weights, and then reenter the iteration. The process converges when the estimate of  $c_2$  stabilizes. This requires seven iterations. The results are shown in Table 9.2. As noted earlier, iteration does not produce any gains here. The second step estimator is already fully efficient. Moreover, this does not produce the MLE, either. That would be obtained by regressing  $[e_{i,t}^2 / \exp(c_1 + c_2 \text{Loadfactor}_{i,t}) - 1]$  on the constant and load factor at each iteration to obtain the new estimates. We will revisit this in Chapter 14.

### 9.7.2 GROUPWISE HETEROSCEDASTICITY

A groupwise heteroscedastic regression has the structural equations

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n,$$

$$E[\varepsilon_i | \mathbf{x}_i] = 0, \quad i = 1, \dots, n.$$

The  $n$  observations are grouped into  $G$  groups, each with  $n_g$  observations. The slope vector is the same in all groups, but within group  $g$

$$\text{Var}[\varepsilon_{ig} | \mathbf{x}_{ig}] = \sigma_g^2, \quad i = 1, \dots, n_g.$$

If the variances are known, then the GLS estimator is

$$\hat{\boldsymbol{\beta}} = \left[ \sum_{g=1}^G \left( \frac{1}{\sigma_g^2} \right) \mathbf{X}'_g \mathbf{X}_g \right]^{-1} \left[ \sum_{g=1}^G \left( \frac{1}{\sigma_g^2} \right) \mathbf{X}'_g \mathbf{y}_g \right]. \quad (9-35)$$

CHAPTER 9 ♦ The Generalized Regression Model **283**

Because  $\mathbf{X}'_g \mathbf{y}_g = \mathbf{X}'_g \mathbf{X}_g \mathbf{b}_g$ , where  $\mathbf{b}_g$  is the OLS estimator in the  $g$ th subset of observations,

$$\hat{\beta} = \left[ \sum_{g=1}^G \left( \frac{1}{\sigma_g^2} \right) \mathbf{X}'_g \mathbf{X}_g \right]^{-1} \left[ \sum_{g=1}^G \left( \frac{1}{\sigma_g^2} \right) \mathbf{X}'_g \mathbf{X}_g \mathbf{b}_g \right] = \left[ \sum_{g=1}^G \mathbf{V}_g \right]^{-1} \left[ \sum_{g=1}^G \mathbf{V}_g \mathbf{b}_g \right] = \sum_{g=1}^G \mathbf{W}_g \mathbf{b}_g.$$

This result is a matrix weighted average of the  $G$  least squares estimators. The weighting matrices are  $\mathbf{W}_g = [\sum_{g=1}^G (\text{Var}[\mathbf{b}_g])^{-1}]^{-1} (\text{Var}[\mathbf{b}_g])^{-1}$ . The estimator with the smaller covariance matrix therefore receives the larger weight. (If  $\mathbf{X}_g$  is the same in every group, then the matrix  $\mathbf{W}_g$  reduces to the simple,  $w_g \mathbf{I} = (h_g / \sum_g h_g) \mathbf{I}$  where  $h_g = 1/\sigma_g^2$ .)

The preceding is a useful construction of the estimator, but it relies on an algebraic result that might be unusable. If the number of observations in any group is smaller than the number of regressors, then the group specific OLS estimator cannot be computed. But, as can be seen in (9-35), that is not what is needed to proceed; what is needed are the weights. As always, pooled least squares is a consistent estimator, which means that using the group specific subvectors of the OLS residuals,

$$\hat{\sigma}_g^2 = \frac{\mathbf{e}'_g \mathbf{e}_g}{n_g} \quad (9-36)$$

provides the needed estimator for the group specific disturbance variance. Thereafter, (9-35) is the estimator and the inverse matrix in that expression gives the estimator of the asymptotic covariance matrix.

Continuing this line of reasoning, one might consider iterating the estimator by returning to (9-36) with the two-step FGLS estimator, recomputing the weights, then returning to (9-35) to recompute the slope vector. This can be continued until convergence. It can be shown [see Oberhofer and Kmenta (1974)] that so long as (9-36) is used without a degrees of freedom correction, then if this does converge, it will do so at the maximum likelihood estimator (with normally distributed disturbances).

For testing the homoscedasticity assumption, both White's test and the LM test are straightforward. The variables thought to enter the conditional variance are simply a set of  $G - 1$  group dummy variables, not including one of them (to avoid the dummy variable trap), which we'll denote  $\mathbf{Z}^*$ . Because the columns of  $\mathbf{Z}^*$  are binary and orthogonal, to carry out White's test, we need only regress the squared least squares residuals on a constant and  $\mathbf{Z}^*$  and compute  $NR^2$  where  $N = \sum_g n_g$ . The LM test is also straightforward. For purposes of this application of the LM test, it will prove convenient to replace the overall constant in  $\mathbf{Z}$  in (9-28), with the remaining group dummy variable. Since the column space of the full set of dummy variables is the same as that of a constant and  $G - 1$  of them, all results that follow will be identical. In (9-28), the vector  $\mathbf{g}$  will now be  $G$  subvectors where each subvector is the  $n_g$  elements of  $[(e_{ig}^2 / \hat{\sigma}^2) - 1]$ , and  $\hat{\sigma}^2 = \mathbf{e}' \mathbf{e} / N$ . By multiplying it out, we find that  $\mathbf{g}' \mathbf{Z}$  is the  $G$  vector with elements  $n_g [(\hat{\sigma}_g^2 / \hat{\sigma}^2) - 1]$ , while  $(\mathbf{Z}' \mathbf{Z})^{-1}$  is the  $G \times G$  matrix with diagonal elements  $1/n_g$ . It follows that

$$LM = \frac{1}{2} \mathbf{g}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{g} = \frac{1}{2} \sum_{g=1}^G n_g \left( \frac{\hat{\sigma}_g^2}{\hat{\sigma}^2} - 1 \right)^2. \quad (9-37)$$

Both statistics have limiting chi squared distributions with  $G - 1$  degrees of freedom under the null hypothesis of homoscedasticity. (There are only  $G - 1$  degrees of freedom because the hypothesis imposes  $G - 1$  restrictions, that the  $G$  variances are all equal to each other. Implicitly, one of the variances is free and the other  $G - 1$  equal to that one.)

## 284 PART II ♦ Generalized Regression Model and Equation Systems

### Example 9.5 Groupwise Heteroscedasticity

Baltagi and Griffin (1983) is a study of gasoline usage in 18 of the 30 OECD countries. The model analyzed in the paper is

$$\ln(\text{Gasoline usage}/\text{car})_{i,t} = \beta_1 + \beta_2 \ln(\text{Per capita income})_{i,t} + \beta_3 \ln(\text{Price}_{i,t}) \\ + \beta_4 \ln(\text{Cars per capita})_{i,t} + \varepsilon_{i,t}$$

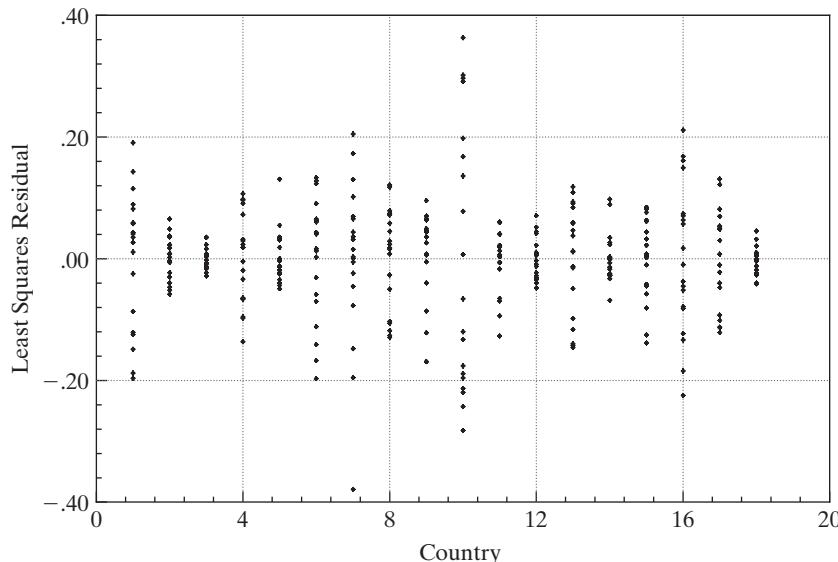
where  $i$  = country and  $t$  = 1960, ..., 1978. This is a balanced panel (see Section 9.2) with 19(18) = 342 observations in total. The data are given in Appendix Table F9.2.

Figure 9.3 displays the OLS residuals using the least squares estimates of the model above with the addition of 18 country dummy variables (18) (and without the overall constant). (The country dummy variables are used so that the country-specific residuals will have mean zero.) The  $F$  statistic for testing the null hypothesis that all the constants are equal is

$$F[(G-1), (\sum_g n_g - K - G)] = \frac{(\mathbf{e}'_0 \mathbf{e}_0 - \mathbf{e}'_1 \mathbf{e}_1)/(G-1)}{(\mathbf{e}'_1 \mathbf{e}_1 / \sum_g n_g - K - G)} \\ = \frac{(14.90436 - 2.73649)/17}{2.73649/(342 - 3 - 18)} = 83.960798,$$

where  $\mathbf{e}_0$  is the vector of residuals in the regression with a single constant term and  $\mathbf{e}_1$  is the regression with country specific constant terms. The critical value from the  $F$  table with 17 and 321 degrees of freedom is 1.655. The regression results are given in Table 9.3. Figure 9.3 does convincingly suggest the presence of groupwise heteroscedasticity. The White and LM statistics are  $342(0.38365) = 131.21$  and 279.588, respectively. The critical value from the chi-squared distribution with 17 degrees of freedom is 27.587. So, we reject the hypothesis of homoscedasticity and proceed to fit the model by feasible GLS. The two-step estimators are shown in Table 9.3. The FGTS estimator is computed by using weighted least squares, where the weights are  $1/\hat{\sigma}_g^2$  for each observation in country  $g$ . Comparing the White standard errors to the two-step estimators, we see that in this instance, there is a substantial gain to using feasible generalized least squares.

**FIGURE 9.3** Plot of OLS Residuals by Country.



CHAPTER 9 ♦ The Generalized Regression Model **285****TABLE 9.3** Estimated Gasoline Consumption Equations

	<b>OLS</b>			<b>FGLS</b>	
	<b>Coefficient</b>	<b>Std. Error</b>	<b>White Std. Err.</b>	<b>Coefficient</b>	<b>Std. Error</b>
In Income	0.66225	0.07339	0.07277	0.57507	0.02927
In Price	-0.32170	0.04410	0.05381	-0.27967	0.03519
Cars/Cap.	-0.64048	0.02968	0.03876	-0.56540	0.01613
Country 1	2.28586	0.22832	0.22608	2.43707	0.11308
Country 2	2.16555	0.21290	0.20983	2.31699	0.10225
Country 3	3.04184	0.21864	0.22479	3.20652	0.11663
Country 4	2.38946	0.20809	0.20783	2.54707	0.10250
Country 5	2.20477	0.21647	0.21087	2.33862	0.10101
Country 6	2.14987	0.21788	0.21846	2.30066	0.10893
Country 7	2.33711	0.21488	0.21801	2.57209	0.11206
Country 8	2.59233	0.24369	0.23470	2.72376	0.11384
Country 9	2.23255	0.23954	0.22973	2.34805	0.10795
Country 10	2.37593	0.21184	0.22643	2.58988	0.11821
Country 11	2.23479	0.21417	0.21311	2.39619	0.10478
Country 12	2.21670	0.20304	0.20300	2.38486	0.09950
Country 13	1.68178	0.16246	0.17133	1.90306	0.08146
Country 14	3.02634	0.39451	0.39180	3.07825	0.20407
Country 15	2.40250	0.22909	0.23280	2.56490	0.11895
Country 16	2.50999	0.23566	0.26168	2.82345	0.13326
Country 17	2.34545	0.22728	0.22322	2.48214	0.10955
Country 18	3.05525	0.21960	0.22705	3.21519	0.11917

**9.8 SUMMARY AND CONCLUSIONS**

This chapter has introduced a major extension of the classical linear model. By allowing for heteroscedasticity and autocorrelation in the disturbances, we expand the range of models to a large array of frameworks. We will explore these in the next several chapters. The formal concepts introduced in this chapter include how this extension affects the properties of the least squares estimator, how an appropriate estimator of the asymptotic covariance matrix of the least squares estimator can be computed in this extended modeling framework and, finally, how to use the information about the variances and covariances of the disturbances to obtain an estimator that is more efficient than ordinary least squares.

We have analyzed in detail one form of the generalized regression model, the model of heteroscedasticity. We first considered least squares estimation. The primary result for least squares estimation is that it retains its consistency and asymptotic normality, but some correction to the estimated asymptotic covariance matrix may be needed for appropriate inference. The White estimator is the standard approach for this computation. After examining two general tests for heteroscedasticity, we then narrowed the model to some specific parametric forms, and considered weighted (generalized) least squares for efficient estimation and maximum likelihood estimation. If the form of the heteroscedasticity is known but involves unknown parameters, then it remains uncertain whether FGLS corrections are better than OLS. Asymptotically, the comparison is clear, but in small or moderately sized samples, the additional variation incorporated by the estimated variance parameters may offset the gains to GLS.

## 286 PART II ♦ Generalized Regression Model and Equation Systems

### Key Terms and Concepts

- Aitken's theorem
- Asymptotic properties
- Autocorrelation
- Breusch–Pagan Lagrange multiplier test
- Efficient estimator
- Feasible generalized least squares (FGLS)
- Finite-sample properties
- Generalized least squares (GLS)
- Generalized linear regression model
- Generalized sum of squares
- Groupwise heteroscedasticity
- Heteroscedasticity
- Kruskal's theorem
- Lagrange multiplier test
- Multiplicative heteroscedasticity
- Nonconstructive test
- Ordinary least squares (OLS)
- Panel data
- Parametric model
- Robust estimation
- Robust estimator
- Robustness to unknown heteroscedasticity
- Semiparametric model
- Specification test
- Spherical disturbances
- Two-step estimator
- Wald test
- Weighted least squares (WLS)
- White heteroscedasticity consistent estimator
- White test

### Exercises

1. What is the covariance matrix,  $\text{Cov}[\hat{\beta}, \hat{\beta} - \mathbf{b}]$ , of the GLS estimator  $\hat{\beta} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y}$  and the difference between it and the OLS estimator,  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ ? The result plays a pivotal role in the development of specification tests in Hausman (1978).
2. This and the next two exercises are based on the test statistic usually used to test a set of  $J$  linear restrictions in the generalized regression model

$$F[J, n - K] = \frac{(\mathbf{R}\hat{\beta} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta} - \mathbf{q})/J}{(\mathbf{y} - \mathbf{X}\hat{\beta})'\Omega^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})/(n - K)},$$

where  $\hat{\beta}$  is the GLS estimator. Show that if  $\Omega$  is known, if the disturbances are normally distributed and if the null hypothesis,  $\mathbf{R}\beta = \mathbf{q}$ , is true, then this statistic is exactly distributed as  $F$  with  $J$  and  $n - K$  degrees of freedom. What assumptions about the regressors are needed to reach this conclusion? Need they be non-stochastic?

3. Now suppose that the disturbances are not normally distributed, although  $\Omega$  is still known. Show that the limiting distribution of previous statistic is  $(1/J)$  times a chi-squared variable with  $J$  degrees of freedom. (*Hint:* The denominator converges to  $\sigma^2$ .) Conclude that in the generalized regression model, the limiting distribution of the Wald statistic

$$W = (\mathbf{R}\hat{\beta} - \mathbf{q})'\{\mathbf{R}(\text{Est. Var}[\hat{\beta}])\mathbf{R}'\}^{-1}(\mathbf{R}\hat{\beta} - \mathbf{q})$$

is chi-squared with  $J$  degrees of freedom, regardless of the distribution of the disturbances, as long as the data are otherwise well behaved. Note that in a finite sample, the true distribution may be approximated with an  $F[J, n - K]$  distribution. It is a bit ambiguous, however, to interpret this fact as implying that the statistic is asymptotically distributed as  $F$  with  $J$  and  $n - K$  degrees of freedom, because the limiting distribution used to obtain our result is the chi-squared, not the  $F$ . In this instance, the  $F[J, n - K]$  is a random variable that tends asymptotically to the chi-squared variate.

CHAPTER 9 ♦ The Generalized Regression Model **287**

4. Finally, suppose that  $\Omega$  must be estimated, but that assumptions (9-16) and (9-17) are met by the estimator. What changes are required in the development of the previous problem?
5. In the generalized regression model, if the  $K$  columns of  $\mathbf{X}$  are characteristic vectors of  $\Omega$ , then ordinary least squares and generalized least squares are identical. (The result is actually a bit broader;  $\mathbf{X}$  may be any linear combination of exactly  $K$  characteristic vectors. This result is **Kruskal's theorem**.)
  - a. Prove the result directly using matrix algebra.
  - b. Prove that if  $\mathbf{X}$  contains a constant term and if the remaining columns are in deviation form (so that the column sum is zero), then the model of Exercise 8 is one of these cases. (The seemingly unrelated regressions model with identical regressor matrices, discussed in Chapter 10, is another.)
6. In the generalized regression model, suppose that  $\Omega$  is known.
  - a. What is the covariance matrix of the OLS and GLS estimators of  $\beta$ ?
  - b. What is the covariance matrix of the OLS residual vector  $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$ ?
  - c. What is the covariance matrix of the GLS residual vector  $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\beta}$ ?
  - d. What is the covariance matrix of the OLS and  $\hat{\mathbf{e}}$  residual vectors?
7. Suppose that  $y$  has the distribution  $f(y | \mathbf{x}) = (1/\mathbf{x}'\boldsymbol{\beta})^{-v/(\boldsymbol{\beta}'\mathbf{x})}$ ,  $y > 0$ .  
Then  $E[y | \mathbf{x}] = \boldsymbol{\beta}'\mathbf{x}$  and  $\text{Var}[y | \mathbf{x}] = (\boldsymbol{\beta}'\mathbf{x})^{-2}$ . For this model, prove that GLS and MLE are the same, even though this distribution involves the same parameters in the conditional mean function and the disturbance variance.
8. Suppose that the regression model is  $y = \mu + \varepsilon$ , where  $\varepsilon$  has a zero mean, constant variance, and equal correlation across observations. Then  $\text{Cov}[\varepsilon_i, \varepsilon_j] = \sigma^2\rho$  if  $i \neq j$ . Prove that the least squares estimator of  $\mu$  is inconsistent. Find the characteristic roots of  $\Omega$  and show that Condition 2 after Theorem 9.2 is violated.
9. Suppose that the regression model is  $y_i = \mu + \varepsilon_i$ , where

$$E[\varepsilon_i | x_i] = 0, \text{Cov}[\varepsilon_i, \varepsilon_j | x_i, x_j] = 0 \quad \text{for } i = j, \text{ but } \text{Var}[\varepsilon_i | x_i] = \sigma^2 x_i^2, x_i > 0.$$

- a. Given a sample of observations on  $y_i$  and  $x_i$ , what is the most efficient estimator of  $\mu$ ? What is its variance?
- b. What is the OLS estimator of  $\mu$ , and what is the variance of the ordinary least squares estimator?
- c. Prove that the estimator in part a is at least as efficient as the estimator in part b.
10. For the model in Exercise 9, what is the probability limit of  $s^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$ ? Note that  $s^2$  is the least squares estimator of the residual variance. It is also  $n$  times the conventional estimator of the variance of the OLS estimator,

$$\text{Est. Var} [\bar{y}] = s^2 (\mathbf{X}'\mathbf{X})^{-1} = \frac{s^2}{n}.$$

How does this equation compare with the true value you found in part b of Exercise 9? Does the conventional estimator produce the correct estimator of the true asymptotic variance of the least squares estimator?

11. For the model in Exercise 9, suppose that  $\varepsilon$  is normally distributed, with mean zero and variance  $\sigma^2[1 + (\gamma x)^2]$ . Show that  $\sigma^2$  and  $\gamma^2$  can be consistently estimated by a regression of the least squares residuals on a constant and  $x^2$ . Is this estimator efficient?

**288 PART II ♦ Generalized Regression Model and Equation Systems**

12. Two samples of 50 observations each produce the following moment matrices. (In each case,  $\mathbf{X}$  is a constant and one variable.)

Sample 1	Sample 2
$\mathbf{X}'\mathbf{X}$	$\mathbf{X}'\mathbf{X}$
$\begin{bmatrix} 50 & 300 \\ 300 & 2100 \end{bmatrix}$	$\begin{bmatrix} 50 & 300 \\ 300 & 2100 \end{bmatrix}$
$\mathbf{y}'\mathbf{X}$	$\mathbf{y}'\mathbf{X}$
$[300 \quad 2000]$	$[300 \quad 2200]$
$\mathbf{y}'\mathbf{y}$	$\mathbf{y}'\mathbf{y}$
[2100]	[2800]

- Compute the least squares regression coefficients and the residual variances  $s^2$  for each data set. Compute the  $R^2$  for each regression.
- Compute the OLS estimate of the coefficient vector assuming that the coefficients and disturbance variance are the same in the two regressions. Also compute the estimate of the asymptotic covariance matrix of the estimate.
- Test the hypothesis that the variances in the two regressions are the same without assuming that the coefficients are the same in the two regressions.
- Compute the two-step FGLS estimator of the coefficients in the regressions, assuming that the constant and slope are the same in both regressions. Compute the estimate of the covariance matrix and compare it with the result of part b.

### **Applications**

1. This application is based on the following data set.

<b>50 Observations on <math>\mathbf{y}</math>:</b>								
-1.42	2.75	2.10	-5.08	1.49	1.00	0.16	-1.11	1.66
-0.26	-4.87	5.94	2.21	-6.87	0.90	1.61	2.11	-3.82
-0.62	7.01	26.14	7.39	0.79	1.93	1.97	-23.17,	-2.52
-1.26	-0.15	3.41	-5.45	1.31	1.52	2.04	3.00	6.31
5.51	-15.22	-1.47	-1.48	6.66	1.78	2.62	-5.16	-4.71
-0.35	-0.48	1.24	0.69	1.91				
<b>50 Observations on <math>\mathbf{x}_1</math>:</b>								
-1.65	1.48	0.77	0.67	0.68	0.23	-0.40	-1.13	0.15
-0.63	0.34	0.35	0.79	0.77	-1.04	0.28	0.58	-0.41
-1.78	1.25	0.22	1.25	-0.12	0.66	1.06	-0.66	-1.18
-0.80	-1.32	0.16	1.06	-0.60	0.79	0.86	2.04	-0.51
0.02	0.33	-1.99	0.70	-0.17	0.33	0.48	1.90	-0.18
-0.18	-1.62	0.39	0.17	1.02				
<b>50 Observations on <math>\mathbf{x}_2</math>:</b>								
-0.67	0.70	0.32	2.88	-0.19	-1.28	-2.72	-0.70	-1.55
-0.74	-1.87	1.56	0.37	-2.07	1.20	0.26	-1.34	-2.10
0.61	2.32	4.38	2.16	1.51	0.30	-0.17	7.82	-1.15
1.77	2.92	-1.94	2.09	1.50	-0.46	0.19	-0.39	1.54
1.87	-3.45	-0.88	-1.53	1.42	-2.70	1.77	-1.89	-1.85
2.01	1.26	-2.02	1.91	-2.23				

CHAPTER 9 ♦ The Generalized Regression Model **289**

- a. Compute the ordinary least squares regression of  $y$  on a constant,  $x_1$ , and  $x_2$ . Be sure to compute the conventional estimator of the asymptotic covariance matrix of the OLS estimator as well.
- b. Compute the White estimator of the appropriate asymptotic covariance matrix for the OLS estimates.
- c. Test for the presence of heteroscedasticity using White's general test. Do your results suggest the nature of the heteroscedasticity?
- d. Use the Breusch-Pagan/Godfrey Lagrange multiplier test to test for heteroscedasticity.
- e. Reestimate the parameters using a two-step FGLS estimator. Use Harvey's formulation,  $\text{Var}[\varepsilon_i | x_{i1}, x_{i2}] = \sigma^2 \exp(\gamma_1 x_{i1} + \gamma_2 x_{i2})$ .

2. In the study of gasoline consumption in Example 9.7 using Baltagi and Griffin's data, we did not use another variable in the data set, LCARPCAP, which is the log of the number of cars per capita in the country. Repeat the analysis after adding this variable to the model. First determine whether this variable "belongs" in the model—that is, using an appropriate standard error, test the significance of the coefficient on this variable in the model.

(We look ahead to our use of maximum likelihood to estimate the models discussed in this chapter in Chapter 14.) In Example 9.7 we computed an iterated FGLS estimator using the airline data and the model  $\text{Var}[\varepsilon_{it} | \text{Loadfactor}] = \exp(\gamma_1 + \gamma_2 \text{Loadfactor})$ . The weights computed at each iteration were computed by estimating  $(\gamma_1, \gamma_2)$  by least squares regression of  $\ln \hat{\varepsilon}_{i,t}^2$  on a constant and Loadfactor. The maximum likelihood estimator would proceed along similar lines, however the weights would be computed by regression of  $[\hat{\varepsilon}_{i,t}^2 / \hat{\sigma}_{i,t}^2 - 1]$  on a constant and Loadfactor instead. Use this alternative procedure to estimate the model. Do you get different results?

# 10

## SYSTEMS OF EQUATIONS

---

### 10.1 INTRODUCTION

There are many settings in which the single equation models of the previous chapters apply to a group of related variables. In these contexts, it makes sense to consider the several models jointly. Some examples follow.

1. Munnell's (1990) model for output by the 48 continental U.S. states is

$$\begin{aligned} \ln GSP_{it} = & \beta_{1i} + \beta_{2i} \ln pc_{it} + \beta_{3i} \ln hwy_{it} + \beta_{4i} \ln water_{it} + \beta_{5i} \ln util_{it} \\ & + \beta_{6i} \ln emp_{it} + \beta_{7i} unemp_{it} + \varepsilon_{it}. \end{aligned}$$

Taken one state at a time, this provides a set of 48 linear regression models. The application develops a model in which the observations are correlated across time within a state. An important question pursued here and in the applications in the next example is whether it is valid to assume that the coefficient vector is the same for all states (individuals) in the sample.

2. The capital asset pricing model of finance specifies that for a given security,

$$r_{it} - r_{ft} = \alpha_i + \beta_i(r_{mt} - r_{ft}) + \varepsilon_{it},$$

where  $r_{it}$  is the return over period  $t$  on security  $i$ ,  $r_{ft}$  is the return on a risk-free security,  $r_{mt}$  is the market return, and  $\beta_i$  is the security's beta coefficient. The disturbances are obviously correlated across securities. The knowledge that the return on security  $i$  exceeds the risk-free rate by a given amount gives some information about the excess return of security  $j$ , at least for some  $j$ 's. It would be useful to estimate the equations jointly rather than ignore this connection.

3. Pesaran and Smith (1995) proposed a dynamic model for wage determination in 38 UK industries. The central equation is of the form

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}_i + \gamma_i y_{i,t-1} + \varepsilon_{it}.$$

Nair-Reichert and Weinhold's (2001) cross-country analysis of growth of developing countries takes the same form. In both cases, each group (industry, country) could be analyzed separately. However, the connections across groups and the interesting question of "poolability"—that is, whether it is valid to assume identical coefficients—is a central part of the analysis. The lagged dependent variable in the model produces a substantial complication.

4. In a model of production, the optimization conditions of economic theory imply that if a firm faces a set of factor prices  $\mathbf{p}$ , then its set of cost-minimizing factor demands for producing output  $Q$  will be a set of equations of the form  $x_m = f_m(Q, \mathbf{p})$ .

CHAPTER 10 ♦ Systems of Equations **291**

The empirical model is

$$x_1 = f_1(Q, \mathbf{p}|\boldsymbol{\theta}) + \varepsilon_1,$$

$$x_2 = f_2(Q, \mathbf{p}|\boldsymbol{\theta}) + \varepsilon_2,$$

...

$$x_M = f_M(Q, \mathbf{p}|\boldsymbol{\theta}) + \varepsilon_M,$$



where  $\boldsymbol{\theta}$  is a vector of parameters that are part of the technology and  $\varepsilon_m$  represents errors in optimization. Once again, the disturbances should be correlated. In addition, the same parameters of the production technology will enter all the demand equations, so the set of equations has cross-equation restrictions. Estimating the equations separately will waste the information that the same set of parameters appears in all the equations.

5. The essential form of a model for equilibrium in a market is

$$Q_{Demand} = \alpha_1 + \alpha_2 Price + \alpha_3 Income + \mathbf{d}'\boldsymbol{\alpha} + \varepsilon_{Demand},$$

$$Q_{Supply} = \beta_1 + \beta_2 Price + \mathbf{s}'\boldsymbol{\beta} + \varepsilon_{Supply},$$

$$Q_{Equilibrium} = Q_{Demand} = Q_{Supply},$$

where  $\mathbf{d}$  and  $\mathbf{s}$  are other variables that influence the equilibrium through their impact on the demand and supply curves, respectively. This model differs from those suggested thus far because the implication of the third equation is that *Price* is not exogenous in the equation system. The equations of this model fit more appropriately in the instrumental variables framework developed in Chapter 8 than in the regression models developed in Chapters 1 to 7. The multiple equations framework developed in this chapter provides additional results for estimating “simultaneous equations models” such as this.

The multiple equations regression model developed in this chapter provides a modeling framework that can be used in many different settings. The models of production and cost developed in Section 10.5 provide the platform for the literature on empirical analysis of firm behavior. At the macroeconomic level, the “vector autoregression models” used in Chapters 21–23 are specific forms of the seemingly unrelated regressions model of Section 10.2. The simultaneous equations model presented in Section 10.6 lies behind the specification of the large variety of specifications considered in Chapter 8.

This chapter will develop the essential theory for sets of related regression equations. Section 10.2 examines the general model in which each equation has its own fixed set of parameters, and it examines efficient estimation techniques. Section 10.2.6 examines the “pooled” model with identical coefficients in all equations. Production and consumer demand models are a special case of the general model in which the equations of the model obey an adding-up constraint that has important implications for specification and estimation. Section 10.3 suggests extensions of the seemingly unrelated regression model to the generalized regression models with heteroscedasticity and autocorrelation that are developed in Chapter 9. Section 10.4 broadens the seemingly unrelated regressions model to nonlinear systems of equations. In Section 10.5, we examine a classic application of the seemingly unrelated regressions model that illustrates the

## 292 PART II ♦ Generalized Regression Model and Equation Systems

interesting features of the current genre of demand studies in the applied literature. The seemingly unrelated regressions model is then extended to the translog specification, which forms the platform for most recent microeconomic studies of production and cost. Finally, Section 10.6 merges the results of Chapter 8 on models with endogenous variables with the development in this chapter of multiple equation systems. In Section 10.6, we will develop **simultaneous equations models**. These are systems of equations that build on the seemingly unrelated regressions model to produce equation systems that include interrelationships among the dependent variables. The supply and demand model suggested in the chapter introduction, of equilibrium in which price and quantity in a market are jointly determined, is an application.

### 10.2 THE SEEMINGLY UNRELATED REGRESSIONS MODEL

All the examples suggested in the chapter introduction have a common multiple equation structure, which we may write as

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1, \\ \mathbf{y}_2 &= \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2, \\ &\dots \\ \mathbf{y}_M &= \mathbf{X}_M\boldsymbol{\beta}_M + \boldsymbol{\varepsilon}_M. \end{aligned} \tag{10-1}$$

There are  $M$  equations and  $T$  observations in the sample of data used to estimate them.<sup>1</sup> The second and third examples embody different types of constraints across equations and different structures of the disturbances. A basic set of principles will apply to them all, however.<sup>2</sup> The **seemingly unrelated regressions** (SUR) model in (10-1) is

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, M. \tag{10-2}$$

Define the  $MT \times 1$  vector of disturbances,

$$\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}'_1, \boldsymbol{\varepsilon}'_2, \dots, \boldsymbol{\varepsilon}'_M]'$$

We assume strict exogeneity of  $\mathbf{X}_i$ ,

$$E[\boldsymbol{\varepsilon} | \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M] = \mathbf{0},$$

and homoscedasticity

$$E[\boldsymbol{\varepsilon}_m \boldsymbol{\varepsilon}'_m | \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M] = \sigma_{mm} \mathbf{I}_T.$$

We assume that a total of  $T$  observations are used in estimating the parameters of the  $M$  equations.<sup>3</sup> Each equation involves  $K_i$  regressors, for a total of  $K = \sum_{i=1}^M K_i$ . We will require  $T > K_i$ . The data are assumed to be well behaved, as described in

<sup>1</sup>The use of  $T$  is not meant to imply any necessary connection to time series. For instance, in the fourth example, above, the data might be cross sectional.

<sup>2</sup>See the surveys by Srivastava and Dwivedi (1979), Srivastava and Giles (1987), and Fiebig (2001).

<sup>3</sup>There are a few results for unequal numbers of observations, such as Schmidt (1977), Baltagi, Garvin, and Kerman (1989), Conniffe (1985), Hwang (1990), and Im (1994). But, the case of fixed  $T$  is the norm in practice.

CHAPTER 10 ♦ Systems of Equations **293**

 Section 4.7.1, and we shall not treat the issue separately here. For the present, we also assume that disturbances are uncorrelated across observations but correlated across equations. Therefore,

$$E[\varepsilon_i \varepsilon_j | \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M] = \sigma_{ij}, \quad \text{if } t = s \text{ and 0 otherwise.}$$

The disturbance formulation is, therefore,

$$E[\varepsilon_i \varepsilon'_j | \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M] = \sigma_{ij} \mathbf{I}_T,$$

or

$$E[\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}' | \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M] = \boldsymbol{\Omega} = \begin{bmatrix} \sigma_{11} \mathbf{I} & \sigma_{12} \mathbf{I} & \cdots & \sigma_{1M} \mathbf{I} \\ \sigma_{21} \mathbf{I} & \sigma_{22} \mathbf{I} & \cdots & \sigma_{2M} \mathbf{I} \\ \vdots & & & \\ \sigma_{M1} \mathbf{I} & \sigma_{M2} \mathbf{I} & \cdots & \sigma_{MM} \mathbf{I} \end{bmatrix}. \quad (10-3)$$

 It will be convenient in the discussion below to have a term for the particular kind of model in which the data matrices are group specific data sets on the same set of variables. The Brunfeld model noted in the introduction is such a case. This special case of the seemingly unrelated regressions model is a **multivariate regression model**. In contrast, the cost function model examined in Section 10.4.1 is not of this type—it consists of a cost function that involves output and prices and a set of cost share equations that have only a set of constant terms. We emphasize, this is merely a convenient term for a specific form of the SUR model, not a modification of the model itself.

### 10.2.1 GENERALIZED LEAST SQUARES

Each equation is, by itself, a classical regression. Therefore, the parameters could be estimated consistently, if not efficiently, one equation at a time by ordinary least squares. The **generalized regression model** applies to the stacked model,

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0} \\ & & \vdots & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_M \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_M \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_M \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \quad (10-4)$$

Therefore, the efficient estimator is generalized least squares.<sup>4</sup> The model has a particularly convenient form. For the  $t$ th observation, the  $M \times M$  covariance matrix of the disturbances is

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2M} \\ \vdots & & & \\ \sigma_{M1} & \sigma_{M2} & \cdots & \sigma_{MM} \end{bmatrix}, \quad (10-5)$$

<sup>4</sup>See Zellner (1962) and Telser (1964).

## 294 PART II ♦ Generalized Regression Model and Equation Systems

so, in (10-3),

$$\Omega = \Sigma \otimes \mathbf{I} \quad (10-6)$$

and

$$\Omega^{-1} = \Sigma^{-1} \otimes \mathbf{I}^5.$$

Denoting the  $ij$ th element of  $\Sigma^{-1}$  by  $\sigma^{ij}$ , we find that the GLS estimator is

$$\hat{\beta} = [\mathbf{X}' \Omega^{-1} \mathbf{X}]^{-1} \mathbf{X}' \Omega^{-1} \mathbf{y} = [\mathbf{X}' (\Sigma^{-1} \otimes \mathbf{I}) \mathbf{X}]^{-1} \mathbf{X}' (\Sigma^{-1} \otimes \mathbf{I}) \mathbf{y}. \quad (10-7)$$

Expanding the **Kronecker products** produces

$$\hat{\beta} = \begin{bmatrix} \sigma^{11} \mathbf{X}'_1 \mathbf{X}_1 & \sigma^{12} \mathbf{X}'_1 \mathbf{X}_2 & \cdots & \sigma^{1M} \mathbf{X}'_1 \mathbf{X}_M \\ \sigma^{21} \mathbf{X}'_2 \mathbf{X}_1 & \sigma^{22} \mathbf{X}'_2 \mathbf{X}_2 & \cdots & \sigma^{2M} \mathbf{X}'_2 \mathbf{X}_M \\ \vdots & & & \vdots \\ \sigma^{M1} \mathbf{X}'_M \mathbf{X}_1 & \sigma^{M2} \mathbf{X}'_M \mathbf{X}_2 & \cdots & \sigma^{MM} \mathbf{X}'_M \mathbf{X}_M \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^M \sigma^{1j} \mathbf{X}'_1 \mathbf{y}_j \\ \sum_{j=1}^M \sigma^{2j} \mathbf{X}'_2 \mathbf{y}_j \\ \vdots \\ \sum_{j=1}^M \sigma^{Mj} \mathbf{X}'_M \mathbf{y}_j \end{bmatrix}.$$

The asymptotic covariance matrix for the GLS estimator is the bracketed inverse matrix in (10-7). All the results of Chapter 8 for the generalized regression model extend to this model (which has both **heteroscedasticity** and **autocorrelation**).

This estimator is obviously different from ordinary least squares. At this point, however, the equations are linked only by their disturbances—hence the name *seemingly unrelated* regressions model—so it is interesting to ask just how much efficiency is gained by using generalized least squares instead of ordinary least squares. Zellner (1962) and Dwivedi and Srivastava (1978) have analyzed some special cases in detail.

1. If the equations are *actually* unrelated—that is, if  $\sigma_{ij} = 0$  for  $i \neq j$ —then there is obviously no payoff to GLS estimation of the full set of equations. Indeed, full GLS is equation by equation OLS.<sup>6</sup>
2. If the equations have **identical explanatory variables**—that is, if  $\mathbf{X}_i = \mathbf{X}_j$ —then OLS and GLS are identical. We will turn to this case in Section 10.2.2.<sup>7</sup>
3. If the regressors in one block of equations are a subset of those in another, then GLS brings no efficiency gain over OLS in estimation of the smaller set of equations; thus, GLS and OLS are once again identical. We will look at an application of this result in Section 21.6.5.<sup>8</sup>

In the more general case, with unrestricted correlation of the disturbances and different regressors in the equations, the results are complicated and dependent on

<sup>5</sup>See Appendix Section A.5.5.

<sup>6</sup>See also Baltagi (1989) and Bartels and Fiebig (1992) for other cases in which OLS = GLS.

<sup>7</sup>An intriguing result, albeit probably of negligible practical significance, is that the result also applies if the  $\mathbf{X}$ 's are all nonsingular, and not necessarily identical, linear combinations of the same set of variables. The formal result which is a corollary of Kruskal's theorem [see Davidson and MacKinnon (1993, p. 294)] is that OLS and GLS will be the same if the  $K$  columns of  $\mathbf{X}$  are a linear combination of exactly  $K$  characteristic vectors of  $\Omega$ . By showing the equality of OLS and GLS here, we have verified the conditions of the corollary. The general result is pursued in the exercises. The intriguing result cited is now an obvious case.

<sup>8</sup>The result was analyzed by Goldberger (1970) and later by Revankar (1974) and Conniffe (1982a, b).

the data. Two propositions that apply generally are as follows:

1. The greater is the correlation of the disturbances, the greater is the efficiency gain accruing to GLS.
2. The less correlation there is between the  $\mathbf{X}$  matrices, the greater is the gain in efficiency in using GLS.<sup>9</sup>

### 10.2.2 SEEMINGLY UNRELATED REGRESSIONS WITH IDENTICAL REGRESSORS

The case of **identical regressors** is quite common, notably in capital asset pricing model in empirical finance—see the chapter introduction and Chapter 21. In this special case, generalized least squares is equivalent to equation by equation ordinary least squares. Impose the assumption that  $\mathbf{X}_i = \mathbf{X}_j = \mathbf{X}$ , so that  $\mathbf{X}'_i \mathbf{X}_j = \mathbf{X}' \mathbf{X}$  for all  $i$  and  $j$  in (10-7). The inverse matrix on the right-hand side now becomes  $[\Sigma^{-1} \otimes \mathbf{X}' \mathbf{X}]^{-1}$ , which, using (A-76), equals  $[\Sigma \otimes (\mathbf{X}' \mathbf{X})^{-1}]$ . Also on the right-hand side, each term  $\mathbf{X}'_i \mathbf{y}_j$  equals  $\mathbf{X}' \mathbf{y}_j$ , which, in turn equals  $\mathbf{X}' \mathbf{X} \mathbf{b}_j$ . With these results, after moving the common  $\mathbf{X}' \mathbf{X}$  out of the summations on the right-hand side, we obtain

$$\hat{\beta} = \begin{bmatrix} \sigma_{11}(\mathbf{X}' \mathbf{X})^{-1} & \sigma_{12}(\mathbf{X}' \mathbf{X})^{-1} & \dots & \sigma_{1M}(\mathbf{X}' \mathbf{X})^{-1} \\ \sigma_{21}(\mathbf{X}' \mathbf{X})^{-1} & \sigma_{22}(\mathbf{X}' \mathbf{X})^{-1} & \dots & \sigma_{2M}(\mathbf{X}' \mathbf{X})^{-1} \\ \vdots & & & \\ \sigma_{M1}(\mathbf{X}' \mathbf{X})^{-1} & \sigma_{M2}(\mathbf{X}' \mathbf{X})^{-1} & \dots & \sigma_{MM}(\mathbf{X}' \mathbf{X})^{-1} \end{bmatrix} \begin{bmatrix} (\mathbf{X}' \mathbf{X}) \sum_{l=1}^M \sigma^{1l} \mathbf{b}_l \\ (\mathbf{X}' \mathbf{X}) \sum_{l=1}^M \sigma^{2l} \mathbf{b}_l \\ \vdots \\ (\mathbf{X}' \mathbf{X}) \sum_{l=1}^M \sigma^{Ml} \mathbf{b}_l \end{bmatrix}. \quad (10-8)$$

Now, we isolate one of the subvectors, say the first, from  $\hat{\beta}$ . After multiplication, the moment matrices cancel, and we are left with

$$\hat{\beta}_1 = \sum_{j=1}^M \sigma_{1j} \sum_{l=1}^M \sigma^{jl} \mathbf{b}_l = \mathbf{b}_1 \left( \sum_{j=1}^M \sigma_{1j} \sigma^{j1} \right) + \mathbf{b}_2 \left( \sum_{j=1}^M \sigma_{1j} \sigma^{j2} \right) + \dots + \mathbf{b}_M \left( \sum_{j=1}^M \sigma_{1j} \sigma^{jM} \right).$$

The terms in parentheses are the elements of the first row of  $\Sigma \Sigma^{-1} = \mathbf{I}$ , so the end result is  $\hat{\beta}_1 = \mathbf{b}_1$ . For the remaining subvectors, which are obtained the same way,  $\hat{\beta}_i = \mathbf{b}_i$ , which is the result we sought.<sup>10</sup>

To reiterate, the important result we have here is that in the SUR model, when all equations have the same regressors, the efficient estimator is single-equation ordinary least squares; OLS is the same as GLS. Also, the asymptotic covariance matrix of  $\hat{\beta}$  for this case is given by the large inverse matrix in brackets in (10-8), which would be estimated by

$$\text{Est. Asy. Cov}[\hat{\beta}_i, \hat{\beta}_j] = \hat{\sigma}_{ij}(\mathbf{X}' \mathbf{X})^{-1}, \quad i, j = 1, \dots, M, \quad \text{where } \hat{\Sigma}_{ij} = \hat{\sigma}_{ij} = \frac{1}{T} \mathbf{e}'_i \mathbf{e}_j.$$

Except in some special cases, this general result is lost if there are any restrictions on  $\beta$ , either within or across equations. We will examine one of those cases, the block of zeros restriction, in Section 21.6.5.

<sup>9</sup>See also Binkley (1982) and Binkley and Nelson (1988).

<sup>10</sup>See Hashimoto and Ohtani (1990) for discussion of hypothesis testing in this case.

## 296 PART II ♦ Generalized Regression Model and Equation Systems

### 10.2.3 FEASIBLE GENERALIZED LEAST SQUARES

The preceding discussion assumes that  $\Sigma$  is known, which, as usual, is unlikely to be the case. FGLS estimators have been devised, however.<sup>11</sup> The least squares residuals may be used (of course) to estimate consistently the elements of  $\Sigma$  with

$$\hat{s}_{ij} = s_{ij} = \frac{\mathbf{e}_i' \mathbf{e}_j}{T}. \quad (10-9)$$

The consistency of  $s_{ij}$  follows from that of  $\mathbf{b}_i$  and  $\mathbf{b}_j$ .<sup>12</sup> A degrees of freedom correction in the divisor is occasionally suggested. Two possibilities that are unbiased when  $i = j$  are

$$s_{ij}^* = \frac{\mathbf{e}_i' \mathbf{e}_j}{[(T - K_i)(T - K_j)]^{1/2}} \quad \text{and} \quad s_{ij}^{**} = \frac{\mathbf{e}_i' \mathbf{e}_j}{T - \max(K_i, K_j)}. \quad (10-10)$$

Whether unbiasedness of the estimator of  $\Sigma$  used for FGLS is a virtue here is uncertain. The asymptotic properties of the **feasible GLS** estimator,  $\hat{\beta}$  do not rely on an unbiased estimator of  $\Sigma$ ; only consistency is required. All our results from Chapters 8 and 9 for FGLS estimators extend to this model, with no modification. We shall use (10-9) in what follows. With

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1M} \\ s_{21} & s_{22} & \cdots & s_{2M} \\ \vdots & & & \\ s_{M1} & s_{M2} & \cdots & s_{MM} \end{bmatrix} \quad (10-11)$$

in hand, FGLS can proceed as usual.

### 10.2.4 TESTING HYPOTHESES

For testing a hypothesis about  $\beta$ , a statistic analogous to the  $F$  ratio in multiple regression analysis is

$$F[J, MT - K] = \frac{(\mathbf{R}\hat{\beta} - \mathbf{q})' [\mathbf{R}(\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta} - \mathbf{q})/J}{\hat{\epsilon}'\hat{\Omega}^{-1}\hat{\epsilon}/(MT - K)}. \quad (10-12)$$

The computation requires the unknown  $\Omega$ . If we insert the FGLS estimate  $\hat{\Omega}$  based on (10-9) and use the result that the denominator converges to one, then, in large samples, the statistic will behave the same as

$$\hat{F} = \frac{1}{J} (\mathbf{R}\hat{\beta} - \mathbf{q})' [\mathbf{R} \widehat{\text{Var}}[\hat{\beta}] \mathbf{R}']^{-1} (\mathbf{R}\hat{\beta} - \mathbf{q}). \quad (10-13)$$

This can be referred to the standard  $F$  table. Because it uses the estimated  $\Sigma$ , even with normally distributed disturbances, the  $F$  distribution is only valid approximately. In general, the statistic  $F[J, n]$  converges to  $1/J$  times a chi-squared  $[J]$  as  $n \rightarrow \infty$ .

<sup>11</sup>See Zellner (1962) and Zellner and Huang (1962). The FGLS estimator for this model is also labeled **Zellner's efficient estimator**, or ZEF, in reference to Zellner (1962) where it was introduced.

<sup>12</sup>Perhaps surprisingly, if it is assumed that the density of  $\epsilon$  is symmetric, as it would be with normality, then  $\mathbf{b}_i$  is also unbiased. See Kakwani (1967).

<sup>13</sup>See, as well, Judge et al. (1985), Theil (1971), and Srivastava and Giles (1987).

## CHAPTER 10 ♦ Systems of Equations 297

Therefore, an alternative test statistic that has a limiting chi-squared distribution with  $J$  degrees of freedom when the null hypothesis is true is

$$J \hat{F} = (\mathbf{R} \hat{\beta} - \mathbf{q})' [\widehat{\mathbf{R} \text{Var}[\hat{\beta}] \mathbf{R}'}]^{-1} (\mathbf{R} \hat{\beta} - \mathbf{q}). \quad (10-14)$$

This can be recognized as a **Wald statistic** that measures the distance between  $\mathbf{R} \hat{\beta}$  and  $\mathbf{q}$ . Both statistics are valid asymptotically, but (10-13) may perform better in a small or moderately sized sample.<sup>14</sup> Once again, the divisor used in computing  $\hat{\sigma}_{ij}$  may make a difference, but there is no general rule.

A hypothesis of particular interest is the **homogeneity restriction** of equal coefficient vectors in the multivariate regression model. That case is fairly common in this setting. The homogeneity restriction is that  $\beta_i = \beta_M, i = 1, \dots, M-1$ . Consistent with (10-13)–(10-14), we would form the hypothesis as

$$\mathbf{R}\beta = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & -\mathbf{I} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} & -\mathbf{I} \\ & & \cdots & & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & -\mathbf{I} \end{bmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{pmatrix} = \begin{pmatrix} \beta_1 - \beta_M \\ \beta_2 - \beta_M \\ \vdots \\ \beta_{M-1} - \beta_M \end{pmatrix} = \mathbf{0}. \quad (10-15)$$

This specifies a total of  $(M-1)K$  restrictions on the  $KM \times 1$  parameter vector. Denote the estimated asymptotic covariance for  $(\hat{\beta}_i, \hat{\beta}_j)$  as  $\hat{\mathbf{V}}_{ij}$ . The bracketed matrix in (10-13) would have typical block

$$[\mathbf{R} \widehat{\text{Var}}[\hat{\beta}] \mathbf{R}']_{ij} = \hat{\mathbf{V}}_{ij} - \hat{\mathbf{V}}_{iM} - \hat{\mathbf{V}}_{Mj} + \hat{\mathbf{V}}_{MM}$$

This may be a considerable amount of computation. The test will be simpler if the model has been fit by maximum likelihood, as we examine in Section 14.9.3. Pesaran and Yamagata (2008) provide an alternative test that can be used when  $M$  is large and  $T$  is relatively small.

#### 10.2.5 A SPECIFICATION TEST FOR THE SUR MODEL

It is of interest to assess statistically whether the off diagonal elements of  $\Sigma$  are zero. If so, then the efficient estimator for the full parameter vector, absent heteroscedasticity or autocorrelation, is equation by equation ordinary least squares. There is no standard test for the general case of the SUR model unless the additional assumption of normality of the disturbances is imposed in (10-2) and (10-3). With normally distributed disturbances, the standard trio of tests, Wald, **likelihood ratio**, and **Lagrange multiplier**, can be used. For reasons we will turn to shortly, the Wald test is likely to be too cumbersome to apply. With normally distributed disturbances, the likelihood ratio statistic for testing the null hypothesis that the matrix  $\Sigma$  in (10-5) is a diagonal matrix against the alternative that  $\Sigma$  is simply an unrestricted positive definite matrix would be

$$\lambda_{LR} = T[\ln |\mathbf{S}_0| - \ln |\mathbf{S}_1|], \quad (10-16)$$

---

<sup>14</sup>See Judge et al. (1985, p. 476). The Wald statistic often performs poorly in the small sample sizes typical in this area. Fiebig (2001, pp. 108–110) surveys a recent literature on methods of improving the power of testing procedures in SUR models.

## 298 PART II ♦ Generalized Regression Model and Equation Systems

where  $\mathbf{S}_0$  is the residual covariance matrix defined in (10-9) (without a degrees of freedom correction). The residuals are computed using maximum likelihood estimates of the parameters, not FGLS.<sup>15</sup> Under the null hypothesis, the model would be efficiently estimated by individual equation OLS, so

$$\ln |\mathbf{S}_0| = \sum_{i=1}^M \ln (\mathbf{e}'_i \mathbf{e}_i / T),$$

where  $\mathbf{e}_i = \mathbf{y}_i - \mathbf{X}_i \mathbf{b}_i$ . The limiting distribution of the likelihood ratio statistic under the null hypothesis would be chi-squared with  $M(M - 1)/2$  degrees of freedom.

The likelihood ratio statistic requires the unrestricted MLE to compute the residual covariance matrix under the alternative, so it is can be cumbersome to compute. A simpler alternative is the Lagrange multiplier statistic developed by Breusch and Pagan (1980) which is

$$\begin{aligned} \lambda_{LM} &= T \sum_{i=2}^M \sum_{j=1}^{i-1} r_{ij}^2 \\ &= (T/2)[\text{trace}(\mathbf{R}' \mathbf{R}) - M], \end{aligned} \quad (10-17)$$

where  $\mathbf{R}$  is the sample correlation matrix of the  $M$  sets of  $T$  OLS residuals. This has the same large sample distribution under the null hypothesis as the likelihood ratio statistic, but is obviously easier to compute, as it only requires the OLS residuals.

The third test statistic in the trio is the Wald statistic. In principle, the Wald statistic for the SUR model would be computed using

$$W = \hat{\sigma}' [\text{Asy. Var}(\hat{\sigma})]^{-1} \hat{\sigma},$$

where  $\hat{\sigma}$  is the  $M(M - 1)/2$  length vector containing the estimates of the off-diagonal (lower triangle) elements of  $\Sigma$ , and the asymptotic covariance matrix of the estimator appears in the brackets. Under normality, the asymptotic covariance matrix contains the corresponding elements of  $2\Sigma \otimes \Sigma/T$ . It would be possible to estimate the covariance term more generally using a moment-based estimator. Because

$$\hat{\sigma}_{ij} = \frac{1}{T} \sum_{t=1}^T e_{it} e_{jt}$$

is a mean of  $T$  observations, one might use the conventional estimator of its variance and its covariance with  $\hat{\sigma}_{lm}$ , which would be

$$f_{ij,lm} = \frac{1}{T} \frac{1}{T-1} \sum_{t=1}^T (e_{it} e_{jt} - \hat{\sigma}_{ij})(e_{lt} e_{mt} - \hat{\sigma}_{lm}). \quad (10-18)$$

The modified Wald statistic would then be

$$W' = \hat{\sigma}' [\mathbf{F}]^{-1} \hat{\sigma}$$

<sup>15</sup>In the SUR model of this chapter, the MLE for normally distributed disturbances can be computed by iterating the FGLS procedure, back and forth between (10-7) and (10-9) until the estimates are no longer changing. We note, this procedure produces the MLE when it converges, but it is not guaranteed to converge, nor is it assured that there is a unique MLE. For our regional data set, the iterated FGLS procedure does not converge after 1,000 iterations. The Oberhofer-Kmenta (1974) result implies that if the iteration converges, it reaches the MLE. It does not guarantee that the iteration will converge, however. The problem with this application may be the very small sample size, 17 observations. One would not normally use the technique of maximum likelihood with a sample this small.

where the elements of  $\mathbf{F}$  are the corresponding values in (10-18). This computation is obviously more complicated than the other two. However, it does have the virtue that it does not require an assumption of normality of the disturbances in the model. What would be required is (a) consistency of the estimators of  $\beta_i$  so that we can assert (b) consistency of the estimators of  $\sigma_{ij}$  and, finally, (c) asymptotic normality of the estimators in (b) so that we can apply Theorem 4.4. All three requirements should be met in the SUR model with well-behaved regressors.

Alternative approaches that have been suggested [see, e.g., Johnson and Wichern (2005, p. 424)] are based on the following general strategy: Under the alternative hypothesis of an unrestricted  $\Sigma$ , the sample estimate of  $\Sigma$  will be  $\hat{\Sigma} = [\hat{\sigma}_{ij}]$  as defined in (10-9). Under any restrictive null hypothesis, the estimator of  $\Sigma$  will be  $\hat{\Sigma}_0$ , a matrix that by construction will be larger than  $\hat{\Sigma}_1$  in the matrix sense defined in Appendix A. Statistics based on the “excess variation,” such as  $T(\hat{\Sigma}_0 - \hat{\Sigma}_1)$  are suggested for the testing procedure. One of these is the likelihood ratio test in (10-16).

#### 10.2.6 THE POOLED MODEL

If the variables in  $\mathbf{X}_i$  are all the same and the coefficient vectors in (10-2) are assumed all to be equal, the **pooled model**,

$$y_{it} = \mathbf{x}'_{it}\beta + \varepsilon_{it}$$

results. This differs from the panel data treatment in Chapter 11, however, in that the correlation across observations is assumed to occur at time  $t$ , not within group  $i$ . (Of course, by a minor rearrangement of the data, the same model results. However, the interpretation differs, so we will maintain the distinction.) Collecting the  $T$  observations for group  $i$ , we obtain

$$\mathbf{y}_i = \mathbf{X}_i\beta + \boldsymbol{\varepsilon}_i$$

or, for all  $n$  groups,

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \beta + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix} = \mathbf{X}\beta + \boldsymbol{\varepsilon}, \quad (10-19)$$

where

$$\begin{aligned} E[\boldsymbol{\varepsilon}_i | \mathbf{X}] &= \mathbf{0}, \\ E[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}'_j | \mathbf{X}] &= \sigma_{ij} \boldsymbol{\Omega}_{ij}. \end{aligned} \quad (10-20)$$

If  $\boldsymbol{\Omega}_{ij} = \mathbf{I}$ , then this is equivalent to the SUR model of (10-2) with identical coefficient vectors. The generalized least squares estimator under this **covariance structures model** assumption is

$$\begin{aligned} \hat{\beta} &= [\mathbf{X}'(\Sigma \otimes \mathbf{I})^{-1}\mathbf{X}]^{-1}[\mathbf{X}'(\Sigma \otimes \mathbf{I})^{-1}\mathbf{y}] \\ &= \left[ \sum_{i=1}^n \sum_{j=1}^n \sigma^{ij} \mathbf{X}'_i \mathbf{X}_j \right]^{-1} \left[ \sum_{i=1}^n \sum_{j=1}^n \sigma^{ij} \mathbf{X}'_i \mathbf{y}_j \right]. \end{aligned} \quad (10-21)$$

## 300 PART II ♦ Generalized Regression Model and Equation Systems

where  $\sigma^{ij}$  denotes the  $ij$ th element of  $\Sigma^{-1}$ . The FGLS estimator can be computed using (10-9), where  $\mathbf{e}_t$  can either be computed using group-specific OLS residuals or it can be a subvector of the pooled OLS residual vector using all  $nT$  observations.

There is an important consideration to note in feasible GLS estimation of this model. The computation requires inversion of the matrix  $\hat{\Sigma}$  where the  $ij$ th element is given by (10-9). This matrix is  $n \times n$ . It is computed from the least squares residuals using

$$\hat{\Sigma} = \frac{1}{T} \sum_{t=1}^T \mathbf{e}_t \mathbf{e}_t' = \frac{1}{T} \mathbf{E}' \mathbf{E}, \quad (10-22)$$

where  $\mathbf{e}_t'$  is a  $1 \times n$  vector containing all  $n$  residuals for the  $n$  groups at time  $t$ , placed as the  $t$ th row of the  $T \times n$  matrix of residuals,  $\mathbf{E}$ . The rank of this matrix cannot be larger than  $T$ . Note what happens if  $n > T$ . In this case, the  $n \times n$  matrix has rank  $T$ , which is less than  $n$ , so it must be singular, and the FGLS estimator cannot be computed. Consider Example 10.1. We aggregated the 48 states into  $n = 9$  regions. It would not be possible to fit a full model for the  $n = 48$  states with only  $T = 17$  observations. This result is a deficiency of the data set, not the model. The population matrix,  $\Sigma$  is positive definite. But, if there are not enough observations, then the data set is too short to obtain a positive definite estimate of the matrix.

### **Example 10.1 A Regional Production Model for Public Capital**

Munnell (1990) proposed a model of productivity of public capital at the state level. The central equation of the analysis that we will extend here is a Cobb-Douglas production function,

$$\ln gsp_{it} = \alpha_0 + \beta_{1i} \ln pc_{it} + \beta_{2i} \ln hwy_{it} + \beta_{3i} \ln water_{it} \\ + \beta_{4i} \ln util_{it} + \beta_{5i} \ln emp_{it} + \beta_{6i} unemp_{it} + \varepsilon_{it},$$

where the variables in the model, measured for the lower 48 U.S. states and years 1970–1986, are

<i>gsp</i>	= gross state product,
<i>pc</i>	= public capital,
<i>hwy</i>	= highway capital,
<i>water</i>	= water utility capital,
<i>util</i>	= utility capital,
<i>p_cap</i>	= private capital,
<i>emp</i>	= employment (labor),
<i>unemp</i>	= unemployment rate.

In Example 9.9, we defined nine regions consisting of groups of the 48 states:

1. GF = Gulf = AL, FL, LA, MS,
2. MW = Midwest = IL, IN, KY, MI, MN, OH, WI,
3. MA = Mid Atlantic = DE, MD, NJ, NY, PA, VA,
4. MT = Mountain = CO, ID, MT, ND, SD, WY,
5. NE = New England = CT, ME, MA, NH, RI, VT,
6. SO = South = GA, NC, SC, TN, WV, R
7. SW = Southwest = AZ, NV, NM, TX, UT,
8. CN = Central = AK, IA, KS, MO, NE, OK,
9. WC = West Coast = CA, OR, WA.

For our application in this chapter, we will use the aggregated data to analyze a nine-region (equation) model. Data on output, the capital stocks, and employment are aggregated simply by summing the values for the individual states (before taking logarithms). The unemployment rate for each region,  $m$ , at time  $t$  is determined by a weighted average of the unemployment

CHAPTER 10 ♦ Systems of Equations **301**

rates for the states in the region, where the weights are

$$w_{it} = \text{emp}_{it} / \sum_j \text{emp}_{jt}.$$

Then, the unemployment rate for region  $m$  at time  $t$  is the following average of the unemployment rates of the states ( $j$ ) in region ( $m$ ) at time  $t$ :

$$\text{unemp}_{mt} = \sum_j w_{jt}(m) \text{unemp}_{jt}(m).$$

We initially estimated the nine equations of the regional productivity model separately by OLS. The OLS estimates are shown in Table 10.1. The correlation matrix for the OLS residuals is as follows:

	<b>GF</b>	<b>MW</b>	<b>MA</b>	<b>MT</b>	<b>NE</b>	<b>SO</b>	<b>SW</b>	<b>CN</b>	<b>WC</b>
<b>GF</b>	1.0000								
<b>MW</b>	0.1036	1.0000							
<b>MA</b>	0.3421	0.0634	1.0000						
<b>MT</b>	0.4243	0.6970	-0.0158	1.0000					
<b>NE</b>	-0.5127	-0.2896	0.1915	-0.5372	1.0000				
<b>SO</b>	0.5897	0.4893	0.2329	0.3434	-0.2411	1.0000			
<b>SW</b>	0.3115	0.1320	0.6514	0.1301	-0.3220	0.2594	1.0000		
<b>CN</b>	0.7958	0.3370	0.3904	0.4957	-0.02980	0.8050	0.3465	1.0000	
<b>WC</b>	0.2340	0.5654	0.2116	0.5736	-0.0576	0.2693	-0.0375	0.3818	1.0000

The values in  $\mathbf{R}$  are large enough to suggest that there is substantial correlation of the disturbances across regions.

Table 10.1 also presents the FGLS estimates of the parameters of the SUR model for regional output. These are computed in two steps, with the first-step OLS results producing the estimate  $\hat{\beta}_1$  for FGLS. (The pooled results that are also presented are discussed in Section 10.2.8.) The correlations listed earlier suggest that there is likely to be considerable benefit to using FGLS in terms of efficiency of the estimator. The individual equation OLS estimators are consistent, but they neglect the cross-equation correlation. The substantially lower estimated standard errors for the FGLS results with each equation appear to confirm that expectation.

We used (10-14) to construct test statistics for two hypotheses. We first tested the hypothesis of constant returns to scale through the system. Constant returns to scale would require that the coefficients on the inputs,  $\beta_2, \dots, \beta_6$  (our capital variables and the labor variable) sum to 1.0. The  $9 \times 9(7)$  matrix,  $\mathbf{R}$ , for (10-14) would have rows equal to

$$\begin{aligned}\mathbf{R}_1 &= (0, 1, 1, 1, 1, 1, 0) \quad \mathbf{0}' \\ \mathbf{R}_2 &= (\mathbf{0}', 0, 1, 1, 1, 1, 1, 0) \quad \mathbf{0}' \quad \mathbf{0}' \quad \mathbf{0}' \quad \mathbf{0}' \quad \mathbf{0}' \quad \mathbf{0}' \quad \mathbf{0}',\end{aligned}$$

and so on. In (10-14), we would have  $\mathbf{q}' = (1, 1, 1, 1, 1, 1, 1, 1, 1)$ . This hypothesis imposes nine restrictions. The computed chi-squared is 102.305. The critical value is 16.919, so this hypothesis is rejected as well. The discrepancy vector for these results is

$$(\mathbf{R}\beta - \mathbf{q})' = (-0.64674, -0.12883, 0.96435, 0.03930, 0.06710, 1.79472, 2.30283, \dots, 2.907, 1.10534).$$

The distance is quite large for some regions, so the hypothesis of constant returns to scale (to the extent it is meaningful at this level of aggregation) does appear to be inconsistent with the data (results).

The “pooling” restriction for the multivariate regression (same variables—not necessarily the same data, as in our example) is formulated as

$$H_0: \beta_1 = \beta_2 = \dots = \beta_M,$$

$$H_1: \text{Not } H_0.$$

**TABLE 10.1** Estimated SUR Model for Regional Output. (standard errors in parentheses)

<i>Region</i>	<i>Estimator</i>	$\alpha$	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$	$\beta_6$	$\sigma_m$	$R^2$
	<b>OLS</b>	12.1458 (3.3154)	-0.007117 (0.01114)	-2.1352 (0.8677)	0.1161 (0.06278)	1.4247 (0.5944)	0.7851 (0.1493)	-0.00742 (0.00316)	0.01075	0.9971
<b>GF</b>	<b>FGLS</b>	10.4792 (1.5912)	-0.003160 (0.005391)	-1.5448 (0.3888)	0.1139 (0.03651)	0.8987 (0.2516)	0.8886 (0.07715)	-0.005299 (0.00182)	0.008745	0.9967
	<b>OLS</b>	3.0282 (1.7834)	0.1635 (0.1660)	-0.07471 (0.2205)	-0.1689 (0.09896)	0.6372 (0.2078)	0.3622 (0.1650)	-0.01736 (0.004741)	0.009942	0.9984
<b>MW</b>	<b>FGLS</b>	4.1206 (1.0091)	0.06370 (0.08739)	-0.1275 (0.1284)	-0.1292 (0.06152)	0.5144 (0.1118)	0.5497 (0.08597)	-0.01545 (0.00252)	0.008608	0.9980
	<b>OLS</b>	-11.2110 (3.5867)	0.4120 (0.2281)	2.1355 (0.5571)	0.5122 (0.1192)	-0.4740 (0.2519)	-0.4620 (0.3529)	-0.03022 (0.00853)	0.01040	0.9950
<b>MA</b>	<b>FGLS</b>	-9.1438 (2.2025)	0.3511 (0.1077)	1.7972 (0.3410)	0.5168 (0.06405)	-0.3616 (0.1294)	-0.3391 (0.1997)	-0.02954 (0.00474)	0.008625	0.9946
	<b>OLS</b>	3.5902 (6.9490)	0.2948 (0.2054)	0.1740 (0.2082)	-0.2257 (0.3840)	-0.2144 (0.9712)	0.9166 (0.3772)	-0.008143 (0.00839)	0.01688	0.9940
<b>MT</b>	<b>FGLS</b>	2.8150 (3.4428)	0.1843 (0.09220)	0.1164 (0.1165)	-0.3811 (0.1774)	0.01648 (0.4654)	1.032 (0.1718)	-0.005507 (0.00422)	0.01321	0.9938
	<b>OLS</b>	6.3783 (2.3823)	-0.1526 (0.08403)	-0.1233 (0.2850)	0.3065 (0.08917)	-0.5326 (0.2375)	1.3437 (0.1876)	0.005098 (0.00517)	0.008601	0.9986
<b>NE</b>	<b>FGLS</b>	3.5331 (1.3388)	-0.1097 (0.04570)	0.1637 (0.1676)	0.2459 (0.04974)	-0.3155 (0.1194)	1.0828 (0.09248)	-0.00664 (0.00263)	0.007249	0.9983
	<b>OLS</b>	-13.7297 (18.0199)	-0.02040 (0.2856)	0.6621 (1.8111)	-0.9693 (0.2843)	-0.1074 (0.5634)	3.3803 (1.1643)	0.03378 (0.02150)	0.02241	0.9852
<b>SO</b>	<b>FGLS</b>	-13.1186 (7.6009)	0.1007 (0.1280)	0.9923 (0.7827)	-0.5851 (0.1373)	-0.3029 (0.2412)	2.5897 (0.4665)	0.02143 (0.00809)	0.01908	0.9817

**TABLE 10.1** (Continued)

<i>Region</i>	<i>Estimator</i>	$\alpha$	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$	$\beta_6$	$\sigma_m$	$R^2$	
	<b>OLS</b>	-22.8553 (4.8739)	-0.3776 (0.1673)	3.3478 (1.8584)	-0.2637 (0.4317)	-1.7783 (1.1757)	2.6732 (1.0325)	0.02592 (0.01727)	0.01293	0.9864	
<b>SW</b>	<b>FGLS</b>	-19.9917 (2.8649)	-0.3386 (0.08943)	3.2821 (0.8894)	-0.1105 (0.1993)	-1.7812 (0.5609)	2.2510 (0.4802)	0.01846 (0.00793)	0.01055	0.9846	
	<b>OLS</b>	3.4425 (1.2571)	0.05040 (0.2662)	-0.5938 (0.3219)	0.06351 (0.3333)	-0.01294 (0.3787)	1.5731 (0.4125)	0.006125 (0.00892)	0.01753	0.9937	
<b>CN</b>	<b>FGLS</b>	2.8172 (0.8434)	0.01412 (0.08833)	-0.5086 (0.1869)	-0.02685 (0.1405)	0.1165 (0.1774)	1.5339 (0.1762)	0.006499 (0.00421)	0.01416	0.9930	
	<b>OLS</b>	-9.1108 (3.9704)	0.2334 (0.2062)	1.6043 (0.7449)	0.7174 (0.1613)	-0.3563 (0.3153)	-0.2592 (0.3029)	-0.03416 (0.00629)	0.01085	0.9895	
<b>WC</b>	<b>FGLS</b>	-10.2989 (2.4189)	0.03734 (0.1107)	1.8176 (0.4503)	0.6572 (0.1011)	-0.4358 (0.1912)	0.02904 (0.1828)	-0.02867 (0.00373)	0.008837	0.9881	
	<b>OLS</b>	3.1567 (0.1377)	0.08692 (0.01058)	-0.02956 (0.03405)	0.4922 (0.04167)	0.06092 (0.03833)	0.3676 (0.04018)	-0.01746 (0.00304)	0.05558	0.9927	
	<b>Pooled</b>	<b>FGLS</b>	3.1089 (0.0208)	0.08076 (0.005148)	-0.01797 (0.006186)	0.3728 (0.01311)	0.1221 (0.00557)	0.4206 (0.01442)	-0.01506 (0.00101)	NA	0.9882 <sup>a</sup>
	<b>FGLS</b>	3.0977 (0.1233)	0.08646 (0.01144)	-0.02141 (0.02830)	0.03874 (0.03529)	0.1215 (0.02805)	0.4032 (0.03410)	-0.01529 (0.00256)	NA	0.9875 <sup>a</sup>	

<sup>a</sup>  $R^2$  for models fit by FGLS is computed using  $1 - 9/\text{tr}(\mathbf{S}^{-1}\mathbf{S}_{yy})$

### 304 PART II ♦ Generalized Regression Model and Equation Systems

For this hypothesis, the  $\mathbf{R}$  matrix is shown in (10-15). The test statistic is in (10-14). For our model with nine equations and seven parameters in each, the null hypothesis imposes  $8(7) = 56$  restrictions. The computed test statistic is 10,554.77, which is far larger than the critical value from the table, 74,468. So, the hypothesis of homogeneity is rejected.

As noted in Section 10.2.7, we do not have a standard test of the specification of the SUR model against the alternative hypothesis of uncorrelated disturbances for the general SUR model without an assumption of normality. The Breusch and Pagan (1980) Lagrange multiplier test based on the correlation matrix does have some intuitive appeal. We used (10-17) to compute the LM statistic for the nine-equation model reported in Table 10.1. For the correlation matrix shown earlier, the chi-squared statistic equals 102.305 with  $8(9)/2 = 36$  degrees of freedom. The critical value from the chi-squared table is 50.998, so the null hypothesis that the seemingly unrelated regressions are actually unrelated is rejected. We conclude that the disturbances in the regional model are not actually unrelated. The null hypothesis that  $\sigma_{ij} = 0$  for all  $i \neq j$  is rejected. To investigate a bit further, we repeated the test with the completely disaggregated (statewide) data. The corresponding chi-squared statistic is 8399.41 with  $48(47)/2 = 1,128$  degrees of freedom. The critical value is 1,207.25, so the null hypothesis is rejected at the state level as well.

### 10.3 SEEMINGLY UNRELATED GENERALIZED REGRESSION MODELS

In principle, the SUR model can accommodate heteroscedasticity as well as autocorrelation. Bartels and Fiebig (1992) suggested the generalized SUR model,  $\Omega = \mathbf{A}[\Sigma \otimes \mathbf{I}]\mathbf{A}'$  where  $\mathbf{A}$  is a block diagonal matrix. Ideally,  $\mathbf{A}$  is made a function of measured characteristics of the individual and a separate parameter vector,  $\theta$ , so that the model can be estimated in stages. In a first step, OLS residuals could be used to form a preliminary estimator of  $\theta$ , and then the data are transformed to homoscedasticity, leaving  $\Sigma$  and  $\beta$  to be estimated at subsequent steps using transformed data. One application along these lines is the random parameters model of Fiebig, Bartels, and Aigner (1991); (9-50) shows how the random parameters model induces heteroscedasticity. Another application is Mandy and Martins-Filho (1993), who specified  $\sigma_{ij}(t) = \mathbf{z}_{ij}(t)' \boldsymbol{\alpha}_{ij}$ . (The linear specification of a variance does present some problems, as a negative value is not precluded.) Kumbhakar and Heshmati (1996) proposed a cost and demand system that combined the translog model of Section 10.4.2 with the complete equation system in 10.4.1. In their application, only the cost equation was specified to include a heteroscedastic disturbance.

Autocorrelation in the disturbances of regression models usually arises as a particular feature of the time-series model. It is among the properties of the time series. (We will explore this aspect of the model specification in detail in Chapter 20.) In the multiple equation models examined in this chapter, the time-series properties of the data are usually not the main focus of the investigation. The main advantage of the SUR specification is its treatment of the correlation *across* observations at a particular point in time. Frequently, panel data specifications, such as those in examples 3 and 4 in the chapter introduction, can also be analyzed in the framework of the SUR model of this chapter. In these cases, there may be persistent effects in the disturbances, but here, again, those effects are often viewed as a consequence of the presence of latent, time invariant heterogeneity. Nonetheless, because the multiple equations models examined in this chapter often do involve moderately long time series, it is appropriate to deal at least somewhat more formally with autocorrelation. Opinions differ on the appropriateness

CHAPTER 10 ♦ Systems of Equations **305**

of “corrections” for autocorrelation. At one extreme is Mizon (1995) who argues forcefully that autocorrelation arises as a consequence of a remediable failure to include dynamic effects in the model. However, in a system of equations, the analysis that leads to this conclusion is going to be far more complex than in a single equation model.<sup>16</sup> Suffice to say, the issue remains to be settled conclusively.

## 10.4 NONLINEAR SYSTEMS OF EQUATIONS

We now consider estimation of nonlinear systems of equations. The underlying theory is essentially the same as that for linear systems. As such, most of the following will describe practical aspects of estimation. Consider estimation of the parameters of the equation system

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{h}_1(\boldsymbol{\beta}, \mathbf{X}) + \boldsymbol{\varepsilon}_1, \\ \mathbf{y}_2 &= \mathbf{h}_2(\boldsymbol{\beta}, \mathbf{X}) + \boldsymbol{\varepsilon}_2, \\ &\vdots \\ \mathbf{y}_M &= \mathbf{h}_M(\boldsymbol{\beta}, \mathbf{X}) + \boldsymbol{\varepsilon}_M. \end{aligned} \tag{10-23}$$

[Note the analogy to (10-19).]

There are  $M$  equations in total, to be estimated with  $t = 1, \dots, T$  observations. There are  $K$  parameters in the model. No assumption is made that each equation has “its own” parameter vector; we simply use some of or all the  $K$  elements in  $\boldsymbol{\beta}$  in each equation. Likewise, there is a set of  $T$  observations on each of  $P$  independent variables  $\mathbf{x}_p$ ,  $p = 1, \dots, P$ , some of or all that appear in each equation. For convenience, the equations are written generically in terms of the full  $\boldsymbol{\beta}$  and  $\mathbf{X}$ . The disturbances are assumed to have zero means and contemporaneous covariance matrix  $\Sigma$ . We will leave the extension to autocorrelation for more advanced treatments.

In the multivariate regression model, if  $\Sigma$  is known, then the generalized least squares estimator of  $\boldsymbol{\beta}$  is the vector that minimizes the generalized sum of squares

$$\boldsymbol{\varepsilon}(\boldsymbol{\beta})' \boldsymbol{\Omega}^{-1} \boldsymbol{\varepsilon}(\boldsymbol{\beta}) = \sum_{i=1}^M \sum_{j=1}^M \sigma^{ij} [\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\beta}, \mathbf{X})]' [\mathbf{y}_j - \mathbf{h}_j(\boldsymbol{\beta}, \mathbf{X})], \tag{10-24}$$

where  $\boldsymbol{\varepsilon}(\boldsymbol{\beta})$  is an  $MT \times 1$  vector of disturbances obtained by stacking the equations,  $\boldsymbol{\Omega} = \Sigma \otimes \mathbf{I}$ , and  $\sigma^{ij}$  is the  $ij$ th element of  $\Sigma^{-1}$ . [See (10-7).] As we did in Section 7.2.3, define the pseudoregressors as the derivatives of the  $\mathbf{h}(\boldsymbol{\beta}, \mathbf{X})$  functions with respect to  $\boldsymbol{\beta}$ . That is, linearize each of the equations. Then the first-order condition for minimizing this sum of squares is

$$\frac{\partial \boldsymbol{\varepsilon}(\boldsymbol{\beta})' \boldsymbol{\Omega}^{-1} \boldsymbol{\varepsilon}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^M \sum_{j=1}^M \sigma^{ij} [2\mathbf{X}_i^0(\boldsymbol{\beta}) \boldsymbol{\varepsilon}_j(\boldsymbol{\beta})] = \mathbf{0}, \tag{10-25}$$

<sup>16</sup>Dynamic SUR models in the spirit of Mizon’s admonition were proposed by Anderson and Blundell (1982). A few recent applications are Kiviet, Phillips, and Schipp (1995) and DesChamps (1998). However, relatively little work has been done with dynamic SUR models. The VAR models in Section 21.6 are an important group of applications, but they come from a different analytical framework. Likewise, the panel data applications noted in the introduction and in Section 9.5 would fit into the modeling framework we are developing here. However, in these applications, the regressions are “actually” unrelated—the authors did not model the cross-unit correlation that is the central focus of this chapter. Related results may be found in Guilkey and Schmidt (1973), Guilekey (1974), Berndt and Savin (1977), Moschino and Moro (1994), McLaren (1996), and Holt (1998).

### 306 PART II ♦ Generalized Regression Model and Equation Systems

where  $\mathbf{X}_i^0(\boldsymbol{\beta})$  is the  $T \times K$  matrix of pseudoregressors from the linearization of the  $i$ th equation. (See Section 7.2.6.) If any of the parameters in  $\boldsymbol{\beta}$  do not appear in the  $i$ th equation, then the corresponding column of  $\mathbf{X}_i^0(\boldsymbol{\beta})$  will be a column of zeros.

This problem of estimation is doubly complex. In almost any circumstance, solution will require an iteration using one of the methods discussed in Appendix E. Second, of course, is that  $\Sigma$  is not known and must be estimated. Remember that efficient estimation in the multivariate regression model does not require an efficient estimator of  $\Sigma$ , only a consistent one. Therefore, one approach would be to estimate the parameters of each equation separately using nonlinear least squares. This method will be inefficient if any of the equations share parameters, since that information will be ignored. But at this step, consistency is the objective, not efficiency. The resulting residuals can then be used to compute

$$\mathbf{S} = \frac{1}{T} \mathbf{E}' \mathbf{E}. \quad (10-26)$$

The second step of FGLS is the solution of (10-25), which will require an iterative procedure once again and can be based on  $\mathbf{S}$  instead of  $\Sigma$ . With well-behaved pseudoregressors, this second-step estimator is fully efficient. Once again, the same theory used for FGLS in the linear, single-equation case applies here.<sup>17</sup> Once the FGLS estimator is obtained, the appropriate asymptotic covariance matrix is estimated with

$$\text{Est. Asy. Var}[\hat{\boldsymbol{\beta}}] = \left[ \sum_{i=1}^M \sum_{j=1}^M s^{ij} \mathbf{X}_i^0(\boldsymbol{\beta})' \mathbf{X}_j^0(\boldsymbol{\beta}) \right]^{-1}. \quad (10-27)$$

There is a possible flaw in the strategy just outlined. It may not be possible to fit all the equations individually by nonlinear least squares. It is conceivable that identification of some of the parameters requires joint estimation of more than one equation. But as long as the full system identifies all parameters, there is a simple way out of this problem. Recall that all we need for our first step is a consistent set of estimators of the elements of  $\boldsymbol{\beta}$ . It is easy to show that the preceding defines a GMM estimator (see Chapter 13.) We can use this result to devise an alternative, simple strategy. The weighting of the sums of squares and cross products in (10-24) by  $\sigma^{ij}$  produces an efficient estimator of  $\boldsymbol{\beta}$ . Any other weighting based on some positive definite  $\mathbf{A}$  would produce consistent, although inefficient, estimates. At this step, though, efficiency is secondary, so the choice of  $\mathbf{A} = \mathbf{I}$  is a convenient candidate. Thus, for our first step, we can find  $\boldsymbol{\beta}$  to minimize

$$\boldsymbol{\varepsilon}(\boldsymbol{\beta})' \boldsymbol{\varepsilon}(\boldsymbol{\beta}) = \sum_{i=1}^M [\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\beta}, \mathbf{X})]' [\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\beta}, \mathbf{X})] = \sum_{i=1}^M \sum_{t=1}^T [y_{it} - h_i(\boldsymbol{\beta}, \mathbf{x}_{it})]^2.$$

(This estimator is just pooled nonlinear least squares, where the regression function varies across the sets of observations.) This step will produce the  $\hat{\boldsymbol{\beta}}$  we need to compute  $\mathbf{S}$ .

<sup>17</sup>Neither the nonlinearity nor the multiple equation aspect of this model brings any new statistical issues to the fore. By stacking the equations, we see that this model is simply a variant of the nonlinear regression model with the added complication of a nonscalar disturbance covariance matrix, which we analyzed in Chapter 8. The new complications are primarily practical.

## 10.5 SYSTEMS OF DEMAND EQUATIONS: SINGULAR SYSTEMS

Most of the recent applications of the multivariate regression model<sup>18</sup> have been in the context of **systems of demand equations**, either commodity demands or factor demands in studies of production.

### **Example 10.2 Stone's Expenditure System**

Stone's expenditure system<sup>19</sup> based on a set of logarithmic commodity demand equations, income  $Y$ , and commodity prices  $p_i$  is

$$\log q_i = \alpha_i + \eta_i \log \left( \frac{Y}{P} \right) + \sum_{j=1}^M \eta_{ij}^* \log \left( \frac{p_j}{P} \right),$$

where  $P$  is a generalized (share-weighted) price index,  $\eta_i$  is an income elasticity, and  $\eta_{ij}^*$  is a compensated price elasticity. We can interpret this system as the demand equation in real expenditure and real prices. The resulting set of equations constitutes an econometric model in the form of a set of seemingly unrelated regressions. In estimation, we must account for a number of restrictions including homogeneity of degree one in income,  $\sum_i S_i \eta_i = 1$ , and symmetry of the matrix of compensated price elasticities,  $\eta_{ij}^* = \eta_{ji}^*$ , where  $S_i$  is the budget share for good  $i$ .

Other examples include the system of factor demands and factor cost shares from production, which we shall consider again later. In principle, each is merely a particular application of the model of the Section 10.2. But some special problems arise in these settings. First, the parameters of the systems are generally constrained across equations. That is, the unconstrained model is inconsistent with the underlying theory.<sup>20</sup> The numerous constraints in the system of demand equations presented earlier give an example. A second intrinsic feature of many of these models is that the disturbance covariance matrix  $\Sigma$  is singular.<sup>21</sup>

### 10.5.1 COBB-DOUGLAS COST FUNCTION

Consider a **Cobb-Douglas** production function,

$$Q = \alpha_0 \prod_{i=1}^M x_i^{\alpha_i}.$$

<sup>18</sup>Note the distinction between the *multivariate* or multiple-equation model discussed here and the *multiple* regression model.

<sup>19</sup>A very readable survey of the estimation of systems of commodity demands is Deaton and Muellbauer (1980). The example discussed here is taken from their Chapter 3 and the references to Stone's (1954a,b) work cited therein. Deaton (1986) is another useful survey. A counterpart for production function modeling is Chambers (1988). Other developments in the specification of systems of demand equations include Chavez and Segerson (1987), Brown and Walker (1995), and Fry, Fry, and McLaren (1996).

<sup>20</sup>This inconsistency does not imply that the theoretical restrictions are not testable or that the unrestricted model cannot be estimated. Sometimes, the meaning of the model is ambiguous without the restrictions, however. Statistically rejecting the restrictions implied by the theory, which were used to derive the econometric model in the first place, can put us in a rather uncomfortable position. For example, in a study of utility functions, Christensen, Jorgenson, and Lau (1975), after rejecting the cross-equation symmetry of a set of commodity demands, stated, "With this conclusion we can terminate the test sequence, since these results invalidate the theory of demand" (p. 380). See Silver and Ali (1989) for discussion of testing symmetry restrictions. The theory and the model may also conflict in other ways. For example, Stone's loglinear expenditure system in Example 10.7 does not conform to any theoretically valid utility function. See Goldberger (1987).

<sup>21</sup>Denton (1978) examines several of these cases.

### 308 PART II ♦ Generalized Regression Model and Equation Systems

Profit maximization with an exogenously determined output price calls for the firm to maximize output for a given cost level  $C$  (or minimize costs for a given output  $Q$ ). The Lagrangean for the maximization problem is

$$\Lambda = \alpha_0 \prod_{i=1}^M x_i^{\alpha_i} + \lambda(C - \mathbf{p}'\mathbf{x}),$$

where  $\mathbf{p}$  is the vector of  $M$  factor prices. The necessary conditions for maximizing this function are

$$\frac{\partial \Lambda}{\partial x_i} = \frac{\alpha_i Q}{x_i} - \lambda p_i = 0 \quad \text{and} \quad \frac{\partial \Lambda}{\partial \lambda} = C - \mathbf{p}'\mathbf{x} = 0.$$

The joint solution provides  $x_i(Q, \mathbf{p})$  and  $\lambda(Q, \mathbf{p})$ . The total cost of production is

$$\sum_{i=1}^M p_i x_i = \sum_{i=1}^M \frac{\alpha_i Q}{\lambda}.$$

The cost share allocated to the  $i$ th factor is

$$\frac{p_i x_i}{\sum_{i=1}^M p_i x_i} = \frac{\alpha_i}{\sum_{i=1}^M \alpha_i} = \beta_i. \quad (10-28)$$

The full model is<sup>22</sup>

$$\begin{aligned} \ln C &= \beta_0 + \beta_q \ln Q + \sum_{i=1}^M \beta_i \ln p_i + \varepsilon_c, \\ s_i &= \beta_i + \varepsilon_i, \quad i = 1, \dots, M. \end{aligned} \quad (10-29)$$

By construction,  $\sum_{i=1}^M \beta_i = 1$  and  $\sum_{i=1}^M s_i = 1$ . (This is the cost function analysis begun in Example 6.6. We will return to that application below.) The cost shares will also sum identically to one in the data. It therefore follows that  $\sum_{i=1}^M \varepsilon_i = 0$  at every data point, so the system is singular. For the moment, ignore the cost function. Let the  $M \times 1$  disturbance vector from the shares be  $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_M]'$ . Because  $\boldsymbol{\varepsilon}'\mathbf{i} = 0$ , where  $\mathbf{i}$  is a column of 1s, it follows that  $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{i}] = \Sigma\mathbf{i} = \mathbf{0}$ , which implies that  $\Sigma$  is singular. Therefore, the methods of the previous sections cannot be used here. (You should verify that the sample covariance matrix of the OLS residuals will also be singular.)

The solution to the singularity problem appears to be to drop one of the equations, estimate the remainder, and solve for the last parameter from the other  $M - 1$ . The constraint  $\sum_{i=1}^M \beta_i = 1$  states that the cost function must be homogeneous of degree one in the prices, a theoretical necessity. If we impose the constraint

$$\beta_M = 1 - \beta_1 - \beta_2 - \cdots - \beta_{M-1}, \quad (10-30)$$

then the system is reduced to a nonsingular one:

$$\begin{aligned} \ln \left( \frac{C}{p_M} \right) &= \beta_0 + \beta_q \ln Q + \sum_{i=1}^{M-1} \beta_i \ln \left( \frac{p_i}{p_M} \right) + \varepsilon_c, \\ s_i &= \beta_i + \varepsilon_i, \quad i = 1, \dots, M-1. \end{aligned}$$

<sup>22</sup>We leave as an exercise the derivation of  $\beta_0$ , which is a mixture of all the parameters, and  $\beta_q$ , which equals  $1/\sum_m \alpha_m$ .

**TABLE 10.2** Regression Estimates (standard errors in parentheses)

	<i>Ordinary Least Squares</i>				<i>Multivariate Regression</i>			
$\beta_0$	-4.686	(0.885)	-3.764	(0.702)	-7.069	(0.107)	-5.707	(0.165)
$\beta_q$	0.721	(0.0174)	0.153	(0.0618)	0.766	(0.0154)	0.238	(0.0587)
$\beta_{qq}$	—		0.0505	(0.00536)	—		0.0451	(0.00508)
$\beta_k$	-0.00847	(0.191)	0.0739	(0.150)	0.424	(0.00946)	0.424	(0.00944)
$\beta_l$	0.594	(0.205)	0.481	(0.161)	0.106	(0.00386)	0.106	(0.00382)
$\beta_f$	0.414	(0.0989)	0.445	(0.0777)	0.470	(0.0101)	0.470	(0.0100)
$R^2$	0.9316		0.9581		—		—	
	—		—					

This system provides estimates of  $\beta_0$ ,  $\beta_q$ , and  $\beta_1, \dots, \beta_{M-1}$ . The last parameter is estimated using (10-30). It is immaterial which factor is chosen as the numeraire. Both FGLS and **maximum likelihood**, which can be obtained by iterating FGLS or by direct maximum likelihood estimation, are **invariant** to which factor is chosen as the numeraire.<sup>23</sup>

Nerlove's (1963) study of the electric power industry that we examined in Example 6.6 provides an application of the Cobb–Douglas cost function model. His ordinary least squares estimates of the parameters were listed in Example 6.6. Among the results are (unfortunately) a negative capital coefficient in three of the six regressions. Nerlove also found that the simple Cobb–Douglas model did not adequately account for the relationship between output and average cost. Christensen and Greene (1976) further analyzed the Nerlove data and augmented the data set with cost share data to estimate the complete **demand system**. Appendix Table F6.2 lists Nerlove's 145 observations with Christensen and Greene's cost share data. Cost is the total cost of generation in millions of dollars, output is in millions of kilowatt-hours, the capital price is an index of construction costs, the wage rate is in dollars per hour for production and maintenance, the fuel price is an index of the cost per Btu of fuel purchased by the firms, and the data reflect the 1955 costs of production. The regression estimates are given in Table 10.2.

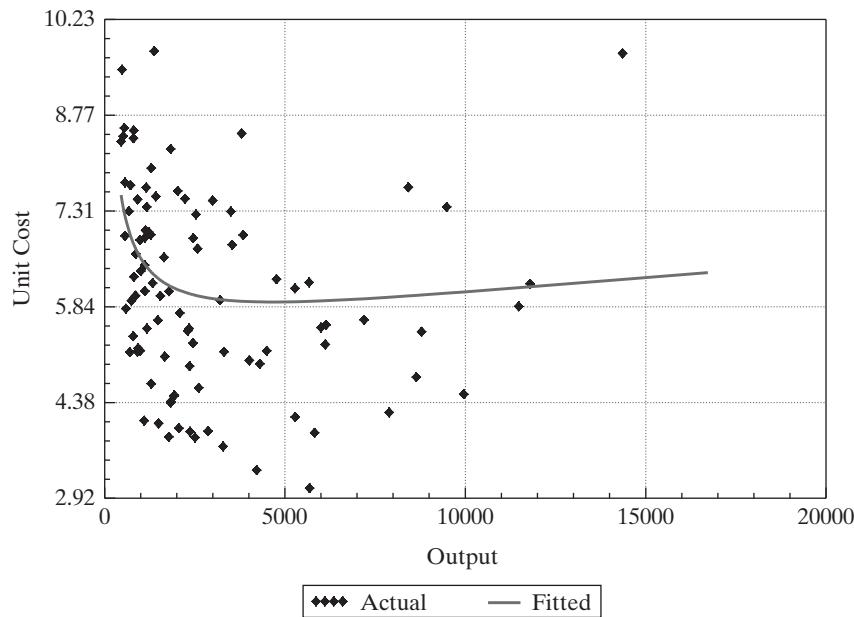
Least squares estimates of the Cobb–Douglas cost function are given in the first column.<sup>24</sup> The coefficient on capital is negative. Because  $\beta_i = \beta_q \partial \ln Q / \partial \ln x_i$ —that is, a positive multiple of the output elasticity of the  $i$ th factor—this finding is troubling. The third column presents the constrained FGLS estimates. To obtain the constrained estimator, we set up the model in the form of the pooled SUR estimator in (10-19);

$$\mathbf{y} = \begin{bmatrix} \ln(\mathbf{C}/\mathbf{P}_f) \\ \mathbf{s}_k \\ \mathbf{s}_l \end{bmatrix} = \begin{bmatrix} \mathbf{i} & \ln \mathbf{Q} & \ln(\mathbf{P}_k/\mathbf{P}_f) & \ln(\mathbf{P}_l/\mathbf{P}_f) \\ \mathbf{0} & \mathbf{0} & \mathbf{i} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{i} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_q \\ \beta_k \\ \beta_l \end{pmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_c \\ \boldsymbol{\varepsilon}_k \\ \boldsymbol{\varepsilon}_l \end{bmatrix}$$

[There are  $3(145) = 435$  observations in the data matrices.] The estimator is then FGLS as shown in (10-21). An additional column is added for the log quadratic model. Two

<sup>23</sup>The invariance result is proved in Barten (1969). Some additional results on the method are given by Revankar (1976), Deaton (1986), Powell (1969), and McGuire et al. (1968).

<sup>24</sup>Results based on Nerlove's full data set are given in Example 6.6.

**310 PART II ♦ Generalized Regression Model and Equation Systems**

**FIGURE 10.1** Predicted and Actual Average Costs.

things to note are the dramatically smaller standard errors and the now positive (and reasonable) estimate of the capital coefficient. The estimates of economies of scale in the basic Cobb–Douglas model are  $1/\beta_q = 1.39$  (column 1) and 1.31 (column 3), which suggest some increasing returns to scale. Nerlove, however, had found evidence that at extremely large firm sizes, economies of scale diminished and eventually disappeared. To account for this (essentially a classical U-shaped average cost curve), he appended a quadratic term in log output in the cost function. The single equation and multivariate regression estimates are given in the second and fourth sets of results.

The quadratic output term gives the cost function the expected U-shape. We can determine the point where average cost reaches its minimum by equating  $\partial \ln C / \partial \ln Q$  to 1. This is  $Q^* = \exp[(1 - \beta_q)/(2\beta_{qq})]$ . For the multivariate regression, this value is  $Q^* = 4665$ . About 85 percent of the firms in the sample had output less than this, so by these estimates, most firms in the sample had not yet exhausted the available economies of scale. Figure 10.1 shows predicted and actual average costs for the sample. (To obtain a reasonable scale, the smallest one third of the firms are omitted from the figure.) Predicted average costs are computed at the sample averages of the input prices. The figure does reveal that that beyond a quite small scale, the economies of scale, while perhaps statistically significant, are economically quite small.

#### **10.5.2 FLEXIBLE FUNCTIONAL FORMS: THE TRANSLOG COST FUNCTION**

The literatures on production and cost and on utility and demand have evolved in several directions. In the area of models of producer behavior, the classic paper by Arrow et al. (1961) called into question the inherent restriction of the Cobb–Douglas model that

CHAPTER 10 ♦ Systems of Equations **311**

all elasticities of factor substitution are equal to 1. Researchers have since developed numerous **flexible functions** that allow substitution to be unrestricted (i.e., not even constant).<sup>25</sup> Similar strands of literature have appeared in the analysis of commodity demands.<sup>26</sup> In this section, we examine in detail a model of production.

Suppose that production is characterized by a production function,  $Q = f(\mathbf{x})$ . The solution to the problem of minimizing the cost of producing a specified output rate given a set of factor prices produces the cost-minimizing set of factor demands  $x_i = x_i(Q, \mathbf{p})$ . The total cost of production is given by the cost function,

$$C = \sum_{i=1}^M p_i x_i(Q, \mathbf{p}) = C(Q, \mathbf{p}). \quad (10-31)$$

If there are **constant returns to scale**, then it can be shown that  $C = Qc(\mathbf{p})$  or

$$C/Q = c(\mathbf{p}),$$

where  $c(\mathbf{p})$  is the unit or average cost function.<sup>27</sup> The cost-minimizing factor demands are obtained by applying **Shephard's (1970) lemma**, which states that if  $C(Q, \mathbf{p})$  gives the minimum total cost of production, then the cost-minimizing set of factor demands is given by

$$x_i^* = \frac{\partial C(Q, \mathbf{p})}{\partial p_i} = \frac{Q \partial c(\mathbf{p})}{\partial p_i}. \quad (10-32)$$

Alternatively, by differentiating logarithmically, we obtain the cost-minimizing factor cost shares:

$$s_i = \frac{\partial \ln C(Q, \mathbf{p})}{\partial \ln p_i} = \frac{p_i x_i}{C}. \quad (10-33)$$

With constant returns to scale,  $\ln C(Q, \mathbf{p}) = \ln Q + \ln c(\mathbf{p})$ , so

$$s_i = \frac{\partial \ln c(\mathbf{p})}{\partial \ln p_i}. \quad (10-34)$$

In many empirical studies, the objects of estimation are the elasticities of factor substitution and the own price elasticities of demand, which are given by

$$\theta_{ij} = \frac{c(\partial^2 c / \partial p_i \partial p_j)}{(\partial c / \partial p_i)(\partial c / \partial p_j)}$$

and

$$\eta_{ii} = s_i \theta_{ii}.$$

<sup>25</sup>See, in particular, Berndt and Christensen (1973). Two useful surveys of the topic are Jorgenson (1983) and Diewert (1974).

<sup>26</sup>See, for example, Christensen, Jorgenson, and Lau (1975) and two surveys, Deaton and Muellbauer (1980) and Deaton (1983). Berndt (1990) contains many useful results.

<sup>27</sup>The Cobb–Douglas function of the previous section gives an illustration. The restriction of constant returns to scale is  $\beta_q = 1$ , which is equivalent to  $C = Qc(\mathbf{p})$ . Nerlove's more general version of the cost function allows nonconstant returns to scale. See Christensen and Greene (1976) and Diewert (1974) for some of the formalities of the cost function and its relationship to the structure of production.

### 312 PART II ♦ Generalized Regression Model and Equation Systems

By suitably parameterizing the cost function (10-31) and the cost shares (10-34), we obtain an  $M$  or  $M + 1$  equation econometric model that can be used to estimate these quantities.<sup>28</sup>

The transcendental logarithmic, or **translog function** is the most frequently used flexible function in empirical work.<sup>29</sup> By expanding  $\ln c(\mathbf{p})$  in a second-order **Taylor series** about the point  $\ln \mathbf{p} = \mathbf{0}$ , we obtain

$$\ln c \approx \beta_0 + \sum_{i=1}^M \left( \frac{\partial \ln c}{\partial \ln p_i} \right) \log p_i + \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^M \left( \frac{\partial^2 \ln c}{\partial \ln p_i \partial \ln p_j} \right) \ln p_i \ln p_j, \quad (10-35)$$

where all derivatives are evaluated at the expansion point. If we treat these derivatives as the coefficients, then the cost function becomes

$$\begin{aligned} \ln c = & \beta_0 + \beta_1 \ln p_1 + \cdots + \beta_M \ln p_M + \delta_{11} \left( \frac{1}{2} \ln^2 p_1 \right) + \delta_{12} \ln p_1 \ln p_2 \\ & + \delta_{22} \left( \frac{1}{2} \ln^2 p_2 \right) + \cdots + \delta_{MM} \left( \frac{1}{2} \ln^2 p_M \right). \end{aligned} \quad (10-36)$$

This is the translog cost function. If  $\delta_{ij}$  equals zero, then it reduces to the Cobb–Douglas function we looked at earlier. The cost shares are given by

$$\begin{aligned} s_1 &= \frac{\partial \ln c}{\partial \ln p_1} = \beta_1 + \delta_{11} \ln p_1 + \delta_{12} \ln p_2 + \cdots + \delta_{1M} \ln p_M, \\ s_2 &= \frac{\partial \ln c}{\partial \ln p_2} = \beta_2 + \delta_{21} \ln p_1 + \delta_{22} \ln p_2 + \cdots + \delta_{2M} \ln p_M, \\ &\vdots \\ s_M &= \frac{\partial \ln c}{\partial \ln p_M} = \beta_M + \delta_{M1} \ln p_1 + \delta_{M2} \ln p_2 + \cdots + \delta_{MM} \ln p_M. \end{aligned} \quad (10-37)$$

The cost shares must sum to 1, which requires,

$$\begin{aligned} \beta_1 + \beta_2 + \cdots + \beta_M &= 1, \\ \sum_{i=1}^M \delta_{ij} &= 0 \quad (\text{column sums equal zero}), \\ \sum_{j=1}^M \delta_{ij} &= 0 \quad (\text{row sums equal zero}). \end{aligned} \quad (10-38)$$

We will also impose the (theoretical) symmetry restriction,  $\delta_{ij} = \delta_{ji}$ .

The system of **share equations** provides a seemingly unrelated regressions model that can be used to estimate the parameters of the model.<sup>30</sup> To make the model

<sup>28</sup>The cost function is only one of several approaches to this study. See Jorgenson (1983) for a discussion.

<sup>29</sup>See Example 2.4. The function was developed by Kmenta (1967) as a means of approximating the CES production function and was introduced formally in a series of papers by Berndt, Christensen, Jorgenson, and Lau, including Berndt and Christensen (1973) and Christensen et al. (1975). The literature has produced something of a competition in the development of exotic functional forms. The translog function has remained the most popular, however, and by one account, Guilkey, Lovell, and Sickles (1983) is the most reliable of several available alternatives. See also Example 5.4.

<sup>30</sup>The cost function may be included, if desired, which will provide an estimate of  $\beta_0$  but is otherwise inessential. Absent the assumption of constant returns to scale, however, the cost function will contain parameters of interest that do not appear in the share equations. As such, one would want to include it in the model. See Christensen and Greene (1976) for an application.

CHAPTER 10 ♦ Systems of Equations **313****TABLE 10.3** Parameter Estimates (standard errors in parentheses)

$\beta_K$	0.05682	(0.00131)	$\delta_{KM}$	-0.02169*	(0.00963)
$\beta_L$	0.25355	(0.001987)	$\delta_{LL}$	0.07488	(0.00639)
$\beta_E$	0.04383	(0.00105)	$\delta_{LE}$	-0.00321	(0.00275)
$\beta_M$	0.64580*	(0.00299)	$\delta_{LM}$	-0.07169*	(0.00941)
$\delta_{KK}$	0.02987	(0.00575)	$\delta_{EE}$	0.02938	(0.00741)
$\delta_{KL}$	0.0000221	(0.00367)	$\delta_{EM}$	-0.01797*	(0.01075)
$\delta_{KE}$	-0.00820	(0.00406)	$\delta_{MM}$	0.11134*	(0.02239)

\*Estimated indirectly using (10-38).

operational, we must impose the restrictions in (10-38) and solve the problem of **singularity of the disturbance covariance matrix** of the share equations. The first is accomplished by dividing the first  $M - 1$  prices by the  $M$ th, thus eliminating the last term in each row and column of the parameter matrix. As in the Cobb–Douglas model, we obtain a nonsingular system by dropping the  $M$ th share equation. We compute maximum likelihood estimates of the parameters to ensure **invariance** with respect to the choice of which share equation we drop. For the translog cost function, the elasticities of substitution are particularly simple to compute once the parameters have been estimated:

$$\theta_{ij} = \frac{\delta_{ij} + s_i s_j}{s_i s_j}, \quad \theta_{ii} = \frac{\delta_{ii} + s_i(s_i - 1)}{s_i^2}. \quad (10-39)$$

These elasticities will differ at every data point. It is common to compute them at some central point such as the means of the data.<sup>31</sup>

### Example 10.3 A Cost Function for U.S. Manufacturing

A number of recent studies using the translog methodology have used a four-factor model, with capital  $K$ , labor  $L$ , energy  $E$ , and materials  $M$ , the factors of production. Among the first studies to employ this methodology was Berndt and Wood's (1975) estimation of a translog cost function for the U.S. manufacturing sector. The three factor shares used to estimate the model are

$$\begin{aligned} s_K &= \beta_K + \delta_{KK} \ln \left( \frac{p_K}{p_M} \right) + \delta_{KL} \ln \left( \frac{p_L}{p_M} \right) + \delta_{KE} \ln \left( \frac{p_E}{p_M} \right), \\ s_L &= \beta_L + \delta_{KL} \ln \left( \frac{p_K}{p_M} \right) + \delta_{LL} \ln \left( \frac{p_L}{p_M} \right) + \delta_{LE} \ln \left( \frac{p_E}{p_M} \right), \\ s_E &= \beta_E + \delta_{KE} \ln \left( \frac{p_K}{p_M} \right) + \delta_{LE} \ln \left( \frac{p_L}{p_M} \right) + \delta_{EE} \ln \left( \frac{p_E}{p_M} \right). \end{aligned}$$

Berndt and Wood's data are reproduced in Appendix Table F10.2. Constrained FGLS estimates of the parameters presented in Table 10.3 were obtained by constructing the "pooled

<sup>31</sup>They will also be highly nonlinear functions of the parameters and the data. A method of computing asymptotic standard errors for the estimated elasticities is presented in Anderson and Thursby (1986). Krinsky and Robb (1986, 1990) (see Section 15.3) proposed their method as an alternative approach to this computation.

**314 PART II ♦ Generalized Regression Model and Equation Systems**
**TABLE 10.4** Estimated Elasticities

	<i>Capital</i>	<i>Labor</i>	<i>Energy</i>	<i>Materials</i>
<i>Cost Shares for 1959</i>				
Fitted shares	0.05646	0.27454	0.04424	0.62476
Actual shares	0.06185	0.27303	0.04563	0.61948
<i>Implied Elasticities of Substitution, 1959</i>				
Capital	-7.34124			
Labor	1.0014	-1.64902		
Energy	-2.34994	0.73556	-6.34994	
Materials	0.34994	0.58205	0.58205	-0.19702
	0.38512		0.34994	-0.31536
<i>Implied Own Price Elasticities</i>				
	-0.41448	-0.45274	-0.29161	-0.19702

regression" in (10-19) with data matrices

$$\mathbf{y} = \begin{bmatrix} \mathbf{s}_K \\ \mathbf{s}_L \\ \mathbf{s}_E \end{bmatrix}, \quad (10-40)$$

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & \ln P_K/P_M & \ln P_L/P_M & \ln P_E/P_M & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \ln P_K/P_M & 0 & \ln P_L/P_M & \ln P_E/P_M & 0 \\ 0 & 0 & 1 & 0 & 0 & \ln P_K/P_M & 0 & \ln P_L/P_M & \ln P_E/P_M \end{bmatrix},$$

$$\boldsymbol{\beta}' = (\beta_K, \beta_L, \beta_E, \delta_{KK}, \delta_{KL}, \delta_{KE}, \delta_{LL}, \delta_{LE}, \delta_{EE}).$$

Estimates are then obtained using the two-step procedure in (10-7) and (10-9).<sup>32</sup> The full set of estimates are given in Table 10.4. The parameters not estimated directly in (10-36) are computed using (10-38).

The implied estimates of the elasticities of substitution and demand for 1959 (the central year in the data) are derived in Table 10.4 using the fitted cost shares and the estimated parameters in (10-39). The departure from the Cobb-Douglas model with unit elasticities is substantial. For example, the results suggest almost no substitutability between energy and labor and some complementarity between capital and energy.<sup>33</sup>

## 10.6 SIMULTANEOUS EQUATIONS MODELS

There is a qualitative difference between the market equilibrium model suggested in the chapter Introduction,

$$Q_{Demand} = \alpha_1 + \alpha_2 Price + \alpha_3 Income + \mathbf{d}'\boldsymbol{\alpha} + \varepsilon_{Demand},$$

$$Q_{Supply} = \beta_1 + \beta_2 Price + \mathbf{s}'\boldsymbol{\beta} + \varepsilon_{Supply},$$

$$Q_{Equilibrium} = Q_{Demand} = Q_{Supply},$$

<sup>32</sup>These estimates do not match those reported by Berndt and Wood. They used an iterative estimator whereas ours is two step GLS. To purge their data of possible correlation with the disturbances, they first regressed the prices on 10 exogenous macroeconomic variables, such as U.S. population, government purchases of labor services, real exports of durable goods and U.S. tangible capital stock, and then based their analysis on the fitted values. The estimates given here are, in general quite close to those given by Berndt and Wood. For example, their estimates of the first five parameters are 0.0564, 0.2539, 0.0442, 0.6455, and 0.0254.

<sup>33</sup>Berndt and Wood's estimate of  $\theta_{EL}$  for 1959 is 0.64.

## CHAPTER 10 ♦ Systems of Equations 315

and the other examples considered thus far. The seemingly unrelated regression model,

$$y_{im} = \mathbf{x}_{im}' \boldsymbol{\beta}_m + \varepsilon_{im},$$

derives from a set of regression equations that are connected through the disturbances. The regressors,  $\mathbf{x}_{im}$  are exogenous and vary autonomously for reasons that are not explained within the model. Thus, the coefficients are directly interpretable as partial effects and can be estimated by least squares or other methods that are based on the conditional mean functions,  $E[y_{im}|\mathbf{x}_{im}] = \mathbf{x}_{im}'\boldsymbol{\beta}$ . In a model such as the preceding equilibrium model, the relationships are explicit and neither of the two market equations is a regression model. As a consequence, the partial equilibrium experiment of changing the price and inducing a change in the equilibrium quantity so as to elicit an estimate of the price elasticity of demand,  $\alpha_2$  (or supply elasticity,  $\beta_2$ ) makes no sense. The model is of the joint determination of quantity and price. Price changes when the market equilibrium changes, but that is induced by changes in other factors, such as changes in incomes or other variables that affect the supply function. (See Figure 8.1 for a graphical treatment.)

As we saw in Example 8.4, least squares regression of observed equilibrium quantities on price and the other factors will compute an ambiguous mixture of the supply and demand functions. The result follows from the endogeneity of Price in either equation. “Simultaneous equations models” arise in settings such as this one, in which the set of equations are interdependent by design. Simultaneous equations models will fit in the framework developed in Chapter 8, where we considered equations in which some of the right-hand-side variables are endogenous—that is, correlated with the disturbances. The substantive difference at this point is the source of the endogeneity. In our treatments in Chapter 8, endogeneity arose, for example, in the models of omitted variables, measurement error, or endogenous treatment effects, essentially as an unintended deviation from the assumptions of the linear regression model. In the simultaneous equations framework, endogeneity is a fundamental part of the specification. This section will consider the issues of specification and estimation in systems of simultaneous equations. We begin in Section 10.6.1 with a development of a general framework for the analysis and a statement of some fundamental issues. Section 10.6.2 presents the simultaneous equations model as an extension of the seemingly unrelated regressions model in Section 10.2. The ultimate objective of the analysis will be to learn about the model coefficients. The issue of whether this is even possible is considered in Section 10.6.3, where we develop the issue of identification. Once the identification question is settled, methods of estimation and inference are presented in Section 10.6.4 and 10.6.5.

### 10.6.1 SYSTEMS OF EQUATIONS

Consider a simplified version of the preceding equilibrium model, above,

$$\text{demand equation: } q_{d,t} = \alpha_1 p_t + \alpha_2 x_t + \varepsilon_{d,t},$$

$$\text{supply equation: } q_{s,t} = \beta_1 p_t + \varepsilon_{s,t},$$

$$\text{equilibrium condition: } q_{d,t} = q_{s,t} = q_t.$$

These equations are **structural equations** in that they are derived from theory and each purports to describe a particular aspect of the economy.<sup>34</sup> Because the model is one

---

<sup>34</sup>The distinction between structural and nonstructural models is sometimes drawn on this basis. See, for example, Cooley and LeRoy (1985).

### 316 PART II ♦ Generalized Regression Model and Equation Systems

of the joint determination of price and quantity, they are labeled **jointly dependent** or **endogenous** variables. Income,  $x$ , is assumed to be determined outside of the model, which makes it **exogenous**. The disturbances are added to the usual textbook description to obtain an **econometric model**. All three equations are needed to determine the equilibrium price and quantity, so the system is **interdependent**. Finally, because an equilibrium solution for price and quantity in terms of income and the disturbances is, indeed, implied (unless  $\alpha_1$  equals  $\beta_1$ ), the system is said to be a **complete system of equations**. *The completeness of the system requires that the number of equations equal the number of endogenous variables.* As a general rule, it is not possible to estimate all the parameters of incomplete systems (although it may be possible to estimate some of them).

Suppose that interest centers on estimating the demand elasticity  $\alpha_1$ . For simplicity, assume that  $\varepsilon_d$  and  $\varepsilon_s$  are well behaved, classical disturbances with

$$\begin{aligned} E[\varepsilon_{d,t} | x_t] &= E[\varepsilon_{s,t} | x_t] = 0, \\ E[\varepsilon_{d,t}^2 | x_t] &= \sigma_d^2, \\ E[\varepsilon_{s,t}^2 | x_t] &= \sigma_s^2, \\ E[\varepsilon_{d,t}\varepsilon_{s,t} | x_t] &= 0. \end{aligned}$$

All variables are mutually uncorrelated with observations at different time periods. Price, quantity, and income are measured in logarithms in deviations from their sample means. Solving the equations for  $p$  and  $q$  in terms of  $x$ ,  $\varepsilon_d$ , and  $\varepsilon_s$  produces the **reduced form** of the model

$$\begin{aligned} p &= \frac{\alpha_2 x}{\beta_1 - \alpha_1} + \frac{\varepsilon_d - \varepsilon_s}{\beta_1 - \alpha_1} = \pi_1 x + v_1, \\ q &= \frac{\beta_1 \alpha_2 x}{\beta_1 - \alpha_1} + \frac{\beta_1 \varepsilon_d - \alpha_1 \varepsilon_s}{\beta_1 - \alpha_1} = \pi_2 x + v_2. \end{aligned} \tag{10-41}$$

(Note the role of the “completeness” requirement that  $\alpha_1$  not equal  $\beta_1$ .)

It follows that  $\text{Cov}[p, \varepsilon_d] = \sigma_d^2 / (\beta_1 - \alpha_1)$  and  $\text{Cov}[p, \varepsilon_s] = -\sigma_s^2 / (\beta_1 - \alpha_1)$  so neither the demand nor the supply equation satisfies the assumptions of the classical regression model. The price elasticity of demand cannot be consistently estimated by least squares regression of  $q$  on  $x$  and  $p$ . This result is characteristic of simultaneous-equations models. Because the endogenous variables are all correlated with the disturbances, the least squares estimators of the parameters of equations with endogenous variables on the right-hand side are inconsistent.<sup>35</sup>

Suppose that we have a sample of  $T$  observations on  $p$ ,  $q$ , and  $x$  such that

$$\text{plim}(1/T)\mathbf{x}'\mathbf{x} = \sigma_x^2.$$

Since least squares is inconsistent, we might instead use an **instrumental variable estimator**.<sup>36</sup> The only variable in the system that is not correlated with the disturbances is  $x$ .

<sup>35</sup>This failure of least squares is sometimes labeled simultaneous equations bias.

<sup>36</sup>See Section 8.3.

CHAPTER 10 ♦ Systems of Equations **317**

Consider, then, the IV estimator,  $\hat{\beta}_1 = \mathbf{q}'\mathbf{x}/\mathbf{p}'\mathbf{x}$ . This estimator has

$$\text{plim } \hat{\beta}_1 = \text{plim} \frac{\mathbf{q}'\mathbf{x}/T}{\mathbf{p}'\mathbf{x}/T} = \frac{\sigma_x^2 \beta_1 \alpha_2 / (\beta_1 - \alpha_1)}{\sigma_x^2 \alpha_2 / (\beta_1 - \alpha_1)} = \beta_1.$$

Evidently, the parameter of the supply curve can be estimated by using an instrumental variable estimator. In the least squares regression of  $\mathbf{p}$  on  $\mathbf{x}$ , the predicted values are  $\hat{\mathbf{p}} = (\mathbf{p}'\mathbf{x} / \mathbf{x}'\mathbf{x})\mathbf{x}$ . It follows that in the instrumental variable regression the instrument is  $\hat{\mathbf{p}}$ . That is,

$$\hat{\beta}_1 = \frac{\hat{\mathbf{p}}'\mathbf{q}}{\hat{\mathbf{p}}'\mathbf{p}}.$$

Because  $\hat{\mathbf{p}}'\mathbf{p} = \hat{\mathbf{p}}'\hat{\mathbf{p}}$ ,  $\hat{\beta}_1$  is also the slope in a regression of  $q$  on these predicted values. This interpretation defines the **two-stage least squares estimator**.

It would be desirable to use a similar device to estimate the parameters of the demand equation, but unfortunately, we have exhausted the information in the sample. Not only does least squares fail to estimate the demand equation, but without some further assumptions, the sample contains no other information that can be used. This example illustrates the **problem of identification** alluded to in the introduction to this section.

The distinction between “exogenous” and “endogenous” variables in a model is a subtle and sometimes controversial complication. It is the subject of a long literature. We have drawn the distinction in a useful economic fashion at a few points in terms of whether a variable in the model could reasonably be expected to vary “autonomously,” independently of the other variables in the model. Thus, in a model of supply and demand, the weather variable in a supply equation seems obviously to be exogenous in a pure sense to the determination of price and quantity, whereas the current price clearly is “endogenous” by any reasonable construction. Unfortunately, this neat classification is of fairly limited use in macroeconomics, where almost no variable can be said to be truly exogenous in the fashion that most observers would understand the term. To take a common example, the estimation of consumption functions by ordinary least squares, as we did in some earlier examples, is usually treated as a respectable enterprise, even though most macroeconomic models (including the examples given here) depart from a consumption function in which income is exogenous. This departure has led analysts, for better or worse, to draw the distinction largely on statistical grounds. The methodological development in the literature has produced some consensus on this subject. As we shall see, the definitions formalize the economic characterization we drew earlier. We will loosely sketch a few results here for purposes of our derivations to follow. The interested reader is referred to the literature (and forewarned of some challenging reading).

Engle, Hendry, and Richard (1983) define a set of variables  $\mathbf{x}_t$  in a parameterized model to be **weakly exogenous** if the full model can be written in terms of a marginal probability distribution for  $\mathbf{x}_t$  and a conditional distribution for  $\mathbf{y}_t | \mathbf{x}_t$  such that estimation of the parameters of the conditional distribution is no less efficient than estimation of the full set of parameters of the joint distribution. This case will be true if none of the parameters in the conditional distribution appears in the marginal distribution for  $\mathbf{x}_t$ . In the present context, we will need this sort of construction to derive reduced forms the way we did previously. With reference to time-series applications (although the notion extends to cross sections as well), variables  $\mathbf{x}_t$  are said to be **predetermined** in the model if  $\mathbf{x}_t$  is independent of all *subsequent* structural disturbances  $\varepsilon_{t+s}$  for  $s \geq 0$ .

### 318 PART II ♦ Generalized Regression Model and Equation Systems

Variables that are predetermined in a model can be treated, at least asymptotically, as if they were exogenous in the sense that consistent estimators can be derived when they appear as regressors. We will use this result in Chapter 21, when we derive the properties of regressions containing lagged values of the dependent variable. A related concept is **Granger 1969-Sims (1977) causality**. Granger causality (a kind of statistical feedback) is absent when  $f(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{y}_{t-1})$  equals  $f(\mathbf{x}_t | \mathbf{x}_{t-1})$ . The definition states that in the conditional distribution, lagged values of  $\mathbf{y}_t$  add no information to explanation of movements of  $\mathbf{x}_t$  beyond that provided by lagged values of  $\mathbf{x}_t$  itself. This concept is useful in the construction of forecasting models. Finally, if  $\mathbf{x}_t$  is weakly exogenous and if  $\mathbf{y}_{t-1}$  does not Granger cause  $\mathbf{x}_t$ , then  $\mathbf{x}_t$  is **strongly exogenous**.

#### 10.6.2 A GENERAL NOTATION FOR LINEAR SIMULTANEOUS EQUATIONS MODELS<sup>37</sup>

The **structural form** of the model is<sup>38</sup>

$$\begin{aligned} \gamma_{11}y_{t1} + \gamma_{21}y_{t2} + \cdots + \gamma_{M1}y_{tM} + \beta_{11}x_{t1} + \cdots + \beta_{K1}x_{tK} &= \varepsilon_{t1}, \\ \gamma_{12}y_{t1} + \gamma_{22}y_{t2} + \cdots + \gamma_{M2}y_{tM} + \beta_{12}x_{t1} + \cdots + \beta_{K2}x_{tK} &= \varepsilon_{t2}, \\ &\vdots \\ \gamma_{1M}y_{t1} + \gamma_{2M}y_{t2} + \cdots + \gamma_{MM}y_{tM} + \beta_{1M}x_{t1} + \cdots + \beta_{KM}x_{tK} &= \varepsilon_{tM}. \end{aligned} \tag{10-42}$$

There are  $M$  equations and  $M$  endogenous variables, denoted  $y_1, \dots, y_M$ . There are  $K$  exogenous variables,  $x_1, \dots, x_K$ , that may include predetermined values of  $y_1, \dots, y_M$  as well. The first element of  $\mathbf{x}_t$  will usually be the constant, 1. Finally,  $\varepsilon_{t1}, \dots, \varepsilon_{tM}$  are the **structural disturbances**. The subscript  $t$  will be used to index observations,  $t = 1, \dots, T$ .

In matrix terms, the system may be written

$$\begin{aligned} [y_1 & y_2 & \cdots & y_M]_t & \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1M} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2M} \\ \vdots & & & \\ \gamma_{M1} & \gamma_{M2} & \cdots & \gamma_{MM} \end{bmatrix} \\ & + [x_1 & x_2 & \cdots & x_K]_t \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1M} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2M} \\ \vdots & & & \\ \beta_{K1} & \beta_{K2} & \cdots & \beta_{KM} \end{bmatrix} = [\varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_M]_t, \end{aligned}$$

<sup>37</sup>We will be restricting our attention to linear models. **Nonlinear systems** occupy another strand of literature in this area. Nonlinear systems bring forth numerous complications beyond those discussed here and are beyond the scope of this text. Gallant (1987), Gallant and Holly (1980), Gallant and White (1988), Davidson and MacKinnon (2004), and Wooldridge (2002a) provide further discussion.

<sup>38</sup>For the present, it is convenient to ignore the special nature of lagged endogenous variables and treat them the same as the strictly exogenous variables.

CHAPTER 10 ♦ Systems of Equations **319**

or

$$\mathbf{y}'_t \boldsymbol{\Gamma} + \mathbf{x}'_t \mathbf{B} = \boldsymbol{\varepsilon}'_t.$$

Each column of the parameter matrices is the vector of coefficients in a particular equation, whereas each row applies to a specific endogenous variable.

The underlying theory will imply a number of restrictions on  $\boldsymbol{\Gamma}$  and  $\mathbf{B}$ . One of the variables in each equation is labeled the *dependent* variable so that its coefficient in the model will be 1. Thus, there will be at least one “1” in each column of  $\boldsymbol{\Gamma}$ . This **normalization** is not a substantive restriction. The relationship defined for a given equation will be unchanged if every coefficient in the equation is multiplied by the same constant. Choosing a “dependent variable” simply removes this indeterminacy. If there are any identities, then the corresponding columns of  $\boldsymbol{\Gamma}$  and  $\mathbf{B}$  will be completely known, and there will be no disturbance for that equation. Because not all variables appear in all equations, some of the parameters will be zero. The theory may also impose other types of restrictions on the parameter matrices.

If  $\boldsymbol{\Gamma}$  is an upper triangular matrix, then the system is said to be **triangular**. In this case, the model is of the form

$$\begin{aligned} y_{t1} &= f_1(\mathbf{x}_t) + \varepsilon_{t1}, \\ y_{t2} &= f_2(y_{t1}, \mathbf{x}_t) + \varepsilon_{t2}, \\ &\vdots \\ y_{tM} &= f_M(y_{t1}, y_{t2}, \dots, y_{t,M-1}, \mathbf{x}_t) + \varepsilon_{tM}. \end{aligned}$$

The joint determination of the variables in this model is **recursive**. The first is completely determined by the exogenous factors. Then, given the first, the second is likewise determined, and so on.

The solution of the system of equations determining  $\mathbf{y}_t$  in terms of  $\mathbf{x}_t$  and  $\boldsymbol{\varepsilon}_t$  is the **reduced form** of the model,

$$\begin{aligned} \mathbf{y}'_t &= [x_1 \quad x_2 \quad \cdots \quad x_K]_t \begin{bmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1M} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2M} \\ \vdots & & & \\ \pi_{K1} & \pi_{K2} & \cdots & \pi_{KM} \end{bmatrix} + [v_1 \quad \cdots \quad v_M]_t \\ &= -\mathbf{x}'_t \mathbf{B} \boldsymbol{\Gamma}^{-1} + \boldsymbol{\varepsilon}'_t \boldsymbol{\Gamma}^{-1} \\ &= \mathbf{x}'_t \boldsymbol{\Pi} + \mathbf{v}'_t. \end{aligned}$$

For this solution to exist, the model must satisfy the **completeness condition** for simultaneous equations systems:  $\boldsymbol{\Gamma}$  must be nonsingular.

#### **Example 10.4 Structure and Reduced Form in a Small Macroeconomic Model**

Consider the model

$$\text{consumption : } c_t = \alpha_0 + \alpha_1 y_t + \alpha_2 c_{t-1} + \varepsilon_{t1},$$

$$\text{investment : } i_t = \beta_0 + \beta_1 r_t + \beta_2 (y_t - y_{t-1}) + \varepsilon_{t2},$$

$$\text{demand : } y_t = c_t + i_t + g_t.$$

## 320 PART II ♦ Generalized Regression Model and Equation Systems

The model contains an autoregressive consumption function based on output,  $y_t$ , and one lagged value, an investment equation based on interest,  $r_t$  and the growth in output, and an equilibrium condition. The model determines the values of the three endogenous variables  $c_t$ ,  $i_t$ , and  $y_t$ . This model is a **dynamic model**. In addition to the exogenous variables  $r_t$  and government spending,  $g_t$ , it contains two **predetermined variables**,  $c_{t-1}$  and  $y_{t-1}$ . These are obviously not exogenous, but with regard to the current values of the endogenous variables, they may be regarded as having already been determined. The deciding factor is whether or not they are uncorrelated with the current disturbances, which we might assume. The reduced form of this model is

$$Ac_t = \alpha_0(1 - \beta_2) + \beta_0\alpha_1 + \alpha_1\beta_1r_t + \alpha_1g_t + \alpha_2(1 - \beta_2)c_{t-1} - \alpha_1\beta_2y_{t-1} + (1 - \beta_2)\varepsilon_{t1} + \alpha_1\varepsilon_{t2},$$

$$Ai_t = \alpha_0\beta_2 + \beta_0(1 - \alpha_1) + \beta_1(1 - \alpha_1)r_t + \beta_2g_t + \alpha_2\beta_2c_{t-1} - \beta_2(1 - \alpha_1)y_{t-1} \\ + \beta_2\varepsilon_{t1} + (1 - \alpha_1)\varepsilon_{t2},$$

$$Ay_t = \alpha_0 + \beta_0 + \beta_1y_{t-1} + g_t + \alpha_2c_{t-1} - \beta_2y_{t-1} + \varepsilon_{t1} + \varepsilon_{t2},$$

where  $A = 1 - \alpha_1 - \beta_2$ . Note that the reduced form preserves the equilibrium condition.

Denote  $\mathbf{y}' = [c, i, y]$ ,  $\mathbf{x}' = [1, r, g, c_{-1}, y_{-1}]$ , and

$$\Gamma = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -\alpha_1 & -\beta_2 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} -\alpha_0 & -\beta_0 & 0 \\ 0 & -\beta_1 & 0 \\ 0 & 0 & -1 \\ -\alpha_2 & 0 & 0 \\ 0 & \beta_2 & 0 \end{bmatrix}, \quad \Gamma^{-1} = \frac{1}{\Delta} \begin{bmatrix} 1 - \beta_2 & \beta & 1 \\ \alpha_1 & 1 - \alpha_1 & 1 \\ \alpha_1 & \beta_2 & 1 \end{bmatrix},$$

$$\Pi' = \frac{1}{\Delta} \begin{bmatrix} \alpha_0(1 - \beta_2 + \beta_0\alpha_1) & \alpha_1\beta_1 & \alpha_1 & \alpha_2(1 - \beta_2) & -\beta_2\alpha_1 \\ \alpha_0\beta_2 + \beta_0(1 - \alpha_1) & \beta_1(1 - \alpha_1) & \beta_2 & \alpha_2\beta_2 & -\beta_2(1 - \alpha_1) \\ \alpha_0 + \beta_0 & \beta_1 & 1 & \alpha_2 & -\beta_2 \end{bmatrix},$$

where  $\Delta = 1 - \alpha_1 - \beta_2$ . The completeness condition is that  $\alpha_1$  and  $\beta_2$  do not sum to one.

There is ambiguity in the interpretation of coefficients in a simultaneous equations model. The effects in the structural form of the model would be labeled "causal," in that they are derived directly from the underlying theory. However, in order to trace through the effects of autonomous changes in the variables in the model, it is necessary to work through the reduced form. For example, the interest rate does not appear in the consumption function. But, that does not imply that changes in  $r_t$  would not "cause" changes in consumption, since changes in  $r_t$  change investment, which impacts demand which, in turn, does appear in the consumption function. Thus, we can see from the reduced form that  $\Delta c_t / \Delta r_t = \alpha_1\beta_1/A$ . Similarly, the "experiment,"  $\Delta c_t / \Delta y_t$  is meaningless without first determining what caused the change in  $y_t$ . If the change were induced by a change in the interest rate, we would find  $(\Delta c_t / \Delta r_t) / (\Delta y_t / \Delta r_t) = (\alpha_1\beta_1/A) / (\beta_1/A) = \alpha_1$ .

The structural disturbances are assumed to be randomly drawn from an  $M$ -variate distribution with

$$E[\boldsymbol{\varepsilon}_t | \mathbf{x}_t] = \mathbf{0} \quad \text{and} \quad E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_t | \mathbf{x}_t] = \boldsymbol{\Sigma}.$$

For the present, we assume that

$$E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_s | \mathbf{x}_t, \mathbf{x}_s] = \mathbf{0}, \quad \forall t, s.$$

Later, we will drop this assumption to allow for heteroscedasticity and autocorrelation. It will occasionally be useful to assume that  $\boldsymbol{\varepsilon}_t$  has a multivariate normal distribution, but we shall postpone this assumption until it becomes necessary. It may be convenient to retain the identities without disturbances as separate equations. If so, then one way to proceed with the stochastic specification is to place rows and columns of zeros in the

CHAPTER 10 ♦ Systems of Equations **321**

appropriate places in  $\Sigma$ . It follows that the **reduced-form disturbances**,  $\mathbf{v}'_t = \boldsymbol{\varepsilon}'_t \boldsymbol{\Gamma}^{-1}$  have

$$E[\mathbf{v}_t | \mathbf{x}_t] = (\boldsymbol{\Gamma}^{-1})' \mathbf{0} = \mathbf{0},$$

$$E[\mathbf{v}_t \mathbf{v}'_t | \mathbf{x}_t] = (\boldsymbol{\Gamma}^{-1})' \boldsymbol{\Sigma} \boldsymbol{\Gamma}^{-1} = \boldsymbol{\Omega}.$$

This implies that

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}' \boldsymbol{\Omega} \boldsymbol{\Gamma}.$$

The preceding formulation describes the model as it applies to an observation  $[\mathbf{y}', \mathbf{x}', \boldsymbol{\varepsilon}']_t$  at a particular point in time or in a cross section. In a sample of data, each joint observation will be one row in a data matrix,

$$[\mathbf{Y} \quad \mathbf{X} \quad \mathbf{E}] = \begin{bmatrix} \mathbf{y}'_1 & \mathbf{x}'_1 & \boldsymbol{\varepsilon}'_1 \\ \mathbf{y}'_2 & \mathbf{x}'_2 & \boldsymbol{\varepsilon}'_2 \\ \vdots \\ \mathbf{y}'_T & \mathbf{x}'_T & \boldsymbol{\varepsilon}'_T \end{bmatrix}.$$

In terms of the full set of  $T$  observations, the structure is

$$\mathbf{Y}\boldsymbol{\Gamma} + \mathbf{X}\mathbf{B} = \mathbf{E},$$

with

$$E[\mathbf{E} | \mathbf{X}] = \mathbf{0} \quad \text{and} \quad E[(1/T)\mathbf{E}'\mathbf{E} | \mathbf{X}] = \boldsymbol{\Sigma}.$$

Under general conditions, we can strengthen this structure to

$$\text{plim}[(1/T)\mathbf{E}'\mathbf{E}] = \boldsymbol{\Sigma}.$$

An important assumption, comparable with the one made in Chapter 4 for the classical regression model, is

$$\text{plim}(1/T)\mathbf{X}'\mathbf{X} = \mathbf{Q}, \quad \text{a finite positive definite matrix.} \quad (10-43)$$

We also assume that

$$\text{plim}(1/T)\mathbf{X}'\mathbf{E} = \mathbf{0}. \quad (10-44)$$

This assumption is what distinguishes the predetermined variables from the endogenous variables. The reduced form is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Pi} + \mathbf{V}, \quad \text{where } \mathbf{V} = \mathbf{E}\boldsymbol{\Gamma}^{-1}. \quad (10-45)$$

Combining the earlier results, we have

$$\text{plim} \frac{1}{T} \begin{bmatrix} \mathbf{Y}' \\ \mathbf{X}' \\ \mathbf{V}' \end{bmatrix} [\mathbf{Y} \quad \mathbf{X} \quad \mathbf{V}] = \begin{bmatrix} \boldsymbol{\Pi}' \mathbf{Q} \boldsymbol{\Pi} + \boldsymbol{\Omega} & \boldsymbol{\Pi}' \mathbf{Q} & \boldsymbol{\Omega} \\ \mathbf{Q} \boldsymbol{\Pi} & \mathbf{Q} & \mathbf{0}' \\ \boldsymbol{\Omega} & \mathbf{0} & \boldsymbol{\Omega} \end{bmatrix}.$$

### 10.6.3 THE PROBLEM OF IDENTIFICATION

Solving the **identification problem** logically precedes estimation. It is a crucial element of the model specification step. The issue is whether there is *any* way to obtain estimates of the parameters of the specified model. We have in hand a certain amount of information

## 322 PART II ♦ Generalized Regression Model and Equation Systems

to use for inference about the underlying structure. If more than one theory is consistent with the same “data,” then the theories are said to be **observationally equivalent** and there is no way of distinguishing them. We have already encountered this problem in Chapter 4, where we examined the issue of *multicollinearity*. The “model,”

$$\text{consumption} = \beta_1 + \beta_2 \text{WageIncome} + \beta_3 \text{NonWageIncome} + \beta_4 \text{TotalIncome} + \varepsilon, \quad (10-46)$$

cannot be distinguished from the alternative model

$$\text{consumption} = \gamma_1 + \gamma_2 \text{WageIncome} + \gamma_3 \text{NonWageIncome} + \gamma_4 \text{TotalIncome} + \omega, \quad (10-47)$$

where  $\gamma_1 = \beta_1$ ,  $\gamma_2 = \beta_2 + a$ ,  $\gamma_3 = \beta_3 + a$ ,  $\gamma_4 = \beta_4 - a$  for some nonzero  $a$ , if the data consist only of consumption and the two income values (and their sum). However, if we know that if  $\beta_4$  equals zero, then, as we saw in Chapter 4,  $\gamma_2$  must equal  $\beta_2$  and  $\gamma_3$  must equal  $\beta_3$ . The additional information serves to rule out the alternative model. The notion of observational equivalence relates to what can be learned from the available information, which consists of the sample data and the restrictions that theory places on the equations of the model. In Chapter 8, where we examined the instrumental variable estimator, we defined identification in terms of sufficient moment equations. Indeed, Figure 8.1 is precisely an application of the principle of observational equivalence. The case of measurement error that we examined in Section 8.5 is likewise about identification. The sample regression coefficient,  $b$ , converges to a function of two underlying parameters,  $\beta$  and  $\sigma_u^2$ ;  $\text{plim } b = \beta/[1 + \sigma_u^2/Q^{**}]$  where  $Q^{**} = \text{plim}(\mathbf{x}'\mathbf{x}^*/n)$ . With no further information about  $\sigma_u^2$ , we cannot recover  $\beta$  from the sample information,  $b$  and  $Q^{**}$  — by setting the differential,  $db = 0$ , you can see that there are different pairs of  $\beta$  and  $\sigma_u^2$  that produce the same  $\text{plim } b$ .

A mathematical statement of the idea can be made in terms of the likelihood function, which embodies the sample information. At this point, it helps to drop the statistical distinction between “y” and “x” and consider, in generic terms, the joint probability distribution for the observed data,  $p(\mathbf{Y}, \mathbf{X}|\boldsymbol{\theta})$ , given the model parameters. Two model structures are observationally equivalent if

$$p(\mathbf{Y}, \mathbf{X}|\boldsymbol{\theta}_1) = p(\mathbf{Y}, \mathbf{X}|\boldsymbol{\theta}_2) \quad \text{for } \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \text{ for all realizations of } (\mathbf{Y}, \mathbf{X}).$$

A structure is said to be *unidentified* if it is observationally equivalent to another structure.<sup>39</sup> (For our preceding consumption example, as will usually be the case when a model is unidentified, there are an infinite number of structures that are all equivalent to (10-46), one for each nonzero value of  $a$  in (10-47)).

The general simultaneous equations model we have specified in (10-42) is not identified. We have implicitly assumed that the marginal distribution of  $\mathbf{X}$  can be separated from the conditional distribution of  $\mathbf{Y}|\mathbf{X}$ . We can write the model as

$$p(\mathbf{Y}, \mathbf{X}|\boldsymbol{\Gamma}, \mathbf{B}, \boldsymbol{\Sigma}, \boldsymbol{\Theta}) = p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Pi}, \boldsymbol{\Omega})p(\mathbf{X}|\boldsymbol{\Theta}) \text{ with } \boldsymbol{\Pi} = -\mathbf{B}\boldsymbol{\Pi}^{-1} \text{ and } \boldsymbol{\Omega} = (\boldsymbol{\Gamma}')^{-1}\boldsymbol{\Sigma}(\boldsymbol{\Gamma})^{-1}.$$

We assume that  $\boldsymbol{\Theta}$  and  $(\boldsymbol{\Gamma}, \mathbf{B}, \boldsymbol{\Sigma})$  have no elements in common. But, let  $\mathbf{F}$  be any non-singular  $M \times M$  matrix and define  $\mathbf{B}_2 = \mathbf{FB}$  and  $\boldsymbol{\Gamma}_2 = \mathbf{F}\boldsymbol{\Gamma}$  and  $\boldsymbol{\Sigma}_2 = \mathbf{F}'\boldsymbol{\Sigma}\mathbf{F}$  (i.e., we just multiply the whole model by  $\mathbf{F}$ ). If  $\mathbf{F}$  is not equal to an identity matrix, then  $\mathbf{B}_2$ ,  $\boldsymbol{\Gamma}_2$ , and

<sup>39</sup>See Hsiao (1983) for a survey of this issue.

## CHAPTER 10 ♦ Systems of Equations 323

$\Sigma_2$  are a different  $\mathbf{B}$ ,  $\Gamma$  and  $\Sigma$  that are consistent with the same data, that is, with the same  $(\mathbf{Y}, \mathbf{X})$  which imply  $(\Pi$  and  $\Omega)$ . This follows because  $\Pi_2 = -\mathbf{B}_2^{-1}\Gamma_2 = -\mathbf{B}^{-1}\Gamma = \Pi$  and likewise for  $\Omega_2$ . To see how this will proceed from here, consider that in each equation, there is one “dependent variable,” that is a variable whose coefficient equals one. Therefore, one specific element of  $\Gamma$  in every equation (column) equals one. That rules out any matrix  $\mathbf{F}$  which does not leave a one in that position in  $\Gamma_2$ . Likewise, in the market equilibrium case in Section 10.6.1, the coefficient on  $x$  in the supply equation is zero. That means there is an element in one of the columns of  $\mathbf{B}$  that equals zero. Any  $\mathbf{F}$  that does not preserve that zero restriction is invalid. Thus, certain restrictions that theory imposes on the model rule out some of the alternative models. With enough restrictions, the only valid  $\mathbf{F}$  matrix will be  $\mathbf{F} = \mathbf{I}$ , and the model becomes identified.

The structural model consists of the equation system

$$\mathbf{y}'\mathbf{T} = -\mathbf{x}'\mathbf{B} + \boldsymbol{\epsilon}'$$

Each column in  $\Gamma$  and  $\mathbf{B}$  are the parameters of a specific equation in the system. The sample information consists of, at the first instance the data,  $(\mathbf{Y}, \mathbf{X})$ , and other nonsample information in the form of restrictions on parameter matrices, such as the normalizations noted in the preceding example. The sample data provide sample moments,  $\mathbf{X}'\mathbf{X}/n$ ,  $\mathbf{X}'\mathbf{Y}/n$  and  $\mathbf{Y}'\mathbf{Y}/n$ . For purposes of identification, which is independent of issues of sample size, suppose we could observe as large a sample as desired. Then, we could observe [from (10-45)]

$$\text{plim}(1/n)\mathbf{X}'\mathbf{X} = \mathbf{Q},$$

$$\text{plim}(1/n)\mathbf{X}'\mathbf{Y} = \text{plim}(1/n)\mathbf{X}'(\mathbf{X}\Pi + \mathbf{V}) = \mathbf{Q}\Pi,$$

$$\text{plim}(1/n)\mathbf{Y}'\mathbf{Y} = \text{plim}(1/n)(\mathbf{X}\Pi + \mathbf{V})'(\mathbf{X}\Pi + \mathbf{V}) = \Pi'\mathbf{Q}\Pi + \Omega.$$

Therefore,  $\Pi$ , the matrix of reduced-form coefficients, is observable:

$$\Pi = [\text{plim}(1/n)\mathbf{X}'\mathbf{Y}]^{-1}[\text{plim}(1/n)\mathbf{X}'\mathbf{X}]$$

This estimator is simply the equation-by-equation least squares regression of  $\mathbf{Y}$  on  $\mathbf{X}$ . Because  $\Pi$  is observable,  $\Omega$  is also:

$$\Omega = [\text{plim}(1/n)\mathbf{Y}'\mathbf{Y}] - [\text{plim}(1/n)\mathbf{Y}'\mathbf{X}][\text{plim}(1/n)\mathbf{X}'\mathbf{X}]^{-1}[\text{plim}(1/n)\mathbf{X}'\mathbf{Y}].$$

This result should be recognized as the matrix of least squares residual variances and covariances. Therefore,

$\Pi$  and  $\Omega$  can be estimated consistently by least squares regression of  $\mathbf{Y}$  on  $\mathbf{X}$ .

The information in hand, therefore, consists of  $\Pi$ ,  $\Omega$ , and whatever other nonsample information we have about the structure.<sup>40</sup>

Thus,  $\Pi$  and  $\Omega$  are “observable.” The ultimate question is whether we can deduce  $\Gamma$ ,  $\mathbf{B}$ ,  $\Sigma$  from  $\Pi$ ,  $\Omega$ . A simple counting exercise immediately reveals that the answer is

<sup>40</sup>We have not necessarily shown that this is *all* the information in the sample. In general, we observe the conditional distribution  $f(\mathbf{y}_i|\mathbf{x}_i)$ , which constitutes the likelihood for the reduced form. With normally distributed disturbances, this distribution is a function of only  $\Pi$  and  $\Omega$ . With other distributions, other or higher moments of the variables might provide additional information. See, for example, Goldberger (1964, p. 311), Hausman (1983, pp. 402–403), and especially Reinsel (1950).

### 324 PART II ♦ Generalized Regression Model and Equation Systems

no—there are  $M^2$  parameters  $\Gamma$ ,  $M(M+1)/2$  in  $\Sigma$  and  $KM$  in  $\mathbf{B}$  to be deduced. The sample data contain  $KM$  elements in  $\boldsymbol{\Pi}$  and  $M(M+1)/2$  elements in  $\boldsymbol{\Omega}$ . By simply counting equations and unknowns, we find that our data are insufficient by  $M^2$  pieces of information. We have (in principle) used the sample information already, so these  $M^2$  additional restrictions are going to be provided by the theory of the model. A small example will help to fix ideas.

#### **Example 10.5 Identification**

Consider a market in which  $q$  is quantity of  $Q$ ,  $p$  is price, and  $z$  is the price of  $Z$ , a related good. We assume that  $z$  enters both the supply and demand equations. For example,  $Z$  might be a crop that is purchased by consumers and that will be grown by farmers instead of  $Q$  if its price rises enough relative to  $p$ . Thus, we would expect  $\alpha_2 > 0$  and  $\beta_2 < 0$ . So,

$$\begin{aligned} q_d &= \alpha_0 + \alpha_1 p + \alpha_2 z + \varepsilon_d && (\text{demand}), \\ q_s &= \beta_0 + \beta_1 p + \beta_2 z + \varepsilon_s && (\text{supply}), \\ q_d &= q_s = q && (\text{equilibrium}). \end{aligned}$$

The reduced form is

$$\begin{aligned} q &= \frac{\alpha_1 \beta_0 - \alpha_0 \beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 \beta_2 - \alpha_2 \beta_1}{\alpha_1 - \beta_1} z + \frac{\alpha_1 \varepsilon_s - \alpha_2 \varepsilon_d}{\alpha_1 - \beta_1} = \pi_{11} + \pi_{21} z + v_q, \\ p &= \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{\beta_2 - \alpha_2}{\alpha_1 - \beta_1} z + \frac{\varepsilon_s - \varepsilon_d}{\alpha_1 - \beta_1} = \pi_{12} + \pi_{22} z + v_p. \end{aligned}$$

With only four reduced-form coefficients and six structural parameters, it is obvious that there will not be a complete solution for all six structural parameters in terms of the four reduced parameters. Suppose, though, that it is known that  $\beta_2 = 0$  (farmers do not substitute the alternative crop for this one). Then the solution for  $\beta_1$  is  $\pi_{21} / \pi_{22}$ . After a bit of manipulation, we also obtain  $\beta_0 = \pi_{11} - \pi_{12} \pi_{21} / \pi_{22}$ . The restriction identifies the supply parameters, but this step is as far as we can go.

Now, suppose that income  $x$ , rather than  $z$ , appears in the demand equation. The revised model is

$$\begin{aligned} q &= \alpha_0 + \alpha_1 p + \alpha_2 x + \varepsilon_1, \\ q &= \beta_0 + \beta_1 p + \beta_2 z + \varepsilon_2. \end{aligned}$$

The structure is now

$$[q \quad p] \begin{bmatrix} 1 & 1 \\ -\alpha_1 & -\beta_1 \end{bmatrix} + [1 \not\propto z] \begin{bmatrix} -\alpha_0 & -\beta_0 \\ -\alpha_2 & 0 \\ 0 & -\beta_2 \end{bmatrix} = [\varepsilon_1 \quad \varepsilon_2].$$

The reduced form is

$$[q \quad p] = [1 \not\propto z] \begin{bmatrix} (\alpha_1 \beta_0 - \alpha_0 \beta_1) / \Delta & (\beta_0 - \alpha_0) / \Delta \\ -\alpha_2 \beta_1 / \Delta & -\alpha_2 / \Delta \\ \alpha_1 \beta_2 / \Delta & \beta_2 / \Delta \end{bmatrix} + [v_1 \quad v_2],$$

where  $\Delta = (\alpha_1 - \beta_1)$ . Every false structure has the same reduced form. But in the coefficient matrix,

$$\tilde{\mathbf{B}} = \mathbf{BF} = \begin{bmatrix} \alpha_0 f_{11} + \beta_0 f_{21} & \alpha_0 f_{12} + \beta_0 f_{22} \\ \alpha_2 f_{11} & \alpha_2 f_{12} \\ \beta_2 f_{21} & \beta_2 f_{22} \end{bmatrix},$$

if  $f_{12}$  is not zero, then the imposter will have income appearing in the supply equation, which our theory has ruled out. Likewise, if  $f_{21}$  is not zero, then  $z$  will appear in the demand

## CHAPTER 10 ♦ Systems of Equations 325

equation, which is also ruled out by our theory. Thus, although all false structures have the same reduced form as the true one, the only one that is consistent with our theory (i.e., is **admissible**) and has coefficients of 1 on  $q$  in both equations (examine  $\Gamma F$ ) is  $F = I$ . This transformation just produces the original structure.

The unique solutions for the structural parameters in terms of the reduced-form parameters are now

$$\begin{aligned}\alpha_0 &= \pi_{11} - \pi_{12} \left( \frac{\pi_{31}}{\pi_{32}} \right), & \beta_0 &= \pi_{11} - \pi_{12} \left( \frac{\pi_{21}}{\pi_{22}} \right), \\ \alpha_1 &= \frac{\pi_{31}}{\pi_{32}}, & \beta_1 &= \frac{\pi_{21}}{\pi_{22}}, \\ \alpha_2 &= \pi_{22} \left( \frac{\pi_{21}}{\pi_{22}} - \frac{\pi_{31}}{\pi_{32}} \right), & \beta_2 &= \pi_{32} \left( \frac{\pi_{31}}{\pi_{32}} - \frac{\pi_{21}}{\pi_{22}} \right).\end{aligned}$$

The conclusion is that some equation systems are identified and others are not. The formal mathematical conditions under which an equation system is identified turns on some intricate results known as the **rank and order conditions**.

The *order condition* is a simple counting rule. In the equation system context, the order condition is that the number of exogenous variables that appear elsewhere in the equation system must be at least as large as the number of endogenous variables in the equation. We used this rule when we constructed the IV estimator in Chapter 8. In that setting, we required our model to be at least “identified” by requiring that the number of instrumental variables not contained in  $\mathbf{X}$  be at least as large as the number of endogenous variables. The correspondence of that single equation application with the condition defined here is that the rest of the equation system is, essentially, the rest of the world (i.e., the source of the instrumental variables).<sup>41</sup> A simple sufficient order condition for an equation system is that each equation must contain “its own” exogenous variable that does not appear elsewhere in the system.

The **order condition** is necessary for identification; the **rank condition** is sufficient. The equation system in (10-42) in structural form is  $\mathbf{y}'\Gamma = -\mathbf{x}'\mathbf{B} + \boldsymbol{\epsilon}'$ . The reduced form is  $\mathbf{y}' = \mathbf{x}'(-\mathbf{B}\Gamma^{-1}) + \boldsymbol{\epsilon}'\Gamma^{-1} = \mathbf{x}'\Pi + \mathbf{v}'$ . The way we are going to deduce the parameters in  $(\Gamma, \mathbf{B}, \Sigma)$  is from the reduced form parameters  $(\Pi, \Omega)$ . For a particular equation, say the  $j$ th, the solution is contained in  $\Pi\Gamma = -\mathbf{B}$ , or for a particular equation,  $\Pi\Gamma_j = -\mathbf{B}_j$  where  $\Gamma_j$  contains all the coefficients in the  $j$ th equation that multiply endogenous variables. One of these coefficients will equal one, usually some will equal zero, and the remainder are the nonzero coefficients on endogenous variables in the equation,  $\mathbf{Y}_j$  [these are denoted  $\gamma_j$  in (10-48) following]. Likewise,  $\mathbf{B}_j$  contains the coefficients in equation  $j$  on all exogenous variables in the model—some of these will be zero and the remainder will multiply variables in  $\mathbf{X}_j$ , the exogenous variables that appear in this equation [these are denoted  $\beta_j$  in (10-48) following]. The empirical counterpart will be

$$[\text{plim}(1/n)\mathbf{X}'\mathbf{X}]^{-1}[\text{plim}(1/n)\mathbf{X}'\mathbf{Y}_j]\Gamma_j - \mathbf{B}_j = \mathbf{0}.$$

The rank condition ensures that there is a unique solution to this set of equations. In practical terms, the rank condition is difficult to establish in large equation systems. Practitioners typically take it as a given. In small systems, such as the 2 or 3 equation

<sup>41</sup>This invokes the perennial question (encountered repeatedly in the applications in Chapter 8), “where do the instruments come from?” See Section 8.8 for discussion.

## 326 PART II ♦ Generalized Regression Model and Equation Systems

systems that dominate contemporary research, it is trivial. We have already used the rank condition in Chapter 8 where it played a role in the “relevance” condition for instrumental variable estimation. In particular, note after the statement of the assumptions for instrumental variable estimation, we assumed  $\text{plim}(1/n)\mathbf{Z}'\mathbf{X}$  is a matrix with rank  $K$ . (This condition is often labeled the “rank condition” in contemporary applications. It is not identical, but it is sufficient for the condition mentioned here).

To add all this up, it is instructive to return to the order condition. We are trying to solve a set of moment equations based on the relationship between the structural parameters and the reduced form. The sample information provides  $KM + M(M + 1)/2$  items in  $\boldsymbol{\Pi}$  and  $\boldsymbol{\Omega}$ . We require  $M^2$  additional **restrictions**, imposed by the theory behind the model. The restrictions come in the form of normalizations, most commonly **exclusion restrictions**, and other relationships among the parameters, such as linear relationships, or specific values attached to coefficients.

The question of identification is a theoretical exercise. It arises in all econometric settings in which the parameters of a model are to be deduced from the combination of sample information and nonsample (theoretical) information. The crucial issue in each of these cases is our ability (or lack of) to deduce the values of structural parameters uniquely from sample information in terms of sample moments coupled with **nonsample information**, mainly restrictions on parameter values. The issue of identification is the subject of a lengthy literature including Working (1927) (which has been adapted to produce Figure 8.1), Gabrielsen (1978), Amemiya (1985), Bekker and Wansbeek (2001), and continuing through the contemporary discussion of natural experiments (Section 8.8 and Angrist and Pischke (2010), with commentary).

### 10.6.4 SINGLE EQUATION ESTIMATION AND INFERENCE

For purposes of estimation and inference, we write the specification of the simultaneous equations model in the form that the researcher would typically formulate it;

$$\begin{aligned}\mathbf{y}_j &= \mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{Y}_j\boldsymbol{\gamma}_j + \boldsymbol{\varepsilon}_j \\ &= \mathbf{Z}_j\boldsymbol{\delta}_j + \boldsymbol{\varepsilon}_j\end{aligned}\tag{10-48}$$

where  $\mathbf{y}_j$  is the “dependent variable” in the equation,  $\mathbf{X}_j$  is the set of exogenous variables that appear in the  $j$ th equation—note that this is not all the variables in the model—and  $\mathbf{Z}_j = (\mathbf{X}_j, \mathbf{Y}_j)$ . The full set of exogenous variables in the model, including  $\mathbf{X}_j$  and variables that appear elsewhere in the model (including a constant term if any equation includes one) is denoted  $\mathbf{X}$ . For example, in the supply/demand model in Example 10.5, the full set of exogenous variables is  $\mathbf{X} = (\mathbf{1}, \mathbf{x}, \mathbf{z})$ , while for the demand equation,  $\mathbf{X}_{\text{Demand}} = (\mathbf{1}, \mathbf{x})$  and  $\mathbf{X}_{\text{Supply}} = (\mathbf{1}, \mathbf{z})$ . Finally,  $\mathbf{Y}_j$  is the endogenous variables that appear on the right-hand side of the  $j$ th equation. Once again, this is likely to be a subset of the endogenous variables in the full model. In Example 10.5,  $\mathbf{Y}_j = (\text{price})$  in both cases.

There are two approaches to estimation and inference for simultaneous equations models. **Limited information estimators** are constructed for each equation individually. The approach is analogous to estimation of the seemingly unrelated regressions model in Section 10.2 by least squares, one equation at a time. **Full information estimators** are used to estimate all equations simultaneously. The counterpart for the seemingly unrelated regressions model is the feasible generalized least squares estimator discussed in

## CHAPTER 10 ♦ Systems of Equations 327

Section 10.2.3. The major difference to be accommodated at this point is the endogeneity of  $\mathbf{Y}_j$  in (10-48).

The equation system in (10-48) is precisely the model developed in Chapter 8. Least squares will generally be unsuitable as it is inconsistent due to the correlation between  $\mathbf{Y}_j$  and  $\boldsymbol{\epsilon}_j$ . The usual approach will be two-stage least squares as developed in Sections 8.3.2 to 8.3.4. The only difference between the case considered here and that in Chapter 8 is the source of the instrumental variables. In our general model in Chapter 8, the source of the instruments remained somewhat ambiguous; the overall rule was “outside the model.” In this setting, the instruments come from elsewhere in the model—that is, “not in the  $j$ th equation.” Thus, for estimating the linear simultaneous equations model, the most common estimator is

$$\begin{aligned}\hat{\boldsymbol{\delta}}_{j,2SLS} &= [\hat{\mathbf{Z}}'_j \hat{\mathbf{Z}}_j]^{-1} \hat{\mathbf{Z}}'_j \mathbf{y}_j \\ &= [(\mathbf{Z}'_j \mathbf{X})(\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{Z}_j)]^{-1} (\mathbf{Z}'_j \mathbf{X})(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}_j,\end{aligned}\tag{10-49}$$

where all columns of  $\hat{\mathbf{Z}}'_j$  are obtained as predictions in a regression of the corresponding column of  $\mathbf{Z}_j$  on  $\mathbf{X}$ . This equation also results in a useful simplification of the estimated asymptotic covariance matrix,

$$\text{Est. Asy. Var}[\hat{\boldsymbol{\delta}}_{j,2SLS}] = \hat{\sigma}_{jj} [\hat{\mathbf{Z}}'_j \hat{\mathbf{Z}}_j]^{-1}.$$

It is important to note that  $\sigma_{jj}$  is estimated by

$$\hat{\sigma}_{jj} = \frac{(\mathbf{y}_j - \mathbf{Z}_j \hat{\boldsymbol{\delta}}_j)' (\mathbf{y}_j - \mathbf{Z}_j \hat{\boldsymbol{\delta}}_j)}{T},\tag{10-50}$$

using the original data, not  $\hat{\mathbf{Z}}_j$ .

Note the role of the order condition for identification in the two-stage least squares estimator. Formally, the order condition requires that the number of exogenous variables that appear elsewhere in the model (not in this equation) be at least as large as the number of endogenous variables that appear in this equation. The implication will be that we are going to predict  $\mathbf{Z}_j = (\mathbf{X}_j, \mathbf{Y}_j)$  using  $\mathbf{X} = (\mathbf{X}_j, \mathbf{X}_j^*)$ . In order for these predictions to be linearly independent, there must be at least as many variables used to compute the predictions as there are variables being predicted. Comparing  $(\mathbf{X}_j, \mathbf{Y}_j)$  to  $(\mathbf{X}_j, \mathbf{X}_j^*)$ , we see that there must be at least as many variables in  $\mathbf{X}_j^*$  as there are in  $\mathbf{Y}_j$ , which is the order condition. The practical rule of thumb that every equation have at least one variable in it that does not appear in any other equation will guarantee this outcome.

Two-stage least squares is used nearly universally in estimation of simultaneous equation models—for precisely the reasons outlined in Chapter 8. However, some applications (and some theoretical treatments) have suggested that the **limited information maximum likelihood (LIML) estimator** based on the normal distribution may have better properties. The technique has also found recent use in the analysis of weak instruments that we consider in Section 10.6.5. A full (lengthy) derivation of the log-likelihood is provided in Davidson and MacKinnon (2004). We will proceed to the practical aspects of this estimator and refer the reader to this source for the background formalities. A result that emerges from the derivation is that the LIML estimator has the same asymptotic distribution as the 2SLS estimator, and the latter does not rely on an assumption

### 328 PART II ♦ Generalized Regression Model and Equation Systems

of normality. This raises the question why one would use the LIML technique given the availability of the more robust (and computationally simpler) alternative. Small sample results are sparse, but they would favor 2SLS as well. [See Phillips (1983).] One significant virtue of LIML is its invariance to the normalization of the equation. Consider an example in a system of equations,

$$y_1 = y_2\gamma_2 + y_3\gamma_3 + x_1\beta_1 + x_2\beta_2 + \varepsilon_1.$$

An equivalent equation would be

$$\begin{aligned} y_2 &= y_1(1/\gamma_2) + y_3(-\gamma_3/\gamma_2) + x_1(-\beta_1/\gamma_2) + x_2(-\beta_2/\gamma_2) + \varepsilon_1(-1/\gamma_2) \\ &= y_1\tilde{\gamma}_1 + y_3\tilde{\gamma}_3 + x_1\tilde{\beta}_1 + x_2\tilde{\beta}_2 + \tilde{\varepsilon}_1. \end{aligned}$$

The parameters of the second equation can be manipulated to produce those of the first. But, as you can easily verify, the 2SLS estimator is not invariant to the normalization of the equation—2SLS would produce numerically different answers. LIML would give the same numerical solutions to both estimation problems suggested earlier. A second virtue is LIML's better performance in the presence of weak instruments.

The LIML, or **least variance ratio** estimator, can be computed as follows.<sup>42</sup> Let

$$\mathbf{W}_j^0 = \mathbf{E}_j^{0'} \mathbf{E}_j^0, \quad (10-51)$$

where

$$\mathbf{Y}_j^0 = [\mathbf{y}_j, \mathbf{Y}_j],$$

and

$$\mathbf{E}_j^0 = \mathbf{M}_j \mathbf{Y}_j^0 = [\mathbf{I} - \mathbf{X}_j (\mathbf{X}'_j \mathbf{X}_j)^{-1} \mathbf{X}'_j] \mathbf{Y}_j^0. \quad (10-52)$$

Each column of  $\mathbf{E}_j^0$  is a set of least squares residuals in the regression of the corresponding column of  $\mathbf{Y}_j^0$  on  $\mathbf{X}_j$ , that is, the exogenous variables that appear in the  $j$ th equation. Thus,  $\mathbf{W}_j^0$  is the matrix of sums of squares and cross products of these residuals. Define

$$\mathbf{W}_j^1 = \mathbf{E}_j^{1'} \mathbf{E}_j^1 = \mathbf{Y}_j^{0'} [\mathbf{I} - \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'] \mathbf{Y}_j^0. \quad (10-53)$$

That is,  $\mathbf{W}_j^1$  is defined like  $\mathbf{W}_j^0$  except that the regressions are on all the  $x$ 's in the model, not just the ones in the  $j$ th equation. Let

$$\lambda_1 = \text{smallest characteristic root of } (\mathbf{W}_j^1)^{-1} \mathbf{W}_j^0. \quad (10-54)$$

This matrix is asymmetric, but all its roots are real and greater than or equal to 1. Depending on the available software, it may be more convenient to obtain the identical smallest root of the symmetric matrix  $\mathbf{D} = (\mathbf{W}_j^1)^{-1/2} \mathbf{W}_j^0 (\mathbf{W}_j^1)^{-1/2}$ . Now partition  $\mathbf{W}_j^0$  into

$\mathbf{W}_j^0 = \begin{bmatrix} \mathbf{w}_{jj}^0 & \mathbf{w}_j^{0'} \\ \mathbf{w}_j^0 & \mathbf{W}_{jj}^0 \end{bmatrix}$  corresponding to  $[\mathbf{y}_j, \mathbf{Y}_j]$ , and partition  $\mathbf{W}_j^1$  likewise. Then, with these

<sup>42</sup>The least variance ratio estimator is derived in Johnston (1984). The LIML estimator was derived by Anderson and Rubin (1949, 1950). The LIML estimator has, since its derivation by Anderson and Rubin in 1949 and 1950, been of largely theoretical interest only. The much simpler and equally efficient two-stage least squares estimator has stood as the estimator of choice. But LIML and the A-R specification test have been rediscovered and reinvigorated with their use in the analysis of weak instruments. See Hahn and Hausman (2002, 2003) and Sections 8.7 and 10.6.6.

parts in hand,

$$\hat{\gamma}_{j,\text{LIML}} = [\mathbf{W}_{jj}^0 - \lambda_1 \mathbf{W}_{jj}^1]^{-1} (\mathbf{w}_j^0 - \lambda_1 \mathbf{w}_j^1) \quad (10-55)$$

and

$$\hat{\beta}_{j,\text{LIML}} = [\mathbf{X}'_j \mathbf{X}_j]^{-1} \mathbf{X}'_j (\mathbf{y}_j - \mathbf{Y}_j \hat{\gamma}_{j,\text{LIML}}).$$

Note that  $\beta_j$  is estimated by a simple least squares regression. [See (3-18).] The asymptotic covariance matrix for the LIML estimator is identical to that for the 2SLS estimator.<sup>43</sup> The implication is that with normally distributed disturbances, 2SLS is fully efficient.

The  **$k$  class** of estimators is defined by the following form

$$\hat{\delta}_{j,k} = \begin{pmatrix} \hat{\gamma}_{j,k} \\ \hat{\beta}_{j,k} \end{pmatrix} \begin{bmatrix} \mathbf{Y}'_j \mathbf{Y}_j - k \mathbf{V}'_j \mathbf{V}_j & \mathbf{Y}'_j \mathbf{X}_j \\ \mathbf{X}'_j \mathbf{Y}_j & \mathbf{X}'_j \mathbf{X}_j \end{bmatrix} \begin{bmatrix} \mathbf{Y}'_j \mathbf{y}_j - k \mathbf{V}'_j \mathbf{v}_j \\ \mathbf{X}'_j \mathbf{y}_j \end{bmatrix}, \quad (10-56)$$

where  $\mathbf{V}_j$  and  $\mathbf{v}_j$  are the reduced form disturbances in (10-45). The feasible estimator is computed using the residuals from the OLS regressions of  $\mathbf{Y}_j$  and  $\mathbf{y}_j$  on  $\mathbf{X}$  (not  $\mathbf{X}_j$ ). We have already considered three members of the class, OLS with  $k = 0$ , 2SLS with  $k = 1$ , and, it can be shown, LIML with  $k = \lambda_1$ . [This last result follows from (10-55).] There have been many other  $k$ -class estimators derived; Davidson and MacKinnon (2004, pp. 537–538 and 548–549) and Mariano (2001) give discussion. It has been shown that all members of the  $k$  class for which  $k$  converges to 1 at a rate faster than  $1/\sqrt{n}$  have the same asymptotic distribution as that of the 2SLS estimator that we examined earlier. These are largely of theoretical interest, given the pervasive use of 2SLS or OLS, save for an important consideration. The large sample properties of all  $k$ -class estimators are the same, but the finite-sample properties are possibly very different. Davidson and MacKinnon (2004, pp. 537–538 and 548–549) and Mariano (1982, 2001) suggest that some evidence favors LIML when the sample size is small or moderate and the number of overidentifying restrictions is relatively large.

### 10.6.5 SYSTEM METHODS OF ESTIMATION

We may formulate the full system of equations as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_M \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \\ \vdots \\ \boldsymbol{\delta}_M \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_M \end{bmatrix} \quad (10-57)$$

or

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon},$$

where

$$E[\boldsymbol{\epsilon} | \mathbf{X}] = \mathbf{0}, \quad \text{and} \quad E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}' | \mathbf{X}] = \bar{\Sigma} = \boldsymbol{\Sigma} \otimes \mathbf{I}. \quad (10-58)$$

<sup>43</sup>This is proved by showing that both estimators are members of the “ $k$  class” of estimators, all of which have the same asymptotic covariance matrix. Details are given in Theil (1971) and Schmidt (1976).

### 330 PART II ♦ Generalized Regression Model and Equation Systems

[See (10-6).] The least squares estimator,

$$\mathbf{d} = [\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{y},$$

is equation-by-equation ordinary least squares and is inconsistent. But even if ordinary least squares were consistent, we know from our results for the seemingly unrelated regressions model that it would be inefficient compared with an estimator that makes use of the cross-equation correlations of the disturbances. For the first issue, we turn once again to an IV estimator. For the second, as we did Section 10.2.1, we use a generalized least squares approach. Thus, assuming that the matrix of instrumental variables,  $\bar{\mathbf{W}}$  satisfies the requirements for an IV estimator, a consistent though inefficient estimator would be

$$\hat{\mathbf{d}}_{IV} = [\bar{\mathbf{W}}'\mathbf{Z}]^{-1}\bar{\mathbf{W}}'\mathbf{y}. \quad (10-59)$$

Analogous to the seemingly unrelated regressions model, a more efficient estimator would be based on the generalized least squares principle,

$$\hat{\mathbf{d}}_{IV,GLS} = [\bar{\mathbf{W}}'(\Sigma^{-1} \otimes \mathbf{I})\mathbf{Z}]^{-1}\bar{\mathbf{W}}'(\Sigma^{-1} \otimes \mathbf{I})\mathbf{y}, \quad (10-60)$$

or, where  $\mathbf{W}_j$  is the set of instrumental variables for the  $j$ th equation,

$$\hat{\mathbf{d}}_{IV,GLS} = \begin{bmatrix} \sigma^{11}\mathbf{W}'_1\mathbf{Z}_1 & \sigma^{12}\mathbf{W}'_1\mathbf{Z}_2 & \dots & \sigma^{1M}\mathbf{W}'_1\mathbf{Z}_M \\ \sigma^{21}\mathbf{W}'_2\mathbf{Z}_1 & \sigma^{22}\mathbf{W}'_2\mathbf{Z}_2 & \dots & \sigma^{2M}\mathbf{W}'_2\mathbf{Z}_M \\ \vdots & & & \vdots \\ \sigma^{M1}\mathbf{W}'_M\mathbf{Z}_1 & \sigma^{M2}\mathbf{W}'_M\mathbf{Z}_2 & \dots & \sigma^{MM}\mathbf{W}'_M\mathbf{Z}_M \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^M \sigma^{1j}\mathbf{W}'_1\mathbf{y}_j \\ \sum_{j=1}^M \sigma^{2j}\mathbf{W}'_2\mathbf{y}_j \\ \vdots \\ \sum_{j=1}^M \sigma^{Mj}\mathbf{W}'_M\mathbf{y}_j \end{bmatrix}.$$

Three IV techniques are generally used for joint estimation of the entire system of equations: three-stage least squares, GMM, and **full information maximum likelihood (FIML)**. We will consider three-stage least squares here. GMM and FIML are discussed in Chapters 13 and 14, respectively.

Consider the IV estimator formed from

$$\hat{\mathbf{W}} = \hat{\mathbf{Z}} = \text{diag}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_1, \dots, \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_M] = \begin{bmatrix} \hat{\mathbf{Z}}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{Z}}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \hat{\mathbf{Z}}_M \end{bmatrix}.$$

The IV estimator,

$$\hat{\mathbf{d}}_{IV} = [\hat{\mathbf{Z}}'\mathbf{Z}]^{-1}\hat{\mathbf{Z}}'\mathbf{y},$$

is simply equation-by-equation 2SLS. We have already established the consistency of 2SLS. By analogy to the seemingly unrelated regressions model of Section 10.2, however, we would expect this estimator to be less efficient than a GLS estimator. A natural candidate would be

$$\hat{\mathbf{d}}_{3SLS} = [\hat{\mathbf{Z}}'(\Sigma^{-1} \otimes \mathbf{I})\mathbf{Z}]^{-1}\hat{\mathbf{Z}}'(\Sigma^{-1} \otimes \mathbf{I})\mathbf{y}.$$

CHAPTER 10 ♦ Systems of Equations **331**

For this estimator to be a valid IV estimator, we must establish that

$$\text{plim} \frac{1}{T} \hat{\mathbf{Z}}' (\Sigma^{-1} \otimes \mathbf{I}) \boldsymbol{\epsilon} = \mathbf{0},$$

which is  $M$  sets of equations, each one of the form

$$\text{plim} \frac{1}{T} \sum_{j=1}^M \sigma^{ij} \hat{\mathbf{Z}}'_i \boldsymbol{\epsilon}_j = \mathbf{0}.$$

Each is the sum of vectors all of which converge to zero, as we saw in the development of the 2SLS estimator. The second requirement, that

$$\text{plim} \frac{1}{T} \hat{\mathbf{Z}}' (\Sigma^{-1} \otimes \mathbf{I}) \mathbf{Z} \neq \mathbf{0},$$

and that the matrix be nonsingular, can be established along the lines of its counterpart for 2SLS. Identification of every equation by the rank condition is sufficient. [But, see Mariano (2001) on the subject of “weak instruments.”]

Once again using the idempotency of  $\mathbf{I} - \mathbf{M}$ , we may also interpret this estimator as a GLS estimator of the form

$$\hat{\boldsymbol{\delta}}_{3SLS} = [\hat{\mathbf{Z}}' (\Sigma^{-1} \otimes \mathbf{I}) \hat{\mathbf{Z}}]^{-1} \hat{\mathbf{Z}}' (\Sigma^{-1} \otimes \mathbf{I}) \mathbf{y}. \quad (10-61)$$

The appropriate asymptotic covariance matrix for the estimator is

$$\text{Asy. Var}[\hat{\boldsymbol{\delta}}_{3SLS}] = [\bar{\mathbf{Z}}' (\Sigma^{-1} \otimes \mathbf{I}) \bar{\mathbf{Z}}]^{-1}, \quad (10-62)$$

where  $\bar{\mathbf{Z}} = \text{diag}[\mathbf{X}\boldsymbol{\Pi}_j, \mathbf{X}_j]$ . This matrix would be estimated with the bracketed inverse matrix in (10-61).

Using sample data, we find that  $\bar{\mathbf{Z}}$  may be estimated with  $\hat{\mathbf{Z}}$ . The remaining difficulty is to obtain an estimate of  $\Sigma$ . In estimation of the multivariate regression model, for efficient estimation, any consistent estimator of  $\Sigma$  will do. The designers of the 3SLS method, Zellner and Theil (1962), suggest the natural choice arising out of the two-stage least estimates. The **three-stage least squares (3SLS) estimator** is thus defined as follows:

1. Estimate  $\boldsymbol{\Pi}$  by ordinary least squares and compute  $\hat{\mathbf{Y}}_j$  for each equation.
2. Compute  $\hat{\boldsymbol{\delta}}_{j,2SLS}$  for each equation; then

$$\hat{\sigma}_{ij} = \frac{(\mathbf{y}_i - \mathbf{Z}_i \hat{\boldsymbol{\delta}}_i)' (\mathbf{y}_j - \mathbf{Z}_j \hat{\boldsymbol{\delta}}_j)}{T}. \quad (10-63)$$

3. Compute the GLS estimator according to (10-61) and an estimate of the asymptotic covariance matrix according to (10-62) using  $\hat{\mathbf{Z}}$  and  $\hat{\Sigma}$ .

It is also possible to iterate the 3SLS computation. Unlike the seemingly unrelated regressions estimator, however, this method does not provide the maximum likelihood estimator, nor does it improve the asymptotic efficiency.<sup>44</sup>

By showing that the 3SLS estimator satisfies the requirements for an IV estimator, we have established its consistency. The question of asymptotic efficiency remains. It can

---

<sup>44</sup>A Jacobian term needed to maximize the log-likelihood is not treated by the 3SLS estimator. See Dhrymes (1973).

### 332 PART II ♦ Generalized Regression Model and Equation Systems

be shown that among all IV estimators that use only the sample information embodied in the system, 3SLS is asymptotically efficient.<sup>45</sup> For normally distributed disturbances, it can also be shown that 3SLS has the same asymptotic distribution as the full information maximum likelihood estimator.

#### **Example 10.6 Klein's Model I**

A widely used example of a simultaneous equations model of the economy is Klein's (1950) *Model I*. The model may be written

$$\begin{aligned}
 C_t &= \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + \alpha_3 (W_t^P + W_t^G) + \varepsilon_{1t} && \text{(consumption),} \\
 I_t &= \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 K_{t-1} &+ \varepsilon_{2t} & \text{(investment),} \\
 W_t^P &= \gamma_0 + \gamma_1 X_t + \gamma_2 X_{t-1} + \gamma_3 A_t &+ \varepsilon_{3t} & \text{(private wages),} \\
 X_t &= C_t + I_t + G_t && \text{(equilibrium demand),} \\
 P_t &= X_t - T_t - W_t^P && \text{(private profits),} \\
 K_t &= K_{t-1} + I_t && \text{(capital stock).}
 \end{aligned}$$

The endogenous variables are each on the left-hand side of an equation and are labeled on the right. The exogenous variables are  $G_t$  = government nonwage spending,  $T_t$  = indirect business taxes plus net exports,  $W_t^G$  = government wage bill,  $A_t$  = time trend measured as years from 1931, and the constant term. There are also three predetermined variables: the lagged values of the capital stock, private profits, and total demand. The model contains three **behavioral equations**, an **equilibrium condition** and two accounting identities. This model provides an excellent example of a small, dynamic model of the economy. It has also been widely used as a test ground for simultaneous equations estimators. Klein estimated the  parameters using yearly data for 1921 to 1941. The data are listed in Appendix Table F10.2. Table 10.5. presents limited and full information estimates for Klein's Model I based on the original data for 1920–1941.<sup>46</sup>

It might seem, in light of the entire discussion, that one of the structural estimators described previously should always be preferred to ordinary least squares, which, alone among the estimators considered here, is inconsistent. Unfortunately, the issue is not so clear. First, it is often found that the OLS estimator is surprisingly close to the structural estimator. It can be shown that, at least in some cases, OLS has a smaller variance about its mean than does 2SLS about its mean, leading to the possibility that OLS might be more precise in a mean-squared-error sense.<sup>47</sup> But this result must be tempered by the finding that the OLS standard errors are, in all likelihood, not useful for inference purposes.<sup>48</sup> Nonetheless, OLS is a frequently used estimator. Obviously, this discussion is relevant only to finite samples. Asymptotically, 2SLS must dominate OLS, and in a correctly specified model, any full information estimator must dominate any limited

<sup>45</sup>See Schmidt (1976) for a proof of its efficiency relative to 2SLS.

<sup>46</sup>The asymptotic covariance matrix for the LIML estimator will differ from that for the 2SLS estimator in a finite sample because the estimator of  $\sigma_{jj}$  that multiplies the inverse matrix will differ and because in computing the matrix to be inverted, the value of "k" [see the equation after (10-55)] is one for 2SLS and the smallest root in (10-54) for LIML. Asymptotically, k equals one and the estimators of  $\sigma_{jj}$  are equivalent.

<sup>47</sup>See Goldberger (1964, pp. 359–360).

<sup>48</sup>Cragg (1967).

CHAPTER 10 ♦ Systems of Equations **333****TABLE 10.5** Estimates of Klein's Model I (Estimated Asymptotic Standard Errors in Parentheses)

<i>Limited Information Estimates</i>				<i>Full Information Estimates</i>				
<b>2SLS</b>				<b>3SLS</b>				
<i>C</i>	16.6 (1.32)	0.017 (0.118)	0.216 (0.107)	0.810 (0.040)	16.4 (1.30)	0.125 (0.108)	0.163 (0.100)	0.790 (0.038)
<i>I</i>	20.3 (7.54)	0.150 (0.173)	0.616 (0.162)	-0.158 (0.036)	28.2 (6.79)	-0.013 (0.162)	0.756 (0.153)	-0.195 (0.033)
<i>W<sup>p</sup></i>	1.50 (1.15)	0.439 (0.036)	0.147 (0.039)	0.130 (0.029)	1.80 (1.12)	0.400 (0.032)	0.181 (0.034)	0.150 (0.028)
<b>LIML</b>				<b>FIML</b>				
<i>C</i>	17.1 (1.84)	-0.222 (0.202)	0.396 (0.174)	0.823 (0.055)	18.3 (2.49)	-0.232 (0.312)	0.388 (0.217)	0.802 (0.036)
<i>I</i>	22.6 (9.24)	0.075 (0.219)	0.680 (0.203)	-0.168 (0.044)	27.3 (7.94)	-0.801 (0.491)	1.052 (0.353)	-0.146 (0.30)
<i>W<sup>p</sup></i>	1.53 (2.40)	0.434 (0.137)	0.151 (0.135)	0.132 (0.065)	5.79 (1.80)	0.234 (0.049)	0.285 (0.045)	0.235 (0.035)
<b>OLS</b>				<b>I3SLS</b>				
<i>C</i>	16.2 (1.30)	0.193 (0.091)	0.090 (0.091)	0.796 (0.040)	16.6 (1.22)	0.165 (0.096)	0.177 (0.090)	0.766 (0.035)
<i>I</i>	10.1 (5.47)	0.480 (0.097)	0.333 (0.101)	-0.112 (0.027)	42.9 (10.6)	-0.356 (0.260)	1.01 (0.249)	-0.260 (0.051)
<i>W<sup>p</sup></i>	1.50 (1.27)	0.439 (0.032)	0.146 (0.037)	0.130 (0.032)	2.62 (1.20)	0.375 (0.031)	0.194 (0.032)	0.168 (0.029)

information one. The finite sample properties are of crucial importance. Most of what we know is asymptotic properties, but most applications are based on rather small or moderately sized samples.

The large difference between the inconsistent OLS and the other estimates suggests the bias discussed earlier. On the other hand, the incorrect sign on the LIML and FIML estimate of the coefficient on  $P$  and the even larger difference of the coefficient on  $P_{-1}$  in the  $C$  equation are striking. Assuming that the equation is properly specified, these anomalies would likewise be attributed to finite sample variation, because LIML and 2SLS are asymptotically equivalent.

Intuition would suggest that systems methods, 3SLS, and FIML, are to be preferred to single-equation methods, 2SLS and LIML. Indeed, if the advantage is so transparent, why would one ever choose a single-equation estimator? The proper analogy is to the use of single-equation OLS versus GLS in the SURE model of Section 10.2. An obvious practical consideration is the computational simplicity of the single-equation methods. But the current state of available software has eliminated this advantage.

Although the system methods of estimation are asymptotically better, they have two problems. First, any specification error in the structure of the model will be propagated throughout the system by 3SLS or FIML. The limited information estimators will, by and large, confine a problem to the particular equation in which it appears. Second, in the same fashion as the SURE model, the finite-sample variation of the estimated covariance matrix is transmitted throughout the system. Thus, the finite-sample variance of 3SLS may well be as large as or larger than that of 2SLS. Although they are only

### 334 PART II ♦ Generalized Regression Model and Equation Systems

single estimates, the results for Klein's Model I give a striking example. The upshot would appear to be that the advantage of the systems estimators in finite samples may be more modest than the asymptotic results would suggest. Monte Carlo studies of the issue have tended to reach the same conclusion.<sup>49</sup>

#### 10.6.6 TESTING IN THE PRESENCE OF WEAK INSTRUMENTS

In Section 8.7, we introduced the problems of estimation and inference with instrumental variables in the presence of **weak instruments**. The first-stage regression method of Staiger and Stock (1997) is often used to detect the condition. Other tests have also been proposed, notably that of Hahn and Hausman (2002, 2003). Consider an equation with a single endogenous variable on the right-hand side,

$$y_1 = \gamma y_2 + \mathbf{x}'_1 \boldsymbol{\beta}_1 + \varepsilon_1.$$

Given the way the model has been developed, the placement of  $y_1$  on the left-hand side of this equation and  $y_2$  on the right represents nothing more than a normalization of the coefficient matrix  $\Gamma$  in (10-42). For the moment, label this the "forward" equation. If we renormalize the model in terms of  $y_2$ , we obtain the completely equivalent equation

$$\begin{aligned} y_2 &= (1/\gamma)y_1 + \mathbf{x}'_1(\boldsymbol{\beta}_1/\gamma) + \varepsilon_1/\gamma \\ &= \theta y_1 + \mathbf{x}'_1 \lambda_1 + v_1, \end{aligned}$$

which we [i.e., Hahn and Hausman (2002)] label the "reverse equation." In principle, for estimation of  $\gamma$ , it should make no difference which form we estimate; we can estimate  $\gamma$  directly in the first equation or indirectly through  $1/\theta$  in the second. However, in practice, of all the  $k$ -class estimators listed in Section 10.6.4 which includes all the estimators we have examined, only the LIML estimator is invariant to this renormalization; certainly the 2SLS estimator is not. If we consider the forward 2SLS estimator,  $\hat{\gamma}$ , and the reverse estimator,  $1/\hat{\theta}$ , we should in principle obtain similar estimates. But there is a bias in the 2SLS estimator that becomes more pronounced as the instruments become weaker. The Hahn and Hausman test statistic is based on the difference between these two estimators (corrected for the known bias of the 2SLS estimator in this case). [Research on this and other tests is ongoing. Hausman, Stock, and Yogo (2005) do report rather disappointing results for the power of this test in the presence of irrelevant instruments.]

The problem of inference remains. The upshot of the development so far is that the usual test statistics are likely to be unreliable. Some useful results have been obtained for devising inference procedures that are more robust than the standard first-order asymptotics that we have employed (for example, in Theorem 8.1 and Section 10.6). Kleibergen (2002) has constructed a class of test statistics based on Anderson and Rubin's (1949, 1950) results that appears to offer some progress. An intriguing aspect of this strand of research is that the Anderson and Rubin test was developed in their 1949 and 1950 studies and predates by several years the development of two-stage least squares by Theil (1953) and Basmann (1957). [See Stock and Trebbi (2003) for discussion of the early development of the method of instrumental variables.] A lengthy description

<sup>49</sup>See Cragg (1967) and the many related studies listed by Judge et al. (1985, pp. 646–653).

## CHAPTER 10 ♦ Systems of Equations 335

of Kleibergen's method and several extensions appears in the survey by Dufour (2003), which we draw on here for a cursory look at the Anderson and Rubin statistic.

The simultaneous equations model in terms of equation 1 is written

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{Y}_1 \boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}_1, \\ \mathbf{Y}_1 &= \mathbf{X}_1 \boldsymbol{\Pi}_1 + \mathbf{X}_1^* \boldsymbol{\Pi}_1^* + \mathbf{V}_1, \end{aligned} \quad (10-64)$$

where  $\mathbf{y}_1$  is the  $n$  observations on the left-hand variable in the equation of interest,  $\mathbf{Y}_1$  is the  $n$  observations on  $M_1$  endogenous variables in this equation,  $\boldsymbol{\gamma}_1$  is the structural parameter vector in this equation, and  $\mathbf{X}_1$  is the  $K_1$  included exogenous variables in equation 1. The second equation is the set of  $M_1$  reduced form equations for the included endogenous variables that appear in equation 1. (Note that  $M_1^*$  endogenous variables,  $\mathbf{Y}_1^*$ , are excluded from equation 1.) The full set of exogenous variables in the model is

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_1^*],$$

where  $\mathbf{X}_1^*$  is the  $K_1^*$  exogenous variables that are excluded from equation 1. (We are changing Dufour's notation slightly to conform to the conventions used in our development of the model.) Note that the second equation represents the first stage of the two-stage least squares procedure.

We are interested in inference about  $\boldsymbol{\gamma}_1$ . We must first assume that the model is identified. We will invoke the rank and order conditions as usual. The order condition is that there must be at least as many excluded exogenous variables as there are included endogenous variables, which is that  $K_1^* \geq M_1$ . For the rank condition to be met, we must have

$$\boldsymbol{\pi}_1^* - \boldsymbol{\Pi}_1^* \boldsymbol{\gamma}_1 = \mathbf{0},$$

where  $\boldsymbol{\pi}_1^*$  is the second part of the coefficient vector in the reduced form equation for  $\mathbf{y}_1$ , that is,

$$\mathbf{y}_1 = \mathbf{X}_1 \boldsymbol{\pi}_1 + \mathbf{X}_1^* \boldsymbol{\pi}_1^* + \mathbf{v}_1.$$

For this result to hold,  $\boldsymbol{\Pi}_1^*$  must have full column rank,  $K_1^*$ . The weak instruments problem is embodied in  $\boldsymbol{\Pi}_1^*$ . If this matrix has short rank, the parameter vector  $\boldsymbol{\gamma}_1$  is not identified. The weak instruments problem arises when  $\boldsymbol{\Pi}_1^*$  is nearly short ranked. The important aspect of that observation is that the weak instruments can be characterized as an identification problem.

Anderson and Rubin (1949, 1950) (AR) proposed a method of testing  $H_0: \boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_1^0$ . The AR statistic is constructed as follows: Combining the two equations in (10-64), we have

$$\mathbf{y}_1 = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_1 \boldsymbol{\Pi}_1 \boldsymbol{\gamma}_1 + \mathbf{X}_1^* \boldsymbol{\Pi}_1^* \boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}_1 + \mathbf{V}_1 \boldsymbol{\gamma}_1.$$

Using (10-64) again, subtract  $\mathbf{Y}_1 \boldsymbol{\gamma}_1^0$  from both sides of this equation to obtain

$$\begin{aligned} \mathbf{y}_1 - \mathbf{Y}_1 \boldsymbol{\gamma}_1^0 &= \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_1 \boldsymbol{\Pi}_1 \boldsymbol{\gamma}_1 + \mathbf{X}_1^* \boldsymbol{\Pi}_1^* \boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}_1 + \mathbf{V}_1 \boldsymbol{\gamma}_1 \\ &\quad - \mathbf{X}_1 \boldsymbol{\Pi}_1 \boldsymbol{\gamma}_1^0 - \mathbf{X}_1^* \boldsymbol{\Pi}_1^* \boldsymbol{\gamma}_1^0 - \mathbf{V}_1 \boldsymbol{\gamma}_1^0 \\ &= \mathbf{X}_1 [\boldsymbol{\beta}_1 + \boldsymbol{\Pi}_1 (\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_1^0)] + \mathbf{X}_1^* [\boldsymbol{\Pi}_1^* (\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_1^0)] + \boldsymbol{\varepsilon}_1 + \mathbf{V}_1 (\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_1^0) \\ &= \mathbf{X}_1 \boldsymbol{\theta}_1 + \mathbf{X}_1^* \boldsymbol{\theta}_1^* + \mathbf{w}_1. \end{aligned}$$

### 336 PART II ♦ Generalized Regression Model and Equation Systems

Under the null hypothesis, this equation reduces to

$$\mathbf{y}_1 - \mathbf{Y}_1 \boldsymbol{\gamma}_1^0 = \mathbf{X}_1 \boldsymbol{\theta}_1 + \mathbf{w}_1,$$

so a test of the null hypothesis can be carried out by testing the hypothesis that  $\boldsymbol{\theta}_1^*$  equals zero in the preceding partial reduced-form equation. Anderson and Rubin proposed a simple  $F$  test,

$$\begin{aligned} AR(\boldsymbol{\gamma}_1^0) &= \frac{[(\mathbf{y}_1 - \mathbf{Y}_1 \boldsymbol{\gamma}_1^0)' \mathbf{M}_1 (\mathbf{y}_1 - \mathbf{Y}_1 \boldsymbol{\gamma}_1^0) - (\mathbf{y}_1 - \mathbf{Y}_1 \boldsymbol{\gamma}_1^0)' \mathbf{M} (\mathbf{y}_1 - \mathbf{Y}_1 \boldsymbol{\gamma}_1^0)] / K_1^*}{(\mathbf{y}_1 - \mathbf{Y}_1 \boldsymbol{\gamma}_1^0)' \mathbf{M} (\mathbf{y}_1 - \mathbf{Y}_1 \boldsymbol{\gamma}_1^0) / (n - K)} \\ &\sim F[K_1^*, n - K], \end{aligned}$$

where  $\mathbf{M}_1 = [\mathbf{I} - \mathbf{X}_1(\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1]$  and  $\mathbf{M} = [\mathbf{I} - \mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}']$ . This is the standard  $F$  statistic for testing the hypothesis that the set of coefficients is zero in the classical linear regression. [See (5-29).] [Dufour (2003) shows how the statistic can be extended to allow more general restrictions that also include  $\boldsymbol{\beta}_1$ .]

There are several striking features of this approach, beyond the fact that it has been available since 1949: (1) its distribution is free of the model parameters in finite samples (assuming normality of the disturbances); (2) *it is robust to the weak instruments problem*; (3) it is robust to the exclusion of other instruments; and (4) it is robust to specification errors in the structural equations for  $\mathbf{Y}_1$ , the other variables in the equation. There are some shortcomings as well, namely: (1) the tests developed by this method are only applied to the full parameter vector; (2) the power of the test may diminish as more (and too many more) instrumental variables are added; (3) it relies on a normality assumption for the disturbances; and (4) there does not appear to be a counterpart for nonlinear systems of equations.

## 10.7 SUMMARY AND CONCLUSIONS

This chapter has surveyed the specification and estimation of multiple equations models. The SUR model is an application of the generalized regression model introduced in Chapter 9. The advantage of the SUR formulation is the rich variety of behavioral models that fit into this framework. We began with estimation and inference with the SUR model, treating it essentially as a generalized regression. The major difference between this set of results and the single equation model in Chapter 9 is practical. While the SUR model is, in principle a single equation GR model with an elaborate covariance structure, special problems arise when we explicitly recognize its intrinsic nature as a set of equations linked by their disturbances. The major result for estimation at this step is the feasible GLS estimator. In spite of its apparent complexity, we can estimate the SUR model by a straightforward two-step GLS approach that is similar to the one we used for models with heteroscedasticity in Chapter 9. We also extended the SUR model to autocorrelation and heteroscedasticity. Once again, the multiple equation nature of the model complicates these applications. Section 10.4 presented a common application of the seemingly unrelated regressions model, the estimation of demand systems. One of the signature features of this literature is the seamless transition from the theoretical models of optimization of consumers and producers to the sets of

## CHAPTER 10 ♦ Systems of Equations 337

empirical demand equations derived from Roy's identity for consumers and Shephard's lemma for producers.

The multiple equations models surveyed in this chapter involve most of the issues that arise in analysis of linear equations in econometrics. Before one embarks on the process of estimation, it is necessary to establish that the sample data actually contain sufficient information to provide estimates of the parameters in question. This is the question of identification. Identification involves both the statistical properties of estimators and the role of theory in the specification of the model. Once identification is established, there are numerous methods of estimation. We considered a number of single-equation techniques, including least squares, instrumental variables, and maximum likelihood. Fully efficient use of the sample data will require joint estimation of all the equations in the system. Once again, there are several techniques—these are extensions of the single-equation methods including three-stage least squares, and full information maximum likelihood. In both frameworks, this is one of those benign situations in which the computationally simplest estimator is generally the most efficient one.

### Key Terms and Concepts

- Admissible
- Autocorrelation
- Balanced panel
- Behavioral equation
- Causality
- Cobb-Douglas model
- Complete system of equations
- Completeness condition
- Consistent estimators
- Constant returns to scale
- Covariance structures model
- Demand system
- Dynamic model
- Econometric model
- Endogenous
- Equilibrium condition
- Equilibrium multipliers
- Exactly identified model
- Exclusion restrictions
- Exogenous
- Feasible GLS
- FIML
- Fixed effects
- Flexible functional form
- Flexible functions
- Full information estimator
- Full information maximum likelihood
- Fully recursive model
- Generalized regression model
- Granger causality
- Heteroscedasticity
- Homogeneity restriction
- Identical explanatory variables
- Identical regressors
- Identification
- Instrumental variable estimator
- Interdependent
- Invariance
- Invariant
- Jointly dependent  $K$  class
- Kronecker product
- Lagrange multiplier test
- Least variance ratio
- Likelihood ratio test
- Limited information estimator
- Limited information maximum likelihood (LIML) estimator
- Maximum likelihood
- Multivariate regression model
- Nonlinear systems
- Nonsample information
- Nonstructural normalization
- Observationally equivalent
- Order condition
- Overidentification
- Pooled model
- Predetermined variable
- Problem of identification
- Projection
- Random effects model
- Rank condition
- Recursive model
- Reduced form
- Reduced form disturbance
- Restrictions
- Seemingly unrelated regressions
- Share equations
- Shephard's lemma
- Simultaneous equations models
- Singularity of the disturbance covariance matrix
- Simultaneous equations bias
- Specification test
- Strongly exogenous
- Structural disturbance
- Structural equation
- Structural form
- System methods of estimation

**338 PART II ♦ Generalized Regression Model and Equation Systems**

- Systems of demand equations
- Taylor series
- Three-stage least squares (3SLS) estimator
- Translog function
- Triangular system
- Two-stage least squares (2SLS) estimator
- Underidentified
- Weak instruments
- Weakly exogenous

**Exercises**

1. A sample of 100 observations produces the following sample data:

$$\begin{aligned}\bar{y}_1 &= 1, \quad \bar{y}_2 = 2, \\ \mathbf{y}'_1 \mathbf{y}_1 &= 150, \\ \mathbf{y}'_2 \mathbf{y}_2 &= 550, \\ \mathbf{y}'_1 \mathbf{y}_2 &= 260.\end{aligned}$$

The underlying bivariate regression model is

$$\begin{aligned}y_1 &= \mu + \varepsilon_1, \\ y_2 &= \mu + \varepsilon_2.\end{aligned}$$

- a. Compute the OLS estimate of  $\mu$ , and estimate the sampling variance of this estimator.  
 b. Compute the FGLS estimate of  $\mu$  and the sampling variance of the estimator.  
 2. Consider estimation of the following two-equation model:

$$\begin{aligned}y_1 &= \beta_1 + \varepsilon_1, \\ y_2 &= \beta_2 x + \varepsilon_2.\end{aligned}$$

A sample of 50 observations produces the following moment matrix:

$$\begin{matrix} & 1 & y_1 & y_2 & x \\ 1 & \left[ \begin{matrix} 50 \\ 150 & 500 \\ y_2 & 50 & 40 & 90 \\ x & 100 & 60 & 50 & 100 \end{matrix} \right] \end{matrix}.$$

- a. Write the explicit formula for the GLS estimator of  $[\beta_1, \beta_2]$ . What is the asymptotic covariance matrix of the estimator?  
 b. Derive the OLS estimator and its sampling variance in this model.  
 c. Obtain the OLS estimates of  $\beta_1$  and  $\beta_2$ , and estimate the sampling covariance matrix of the two estimates. Use  $n$  instead of  $(n - 1)$  as the divisor to compute the estimates of the disturbance variances.  
 d. Compute the FGLS estimates of  $\beta_1$  and  $\beta_2$  and the estimated sampling covariance matrix.  
 e. Test the hypothesis that  $\beta_2 = 1$ .
3. The model

$$\begin{aligned}y_1 &= \beta_1 x_1 + \varepsilon_1, \\ y_2 &= \beta_2 x_2 + \varepsilon_2\end{aligned}$$

CHAPTER 10 ♦ Systems of Equations **339**

satisfies all the assumptions of the classical multivariate regression model. All variables have zero means. The following sample second-moment matrix is obtained from a sample of 20 observations:

$$\begin{array}{cccc} y_1 & y_2 & x_1 & x_2 \\ y_1 & \left[ \begin{matrix} 20 & 6 & 4 & 3 \\ 6 & 10 & 3 & 6 \\ 4 & 3 & 5 & 2 \\ 3 & 6 & 2 & 10 \end{matrix} \right] \\ y_2 & \\ x_1 & \\ x_2 & \end{array}$$

- a. Compute the FGLS estimates of  $\beta_1$  and  $\beta_2$ .
  - b. Test the hypothesis that  $\beta_1 = \beta_2$ .
  - c. Compute the maximum likelihood estimates of the model parameters.
  - d. Use the likelihood ratio test to test the hypothesis in part b.
4. Prove that in the model

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1, \\ \mathbf{y}_2 &= \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2, \end{aligned}$$

generalized least squares is equivalent to equation-by-equation ordinary least squares if  $\mathbf{X}_1 = \mathbf{X}_2$ . Does your result hold if it is also known that  $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ ?

5. Consider the two-equation system

$$\begin{aligned} y_1 &= \beta_1 x_1 + \varepsilon_1, \\ y_2 &= \beta_2 x_2 + \beta_3 x_3 + \varepsilon_2. \end{aligned}$$

Assume that the disturbance variances and covariance are known. Now suppose that the analyst of this model applies GLS but erroneously omits  $x_3$  from the second equation. What effect does this specification error have on the consistency of the estimator of  $\beta_1$ ?

6. Consider the system

$$\begin{aligned} y_1 &= \alpha_1 + \beta x + \varepsilon_1, \\ y_2 &= \alpha_2 + \varepsilon_2. \end{aligned}$$

The disturbances are freely correlated. Prove that GLS applied to the system leads to the OLS estimates of  $\alpha_1$  and  $\alpha_2$  but to a mixture of the least squares slopes in the regressions of  $y_1$  and  $y_2$  on  $x$  as the estimator of  $\beta$ . What is the mixture? To simplify the algebra, assume (with no loss of generality) that  $\bar{x} = 0$ .

7. For the model

$$\begin{aligned} y_1 &= \alpha_1 + \beta x + \varepsilon_1, \\ y_2 &= \alpha_2 + \varepsilon_2, \\ y_3 &= \alpha_3 + \varepsilon_3, \end{aligned}$$

assume that  $y_{i2} + y_{i3} = 1$  at every observation. Prove that the sample covariance matrix of the least squares residuals from the three equations will be singular, thereby precluding computation of the FGLS estimator. How could you proceed in this case?

**340 PART II ♦ Generalized Regression Model and Equation Systems**

8. Consider the following two-equation model:

$$\begin{aligned}y_1 &= \gamma_1 y_2 + \beta_{11} x_1 + \beta_{21} x_2 + \beta_{31} x_3 + \varepsilon_1, \\y_2 &= \gamma_2 y_1 + \beta_{12} x_1 + \beta_{22} x_2 + \beta_{32} x_3 + \varepsilon_2.\end{aligned}$$

- a. Verify that, as stated, neither equation is identified.
- b. Establish whether or not the following restrictions are sufficient to identify (or partially identify) the model:
  - (1)  $\beta_{21} = \beta_{32} = 0$ ,
  - (2)  $\beta_{12} = \beta_{22} = 0$ ,
  - (3)  $\gamma_1 = 0$ ,
  - (4)  $\gamma_1 = \gamma_2$  and  $\beta_{32} = 0$ ,
  - (5)  $\sigma_{12} = 0$  and  $\beta_{31} = 0$ ,
  - (6)  $\gamma_1 = 0$  and  $\sigma_{12} = 0$ ,
  - (7)  $\beta_{21} + \beta_{22} = 1$ ,
  - (8)  $\sigma_{12} = 0$ ,  $\beta_{21} = \beta_{22} = \beta_{31} = \beta_{32} = 0$ ,
  - (9)  $\sigma_{12} = 0$ ,  $\beta_{11} = \beta_{21} = \beta_{22} = \beta_{31} = \beta_{32} = 0$ .
- 9. Obtain the reduced form for the model in Exercise 8 under each of the assumptions made in parts a and in parts b1 and b9.
- 10. The following model is specified:

$$\begin{aligned}y_1 &= \gamma_1 y_2 + \beta_{11} x_1 + \varepsilon_1, \\y_2 &= \gamma_2 y_1 + \beta_{22} x_2 + \beta_{32} x_3 + \varepsilon_2.\end{aligned}$$

All variables are measured as deviations from their means. The sample of 25 observations produces the following matrix of sums of squares and cross products:

$$\begin{array}{ccccc} & y_1 & y_2 & x_1 & x_2 & x_3 \\ y_1 & \left[ \begin{array}{ccccc} 20 & 6 & 4 & 3 & 5 \end{array} \right] \\ y_2 & \left[ \begin{array}{ccccc} 6 & 10 & 3 & 6 & 7 \end{array} \right] \\ x_1 & \left[ \begin{array}{ccccc} 4 & 3 & 5 & 2 & 3 \end{array} \right] \\ x_2 & \left[ \begin{array}{ccccc} 3 & 6 & 2 & 10 & 8 \end{array} \right] \\ x_3 & \left[ \begin{array}{ccccc} 5 & 7 & 3 & 8 & 15 \end{array} \right] \end{array}.$$

- a. Estimate the two equations by OLS.
- b. Estimate the parameters of the two equations by 2SLS. Also estimate the asymptotic covariance matrix of the 2SLS estimates.
- c. Obtain the LIML estimates of the parameters of the first equation.
- d. Estimate the two equations by 3SLS.
- e. Estimate the reduced form coefficient matrix by OLS and indirectly by using your structural estimates from part b.
- 11. For the model

$$\begin{aligned}y_1 &= \gamma_1 y_2 + \beta_{11} x_1 + \beta_{21} x_2 + \varepsilon_1, \\y_2 &= \gamma_2 y_1 + \beta_{32} x_3 + \beta_{42} x_4 + \varepsilon_2,\end{aligned}$$

CHAPTER 10 ♦ Systems of Equations **341**

show that there are two restrictions on the reduced form coefficients. Describe a procedure for estimating the model while incorporating the restrictions.

12. Prove that

$$\text{plim } \frac{\mathbf{Y}'_j \boldsymbol{\varepsilon}_j}{T} = \boldsymbol{\omega}_j - \boldsymbol{\Omega}_{jj} \boldsymbol{\gamma}_j.$$

13. Prove that an underidentified equation cannot be estimated by 2SLS.

## Applications

1. Continuing the analysis of Section 10.5.2, we find that a translog cost function for one output and three factor inputs that does not impose constant returns to scale is

$$\begin{aligned} \ln C = & \alpha + \beta_1 \ln p_1 + \beta_2 \ln p_2 + \beta_3 \ln p_3 + \delta_{11} \frac{1}{2} \ln^2 p_1 + \delta_{12} \ln p_1 \ln p_2 \\ & + \delta_{13} \ln p_1 \ln p_3 + \delta_{22} \frac{1}{2} \ln^2 p_2 + \delta_{23} \ln p_2 \ln p_3 + \delta_{33} \frac{1}{2} \ln^2 p_3 \\ & + \gamma_{q1} \ln Q \ln p_1 + \gamma_{q2} \ln Q \ln p_2 + \gamma_{q3} \ln Q \ln p_3 \\ & + \beta_q \ln Q + \beta_{qq} \frac{1}{2} \ln^2 Q + \varepsilon_c. \end{aligned}$$

The factor share equations are

$$\begin{aligned} S_1 &= \beta_1 + \delta_{11} \ln p_1 + \delta_{12} \ln p_2 + \delta_{13} \ln p_3 + \gamma_{q1} \ln Q + \varepsilon_1, \\ S_2 &= \beta_2 + \delta_{12} \ln p_1 + \delta_{22} \ln p_2 + \delta_{23} \ln p_3 + \gamma_{q2} \ln Q + \varepsilon_2, \\ S_3 &= \beta_3 + \delta_{13} \ln p_1 + \delta_{23} \ln p_2 + \delta_{33} \ln p_3 + \gamma_{q3} \ln Q + \varepsilon_3. \end{aligned}$$

[See Christensen and Greene (1976) for analysis of this model.]

- a. The three factor shares must add identically to 1. What restrictions does this requirement place on the model parameters?
- b. Show that the adding-up condition in (10-38) can be imposed directly on the model by specifying the translog model in  $(C/p_3)$ ,  $(p_1/p_3)$ , and  $(p_2/p_3)$  and dropping the third share equation. (See Example 10.3.) Notice that this reduces the number of free parameters in the model to 10.
- c. Continuing part b, the model as specified with the symmetry and equality restrictions has 15 parameters. By imposing the constraints, you reduce this number to 10 in the estimating equations. How would you obtain estimates of the parameters not estimated directly?

The remaining parts of this exercise will require specialized software. The **E-Views**, **TSP**, **Stata** or **LIMDEP**, programs noted in the preface are four that could be used.

All estimation is to be done using the data used in Section 10.5.1

- d. Estimate each of the three equations you obtained in part b by ordinary least squares. Do the estimates appear to satisfy the cross-equation equality and symmetry restrictions implied by the theory?
- e. Using the data in Section 10.5.1, estimate the full system of three equations (cost and the two independent shares), imposing the symmetry and cross-equation equality constraints.

### 342 PART II ♦ Generalized Regression Model and Equation Systems

- f. Using your parameter estimates, compute the estimates of the elasticities in (10-39) at the means of the variables.
- g. Use a likelihood ratio statistic to test the joint hypothesis that  $\gamma_{qi} = 0, i = 1, 2, 3$ .  
[Hint: Just drop the relevant variables from the model.]
- 2. The Grunfeld investment data in Appendix Table 10.5 are a classic data set that have been used for decades to develop and demonstrate estimators for seemingly unrelated regressions.<sup>50</sup> Although somewhat dated at this juncture, they remain an ideal application of the techniques presented in this chapter.<sup>51</sup> The data consist of time series of 20 yearly observations on 10 firms. The three variables are

$I_{it}$  = gross investment,

$F_{it}$  = market value of the firm at the end of the previous year,

$C_{it}$  = value of the stock of plant and equipment at the end of the previous year.

The main equation in the studies noted is

$$I_{it} = \beta_1 + \beta_2 F_{it} + \beta_3 C_{it} + \varepsilon_{it}.^{52}$$

- a. Fit the 10 equations separately by ordinary least squares and report your results.
- b. Use a Wald (Chow) test to test the “aggregation” restriction that the 10 coefficient vectors are the same.
- c. Use the seemingly unrelated regressions (FGLS) estimator to reestimate the parameters of the model, once again, allowing the coefficients to differ across the 10 equations. Now, use the pooled model and, again, FGLS to estimate the constrained equation with equal parameter vectors, and test the aggregation hypothesis.
- d. Using the OLS residuals from the separate regression, use the LM statistic in (10-17) to test for the presence of cross-equation correlation.
- e. An alternative specification to the model in part c that focuses on the variances rather than the means is a groupwise heteroscedasticity model. For the current application, you can fit this model using (10-19), (10-20), and (10-21) while imposing the much simpler model with  $\sigma_{ij} = 0$  when  $i \neq j$ . Do the results of the pooled model differ in the three cases considered, simple OLS, groupwise heteroscedasticity, and full unrestricted covariances [which would be (10-20)] with  $\Omega_{ij} = I$ ?
- 3. The data in Appendix Table F5.2 may be used to estimate a small macroeconomic model. Use these data to estimate the model in Example 10.4. Estimate the parameters of the two equations by two-stage and three-stage least squares.

<sup>50</sup>See Grunfeld (1958), Grunfeld and Griliches (1960), and Boot and de Witt (1960).

<sup>51</sup>See, in particular, Zellner (1962, 1963) and Zellner and Huang (1962).

<sup>52</sup>Note that the model specifies investment, a flow, as a function of two stocks. This could be a theoretical misspecification. It might be preferable to specify the model in terms of planned investment. But, 50 years after the fact, we'll take the specified model as it is.

## 11

# MODELS FOR PANEL DATA

---

## 11.1 INTRODUCTION

Data sets that combine time series and cross sections are common in economics. The published statistics of the OECD contain numerous series of economic aggregates observed yearly for many countries. The Penn World Tables [CIC (2010)] is a data bank that contains national income data on 188 countries for over 50 years. Recently constructed **longitudinal data sets** contain observations on thousands of individuals or families, each observed at several points in time. Other empirical studies have examined time-series data on sets of firms, states, countries, or industries simultaneously. These data sets provide rich sources of information about the economy. The analysis of panel data allows the model builder to learn about economic processes while accounting for both heterogeneity across individuals, firms, countries, and so on and for dynamic effects that are not visible in cross sections. Modeling in this context often calls for complex stochastic specifications. In this chapter, we will survey the most commonly used techniques for time-series—cross section (e.g., cross country) and panel (e.g., longitudinal) data. The methods considered here provide extensions to most of the models we have examined in the preceding chapters. Section 11.2 describes the specific features of panel data. Most of this analysis is focused on individual data, rather than cross-country aggregates. We will examine some aspects of aggregate data modeling in Section 11.11. Sections 11.3, 11.4, and 11.5 consider in turn the three main approaches to regression analysis with panel data, pooled regression, the fixed effects model, and the random effects model. Section 11.6 considers robust estimation of covariance matrices for the panel data estimators, including a general treatment of “cluster” effects. Sections 11.7–11.11 examine some specific applications and extensions of panel data methods. Spatial autocorrelation is discussed in Section 11.7. In Section 11.8, we consider sources of endogeneity in the random effects model, including a model of the sort considered in Chapter 8 with an endogenous right-hand-side variable and then two approaches to dynamic models. Section 11.9 builds the fixed and random effects models into nonlinear regression models. In Section 11.10, the random effects model is extended to the multiple equation systems developed in Chapter 10. Finally, Section 11.11 examines random parameter models. The random parameters approach is an extension of the fixed and random effects model in which the heterogeneity that the FE and RE models build into the constant terms is extended to other parameters as well.

Panel data methods are used throughout the remainder of this book. We will develop several extensions of the fixed and random effects models in Chapter 14 on maximum likelihood methods, and in Chapter 15 where we will continue the development of random parameter models that is begun in Section 11.11. Chapter 14 will also present methods for handling discrete distributions of random parameters under the heading of latent class models. In Chapter 23, we will return to the models of nonstationary panel

**344 PART II ♦ Generalized Regression Model and Equation Systems**

data that are suggested in Section 11.8.4. The fixed and random effects approaches will be used throughout the applications of discrete and limited dependent variables models in microeconomics in Chapters 17, 18, and 19.

## 11.2 PANEL DATA MODELS

Many recent studies have analyzed **panel**, or longitudinal, data sets. Two very famous ones are the National Longitudinal Survey of Labor Market Experience (NLS, <http://www.bls.gov/nls/nlsdoc.htm>) and the Michigan Panel Study of Income Dynamics (PSID, <http://psidonline.isr.umich.edu/>). In these data sets, very large cross sections, consisting of thousands of microunits, are followed through time, but the number of periods is often quite small. The PSID, for example, is a study of roughly 6,000 families and 15,000 individuals who have been interviewed periodically from 1968 to the present. An ongoing study in the United Kingdom is the British Household Panel Survey (BHPS, <http://www.iser.essex.ac.uk/ulsc/bhps/>) which was begun in 1991 and is now in its 18th wave. The survey follows several thousand households (currently over 5,000) for several years. Many very rich data sets have recently been developed in the area of health care and health economics, including the German Socioeconomic Panel (GSOEP, [http://dpls.dacc.wisc.edu/apdu/GSOEP/gsoep\\_cd\\_data.html](http://dpls.dacc.wisc.edu/apdu/GSOEP/gsoep_cd_data.html)) and the Medical Expenditure Panel Survey (MEPS, <http://www.meps.ahrq.gov/>). Constructing long, evenly spaced time series in contexts such as these would be prohibitively expensive, but for the purposes for which these data are typically used, it is unnecessary. Time effects are often viewed as “transitions” or discrete changes of state. The Current Population Survey (CPS, <http://www.census.gov/cps/>), for example, is a monthly survey of about 50,000 households that interviews households monthly for four months, waits for eight months, then reinterviews. This two-wave, **rotating panel** format allows analysis of short-term changes as well as a more general analysis of the U.S. national labor market. They are typically modeled as specific to the period in which they occur and are not carried across periods within a cross-sectional unit.<sup>1</sup> Panel data sets are more oriented toward cross-section analyses; they are wide but typically short. Heterogeneity across units is an integral part—indeed, often the central focus—of the analysis.

The analysis of panel or longitudinal data is the subject of one of the most active and innovative bodies of literature in econometrics,<sup>2</sup> partly because panel data provide such a rich environment for the development of estimation techniques and theoretical results. In more practical terms, however, researchers have been able to use time-series cross-sectional data to examine issues that could not be studied in either cross-sectional or time-series settings alone. Two examples are as follows.

1. In a widely cited study of labor supply, Ben-Porath (1973) observes that at a certain point in time, in a cohort of women, 50 percent may appear to be working. It is

<sup>1</sup> Formal time-series modeling for panel data is briefly examined in Section 23.5.

<sup>2</sup> The panel data literature rivals the received research on unit roots and cointegration in econometrics in its rate of growth. A compendium of the earliest literature is Maddala (1993). Book-length surveys on the econometrics of panel data include Hsiao (2003), Dielman (1989), M<sup>2</sup> and Sevestre (1996), Raj and Baltagi (1992), Nerlove (2002), Arellano (2003), and Baltagi (2001, 2005). There are also lengthy surveys devoted to specific topics, such as limited dependent variable models [Hsiao, Lahiri, Lee, and Bararan (1999)] and semiparametric methods [Lee (1998)]. An extensive bibliography is given in Baltagi (2005).

CHAPTER 11 ♦ Models for Panel Data **345**

ambiguous whether this finding implies that, in this cohort, one-half of the women on average will be working or that the same one-half will be working in every period. These have very different implications for policy and for the interpretation of any statistical results. Cross-sectional data alone will not shed any light on the question.

2. A long-standing problem in the analysis of production functions has been the inability to separate economies of scale and technological change.<sup>3</sup> Cross-sectional data provide information only about the former, whereas time-series data muddle the two effects, with no prospect of separation. It is common, for example, to assume constant returns to scale so as to reveal the technical change.<sup>4</sup> Of course, this practice assumes away the problem. A panel of data on costs or output for a number of firms each observed over several years can provide estimates of both the rate of technological change (as time progresses) and economies of scale (for the sample of different sized firms at each point in time).

Recent applications have allowed researchers to study the impact of health policy changes [e.g., Riphahn et al.'s (2003) analysis of reforms in German public health insurance regulations] and more generally the dynamics of labor market behavior. In principle, the methods of Chapters 6 and 21 can be applied to longitudinal data sets. In the typical panel, however, there are a large number of cross-sectional units and only a few periods. Thus, the time-series methods discussed there may be somewhat problematic. Recent work has generally concentrated on models better suited to these short and wide data sets. The techniques are focused on cross-sectional variation, or heterogeneity. In this chapter, we shall examine in detail the most widely used models and look briefly at some extensions.

### **11.2.1 GENERAL MODELING FRAMEWORK FOR ANALYZING PANEL DATA**

The fundamental advantage of a panel data set over a cross section is that it will allow the researcher great flexibility in modeling differences in behavior across individuals. The basic framework for this discussion is a regression model of the form

$$\begin{aligned} y_{it} &= \mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \boldsymbol{\alpha} + \varepsilon_{it} \\ &= \mathbf{x}'_i \boldsymbol{\beta} + c_i + \varepsilon_{it}. \end{aligned} \tag{11-1}$$

There are  $K$  regressors in  $\mathbf{x}_{it}$ , *not including a constant term*. The **heterogeneity**, or **individual effect** is  $\mathbf{z}'_i \boldsymbol{\alpha}$  where  $\mathbf{z}_i$  contains a constant term and a set of individual or group-specific variables, which may be observed, such as race, sex, location, and so on, or unobserved, such as family specific characteristics, individual heterogeneity in skill or

---

<sup>3</sup>The distinction between these two effects figured prominently in the policy question of whether it was appropriate to break up the AT&T Corporation in the 1980s and, ultimately, to allow competition in the provision of long-distance telephone service.

<sup>4</sup>In a classic study of this issue, Solow (1957) states: "From time series of  $\Delta Q/Q$ ,  $w_K$ ,  $\Delta K/K$ ,  $w_L$  and  $\Delta L/L$  or their discrete year-to-year analogues, we could estimate  $\Delta A/A$  and thence  $A(t)$  itself. Actually an amusing thing happens here. Nothing has been said so far about returns to scale. But if all factor inputs are classified either as  $K$  or  $L$ , then the available figures always show  $w_K$  and  $w_L$  adding up to one. Since we have assumed that factors are paid their marginal products, this amounts to assuming the hypothesis of Euler's theorem. The calculus being what it is, we might just as well assume the conclusion, namely, the  $F$  is homogeneous of degree one."

### 346 PART II ♦ Generalized Regression Model and Equation Systems

preferences, and so on, all of which are taken to be constant over time  $t$ . As it stands, this model is a classical regression model. If  $\mathbf{z}_i$  is observed for all individuals, then the entire model can be treated as an ordinary linear model and fit by least squares. The complications arise when  $c_i$  is unobserved, which will be the case in most applications. Consider, for example, analyses of the effect of education and experience on earnings from which “ability” will always be a missing and unobservable variable. In health care studies, for example, of usage of the health care system, “health” and “health care” will be unobservable factors in the analysis.

The main objective of the analysis will be consistent and efficient estimation of the **partial effects**,

$$\boldsymbol{\beta} = \partial E[y_{it} | \mathbf{x}_{it}] / \partial \mathbf{x}_{it}.$$

Whether this is possible depends on the assumptions about the unobserved effects. We begin with a **strict exogeneity** assumption for the independent variables,

$$E[\varepsilon_{it} | \mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots] = 0.$$

That is, the current disturbance is uncorrelated with the independent variables in every period, past, present, and future. The crucial aspect of the model concerns the heterogeneity. A particularly convenient assumption would be **mean independence**,

$$E[c_i | \mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots] = \alpha.$$

If the missing variable(s) are uncorrelated with the included variables, then, as we shall see, they may be included in the disturbance of the model. This is the assumption that underlies the random effects model, as we will explore later. It is, however, a particularly strong assumption—it would be unlikely in the labor market and health care examples mentioned previously. The alternative would be

$$\begin{aligned} E[c_i | \mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots] &= h(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots) \\ &= h(\mathbf{X}_i). \end{aligned}$$

This formulation is more general, but at the same time, considerably more complicated, the more so since it may require yet further assumptions about the nature of the function.

#### 11.2.2 MODEL STRUCTURES

We will examine a variety of different models for panel data. Broadly, they can be arranged as follows:

**1. Pooled Regression:** If  $\mathbf{z}_i$  contains only a constant term, then ordinary least squares provides consistent and efficient estimates of the common  $\alpha$  and the slope vector  $\boldsymbol{\beta}$ .

**2. Fixed Effects:** If  unobserved, but correlated with  $\mathbf{x}_{it}$ , then the least squares estimator of  $\boldsymbol{\beta}$  is biased and inconsistent as a consequence of an omitted variable. However, in this instance, the model

$$y_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + \alpha_i + \varepsilon_{it},$$

where  $\alpha_i = \mathbf{z}'_i \boldsymbol{\alpha}$ , embodies all the observable effects and specifies an estimable conditional mean. This **fixed effects** approach takes  $\alpha_i$  to be a group-specific constant term in the regression model. It should be noted that the term “fixed” as used here signifies the correlation of  $c_i$  and  $\mathbf{x}_{it}$ , not that  $c_i$  is nonstochastic.

**3. Random Effects:** If the unobserved individual heterogeneity, however formulated, can be assumed to be uncorrelated with the included variables, then the model may be formulated as

$$\begin{aligned} y_{it} &= \mathbf{x}'_{it}\beta + E[\mathbf{z}'_i\alpha] + \{\mathbf{z}'_i\alpha - E[\mathbf{z}'_i\alpha]\} + \varepsilon_{it} \\ &= \mathbf{x}'_{it}\beta + \alpha + u_i + \varepsilon_{it}, \end{aligned}$$

that is, as a linear regression model with a compound disturbance that may be consistently, albeit inefficiently, estimated by least squares. This **random effects** approach specifies that  $u_i$  is a group-specific random element, similar to  $\varepsilon_{it}$  except that for each group, there is but a single draw that enters the regression identically in each period. Again, the crucial distinction between fixed and random effects is whether the unobserved individual effect embodies elements that are correlated with the regressors in the model, not whether these effects are stochastic or not. We will examine this basic formulation, then consider an extension to a dynamic model.

**4. Random Parameters:** The random effects model can be viewed as a regression model with a random constant term. With a sufficiently rich data set, we may extend this idea to a model in which the other coefficients vary randomly across individuals as well. The extension of the model might appear as

$$y_{it} = \mathbf{x}'_{it}(\beta + \mathbf{h}_i) + (\alpha + u_i) + \varepsilon_{it},$$

where  $\mathbf{h}_i$  is a random vector that induces the variation of the parameters across individuals. This random parameters model was proposed quite early in this literature, but has only fairly recently enjoyed widespread attention in several fields. It represents a natural extension in which researchers broaden the amount of heterogeneity across individuals while retaining some commonalities—the parameter vectors still share a common mean. Some recent applications have extended this yet another step by allowing the mean value of the parameter distribution to be person specific, as in

$$y_{it} = \mathbf{x}'_{it}(\beta + \Delta \mathbf{z}_i + \mathbf{h}_i) + (\alpha + u_i) + \varepsilon_{it},$$

where  $\mathbf{z}_i$  is a set of observable, person specific variables, and  $\Delta$  is a matrix of parameters to be estimated. As we will examine in Chapter 17, this **hierarchical model** is extremely versatile.

### 11.2.3 EXTENSIONS

The short list of model types provided earlier only begins to suggest the variety of applications of panel data methods in econometrics. We will begin in this chapter to study some of the formulations and uses of linear models. The random and fixed effects models and random parameters models have also been widely used in models of censoring, binary, and other discrete choices, and models for event counts. We will examine all of these in the chapters to follow. In some cases, such as the models for count data in Chapter 16, the extension of random and fixed effects models is straightforward, if somewhat more complicated computationally. In others, such as in binary choice models in Chapter 17 and censoring models in Chapter 18, these panel data models have been used, but not before overcoming some significant methodological and computational obstacles.

## 348 PART II ♦ Generalized Regression Model and Equation Systems

### 11.2.4 BALANCED AND UNBALANCED PANELS

By way of preface to the analysis to follow, we note an important aspect of panel data analysis. As suggested by the preceding discussion, a “panel” data set will consist of  $n$  sets of observations on individuals to be denoted  $i = 1, \dots, n$ . If each individual in the data set is observed the same number of times, usually denoted  $T$ , the data set is a **balanced panel**. An **unbalanced panel** data set is one in which individuals may be observed different numbers of times. We will denote this  $T_i$ . A **fixed panel** is one in which the same set of individuals is observed for the duration of the study. The data sets we will examine in this chapter, while not all balanced, are fixed. A rotating panel is one in which the cast of individuals changes from one period to the next. For example, Gonzalez and Maloney (1999) examined self-employment decisions in Mexico using the National Urban Employment Survey. This is a quarterly data set drawn from 1987 to 1993 in which individuals are interviewed five times. Each quarter, one-fifth of the individuals is rotated out of the data set. We will not treat rotating panels in this text. Some discussion and numerous references may be found in Baltagi (2001).

The development to follow is structured so that the distinction between balanced and unbalanced panels will entail nothing more than a trivial change in notation—where for convenience we write  $T$  suggesting a balanced panel, merely changing  $T$  to  $T_i$  generalizes the results. We will note specifically when this is not the case, such as in Breusch and Pagan’s (1980) LM statistic.

### 11.2.5 WELL-BEHAVED PANEL DATA

The asymptotic properties of the estimators in the classical regression model were established in Section 4.4 under the following assumptions:

- A.1. **Linearity:**  $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \dots + x_{iK}\beta_K$ .
- A.2. **Full rank:** The  $n \times K$  sample data matrix,  $\mathbf{X}$ , has full column rank.
- A.3. **Exogeneity of the independent variables:**  $E[\varepsilon_i | x_{j1}, x_{j2}, \dots, x_{jK}] = 0$ ,  $i, j = 1, \dots, n$ .
- A.4. **Homoscedasticity and nonautocorrelation.**
- A.5. **Data generating mechanism-independent observations.**

The following are the crucial results needed: For consistency of  $\mathbf{b}$ , we need

$$\text{plim}(1/n)\mathbf{X}'\mathbf{X} = \text{plim } \bar{\mathbf{Q}}_n = \mathbf{Q}, \quad \text{a positive definite matrix,}$$

$$\text{plim}(1/n)\mathbf{X}'\varepsilon = \text{plim } \bar{\mathbf{w}}_n = E[\bar{\mathbf{w}}_n] = \mathbf{0}.$$

(For consistency of  $s^2$ , we added a fairly weak assumption about the moments of the disturbances.) To establish asymptotic normality, we required consistency and

$$\sqrt{n}\bar{\mathbf{w}}_n \xrightarrow{d} N[0, \sigma^2\mathbf{Q}].$$

With these in place, the desired characteristics are then established by the methods of Sections 4.4.1 and 4.4.2.

Exceptions to the assumptions are likely to arise in a **panel data** set. The sample will consist of multiple observations on each of many observational units. For example, a study might consist of a set of observations made at different points in time on a large number of families. In this case, the  $\mathbf{x}$ ’s will surely be correlated across observations, at

least within observational units. They might even be the same for all the observations on a single family.

The panel data set could be treated as follows. Assume for a moment that the data consist of a fixed number of observations, say  $T$ , on a set of  $N$  families, so that the total number of rows in  $\mathbf{X}$  is  $NT$ . The matrix

$$\bar{\mathbf{Q}}_n = \sum_{i=1}^N \mathbf{Q}_i$$

in which  $n$  is all the observations in the sample, could be viewed as

$$\bar{\mathbf{Q}}_n = \frac{1}{N} \sum_i \frac{1}{T} \sum_{\text{observations for family } i} \mathbf{Q}_{it} = \frac{1}{N} \sum_{i=1}^N \bar{\mathbf{Q}}_i,$$

where  $\bar{\mathbf{Q}}_i$  = average  $\mathbf{Q}_{it}$  for family  $i$ . We might then view the set of observations on the  $i$ th unit as if they were a single observation and apply our convergence arguments to the number of families increasing without bound. The point is that the conditions that are needed to establish convergence will apply with respect to the number of observational units. The number of observations taken for each observation unit might be fixed and could be quite small.

This chapter will contain relatively little development of the properties of estimators as was done in Chapter 4. We will rely on earlier results in Chapters 4, 8, and 9 and focus instead on a variety of models and specifications.

### 11.3 THE POOLED REGRESSION MODEL

We begin the analysis by assuming the simplest version of the model, the **pooled model**,

$$\begin{aligned} y_{it} &= \alpha + \mathbf{x}'_{it} \beta + \varepsilon_{it}, \quad i = 1, \dots, n, t = 1, \dots, T_i, \\ E[\varepsilon_{it} | \mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT_i}] &= 0, \\ Var[\varepsilon_{it} | \mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT_i}] &= \sigma_\varepsilon^2, \\ Cov[\varepsilon_{it}, \varepsilon_{js} | \mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT_i}] &= 0 \text{ if } i \neq j \text{ or } t \neq s. \end{aligned} \tag{11-2}$$

(In the panel data context, this is also called the **population averaged model** under the assumption that any latent heterogeneity has been averaged out.) In this form, if the remaining assumptions of the classical model are met (zero conditional mean of  $\varepsilon_{it}$ , homoscedasticity, independence across observations,  $i$ , and strict exogeneity of  $\mathbf{x}_{it}$ ), then no further analysis beyond the results of Chapter 4 is needed. Ordinary least squares is the efficient estimator and inference can reliably proceed along the lines developed in Chapter 5.

#### 11.3.1 LEAST SQUARES ESTIMATION OF THE POOLED MODEL

The crux of the panel data analysis in this chapter is that the assumptions underlying ordinary least squares estimation of the pooled model are unlikely to be met. The

### 350 PART II ♦ Generalized Regression Model and Equation Systems

question, then, is what can be expected of the estimator when the heterogeneity does differ across individuals? The fixed effects case is obvious. As we will examine later, omitting (or ignoring) the heterogeneity when the fixed effects model is appropriate renders the least squares estimator inconsistent—sometimes wildly so. In the random effects case, in which the true model is

$$y_{it} = c_i + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it},$$

where  $E[c_i | \mathbf{X}_i] = \alpha$ , we can write the model

$$\begin{aligned} y_{it} &= \alpha + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it} + (c_i - E[c_i | \mathbf{X}_i]) \\ &= \alpha + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it} + u_i \\ &= \alpha + \mathbf{x}'_{it}\boldsymbol{\beta} + w_{it}. \end{aligned}$$

In this form, we can see that the unobserved heterogeneity induces **autocorrelation**;  $E[w_{it}w_{is}] = \sigma_u^2$  when  $t \neq s$ . As we explored in Chapter 9—we will revisit it in Chapter 20—the ordinary least squares estimator in the generalized regression model may be consistent, but the conventional estimator of its asymptotic variance is likely to underestimate the true variance of the estimator.

#### 11.3.2 ROBUST COVARIANCE MATRIX ESTIMATION

Suppose we consider the model more generally than this. Stack the  $T_i$  observations for individual  $i$  in a single equation,

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{w}_i,$$

where  $\boldsymbol{\beta}$  now includes the constant term. In this setting, there may be heteroscedasticity across individuals. However, in a panel data set, the more substantive problem is cross-observation correlation, or autocorrelation. In a longitudinal data set, the group of observations may all pertain to the same individual, so any latent effects left out of the model will carry across all periods. Suppose, then, we assume that the disturbance vector consists of  $\varepsilon_{it}$  plus these omitted components. Then,

$$\begin{aligned} \text{Var}[\mathbf{w}_i | \mathbf{X}_i] &= \sigma_\varepsilon^2 \mathbf{I}_{T_i} + \boldsymbol{\Sigma}_i \\ &= \boldsymbol{\Omega}_i. \end{aligned}$$

The ordinary least squares estimator of  $\boldsymbol{\beta}$  is

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \left[ \sum_{i=1}^n \mathbf{X}'_i \mathbf{X}_i \right]^{-1} \sum_{i=1}^n \mathbf{X}'_i \mathbf{y}_i \\ &= \left[ \sum_{i=1}^n \mathbf{X}'_i \mathbf{X}_i \right]^{-1} \sum_{i=1}^n \mathbf{X}'_i (\mathbf{X}_i \boldsymbol{\beta} + \mathbf{w}_i) \\ &= \boldsymbol{\beta} + \left[ \sum_{i=1}^n \mathbf{X}'_i \mathbf{X}_i \right]^{-1} \sum_{i=1}^n \mathbf{X}'_i \mathbf{w}_i. \end{aligned}$$

CHAPTER 11 ♦ Models for Panel Data **351**

Consistency can be established along the lines developed in Chapter 4. The true asymptotic covariance matrix would take the form we saw for the generalized regression model in (9-10),

$$\begin{aligned}\text{Asy. Var}[\mathbf{b}] &= \frac{1}{n} \text{plim} \left[ \frac{1}{n} \sum_{i=1}^n \mathbf{X}'_i \mathbf{X}_i \right]^{-1} \text{plim} \left[ \frac{1}{n} \sum_{i=1}^n \mathbf{X}'_i \mathbf{w}_i \mathbf{w}'_i \mathbf{X}_i \right] \text{plim} \left[ \frac{1}{n} \sum_{i=1}^n \mathbf{X}'_i \mathbf{X}_i \right]^{-1} \\ &= \frac{1}{n} \text{plim} \left[ \frac{1}{n} \sum_{i=1}^n \mathbf{X}'_i \mathbf{X}_i \right]^{-1} \text{plim} \left[ \frac{1}{n} \sum_{i=1}^n \mathbf{X}'_i \Omega_i \mathbf{X}_i \right] \text{plim} \left[ \frac{1}{n} \sum_{i=1}^n \mathbf{X}'_i \mathbf{X}_i \right]^{-1}.\end{aligned}$$

This result provides the counterpart to (9-26). As before, the center matrix must be estimated. In the same spirit as the White estimator, we can estimate this matrix with

$$\text{Est. Asy. Var}[\mathbf{b}] = \frac{1}{n} \left[ \frac{1}{n} \sum_{i=1}^n \mathbf{X}'_i \mathbf{X}_i \right]^{-1} \left[ \frac{1}{n} \sum_{i=1}^n \mathbf{X}'_i \hat{\mathbf{w}}_i \hat{\mathbf{w}}'_i \mathbf{X}_i \right] \left[ \frac{1}{n} \sum_{i=1}^n \mathbf{X}'_i \mathbf{X}_i \right]^{-1}, \quad (11-3)$$

where  $\hat{\mathbf{w}}$  is the vector of  $T_i$  residuals for individual  $i$ . In fact, the logic of the White estimator *does* carry over to this estimator. Note, however, this is not quite the same as (9-27). It is quite likely that the more important issue for appropriate estimation of the asymptotic covariance matrix is the correlation across observations, not heteroscedasticity. As such, it is quite likely that the White estimator in (9-27) is not the solution to the inference problem here. Example 11.1 shows this effect at work.

**Example 11.1 Wage Equation**

Cornwell and Rupert (1988) analyzed the returns to schooling in a (balanced) panel of 595 observations on heads of households. The sample data are drawn from years 1976–1982 from the “Non-Survey of Economic Opportunity” from the Panel Study of Income Dynamics. The estimating equation is

$$\begin{aligned}\ln \text{Wage}_{it} &= \beta_1 + \beta_2 \text{Exp}_{it} + \beta_3 \text{Exp}_{it}^2 + \beta_4 \text{Wks}_{it} + \beta_5 \text{Occ}_{it} \\ &\quad + \beta_6 \text{Ind}_{it} + \beta_7 \text{South}_{it} + \beta_8 \text{SMSA}_{it} + \beta_9 \text{MS}_{it} \\ &\quad + \beta_{10} \text{Union}_{it} + \beta_{11} \text{Ed}_i + \beta_{12} \text{Fem}_i + \beta_{13} \text{Blk}_i + \varepsilon_{it}\end{aligned}$$

where the variables are

- $\text{Exp}$  = years of full time work experience, 0 if not,
- $\text{Wks}$  = weeks worked, 0 if not,
- $\text{Occ}$  = 1 if blue-collar occupation, 0 if not,
- $\text{Ind}$  = 1 if the individual works in a manufacturing industry, 0 if not,
- $\text{South}$  = 1 if the individual resides in the south, 0 if not,
- $\text{SMSA}$  = 1 if the individual resides in an SMSA, 0 if not,
- $\text{MS}$  = 1 if the individual is married, 0 if not,
- $\text{Union}$  = 1 if the individual wage is set by a union contract, 0 if not,
- $\text{Ed}$  = years of education,
- $\text{Fem}$  = 1 if the individual is female, 0 if not,
- $\text{Blk}$  = 1 if the individual is black, 0 if not.

Note that  $\text{Ed}$ ,  $\text{Fem}$ , and  $\text{Blk}$  are **time invariant**. See Appendix Table F1 for the data source. The main interest of the study, beyond comparing various estimation methods, is  $\beta_{11}$ , the return to education. Table 11.1 reports the least squares estimates based on the full sample

**352 PART II ♦ Generalized Regression Model and Equation Systems**
**TABLE 11.1** Wage Equation Estimated by OLS

<i>Coefficient</i>	<i>Estimated Coefficient</i>	<i>OLS Standard Error</i>	<i>Panel Robust Standard Error</i>	<i>White Hetero. Consistent Std. Error</i>
$\beta_1$ : Constant	5.2511	0.07129	0.1233	0.07435
$\beta_2$ : Exp	0.04010	0.002159	0.004067	0.002158
$\beta_3$ : Exp <sup>2</sup>	-0.0006734	0.00004744	0.00009111	0.00004789
$\beta_4$ : Wks	0.004216	0.001081	0.001538	0.001143
$\beta_5$ : Occ	-0.1400	0.01466	0.02718	0.01494
$\beta_6$ : Ind	0.04679	0.01179	0.02361	0.01199
$\beta_7$ : South	-0.05564	0.01253	0.02610	0.01274
$\beta_8$ : SMSA	0.1517	0.01207	0.02405	0.01208
$\beta_9$ : MS	0.04845	0.02057	0.04085	0.02049
$\beta_{10}$ : Union	0.09263	0.01280	0.02362	0.01233
$\beta_{11}$ : Ed	0.05670	0.002613	0.005552	0.002726
$\beta_{12}$ : Fem	-0.3678	0.02510	0.04547	0.02310
$\beta_{13}$ : Blk	-0.1669	0.02204	0.04423	0.02075

of 4,165 observations. [The authors do not report OLS estimates. However, they do report linear least squares estimates of the fixed effects model, which are simple least squares using deviations from individual means. (See Section 11.4.) It was not possible to match their reported results for these or any of their other reported results. Because our purpose is to compare the various estimators to each other, we have not attempted to resolve the discrepancy.] The conventional OLS standard errors are given in the second column of results. The third column gives the robust standard errors computed using (11-3). For these data, the computation is

$$\text{Est. Asy. Var}[\mathbf{b}] = \left[ \sum_{i=1}^{595} \mathbf{X}_i' \mathbf{X}_i \right]^{-1} \left[ \sum_{i=1}^{595} \left( \sum_{t=1}^7 \mathbf{x}_{it} \mathbf{e}_{it} \right) \left( \sum_{t=1}^7 \mathbf{x}_{it} \mathbf{e}_{it} \right)' \right] \left[ \sum_{i=1}^{595} \mathbf{X}_i' \mathbf{X}_i \right]^{-1}.$$

The robust standard errors are generally about twice the uncorrected ones. In contrast, the White robust standard errors are almost the same as the uncorrected ones. This suggests that for this model, ignoring the within group correlations does, indeed, substantially affect the inferences one would draw.

**11.3.3 CLUSTERING AND STRATIFICATION**

Many recent studies have analyzed survey data sets, such as the Current Population Survey (CPS). Survey data are often drawn in “clusters,” partly to reduce costs. For example, interviewers might visit all the families in a particular block. In other cases, effects that resemble the common random effects in panel data treatments might arise naturally in the sampling setting. Consider, for example, a study of student test scores across several states. Common effects could arise at many levels in such a data set. Education curriculum or funding policies in a state could cause a “state effect;” there could be school district effects, school effects within districts, and even teacher effects within a particular school. Each of these is likely to induce correlation across observations that resembles the random (or fixed) effects we have identified. One might be reluctant to assume that a tightly structured model such as the simple random effects specification is at work. But, as we saw in Example 11.1, ignoring common effects can lead to serious inference errors. The robust estimator suggested in Section 11.3.2 provides a useful approach.

For a two-level model, such as might arise in a sample of firms that are grouped by industry, or students who share teachers in particular schools, a natural approach to this “clustering” would be the robust common effects approach shown earlier. The

CHAPTER 11 ♦ Models for Panel Data **353**

resemblance of the now standard **cluster estimator** for a one-level model to the common effects panel model considered earlier is more than coincidental. However, there is a difference in the data generating mechanism in that in this setting, the individuals in the group are generally observed once, and their association, that is, common effect, is likely to be less clearly defined than in a panel such as the one analyzed in Example 11.1. A refinement to (11-3) is often employed to account for small-sample effects when the number of clusters is likely to be a significant proportion of a finite total, such as the number of school districts in a state. A degrees of freedom correction as shown in (11-4) is often employed for this purpose. The robust covariance matrix estimator would be

$$\begin{aligned}
 & \text{Est.Asy.Var}[\mathbf{b}] \\
 &= \left[ \sum_{g=1}^G \mathbf{X}'_g \mathbf{X}_g \right]^{-1} \left[ \frac{G}{G-1} \sum_{g=1}^G \left( \sum_{i=1}^{n_g} \mathbf{x}_{ig} \hat{\mathbf{w}}_{ig} \right) \left( \sum_{i=1}^{n_g} \mathbf{x}_{ig} \hat{\mathbf{w}}_{ig} \right)' \right] \left[ \sum_{g=1}^G \mathbf{X}'_g \mathbf{X}_g \right]^{-1} \\
 &= \left[ \sum_{g=1}^G \mathbf{X}'_g \mathbf{X}_g \right]^{-1} \left[ \frac{G}{G-1} \sum_{g=1}^G (\mathbf{X}'_g \hat{\mathbf{w}}_g) (\hat{\mathbf{w}}'_g \mathbf{X}_g) \right] \left[ \sum_{g=1}^G \mathbf{X}'_g \mathbf{X}_g \right]^{-1}, \quad (11-4)
 \end{aligned}$$

where  $G$  is the number of clusters in the sample and each cluster consists of  $n_g$ ,  $g = 1, \dots, G$  observations. [Note that this matrix is simply  $G/(G-1)$  times the matrix in (11-3).] A further correction (without obvious formal motivation) sometimes employed is a “degrees of freedom correction,”  $\Sigma_g n_g / [(\Sigma_g n_g) - K]$ .

Many further refinements for more complex samples—consider the test scores example—have been suggested. For a detailed analysis, see Cameron and Trivedi (2005, Chapter 24). Several aspects of the computation are discussed in Wooldridge (2003) as well. An important question arises concerning the use of asymptotic distributional results in cases in which the number of clusters might be relatively small. Angrist and Lavy (2001) find that the clustering correction after pooled OLS, as we have done in Example 11.1, is not as helpful as might be hoped for (though our correction with 595 clusters each of size 7 would be “safe” by these standards). But, the difficulty might arise, at least in part, from the use of OLS in the presence of the common effects. Kezde (2001) and Bertrand, Dufflo, and Mullainathan (2002) find more encouraging results when the correction is applied after estimation of the fixed effects regression. Yet another complication arises when the groups are very large and the number of groups is relatively small, for example, when the panel consists of many large samples from a subset (or even all) of the U.S. states. Since the asymptotic theory we have used to this point assumes the opposite, the results will be less reliable in this case. Donald and Lang (2007) find that this case gravitates toward analysis of group means, rather than the individual data. Wooldridge (2003) provides results that help explain this finding. Finally, there is a natural question as to whether the correction is even called for if one has used a random effects, generalized least squares procedure (see Section 11.5) to do the estimation at the first step. If the data generating mechanism were strictly consistent with the random effects model, the answer would clearly be negative. Under the view that the random effects specification is only an approximation to the correlation across observations in a cluster, then there would remain “residual correlation” that would be accommodated by the correction in (11-4) (or some GLS counterpart). (This would call the specific random effects correction in Section 11.5 into question, however.) A similar

**354 PART II ♦ Generalized Regression Model and Equation Systems**
**TABLE 11.2** Sale Price Equation

<i>Variable</i>	<i>Estimated Coefficient</i>	<i>OLS Standard Error</i>	<i>Corrected Standard Error</i>
<i>Constant</i>	-9.7068	0.5661	0.6791
<i>In Area</i>	1.3473	0.0822	0.1030
<i>Signature</i>	1.3614	0.1251	0.1281
<i>In Aspect Ratio</i>	-0.0225	0.1479	0.1661

argument would motivate the correction after fitting the fixed effects model as well. We will pursue these possibilities in Section 11.6.4 after we develop the fixed and random effects estimator in detail.

**Example 11.2 Repeated Sales of Monet Paintings**

We examined in Examples 4.9, 4.10, and 6.2 the relationship between the sale price and the surface area of a sample of 430 sales of Monet paintings. In fact, these were not sales of 430 paintings. Many of them were repeat sales of the same painting at different points in time. The sample actually contains 376 paintings. The numbers of sales per painting were one, 333; two, 34; three, 7; and four, 2. If the sale price of the painting is motivated at least partly by intrinsic features of the painting, then this would motivate a correction of the least squares standard errors as suggested in (11-4). Table 11.2 displays the OLS regression results with the conventional and with the corrected standard errors. Even with the quite modest amount of grouping in the data, the impact of the correction, in the expected direction of larger standard errors, is evident.

**11.3.4 ROBUST ESTIMATION USING GROUP MEANS**

The pooled regression model can be estimated using the sample means of the data. The implied regression model is obtained by premultiplying each group by  $(1/T)\mathbf{i}'$  where  $\mathbf{i}'$  is a row vector of ones;

$$(1/T)\mathbf{i}'\mathbf{y}_i = (1/T)\mathbf{i}'\mathbf{X}_i\beta + (1/T)\mathbf{i}'\mathbf{w}_i$$

or

$$\bar{y}_{i \cdot} = \bar{\mathbf{x}}_{i \cdot}'\beta + \bar{\mathbf{w}}_{i \cdot}$$

In the transformed linear regression, the disturbances continue to have zero conditional means but heteroscedastic variances  $\sigma_i^2 = (1/T^2)\mathbf{i}'\Omega_i\mathbf{i}$ . With  $\Omega_i$  unspecified, this is a heteroscedastic regression for which we would use the White estimator for appropriate inference. Why might one want to use this estimator when the full data set is available? If the classical assumptions are met, then it is straightforward to show that the asymptotic covariance matrix for the group means estimator is unambiguously larger, and the answer would be that there is no benefit. But, failure of the classical assumptions is what brought us to this point, and then the issue is less clear-cut. In the presence of unstructured cluster effects the efficiency of least squares can be considerably diminished, as we saw in the preceding example. The loss of information that occurs through the averaging might be relatively small, though in principle, the disaggregated data should still be better.

We emphasize, using **group means** does not solve the problem that is addressed by the fixed effects estimator. Consider the general model,

$$\mathbf{y}_i = \mathbf{X}_i\beta + c_i\mathbf{i} + \mathbf{w}_i,$$

CHAPTER 11 ♦ Models for Panel Data **355**

where as before,  $c_i$  is the latent effect. If the mean independence assumption,  $E[c_i | \mathbf{X}_i] = \alpha$ , is not met, then, the effect will be transmitted to the group means as well. In this case,  $E[c_i | \mathbf{X}_i] = h(\mathbf{X}_i)$ . A common specification is Mundlak's (1978),

$$E[c_i | \mathbf{X}_i] = \bar{\mathbf{x}}_i' \boldsymbol{\gamma}.$$

(We will revisit this specification in Section 11.5.6.) Then,

$$\begin{aligned} y_{it} &= \mathbf{x}'_{it} \boldsymbol{\beta} + c_i + \varepsilon_{it} \\ &= \mathbf{x}'_{it} \boldsymbol{\beta} + \bar{\mathbf{x}}_i' \boldsymbol{\gamma} + [\varepsilon_{it} + c_i - E[c_i | \mathbf{X}_i]] \\ &= \mathbf{x}'_{it} \boldsymbol{\beta} + \bar{\mathbf{x}}_i' \boldsymbol{\gamma} + u_{it} \end{aligned}$$

where  construction,  $E[u_{it} | \mathbf{X}_i] = 0$ . Taking means as before,

$$\begin{aligned} \bar{y}_{i\cdot} &= \bar{\mathbf{x}}_i' \boldsymbol{\beta} + \bar{\mathbf{x}}_i' \boldsymbol{\gamma} + \bar{u}_{i\cdot} \\ &= \bar{\mathbf{x}}_i' (\boldsymbol{\beta} + \boldsymbol{\gamma}) + \bar{u}_{i\cdot}. \end{aligned}$$

The implication is that the group means estimator estimates not  $\boldsymbol{\beta}$ , but  $\boldsymbol{\beta} + \boldsymbol{\gamma}$ . Averaging the observations in the group collects the entire set of effects, observed and latent, in the group means.

One consideration that remains, which, unfortunately, we cannot resolve analytically, is the possibility of **measurement error**. If the regressors are measured with error, then, as we examined in Section 8.5, the least squares estimator is inconsistent and, as a consequence, efficiency is a moot point. In the panel data setting, if the measurement error is random, then using group means would work in the direction of averaging it out—indeed, in this instance, assuming the benchmark case  $\mathbf{x}_{itk} = \mathbf{x}_{itk}^* + u_{itk}$ , one could show that the group means estimator would be consistent as  $T \rightarrow \infty$  while the OLS estimator would not.

#### **Example 11.3 Robust Estimators of the Wage Equation**

Table 11.3 shows the group means estimator of the wage equation shown in Example 11.1 with the original least squares estimates. In both cases, a robust estimator is used for the covariance matrix of the estimator. It appears that similar results are obtained with the means.

##### **11.3.5 ESTIMATION WITH FIRST DIFFERENCES**

First differencing is another approach to estimation. Here, the intent would explicitly be to transform latent heterogeneity out of the model. The base case would be

$$y_{it} = c_i + \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it},$$

which implies the first differences equation

$$\Delta y_{it} = \Delta c_i + (\Delta \mathbf{x}_{it})' \boldsymbol{\beta} + \Delta \varepsilon_{it},$$

or

$$\begin{aligned} \Delta y_{it} &= (\Delta \mathbf{x}_{it})' \boldsymbol{\beta} + \varepsilon_{it} - \varepsilon_{i,t-1} \\ &= (\Delta \mathbf{x}_{it})' \boldsymbol{\beta} + u_{it}. \end{aligned}$$

**356 PART II ♦ Generalized Regression Model and Equation Systems**
**TABLE 11.3** Wage Equation Estimated by OLS

<i>Coefficient</i>	<i>OLS Estimated Coefficient</i>	<i>Panel Robust Standard Error</i>	<i>Group Means Estimates</i>	<i>White Robust Standard Error</i>
$\beta_1$ : Constant	5.2511	0.1233	5.1214	0.2078
$\beta_2$ : Exp	0.04010	0.004067	0.03190	0.004597
$\beta_3$ : Exp <sup>2</sup>	-0.0006734	0.00009111	-0.0005656	0.0001020
$\beta_4$ : Wks	0.004216	0.001538	0.009189	0.003578
$\beta_5$ : Occ	-0.1400	0.02718	-0.1676	0.03338
$\beta_6$ : Ind	0.04679	0.02361	0.05792	0.02636
$\beta_7$ : South	-0.05564	0.02610	-0.05705	0.02660
$\beta_8$ : SMSA	0.1517	0.02405	0.1758	0.02541
$\beta_9$ : MS	0.04845	0.04085	0.1148	0.04989
$\beta_{10}$ : Union	0.09263	0.02362	0.1091	0.02830
$\beta_{11}$ : Ed	0.05670	0.005552	0.05144	0.005862
$\beta_{12}$ : Fem	-0.3678	0.04547	-0.3171	0.05105
$\beta_{13}$ : Blk	-0.1669	0.04423	-0.1578	0.04352

The advantage of the **first difference** approach is that it removes the latent heterogeneity from the model whether the fixed or random effects model is appropriate. The disadvantage is that the differencing also removes any time-invariant variables from the model. In our example, we had three, *Ed*, *Fem*, and *Blk*. If the time-invariant variables in the model are of no interest, then this is a robust approach that can estimate the parameters of the time-varying variables consistently. Of course, this is not helpful for the application in the example, because the impact of *Ed* on *In Wage* was the primary object of the analysis. Note, as well, that the differencing procedure trades the cross-observation correlation in  $c_i$  for a moving average (MA) disturbance,  $u_{i,t} = \varepsilon_{i,t} - \varepsilon_{i,t-1}$ . The new disturbance,  $u_{i,t}$  is autocorrelated, though across only one period. Procedures are available for using two-step feasible GLS for an MA disturbance (see Chapter 1).

Alternatively, this model is a natural candidate for OLS with the Newey-West robust covariance estimator, since the right number of lags (one) is known. (See Section 20.5.2.)

As a general observation, with a variety of approaches available, the first difference estimator does not have much to recommend it, save for one very important application. Many studies involve two period “panels,” a before and after treatment. In these cases, as often as not, the phenomenon of interest may well specifically be the change in the outcome variable—the “treatment effect.” Consider the model

$$y_{it} = c_i + \mathbf{x}'_{it}\boldsymbol{\beta} + \theta S_{it} + \varepsilon_{it},$$

where  $t = 1, 2$  and  $S_{it} = 0$  in period 1 and 1 in period 2;  $S_{it}$  indicates a “treatment” that takes place between the two observations. The “treatment effect” would be

$$E[\Delta y_i | (\Delta \mathbf{x}_i = 0)] = \theta,$$

which is precisely the constant term in the first difference regression,

$$\Delta y_i = \theta + (\Delta \mathbf{x}_i)' \boldsymbol{\beta} + u_i.$$

We will examine cases like these in detail in Section 18.5.

### 11.3.6 THE WITHIN- AND BETWEEN-GROUPS ESTIMATORS

We can formulate the pooled regression model in three ways. First, the original formulation is

$$y_{it} = \alpha + \mathbf{x}'_{it}\beta + \varepsilon_{it}. \quad (11-5a)$$

In terms of the group means,

$$\bar{y}_{i\cdot} = \alpha + \bar{\mathbf{x}}'_i\beta + \bar{\varepsilon}_{i\cdot}, \quad (11-5b)$$

while in terms of deviations from the group means,

$$y_{it} - \bar{y}_{i\cdot} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \beta + \varepsilon_{it} - \bar{\varepsilon}_{i\cdot}. \quad (11-5c)$$

[We are assuming there are no time-invariant variables, such as *Ed* in Example 11.1, in  $\mathbf{x}_{it}$ . These would become all zeros in (11-5c).] All three are classical regression models, and in principle, all three could be estimated, at least consistently if not efficiently, by ordinary least squares. [Note that (11-5b) defines only  $n$  observations, the group means.] Consider then the matrices of sums of squares and cross products that would be used in each case, where we focus only on estimation of  $\beta$ . In (11-5a), the moments would accumulate variation about the overall means,  $\bar{y}$  and  $\bar{\mathbf{x}}$ , and we would use the total sums of squares and cross products,

$$\mathbf{S}_{xx}^{total} = \sum_{i=1}^n \sum_{t=1}^T (\mathbf{x}_{it} - \bar{\mathbf{x}})(\mathbf{x}_{it} - \bar{\mathbf{x}})' \quad \text{and} \quad \mathbf{S}_{xy}^{total} = \sum_{i=1}^n \sum_{t=1}^T (\mathbf{x}_{it} - \bar{\mathbf{x}})(y_{it} - \bar{y}). \quad (11-6)$$

For (11-5c), because the data are in deviations already, the means of  $(y_{it} - \bar{y}_{i\cdot})$  and  $(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$  are zero. The moment matrices are **within-groups** (i.e., variation around group means) sums of squares and cross products,

$$\mathbf{S}_{xx}^{within} = \sum_{i=1}^n \sum_{t=1}^T (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \quad \text{and} \quad \mathbf{S}_{xy}^{within} = \sum_{i=1}^n \sum_{t=1}^T (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(y_{it} - \bar{y}_{i\cdot}).$$

Finally, for (11-5b), the mean of group means is the overall mean. The moment matrices are the **between-groups** sums of squares and cross products—that is, the variation of the group means around the overall means;

$$\mathbf{S}_{xx}^{between} = \sum_{i=1}^n T(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \quad \text{and} \quad \mathbf{S}_{xy}^{between} = \sum_{i=1}^n T(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{y}_i - \bar{y}).$$

It is easy to verify that

$$\mathbf{S}_{xx}^{total} = \mathbf{S}_{xx}^{within} + \mathbf{S}_{xx}^{between} \quad \text{and} \quad \mathbf{S}_{xy}^{total} = \mathbf{S}_{xy}^{within} + \mathbf{S}_{xy}^{between}.$$

Therefore, there are three possible least squares estimators of  $\beta$  corresponding to the decomposition. The least squares estimator is

$$\mathbf{b}^{total} = [\mathbf{S}_{xx}^{total}]^{-1} \mathbf{S}_{xy}^{total} = [\mathbf{S}_{xx}^{within} + \mathbf{S}_{xx}^{between}]^{-1} [\mathbf{S}_{xy}^{within} + \mathbf{S}_{xy}^{between}]. \quad (11-7)$$

The **within-groups estimator** is

$$\mathbf{b}^{within} = [\mathbf{S}_{xx}^{within}]^{-1} \mathbf{S}_{xy}^{within}. \quad (11-8)$$

## 358 PART II ♦ Generalized Regression Model and Equation Systems

This is the dummy variable estimator developed in Section 11.4. An alternative estimator would be the **between-groups estimator**,

$$\mathbf{b}^{between} = [\mathbf{S}_{xx}^{between}]^{-1} \mathbf{S}_{xy}^{between}. \quad (11-9)$$

This is the **group means estimator**. This least squares estimator of (11-5b) is based on the  $n$  sets of groups means. (Note that we are assuming that  $n$  is at least as large as  $K$ .) From the preceding expressions (and familiar previous results),

$$\mathbf{S}_{xy}^{within} = \mathbf{S}_{xx}^{within} \mathbf{b}^{within} \quad \text{and} \quad \mathbf{S}_{xy}^{between} = \mathbf{S}_{xx}^{between} \mathbf{b}^{between}.$$

Inserting these in (11-7), we see that the least squares estimator is a **matrix weighted average** of the within- and between-groups estimators:

$$\mathbf{b}^{total} = \mathbf{F}^{within} \mathbf{b}^{within} + \mathbf{F}^{between} \mathbf{b}^{between}, \quad (11-10)$$

where

$$\mathbf{F}^{within} = [\mathbf{S}_{xx}^{within} + \mathbf{S}_{xx}^{between}]^{-1} \mathbf{S}_{xx}^{within} = \mathbf{I} - \mathbf{F}^{between}.$$

The form of this result resembles the Bayesian estimator in the classical model discussed in Chapter 14. The resemblance is more than passing; it can be shown [see, e.g., Judge et al. (1985)] that

$$\mathbf{F}^{within} = \{[\text{Asy. Var}(\mathbf{b}^{within})]^{-1} + [\text{Asy. Var}(\mathbf{b}^{between})]^{-1}\}^{-1} [\text{Asy. Var}(\mathbf{b}^{within})]^{-1},$$

which is essentially the same mixing result we have for the Bayesian estimator. In the weighted average, the estimator with the smaller variance receives the greater weight.

### Example 11.4 Analysis of Covariance and the World Health Organization Data

The decomposition of the total variation in Section 11.3.6 extends to the linear regression model the familiar “analysis of variance,” or ANOVA, that is often used to decompose the variation in a variable in a clustered or stratified sample, or in a panel data set. One of the useful features of panel data analysis as we are doing here is the ability to analyze the between-groups variation (heterogeneity) to learn about the main regression relationships and the within-groups variation to learn about dynamic effects.

The World Health Organization data used in Example 6.10 is an unbalanced panel data set—we used only one year of the data in Example 6.10. Of the 191 countries in the sample, 140 are observed in the full five years, one is observed four times, and 50 are observed only once. The original WHO studies (2000a, 2000b) analyzed these data using the fixed effects model developed in the next section. The estimator is that in (11-5c). It is easy to see that groups with one observation will fall out of the computation, because if  $T_i = 1$ , then the observation equals the group mean. These data have been used by many researchers in similar panel data analyses. [See, e.g., Greene (2004c) and several references.] Gravelle et al. (2002a) have strongly criticized these analyses, arguing that the WHO data are much more like a cross section than a panel data set.

From Example 6.10, the model used by the researchers at WHO was

$$\ln DALE_{it} = \alpha_i + \beta_1 \ln \text{Health Expenditure}_{it} + \beta_2 \ln \text{Education}_{it} + \beta_3 \ln^2 \text{Education}_{it} + \varepsilon_{it}.$$

Additional models were estimated using WHO’s composite measure of health care attainment, *COMP*. The analysis of variance for a variable  $x_{it}$  is based on the decomposition

$$\sum_{i=1}^n \sum_{t=1}^{T_i} (x_{it} - \bar{x})^2 = \sum_{i=1}^n \sum_{t=1}^{T_i} (x_{it} - \bar{x}_{i.})^2 + \sum_{t=1}^n T_i (\bar{x}_{i.} - \bar{x})^2.$$

**TABLE 11.4** Analysis of Variance for WHO Data on Health Care Attainment

<i>Variable</i>	<i>Within-Groups Variation</i>	<i>Between-Groups Variation</i>
<i>COMP</i>	0.150%	99.850%
<i>DALE</i>	5.645%	94.355%
<i>Expenditure</i>	0.635%	99.365%
<i>Education</i>	0.178%	99.822%

Dividing both sides of the equation by the left-hand side produces the decomposition:

$$1 = \text{Within-groups proportion} + \text{Between-groups proportion}.$$

The first term on the right-hand side is the within-group variation that differentiates a panel data set from a cross section (or simply multiple observations on the same variable). Table 11.4 lists the decomposition of the variation in the variables used in the WHO studies.

The results suggest the reasons for the authors' concern about the data. For all but COMP, virtually all the variation in the data is between groups—that is cross-sectional variation. As the authors argue, these data are only slightly different from a cross section.

## 11.4 THE FIXED EFFECTS MODEL

The fixed effects model arises from the assumption that the omitted effects,  $c_i$ , in the general model,

$$y_{it} = \mathbf{x}'_{it}\beta + c_i + \varepsilon_{it},$$

are correlated with the included variables. In a general form,

$$E[c_i | \mathbf{X}_i] = h(\mathbf{X}_i). \quad (11-11)$$

Because the conditional mean is the same in every period, we can write the model as

$$\begin{aligned} y_{it} &= \mathbf{x}'_{it}\beta + h(\mathbf{X}_i) + \varepsilon_{it} + [c_i - h(\mathbf{X}_i)] \\ &= \mathbf{x}'_{it}\beta + \alpha_i + \varepsilon_{it} + [c_i - h(\mathbf{X}_i)]. \end{aligned}$$

By construction, the bracketed term is uncorrelated with  $\mathbf{X}_i$ , so we may absorb it in the disturbance, and write the model as

$$y_{it} = \mathbf{x}'_{it}\beta + \alpha_i + \varepsilon_{it}. \quad (11-12)$$

A further assumption (usually unstated) is that  $\text{Var}[c_i | \mathbf{X}_i]$  is constant. With this assumption, (11-12) becomes a classical linear regression model. (We will reconsider the homoscedasticity assumption shortly.) We emphasize, it is (11-11) that signifies the “fixed effects” model, not that any variable is “fixed” in this context and random elsewhere. The fixed effects formulation implies that differences across groups can be captured in differences in the constant term.<sup>5</sup> Each  $\alpha_i$  is treated as an unknown parameter to be estimated.

<sup>5</sup>It is also possible to allow the slopes to vary across  $i$ , but this method introduces some new methodological issues, as well as considerable complexity in the calculations. A study on the topic is Cornwell and Schmidt (1984).

## 360 PART II ♦ Generalized Regression Model and Equation Systems

Before proceeding, we note once again a major shortcoming of the fixed effects approach. Any **time-invariant** variables in  $\mathbf{x}_{it}$  will mimic the individual specific constant term. Consider the application of Examples 11.1 and 11.3. We could write the fixed effects formulation as

$$\ln Wage_{it} = \mathbf{x}'_{it}\beta + [\beta_{10}Ed_i + \beta_{11}Fem_i + \beta_{12}Blk_i + c_i] + \varepsilon_{it}.$$

The fixed effects formulation of the model will absorb the last four terms in the regression in  $\alpha_i$ . The coefficients on the time-invariant variables cannot be estimated. This lack of identification is the price of the robustness of the specification to unmeasured correlation between the common effect and the exogenous variables.

### 11.4.1 LEAST SQUARES ESTIMATION

Let  $\mathbf{y}_i$  and  $\mathbf{X}_i$  be the  $T$  observations for the  $i$ th unit,  $\mathbf{i}$  be a  $T \times 1$  column of ones, and let  $\boldsymbol{\varepsilon}_i$  be the associated  $T \times 1$  vector of disturbances.<sup>6</sup> Then,

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{i}\alpha_i + \boldsymbol{\varepsilon}_i.$$

Collecting these terms gives

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \beta + \begin{bmatrix} i & 0 & \cdots & 0 \\ 0 & i & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & i \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix}$$

or

$$\mathbf{y} = [\mathbf{X} \quad \mathbf{d}_1 \quad \mathbf{d}_2, \dots, \mathbf{d}_n] \begin{bmatrix} \beta \\ \alpha \end{bmatrix} + \boldsymbol{\varepsilon}, \quad (11-13)$$

where  $\mathbf{d}_i$  is a dummy variable indicating the  $i$ th unit. Let the  $nT \times n$  matrix  $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n]$ . Then, assembling all  $nT$  rows gives

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{D}\alpha + \boldsymbol{\varepsilon}.$$

This model is usually referred to as the **least squares dummy variable (LSDV) model** (although the “least squares” part of the name refers to the technique usually used to estimate it, not to the model itself).

This model is a classical regression model, so no new results are needed to analyze it. If  $n$  is small enough, then the model can be estimated by ordinary least squares with  $K$  regressors in  $\mathbf{X}$  and  $n$  columns in  $\mathbf{D}$ , as a multiple regression with  $K + n$  parameters. Of course, if  $n$  is thousands, as is typical, then this model is likely to exceed the storage capacity of any computer. But, by using familiar results for a partitioned regression, we can reduce the size of the computation.<sup>7</sup> We write the least squares estimator of  $\beta$  as

$$\mathbf{b} = [\mathbf{X}'\mathbf{M}_D\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{M}_D\mathbf{y}] = \mathbf{b}^{within}, \quad (11-14)$$

<sup>6</sup>The assumption of a fixed group size,  $T$ , at this point is purely for convenience. As noted in Section 11.2.4, the unbalanced case is a minor variation.

<sup>7</sup>See Theorem 3.3.

CHAPTER 11 ♦ Models for Panel Data **361**

where

$$\mathbf{M}_D = \mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'.$$

This amounts to a least squares regression using the transformed data  $\mathbf{X}_* = \mathbf{M}_D \mathbf{X}$  and  $\mathbf{y}_* = \mathbf{M}_D \mathbf{y}$ . The structure of  $\mathbf{D}$  is particularly convenient; its columns are orthogonal, so

$$\mathbf{M}_D = \begin{bmatrix} \mathbf{M}^0 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^0 & \mathbf{0} & \cdots & \mathbf{0} \\ & & \ddots & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{M}^0 \end{bmatrix}.$$

Each matrix on the diagonal is

$$\mathbf{M}^0 = \mathbf{I}_T - \frac{1}{T} \mathbf{i} \mathbf{i}' \quad (11-15)$$

Premultiplying any  $T \times 1$  vector  $\mathbf{z}_i$  by  $\mathbf{M}^0$  creates  $\mathbf{M}^0 \mathbf{z}_i = \mathbf{z}_i - \bar{z} \mathbf{i}$ . (Note that the mean is taken over only the  $T$  observations for unit  $i$ .) Therefore, the least squares regression of  $\mathbf{M}_D \mathbf{y}$  on  $\mathbf{M}_D \mathbf{X}$  is equivalent to a regression of  $[y_{it} - \bar{y}_{i.}]$  on  $[\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.}]$ , where  $\bar{y}_{i.}$  and  $\bar{\mathbf{x}}_{i.}$  are the scalar and  $K \times 1$  vector of means of  $y_{it}$  and  $\mathbf{x}_{it}$  over the  $T$  observations for group  $i$ .<sup>8</sup> The dummy variable coefficients can be recovered from the other normal equation in the partitioned regression:

$$\mathbf{D}' \mathbf{D} \mathbf{a} + \mathbf{D}' \mathbf{X} \mathbf{b} = \mathbf{D}' \mathbf{y}$$

or

$$\mathbf{a} = [\mathbf{D}' \mathbf{D}]^{-1} \mathbf{D}' (\mathbf{y} - \mathbf{X} \mathbf{b}).$$

This implies that for each  $i$ ,

$$a_i = \bar{y}_{i.} - \bar{\mathbf{x}}_{i.}' \mathbf{b}. \quad (11-16)$$

The appropriate estimator of the asymptotic covariance matrix for  $\mathbf{b}$  is

$$\text{Est. Asy. Var}[\mathbf{b}] = s^2 [\mathbf{X}' \mathbf{M}_D \mathbf{X}]^{-1} = s^2 [\mathbf{S}_{xx}^{within}]^{-1}, \quad (11-17)$$

which uses the second moment matrix with  $\mathbf{x}$ 's now expressed as deviations from their respective group means. The disturbance variance estimator is

$$s^2 = \frac{\sum_{i=1}^n \sum_{t=1}^T (y_{it} - \mathbf{x}_{it}' \mathbf{b} - a_i)^2}{nT - n - K} = \frac{(\mathbf{M}_D \mathbf{y} - \mathbf{M}_D \mathbf{X} \mathbf{b})' (\mathbf{M}_D \mathbf{y} - \mathbf{M}_D \mathbf{X} \mathbf{b})}{nT - n - K}. \quad (11-18)$$

The  $it$ th residual used in this computation is

$$e_{it} = y_{it} - \mathbf{x}_{it}' \mathbf{b} - a_i = y_{it} - \mathbf{x}_{it}' \mathbf{b} - (\bar{y}_{i.} - \bar{\mathbf{x}}_{i.}' \mathbf{b}) = (y_{it} - \bar{y}_{i.}) - (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})' \mathbf{b}.$$

Thus, the numerator in  $s^2$  is exactly the sum of squared residuals using the least squares slopes and the data in group mean deviation form. But, done in this fashion, one might then use  $nT - K$  instead of  $nT - n - K$  for the denominator in computing  $s^2$ , so a

<sup>8</sup>An interesting special case arises if  $T = 2$ . In the two-period case, you can show—we leave it as an exercise—that this least squares regression is done with  $nT/2$  first difference observations, by regressing observation  $(y_{i2} - y_{i1})$  (and its negative) on  $(\mathbf{x}_{i2} - \mathbf{x}_{i1})$  (and its negative).

## 362 PART II ♦ Generalized Regression Model and Equation Systems

correction would be necessary. For the individual effects,

$$\text{Asy. Var}[a_i] = \frac{\sigma_e^2}{T} + \bar{x}'_{i.} \{ \text{Asy. Var}[\mathbf{b}] \} \bar{x}_{i.}, \quad (11-19)$$

so a simple estimator based on  $s^2$  can be computed.

### 11.4.2 SMALL T ASYMPTOTICS

From (11-17), we find

$$\begin{aligned} \text{Asy. Var}[\mathbf{b}] &= \sigma_e^2 [\mathbf{X}' \mathbf{M}_D \mathbf{X}]^{-1} \\ &= \frac{\sigma_e^2}{n} \left[ \frac{1}{n} \sum_{i=1}^n \mathbf{X}'_i \mathbf{M}^0 \mathbf{X}_i \right]^{-1} \\ &= \frac{\sigma_e^2}{n} \left[ \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})' \right]^{-1} \\ &= \frac{\sigma_e^2}{n} \left[ T \frac{1}{n} \sum_{i=1}^n \frac{1}{T} \sum_{t=1}^T (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})' \right]^{-1} \\ &= \frac{\sigma_e^2}{n} [T \bar{S}_{xx,i}]^{-1}. \end{aligned} \quad (11-20)$$

Since least squares is unbiased in this model, the question of (mean square) consistency turns on the covariance matrix. Does the matrix above converge to zero? It is necessary to be specific about what is meant by convergence. In this setting, increasing sample size refers to increasing  $n$ , that is, increasing the number of groups. The group size,  $T$ , is assumed fixed. The leading scalar clearly vanishes with increasing  $n$ . The matrix in the square brackets is  $T$  times the average over the  $n$  groups of the within-groups covariance matrices of the variables in  $\mathbf{X}_i$ . So long as the data are well behaved, we can assume that the bracketed matrix does not converge to a zero matrix (or a matrix with zeros on the diagonal). On this basis, we can expect consistency of the least squares estimator. In practical terms, this requires within-groups variation of the data. Notice that the result falls apart if there are time invariant variables in  $\mathbf{X}_i$ , because then there are zeros on the diagonals of the bracketed matrix. This result also suggests the nature of the problem of the WHO data in Example 11.4 as analyzed by Gravelle et al. (2002).

Now, consider the result in (11-19) for the asymptotic variance of  $a_i$ . Assume that  $\mathbf{b}$  is consistent, as shown previously. Then, with increasing  $n$ , the asymptotic variance of  $a_i$  declines to a lower bound of  $\sigma_e^2/T$  which does not converge to zero. The constant term estimators in the fixed effects model are not consistent estimators of  $\alpha_i$ . They are not inconsistent because they gravitate toward the wrong parameter. They are so because their asymptotic variances do not converge to zero, even as the sample size grows. It is easy to see why this is the case. From (11-16), we see that each  $a_i$  is estimated using only  $T$  observations—assume  $n$  were infinite, so that  $\beta$  were known. Because  $T$  is not assumed to be increasing, we have the surprising result. The constant terms are inconsistent unless  $T \rightarrow \infty$ , which is not part of the model.

### 11.4.3 TESTING THE SIGNIFICANCE OF THE GROUP EFFECTS

The  $t$  ratio for  $\alpha_i$  can be used for a test of the hypothesis that  $\alpha_i$  equals zero. This hypothesis about one specific group, however, is typically not useful for testing in this regression context. If we are interested in differences across groups, then we can test the hypothesis that the constant terms are all equal with an  $F$  test. Under the null hypothesis of equality, the efficient estimator is pooled least squares. The  $F$  ratio used for this test is

$$F(n-1, nT-n-K) = \frac{(R_{LSDV}^2 - R_{Pooled}^2)/(n-1)}{(1-R_{LSDV}^2)/(nT-n-K)}, \quad (11-21)$$

where  $LSDV$  indicates the dummy variable model and  $Pooled$  indicates the pooled or restricted model with only a single overall constant term. Alternatively, the model may have been estimated with an overall constant and  $n-1$  dummy variables instead. All other results (i.e., the least squares slopes,  $s^2$ ,  $R^2$ ) will be unchanged, but rather than estimate  $\alpha_i$ , each dummy variable coefficient will now be an estimate of  $\alpha_i - \alpha_1$  where group “1” is the omitted group. The  $F$  test that the coefficients on these  $n-1$  dummy variables are zero is identical to the one above. It is important to keep in mind, however, that although the statistical results are the same, the interpretation of the dummy variable coefficients in the two formulations is different.<sup>9</sup>

### 11.4.4 FIXED TIME AND GROUP EFFECTS

The least squares dummy variable approach can be extended to include a time-specific effect as well. One way to formulate the extended model is simply to add the time effect, as in

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i + \delta_t + \varepsilon_{it}. \quad (11-22)$$

This model is obtained from the preceding one by the inclusion of an additional  $T-1$  dummy variables. (One of the time effects must be dropped to avoid perfect collinearity—the group effects and time effects both sum to one.) If the number of variables is too large to handle by ordinary regression, then this model can also be estimated by using the partitioned regression.<sup>10</sup> There is an asymmetry in this formulation, however, since each of the group effects is a group-specific intercept, whereas the time effects are **contrasts**—that is, comparisons to a base period (the one that is excluded). A symmetric form of the model is

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \mu + \alpha_i + \delta_t + \varepsilon_{it}, \quad (11-23)$$

where a full  $n$  and  $T$  effects are included, but the restrictions

$$\sum_i \alpha_i = \sum_t \delta_t = 0$$

<sup>9</sup>For a discussion of the differences, see Suits (1984).

<sup>10</sup>The matrix algebra and theoretical development of two-way effects in panel data models are complex. See, for example, Baltagi (2001). Fortunately, the practical application is much simpler. The number of periods analyzed in most panel data sets is rarely more than a handful. Because modern computer programs uniformly allow dozens (or even hundreds) of regressors, almost any application involving a second fixed effect can be handled just by literally including the second effect as a set of actual dummy variables.

### 364 PART II ♦ Generalized Regression Model and Equation Systems

are imposed. Least squares estimates of the slopes in this model are obtained by regression of

$$y_{*it} = y_{it} - \bar{y}_{i\cdot} - \bar{y}_{\cdot t} + \bar{\bar{y}} \quad (11-24)$$

on

$$\mathbf{x}_{*it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot t} + \bar{\bar{\mathbf{x}}},$$

where the period-specific and overall means are

$$\bar{y}_{\cdot t} = \frac{1}{n} \sum_{i=1}^n y_{it} \quad \text{and} \quad \bar{\bar{y}} = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T y_{it},$$

and likewise for  $\bar{\mathbf{x}}_{i\cdot}$  and  $\bar{\bar{\mathbf{x}}}$ . The overall constant and the dummy variable coefficients can then be recovered from the normal equations as

$$\begin{aligned} \hat{\mu} &= m = \bar{\bar{y}} - \bar{\bar{\mathbf{x}}} \mathbf{b}, \\ \hat{\alpha}_i &= a_i = (\bar{y}_{i\cdot} - \bar{\bar{y}}) - (\bar{\mathbf{x}}_{i\cdot} - \bar{\bar{\mathbf{x}}})' \mathbf{b}, \\ \hat{\delta}_t &= d_t = (\bar{y}_{\cdot t} - \bar{\bar{y}}) - (\bar{\mathbf{x}}_{\cdot t} - \bar{\bar{\mathbf{x}}})' \mathbf{b}. \end{aligned} \quad (11-25)$$

The estimated asymptotic covariance matrix for  $\mathbf{b}$  is computed using the sums of squares and cross products of  $\mathbf{x}_{*it}$  computed in (11-22) and

$$s^2 = \frac{\sum_{i=1}^n \sum_{t=1}^T (y_{it} - \mathbf{x}'_{it} \mathbf{b} - m - a_i - d_t)^2}{nT - (n-1) - (T-1) - K - 1} \quad (11-26)$$

If one of  $n$  or  $T$  is small and the other is large, then it may be simpler just to treat the smaller set as an ordinary set of variables and apply the previous results to the one-way fixed effects model defined by the larger set. Although more general, this model is infrequently used in practice. There are two reasons. First, the cost in terms of degrees of freedom is often not justified. Second, in those instances in which a model of the timewise evolution of the disturbance is desired, a more general model than this simple dummy variable formulation is usually used.

#### 11.4.5 TIME-INVARIANT VARIABLES AND FIXED EFFECTS VECTOR DECOMPOSITION

The presence of time-invariant variables (TIVs) in the common effects regression presents a vexing problem for the model builder. The significant problem for the *fixed effects model* (FEM) is that the estimator cannot accommodate TIVs. Thus, in the wage equation in Example 11.5, we have omitted three variables of considerable interest from the fixed effects model, *Ed*, *Fem*, and *Blk*. If we write the FEM with a set of time-invariant variables in it as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{D}\boldsymbol{\alpha} + \boldsymbol{\varepsilon},$$

with  $\mathbf{Z}$  being the matrix of  $M$  TIVs, then the problem becomes one of multicollinearity. Since the columns of matrix  $\mathbf{D}$  are a complete set of  $n$  dummy variables, any time-invariant variable in  $\mathbf{Z}$  can be written as a linear combination of the columns of  $\mathbf{D}$ . Let the  $m$ th column of  $\mathbf{Z}$  be the TIV,  $\mathbf{Z}(m) = (z_{m1}, z_{m1}, \dots, z_{m2}, z_{m2}, \dots, \dots, z_{mn}, z_{mn}, \dots)'$ ; each specific value,  $z_{mi}$ , is repeated  $T_i$  times. Then  $\mathbf{Z}(m)$  equals

$\mathbf{Dz}_m$  where  $\mathbf{z}_m$  is the  $n \times 1$  vector  $(z_{m1}, z_{m2}, \dots, z_{mn})'$ . Collecting all  $M$  columns, we have  $\mathbf{Z} = \mathbf{DZ}_n$  where  $\mathbf{Z}_n$  is the  $n \times m$  matrix  $(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m)$ .

We attempt to compute the LSDV estimator of  $(\beta', \gamma')'$  of (11-14) using the transformed variables  $\mathbf{M}_D[\mathbf{X}, \mathbf{Z}]$ , the columns of  $\mathbf{Z}$  are transformed to deviations from group means, which are columns of zeros, since  $\mathbf{Z}$  is already the period means, and the transformed data matrix becomes  $(\mathbf{M}_D\mathbf{X}, \mathbf{0})$ —since  $\mathbf{Z}$  is already in the form of group means, deviations from group means are zero. The LSDV regression cannot be computed with TIVs. In theoretical terms, the problem is that  $\gamma$  is not identified. No amount of data can disentangle  $\gamma$  from  $\alpha$ . The model would be

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{D}(\mathbf{Z}_n\gamma) + \mathbf{D}\alpha + \varepsilon = \mathbf{X}\beta + \mathbf{D}[\mathbf{Z}_n\gamma + \alpha] + \varepsilon.$$

In the fixed effects case, the identifying restriction is  $\gamma = \mathbf{0}$ . That is, in a fixed effects model, the coefficients on TIVs are not identified in terms of the moments of the data so their coefficients are fixed at zero, so as to identify  $\alpha$ .

Plümper and Troeger (2007) have proposed a three-step procedure that they label **Fixed effects vector decomposition** (FEVD) that suggests a solution to the problem of estimating coefficients on TIVs in a fixed effects model and, at the same time, brings noticeable gains in the efficiency of estimation of the parameters. The three steps are

**Step 1:** Linear regression of  $\mathbf{y}$  on  $\mathbf{X} \setminus \mathbf{D}$  to estimate  $\alpha$ . That is, compute the LSDV estimator of  $\beta$  in (11-14) and use (11-17) to compute estimates of the individual constant terms.

**Step 2:** Linear regression of the  $n$  estimated constant terms,  $a_i, i = 1, \dots, n$ , on a constant term and  $\mathbf{Z}_n$ . From this regression, we compute the  $n$  residuals,  $\mathbf{h}_n$ . We then expand this vector to the full sample length using  $\mathbf{h} = \mathbf{D}\mathbf{h}_n$ .

**Step 3:** Linear regression of  $\mathbf{y}$  on  $\mathbf{X} \setminus (\mathbf{i}, \mathbf{Z}), \mathbf{h}$ , where  $\mathbf{i}$  is an overall constant term, to estimate  $(\beta, \alpha, \gamma, \delta)$ :  $\mathbf{y} = \mathbf{X}\beta + \alpha + \mathbf{Z}\gamma + \mathbf{h}\delta + \varepsilon$ .

The suggestion produces some interesting algebraic results that will be instructive for the analysis of this chapter. The surprising result that has apparently gone unnoticed in dozens of recent applications of the technique, but not in several recent comments including Breusch, Ward, Nguyen, and Kompas (2010), Chatelain and Ralf (2010), and Greene (2010), is that step 3 simply reproduces the results in steps 1 and 2, but the covariance matrix computed for the estimator of  $\beta$  at step 3 is not identical and is unambiguously too small. It is instructive to work through a derivation in detail.

We will prove the following results:

FEVD.1 The estimated coefficients on  $\mathbf{X}$  at step 3 are identical to those at step 1.

FEVD.2 The estimated coefficients on  $(\mathbf{i}, \mathbf{Z})$  at step 3 are identical to those at step 2.

FEVD.3 The estimated coefficient on  $\mathbf{h}$  at step 3 is equal to 1.0.

FEVD.4 The sum of squared residuals in the regression at step 3 is identical to that at step 1.

FEVD.5 The  $s^2$  computed at step 3 is less than that at step 1.

FEVD.6 The asymptotic covariance matrix computed for the estimator of  $\beta$  at step 3 is smaller than that at step 1 (even though the estimates are algebraically identical) because of FEVD.5 and because the matrix used is smaller.

(Note there are much more compact proofs of these results. The following approaches are used to demonstrate the tools we have developed in this and the preceding chapters.)

## 366 PART II ♦ Generalized Regression Model and Equation Systems

Proofs of results: Write the results of the three least squares regressions as

$$(Step\ 1)\ \mathbf{y} = \mathbf{X}\mathbf{b}_{LSDV} + \mathbf{D}\mathbf{a}_{LSDV} + \mathbf{e}_{LSDV},$$

$$(Step\ 2)\ \mathbf{a}_{LSDV} = \mathbf{W}_n\mathbf{c}_{LSDV} + \mathbf{h}_n \text{ where } \mathbf{W}_n = (\mathbf{i}_n, \mathbf{Z}_n),$$

$$(Step\ 3)\ \mathbf{y} = \mathbf{X}\mathbf{b}_{FEVD} + \mathbf{W}\mathbf{c}_{FEVD} + \mathbf{h}_{FEVD} + \mathbf{e}_{FEVD}, \text{ where } \mathbf{W} = (\mathbf{i}, \mathbf{Z}).$$

Thus,  $\mathbf{W}$  at step 3 includes the  $M$  time-invariant variables and an overall constant. To begin, we will establish that  $\mathbf{e}_{LSDV} = \mathbf{e}_{FEVD}$ . Recall that  $\mathbf{Z} = \mathbf{D}\mathbf{Z}_n$  and  $\mathbf{i} = \mathbf{D}\mathbf{i}_n$ , where  $\mathbf{i}_n$  is an  $n \times 1$  column vector of ones. The residuals in (step 2) are  $\mathbf{h}_n = \mathbf{a}_{LSDV} - \mathbf{W}_n\mathbf{c}_{LSDV}$  and  $\mathbf{h} = \mathbf{D}\mathbf{h}_n$ . Therefore, the result at step 3) is equivalent to

$$\mathbf{y} = \mathbf{X}\mathbf{b}_{FEVD} + \mathbf{DW}_n\mathbf{c}_{FEVD} + \mathbf{D}(\mathbf{a}_{LSDV} - \mathbf{W}_n\mathbf{c}_{LSDV})d_{FEVD} + \mathbf{e}_{FEVD}.$$

Rearranging it slightly,

$$\mathbf{y} = \mathbf{X}\mathbf{b}_{FEVD} + \mathbf{Da}_{LSDV} + \mathbf{DW}_n\mathbf{c}_{FEVD} - \mathbf{DW}_n\mathbf{c}_{LSDV}(d_{FEVD}) + \mathbf{e}_{FEVD}. \quad (11-27)$$

The first two terms are the predictions from the linear regression of  $\mathbf{y}$  on  $\mathbf{X}$  and  $\mathbf{D}$  and the third and fourth terms simply add more linear combinations of the columns of  $\mathbf{D}$ . Since  $(\mathbf{X}, \mathbf{D})$  has (we have assumed) full column rank, least squares regression (11-27) must provide the same results as step 1. The residuals must be identical; that is  $\mathbf{e}_{FEVD} = \mathbf{e}_{LSDV}$ . Now, premultiply (11-27) by  $\mathbf{X}'\mathbf{M}_D$ . Since  $\mathbf{M}_D\mathbf{D} = \mathbf{0}$  and  $\mathbf{M}_D\mathbf{e}_{LSDV} = \mathbf{e}_{LSDV}$ , we find

$$\mathbf{X}'\mathbf{M}_D\mathbf{y} = \mathbf{X}'\mathbf{M}_D\mathbf{X}\mathbf{b}_{FEVD} + \mathbf{X}'\mathbf{e}_{LSDV}.$$

Since  $\mathbf{X}'\mathbf{e}_{LSDV} = \mathbf{0}$  (from step 1), we have  $\mathbf{b}_{FEVD} = (\mathbf{X}'\mathbf{M}_D\mathbf{X})^{-1}(\mathbf{X}'\mathbf{M}_D\mathbf{y}) = \mathbf{b}_{LSDV}$  which proves FEVD.1.

To compute  $\mathbf{c}_{FEVD}$ , at step 3, we have at the solution (using  $\mathbf{b}_{FEVD} = \mathbf{b}_{LSDV}$  and  $\mathbf{e}_{FEVD} = \mathbf{e}_{LSDV}$ )

$$\mathbf{y} - \mathbf{X}\mathbf{b}_{LSDV} = \mathbf{W}\mathbf{c}_{FEVD} + \mathbf{h}_{FEVD} + \mathbf{e}_{LSDV}.$$

Premultiply this expression by  $\mathbf{W}'$ . From step 2,  $\mathbf{W}'\mathbf{h} = \mathbf{W}_n'\mathbf{D}'\mathbf{D}\mathbf{h}_n = \mathbf{0}$ . This is true because  $\mathbf{D}'\mathbf{D}$  is a diagonal matrix with  $T_i$  on the diagonals. Thus, each element in  $\mathbf{W}'\mathbf{h}$  is  $T_i\mathbf{W}(m)'h_n = 0$ , where  $\mathbf{W}(m)$  is the  $m$ th column of  $\mathbf{W}_n$ . From step 3,  $\mathbf{W}'\mathbf{e}_{FEVD} = \mathbf{W}'\mathbf{e}_{LSDV} = \mathbf{0}$ . Thus,

$$\mathbf{W}'(\mathbf{y} - \mathbf{X}\mathbf{b}_{LSDV}) = \mathbf{W}'\mathbf{W}\mathbf{c}_{FEVD}$$

so

$$\mathbf{c}_{FEVD} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'(\mathbf{y} - \mathbf{X}\mathbf{b}_{LSDV}).$$

From step 1,  $\mathbf{y} - \mathbf{X}\mathbf{b}_{LSDV} = \mathbf{D}\mathbf{a}_{LSDV} + \mathbf{e}_{LSDV}$ . Since  $\mathbf{W}'\mathbf{e}_{FEVD} = \mathbf{W}'\mathbf{e}_{LSDV} = \mathbf{0}$ , from step 3,

$$\mathbf{c}_{FEVD} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{D}\mathbf{a}_{LSDV}.$$

But, by premultiplying step 2 by  $\mathbf{D}$ , we find  $\mathbf{D}\mathbf{a}_{LSDV} = \mathbf{DW}_n\mathbf{c}_{LSDV} + \mathbf{D}\mathbf{h}_n$ . It follows that the solution is

$$\mathbf{c}_{LSDV} = (\mathbf{W}_n'\mathbf{D}'\mathbf{D}\mathbf{W}_n)^{-1}\mathbf{W}_n'\mathbf{D}'\mathbf{D}\mathbf{a}_{LSDV} + (\mathbf{W}_n'\mathbf{D}'\mathbf{D}\mathbf{W}_n)^{-1}\mathbf{W}_n'\mathbf{D}'\mathbf{D}\mathbf{h}_n.$$

The second term is zero as shown earlier. The end result is  $\mathbf{c}_{LSDV} = \mathbf{c}_{FEVD}$  which is FEVD.2.

CHAPTER 11 ♦ Models for Panel Data **367**

Once again using step 3, we now solve for  $d_{FEVD}$  using what we already have. The solution is in

$$\mathbf{y} - \mathbf{X}\mathbf{b}_{LSDV} - \mathbf{W}\mathbf{c}_{LSDV} = \mathbf{h}d_{FEVD} + \mathbf{e}_{LSDV}.$$

But,  $\mathbf{y} - \mathbf{X}\mathbf{b}_{LSDV} = \mathbf{a} + \mathbf{e}_{LSDV} = \mathbf{D}\mathbf{a}_{LSDV} + \mathbf{e}_{LSDV}$  and  $\mathbf{W}\mathbf{c}_{LSDV} = \mathbf{a} - \mathbf{h} = \mathbf{D}\mathbf{a}_{LSDV} - \mathbf{h}$ . Inserting these,

$$\mathbf{D}\mathbf{a}_{LSDV} + \mathbf{e}_{LSDV} - \mathbf{D}\mathbf{a}_{LSDV} + \mathbf{h} = \mathbf{h}d_{FEVD} + \mathbf{e}_{LSDV}$$

or

$$\mathbf{h} + \mathbf{e}_{LSDV} = \mathbf{h}d_{FEVD} + \mathbf{e}_{LSDV},$$

from which it follows that  $d_{FEVD} = 1$ . This proves FEVD.3.

FEVD.4 has already been shown since  $\mathbf{e}_{FEVD} = \mathbf{e}_{LSDV}$ . The  $R^2$ 's in the two regressions are the same as well, as  $R_{FEVD}^2 = 1 - (\mathbf{e}_{FEVD}'\mathbf{e}_{FEVD}/\mathbf{y}'\mathbf{M}^0\mathbf{y}) = R_{LSDV}^2$  since the residual vectors are identical. [See (3-26).] But,

$$s_{FEVD}^2 = \mathbf{e}_{FEVD}'\mathbf{e}_{FEVD}/(\Sigma K - M - 1 - 1) < s_{LSDV}^2 = \mathbf{e}_{LSDV}'\mathbf{e}_{LSDV}/(\Sigma_i T_i - K - n).$$

The difference is the degrees of freedom correction, which can be large. In our example to follow,  $DF_{FEVD} = 4165 - 9 - 3 - 1 - 1 = 4151$  while  $DF_{LSDV} = 4165 - 9 - 595 = 3561$ . For the example, then,  $s_{FEVD}^2/s_{LSDV}^2 = 0.85787$ . This establishes FEVD.5.

To establish FEVD.6, based on (11-17), we are going to compare

$$\text{Est.Asy.Var}[\mathbf{b}_{FEVD}] = s_{FEVD}^2 (\mathbf{X}'\mathbf{M}_{W,h}\mathbf{X})^{-1}$$

to

$$\text{Est.Asy.Var}[\mathbf{b}_{LSDV}] = s_{LSDV}^2 (\mathbf{X}'\mathbf{M}_D\mathbf{X})^{-1}.$$

We have already established that  $s_{LSDV}^2 > s_{FEVD}^2$ . To compare the matrices, we will compare their inverses, and show that the difference matrix

$$\mathbf{A} = \mathbf{X}'\mathbf{M}_{W,h}\mathbf{X} - \mathbf{X}'\mathbf{M}_D\mathbf{X}$$

is positive definite. This will imply that the inverse matrix in  $\text{Est.Asy.Var}[\mathbf{b}_{FEVD}]$  is smaller than that in  $\text{Est.Asy.Var}[\mathbf{b}_{LSDV}]$ . To show this, we note that  $\mathbf{R} = (\mathbf{W}, \mathbf{h}) = \mathbf{D}(\mathbf{W}_n, \mathbf{h}_n)$  is  $M$  linear combinations of the columns of  $\mathbf{D}$  while  $\mathbf{D}$  is all  $n$  columns of  $\mathbf{D}$ . For convenience, let  $\mathbf{R} = (\mathbf{W}, \mathbf{h})$ . The residuals defined by  $\mathbf{M}_D\mathbf{X}$  [see (3-15)] are obtained by regressions of  $\mathbf{X}$  on all  $n$  columns of  $\mathbf{D}$ . They will be identical to the residuals obtained by regression of  $\mathbf{X}$  on any  $n$  linearly independent combinations of the columns of  $\mathbf{D}$ . For these, we will use  $[\mathbf{R}, \mathbf{Q}]$  where  $\mathbf{Q}$  is orthogonal to  $\mathbf{R}$ . Therefore  $\mathbf{X}'\mathbf{M}_D\mathbf{X} = \mathbf{X}'\mathbf{M}_{R,Q}\mathbf{X}$ . Expanding this, we have

$$\mathbf{A} = \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\mathbf{X} - \mathbf{X}'\mathbf{X} + \mathbf{X}'(\mathbf{R}'\mathbf{Q}) \left[ \begin{pmatrix} \mathbf{R}' \\ \mathbf{Q}' \end{pmatrix} (\mathbf{R}'\mathbf{Q}) \right]^{-1} \begin{pmatrix} \mathbf{R}' \\ \mathbf{Q}' \end{pmatrix} \mathbf{X}.$$

The inverse matrix is simplified by  $\mathbf{R}'\mathbf{Q} = \mathbf{0}$ , so the bracketed matrix and its inverse are block diagonal. Multiplying it out, we find

$$\mathbf{A} = \mathbf{X}'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{X} = \mathbf{X}'(\mathbf{I} - \mathbf{M}_Q)\mathbf{X}.$$

### 368 PART II ♦ Generalized Regression Model and Equation Systems

Since  $\mathbf{I} - \mathbf{M}_Q$  is idempotent,  $\mathbf{A} = \mathbf{X}'(\mathbf{I} - \mathbf{M}_Q)'(\mathbf{I} - \mathbf{M}_Q)\mathbf{X} = \mathbf{X}^*\mathbf{X}^*$  is positive definite. This establishes that the computed covariance matrix for  $\mathbf{b}_{FEVD}$  will always be strictly smaller than that for  $\mathbf{b}_{LSDV}$ , which is FEVD.6.

This leaves what should appear to be a loose end in the analysis. How was it possible to estimate  $\gamma$  (in step 2 or step 3) given that it is unidentified in the original model? The answer is the crucial assumption previously noted at several points. From the original specification  $\mathbf{Z}$  is uncorrelated with  $\epsilon$ . But, for the regression (in step 2) estimate a nonzero  $\gamma$ , *it must be further assumed that  $\mathbf{z}_i$  is uncorrelated with  $u_i$* . This restricts the original fixed effects model—it is a hybrid in which the time-varying variables are allowed to be correlated with  $u_i$  but the time-invariant variables are not. The authors note this on page 6 and in their footnote 7 where they state, “If the time-invariant variables are assumed to be orthogonal to the unobserved unit effects—i.e., if the assumption underlying our estimator is correct—the estimator is consistent. If this assumption is violated, the estimated coefficients for the time-invariant variables are biased.... Note that the estimated coefficients of the time-varying variables remain unbiased even in the presence of correlated unit effects. However, the assumption underlying a FE model must be satisfied (*no correlated time-varying variables may exist*).” (Emphasis added—it seems that “varying” should be “invariant”) There are other estimators that would consistently be  $\beta$  and  $\gamma$  in this revised model, including the Hausman and Taylor estimator discussed in Section 11.8.1 and instrumental variables estimators suggested by Breusch et al. (2010) and by Chatelain and Ralf (2010).

The problem of primary interest in Plümper and Troeger was an intermediate case somewhat different from what we have examined here. There are two directions of the work. If only some of the elements of  $\mathbf{Z}$  but not all of them, are correlated with  $u_i$ , then we obtain the setting analyzed by Hausman and Taylor that is examined in Section 11.8.1. Plümper and Troeger’s FEVD estimator will, in that instance, be an inconsistent estimator that may have a smaller variance than the IV estimator proposed by Hausman and Taylor. The second case the authors are interested in is when  $\mathbf{Z}$  is not strictly time invariant but is “slowly changing.” When there is very little within-groups variation, for example, as shown for the World Health Organization data in Example 11.4, then, once again, the estimator suggested here may have some advantages over instrumental variables and other treatments. In that case, when there are no strictly time-invariant variables in the model, then the analysis is governed by the random effects model discussed in the next section.

#### **Example 11.5 Fixed Effects Wage Equation**

Table 11.5 presents the estimated wage equation with individual effects for the Cornwell and Rupert data used in Examples 11.1 and 11.3. The model includes three time-invariant variables, *Ed*, *Fem*, *Blk*, that must be dropped from the equation. As a consequence, the fixed effects estimates computed here are not comparable to the results for the pooled model already examined. For comparison, the least squares estimates with panel robust standard errors are also presented. We have also added a set of time dummy variables to the model. The *F* statistic for testing the significance of the individual effects based on the, *R*<sup>2</sup>'s for the equations is

$$F[594, 3561] = \frac{(0.9072422 - 0.3154548)/594}{(1 - 0.9072422)/(4165 - 9 - 595)} = 38.247$$

The critical value for the *F* table with 594 and 3561 degrees of freedom is 1.106, so the evidence is strongly in favor of an individual-specific effect. As often happens, the fit of the

**TABLE 11.5** Fixed Effects Estimates of the Cornwell and Rupert Wage Equation

Variable	Estimate	Std.Error*	Time Effects		Individual Effects		Time and Ind. Effects		FEVD Step 3	
			Pooled	Estimate	Std.Error*	Estimate	Std.Error	Estimate	Std.Err	Estimate
Constant	5.8802	0.09654	5.6963	0.09425	0.1132	0.002471 (0.00437)	0.1114	0.002618	0.1132	0.00100
Exp	0.03611	0.0045241	0.02738	0.004556	-0.00042	0.00055 (0.00089)	-0.00004	0.000054	-0.00042	0.000192
Exp <sup>2</sup>	-0.00066	0.0001013	-0.00053	0.000101	0.00084	0.000600 (0.00094)	0.00068	0.0005991	0.00084	0.00044
Wks	0.00446	0.001725	0.00409	0.001694	-0.02148	0.01378 (0.02052)	-0.01916	0.01275	-0.02148	0.00596
Occ	-0.3176	0.02721	-0.3045	0.02684	0.01921	0.01545 (0.02450)	0.02076	0.1540	0.01921	0.00476
Ind	0.03213	0.02521	0.04010	0.02489	-0.00186	0.03430 (0.09646)	0.00309	0.03419	-0.00186	0.00506
South	-0.1137	0.028626	-0.1157	0.02834	-0.04247	0.01942 (0.03185)	-0.04188	0.01937	-0.04247	0.00504
SMSA	0.1586	0.025967	0.1722	0.02566	-0.02973	0.01898 (0.02902)	-0.02856	0.018918	-0.02973	0.00831
MS	0.3203	0.03487	0.3425	0.03459	0.03278	0.01492 (0.02708)	0.02952	0.01488	0.03278	0.00517
Union	0.06975	0.026618	0.06272	0.02578	2.8286 -0.13003	0.18599 0.12557	2.8286 -0.13003	0.03315	2.8286 -0.13003	0.03315
Constant					Ed	0.14438 0.01403	Ed	0.01024	Blk	0.14438 0.00121
Fem					Blk	-0.27507 0.15440	Blk	0.00891		0.00891
Year 1	0.0000	0.0000					0.00000	0.00000		0.00683
Year 2	0.07812	0.006860					-0.00775	0.008167		
Year 3	0.2050	0.01072					0.02557	0.007769		
Year 4	0.2926	0.01125					0.02845	0.007639		
Year 5	0.3724	0.01095					0.02418	0.007772		
Year 6	0.4498	0.01245					0.00737	0.008161		
Year 7	0.5422	0.013015					0.00000	0.00000		
e <sup>e</sup>	607.1265	475.6659					82.26732	82.26732		
Deg.Free	4155	4149					3561	3557		
s	0.3822588	0.3385940					0.1514089	0.1514089		
R <sup>2</sup>	0.3154548	0.4636788					0.9072422	0.9072422		

\*Robust standard errors using (11-3) including finite population correction  $[(\sum_i T_i) - 1]/[(\sum_i T_i) - K - m] \times n/(n - 1)$ .

## 370 PART II ♦ Generalized Regression Model and Equation Systems

model increases greatly when the individual effects are added. We have also added time effects to the model. The model with time effects without the individual effects is in the second column results. The  $F$  statistic for testing the significance of the time effects (in the absence of the individual effects) is

$$F[6, 4149] = \frac{(0.4636788 - 0.3154548)/6}{(1 - 0.4636788)/(4165 - 10 - 6)} = 191.11,$$

The critical value from the  $F$  table is 2.101, so the hypothesis that the time effects are zero is also rejected. The last column of results shows the model with both time and individual effects. For this model it is necessary to drop a second time effect because the experience variable,  $Exp$ , is an individual specific time trend. The  $Exp$  variable can be expressed as

$$Exp_{i,t} = E_{i,0} + (t - 1), t = 1, \dots, 7,$$

which can be expressed as a linear combination of the individual dummy variable and the six time variables. For the last model, we have dropped the first and last of the time effects. In this model, the  $F$  statistic for testing the significance of the time effects is

$$F[5, 3556] = \frac{(0.9080847 - 0.9072422)/5}{(1 - 0.9080847)/(4165 - 9 - 5 - 5595)} = 6.519.$$

The time effects remain significant—the critical value is 2.217—but the test statistic is considerably reduced. The time effects reveal a striking pattern. In the equation without the individual effects, we find a steady increase in wages of 7–9 percent per year. But, when the individual effects are added to the model, this progression disappears.

It might seem appropriate to compute the robust standard errors for the fixed effects estimator as well as for the pooled estimator. However, in principle, that should be unnecessary. If the model is correct and completely specified, then the individual effects should be capturing the omitted heterogeneity, and what remains is a classical, homoscedastic, nonautocorrelated disturbance. This does suggest a rough indicator of the appropriateness of the model specification. If the conventional asymptotic covariance matrix in (11-17) and the robust estimator in (11-3), with  $\mathbf{X}_i$  replaced with the data in group mean deviations form, give very different estimates, one might question the model specification. [This is the logic that underlies White's (1982a) information matrix test (and the extensions by Newey (1985a) and Tauchen (1985).] The robust standard errors are shown in parentheses under those for the fixed effects estimates in the sixth column of Table 11.5. They are considerably higher than the uncorrected standard errors—50 percent to 100 percent—which might suggest that the fixed effects specification should be reconsidered.

The FEVD computations are shown in Table 11.5 as well. The third set of results, marked “Individual Effects,” shows the step 1 and step 2 results. Note that these are computed in two least squares regressions. The nd step is indicated by the heavy box. The fit measures are not shown for this intermediate step. The step 3 results are shown in the last two columns of the table. As anticipated, the estimated coefficients match the first and second step regressions. For  $\mathbf{b}_{LSDV}$ , the standard errors have fallen by a factor of 2 to 4. For  $\mathbf{c}_{LSDV}$ , the estimators of  $\gamma$ , they have fallen by a factor of 7 to 10. In view of the previous analytic results, the estimates in the last column of Table 11.5 would be viewed as overly optimistic.

### 11.5 RANDOM EFFECTS

The fixed effects model allows the unobserved individual effects to be correlated with the included variables. We then modeled the differences between units strictly as parametric shifts of the regression function. This model might be viewed as applying only to the

CHAPTER 11 ♦ Models for Panel Data **371**

cross-sectional units in the study, not to additional ones outside the sample. For example, an intercountry comparison may well include the full set of countries for which it is reasonable to assume that the model is constant. If the individual effects are strictly uncorrelated with the regressors, then it might be appropriate to model the individual specific constant terms as randomly distributed across cross-sectional units. This view would be appropriate if we believed that sampled cross-sectional units were drawn from a large population. It would certainly be the case for the longitudinal data sets listed in the introduction to this chapter.<sup>11</sup> The payoff to this form is that it greatly reduces the number of parameters to be estimated. The cost is the possibility of inconsistent estimates, should the assumption turn out to be inappropriate.

Consider, then, a reformulation of the model

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + (\alpha + u_i) + \varepsilon_{it}, \quad (11-28)$$

where there are  $K$  regressors including a constant and now the single constant term is the mean of the unobserved heterogeneity,  $E[\mathbf{z}'_i\boldsymbol{\alpha}]$ . The component  $u_i$  is the random heterogeneity specific to the  $i$ th observation and is constant through time; recall from Section 11.2.1,  $u_i = \{\mathbf{z}'_i\boldsymbol{\alpha} - E[\mathbf{z}'_i\boldsymbol{\alpha}]\}$ . For example, in an analysis of families, we can view  $u_i$  as the collection of factors,  $\mathbf{z}'_i\boldsymbol{\alpha}$ , not in the regression that are specific to that family. We continue to assume strict exogeneity:

$$\begin{aligned} E[\varepsilon_{it} | \mathbf{X}] &= E[u_i | \mathbf{X}] = 0, \\ E[\varepsilon_{it}^2 | \mathbf{X}] &= \sigma_\varepsilon^2, \\ E[u_i^2 | \mathbf{X}] &= \sigma_u^2, \\ E[\varepsilon_{it}u_j | \mathbf{X}] &= 0 \quad \text{for all } i, t, \text{ and } j, \\ E[\varepsilon_{it}\varepsilon_{js} | \mathbf{X}] &= 0 \quad \text{if } t \neq s \text{ or } i \neq j, \\ E[u_iu_j | \mathbf{X}] &= 0 \quad \text{if } i \neq j. \end{aligned} \quad (11-29)$$

As before, it is useful to view the formulation of the model in blocks of  $T$  observations for group  $i$ ,  $\mathbf{y}_i$ ,  $\mathbf{X}_i$ ,  $u_i\mathbf{i}$ , and  $\boldsymbol{\varepsilon}_i$ . For these  $T$  observations, let

$$\eta_{it} = \varepsilon_{it} + u_i$$

and

$$\boldsymbol{\eta}_i = [\eta_{i1}, \eta_{i2}, \dots, \eta_{iT}]'$$

In view of this form of  $\eta_{it}$ , we have what is often called an **error components model**. For this model,

$$\begin{aligned} E[\eta_{it}^2 | \mathbf{X}] &= \sigma_\varepsilon^2 + \sigma_u^2, \\ E[\eta_{it}\eta_{is} | \mathbf{X}] &= \sigma_u^2, \quad t \neq s \\ E[\eta_{it}\eta_{js} | \mathbf{X}] &= 0 \quad \text{for all } t \text{ and } s \text{ if } i \neq j. \end{aligned} \quad (11-30)$$

<sup>11</sup>This distinction is not hard and fast; it is purely heuristic. We shall return to this issue later. See Mundlak (1978) for methodological discussion of the distinction between fixed and random effects.

## 372 PART II ♦ Generalized Regression Model and Equation Systems

For the  $T$  observations for unit  $i$ , let  $\Sigma = E[\eta_i \eta'_i | \mathbf{X}]$ . Then

$$\Sigma = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \vdots & & & & \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_\varepsilon^2 + \sigma_u^2 \end{bmatrix} = \sigma_\varepsilon^2 \mathbf{I}_T + \sigma_u^2 \mathbf{i}_T \mathbf{i}'_T, \quad (11-31)$$

where  $\mathbf{i}_T$  is a  $T \times 1$  column vector of 1s. Because observations  $i$  and  $j$  are independent, the disturbance covariance matrix for the full  $nT$  observations is

$$\Omega = \begin{bmatrix} \Sigma & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Sigma & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \Sigma \end{bmatrix} = \mathbf{I}_n \otimes \Sigma. \quad (11-32)$$

### 11.5.1 LEAST SQUARES ESTIMATION

The model defined by (11-28),

$$y_{it} = \alpha + \mathbf{x}_{it}'\beta + u_i + \varepsilon_{it},$$

with the strict exogeneity assumptions in (11-29) and the covariance matrix detailed in (11-31) and (11-32) is a generalized regression model that fits into the framework we developed in Chapter 9. The disturbances are autocorrelated in that observations are correlated across time within a group, though not across groups. All the implications of Section 9.2.1 would apply here. In particular, the parameters of the random effects model can be estimated consistently,  not efficiently, by ordinary least squares (OLS). An appropriate robust asymptotic covariance matrix for the OLS estimator would be given by (11-3).

There are other consistent estimators available as well. By taking deviations from group means, we obtain

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \beta + \bar{\varepsilon}_i.$$

This implies that (assuming there are no time-invariant regressors in  $\mathbf{x}_{it}$ ), the LSDV estimator of (11-14) is a consistent estimator of  $\beta$ . (Note that alone among the four estimators to be suggested here, the LSDV estimator is robust to whether the correct specification is actually a random or a fixed  model.) As is OLS, LSDV is inefficient since, as we will show in Section 11.5.2, there is an efficient GLS estimator that is not equal to  $\mathbf{b}_{LSDV}$ . The group means (between groups) regression model,

$$\bar{y}_i = \alpha + \bar{\mathbf{x}}'_{it} \beta + u_i + \bar{\varepsilon}_i, i = 1, \dots, n,$$

provides a third method of consistently estimating the coefficients  $\beta$ . None of these is the preferred estimator in this setting, since the GLS estimator will be more efficient than any of them. However, as we saw in Chapters 9 and 10, many generalized regression models are estimated in two steps, with the first step being a robust least squares regression that is used to produce a first round estimate of the variance parameters of the model. That would be the case here as well. To suggest where this logic will lead in Section 11.5.3, note that for the three cases noted, the mean squared residuals would

produce the following consistent estimators of functions of the variances:

$$\begin{array}{ll} \text{(Pooled)} & \text{plim } [\mathbf{e}_{\text{pooled}}' \mathbf{e}_{\text{pooled}} / (nT)] = \sigma_u^2 + \sigma_\varepsilon^2, \\ \text{(LSDV)} & \text{plim } [\mathbf{e}_{\text{LSDV}}' \mathbf{e}_{\text{LSDV}} / (nT)] = \sigma_\varepsilon^2 [1 - 1/T], \\ \text{(Means)} & \text{plim } [\mathbf{e}_{\text{means}}' \mathbf{e}_{\text{means}} / (nT)] = \sigma_u^2 + \sigma_\varepsilon^2 / T. \end{array}$$

Any pair of these estimators would provide a two-equation method of moments estimator of  $(\sigma_u^2, \sigma_\varepsilon^2)$ . With these in mind, we will now develop an efficient generalized least squares estimator.

### 11.5.2 GENERALIZED LEAST SQUARES

The generalized least squares estimator of the slope parameters is

$$\hat{\beta} = (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{y} = \left( \sum_{i=1}^n \mathbf{X}_i' \boldsymbol{\Sigma}^{-1} \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{X}_i' \boldsymbol{\Sigma}^{-1} \mathbf{y}_i \right).$$

To compute this estimator as we did in Chapter 9 by transforming the data and using ordinary least squares with the transformed data, we will require  $\boldsymbol{\Omega}^{-1/2} = [\mathbf{I}_n \otimes \boldsymbol{\Sigma}]^{-1/2}$ . We need only find  $\boldsymbol{\Sigma}^{-1/2}$ , which is

$$\boldsymbol{\Sigma}^{-1/2} = \frac{1}{\sigma_\varepsilon} \left[ \mathbf{I} - \frac{\theta}{T} \mathbf{i}_T \mathbf{i}_T' \right],$$

where

$$\theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_u^2}}.$$

The transformation of  $\mathbf{y}_i$  and  $\mathbf{X}_i$  for GLS is therefore

$$\boldsymbol{\Sigma}^{-1/2} \mathbf{y}_i = \frac{1}{\sigma_\varepsilon} \begin{bmatrix} y_{i1} - \theta \bar{y}_i \\ y_{i2} - \theta \bar{y}_i \\ \vdots \\ y_{iT} - \theta \bar{y}_i \end{bmatrix}, \quad (11-33)$$

and likewise for the rows of  $\mathbf{X}_i$ .<sup>12</sup> For the data set as a whole, then, generalized least squares is computed by the regression of these partial deviations of  $\mathbf{y}_it$  on the same transformations of  $\mathbf{x}_{it}$ . Note the similarity of this procedure to the computation in the LSDV model, which uses  $\theta = 1$  in (11-15). (One could interpret  $\theta$  as the effect that would remain if  $\sigma_\varepsilon$  were zero, because the only effect would then be  $u_i$ . In this case, the fixed and random effects models would be indistinguishable, so this result makes sense.)

It can be shown that the GLS estimator is, like the pooled OLS estimator, a matrix weighted average of the within- and between-units estimators:

$$\hat{\beta} = \hat{\mathbf{F}}^{\text{within}} \mathbf{b}^{\text{within}} + (\mathbf{I} - \hat{\mathbf{F}}^{\text{within}}) \mathbf{b}^{\text{between}}, \quad (11-34)$$

<sup>12</sup>This transformation is a special case of the more general treatment in Nerlove (1971b).

<sup>13</sup>An alternative form of this expression, in which the weighting matrices are proportional to the covariance matrices of the two estimators, is given by Judge et al. (1985).

### 374 PART II ♦ Generalized Regression Model and Equation Systems

where now,

$$\hat{\mathbf{F}}^{within} = [\mathbf{S}_{xx}^{within} + \lambda \mathbf{S}_{xx}^{between}]^{-1} \mathbf{S}_{xx}^{within},$$

$$\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_u^2} = (1 - \theta)^2.$$

To the extent that  $\lambda$  differs from one, we see that the inefficiency of ordinary least squares will follow from an inefficient weighting of the two estimators. Compared with generalized least squares, ordinary least squares places too much weight on the between-units variation. It includes it all in the variation in  $\mathbf{X}$ , rather than apportioning some of it to random variation across groups attributable to the variation in  $u_i$  across units.

Unbalanced panels add a layer of difficulty in the random effects model. The first problem can be seen in (11-32). The matrix  $\Omega$  is no longer  $\mathbf{I}_n \otimes \Sigma$  because the diagonal blocks in  $\Omega$  are of different sizes. There is also groupwise heteroscedasticity in (11-33), because the  $i$ th diagonal block in  $\Omega^{-1/2}$  is

$$\Sigma_i^{-1/2} = \mathbf{I}_{T_i} - \frac{\theta_i}{T_i} \mathbf{i}_{T_i} \mathbf{i}_{T_i}', \quad \theta_i = 1 - \frac{\sigma_\varepsilon^2}{\sqrt{\sigma_\varepsilon^2 + T_i \sigma_u^2}}.$$

In principle, estimation is still straightforward, because the source of the groupwise heteroscedasticity is only the unequal group sizes. Thus, for GLS, or FGLS with estimated variance components, it is necessary only to use the group-specific  $\theta_i$  in the transformation in (11-33).

#### 11.5.3 FEASIBLE GENERALIZED LEAST SQUARES WHEN $\Sigma$ IS UNKNOWN

If the variance components are known, generalized least squares can be computed as shown earlier. Of course, this is unlikely, so as usual, we must first estimate the disturbance variances and then use an FGLS procedure. A heuristic approach to estimation of the variance components is as follows:

$$y_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it} + u_i \quad (11-35)$$

and

$$\bar{y}_{i.} = \bar{\mathbf{x}}'_{i.} \boldsymbol{\beta} + \bar{\varepsilon}_{i.} + u_i.$$

Therefore, taking deviations from the group means removes the heterogeneity:

$$y_{it} - \bar{y}_{i.} = [\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.}]' \boldsymbol{\beta} + [\varepsilon_{it} - \bar{\varepsilon}_{i.}] \quad (11-36)$$

Because

$$E \left[ \sum_{t=1}^T (\varepsilon_{it} - \bar{\varepsilon}_{i.})^2 \right] = (T-1)\sigma_\varepsilon^2,$$

if  $\boldsymbol{\beta}$  were observed, then an unbiased estimator of  $\sigma_\varepsilon^2$  based on  $T$  observations in group  $i$  would be

$$\hat{\sigma}_\varepsilon^2(i) = \frac{\sum_{t=1}^T (\varepsilon_{it} - \bar{\varepsilon}_{i.})^2}{T-1}. \quad (11-37)$$

CHAPTER 11 ♦ Models for Panel Data **375**

Because  $\beta$  must be estimated—(11-33) implies that the LSDV estimator is consistent, indeed, unbiased in general—we make the degrees of freedom correction and use the LSDV residuals in

$$s_e^2(i) = \frac{\sum_{t=1}^T (e_{it} - \bar{e}_{i\cdot})^2}{T - K - 1}. \quad (11-38)$$

(Note that based on the LSDV estimates,  $\bar{e}_{i\cdot}$  is actually zero. We will carry it through nonetheless to maintain the analogy to (11-34) where  $\bar{\varepsilon}_{i\cdot}$  is not zero but is an estimator of  $E[\varepsilon_{it}] = 0$ .) We have  $n$  such estimators, so we average them to obtain

$$\bar{s}_e^2 = \frac{1}{n} \sum_{i=1}^n s_e^2(i) = \frac{1}{n} \sum_{i=1}^n \left[ \frac{\sum_{t=1}^T (e_{it} - \bar{e}_{i\cdot})^2}{T - K - 1} \right] = \frac{\sum_{i=1}^n \sum_{t=1}^T (e_{it} - \bar{e}_{i\cdot})^2}{nT - nK - n}. \quad (11-39)$$

The degrees of freedom correction in  $\bar{s}_e^2$  is excessive because it assumes that  $\beta$  are reestimated for each  $i$ . The estimated parameters are the  $n$  means  $\bar{y}_{i\cdot}$  and the  $K$  slopes. Therefore, we propose the unbiased estimator<sup>14</sup>

$$\hat{\sigma}_e^2 = s_{LSDV}^2 = \frac{\sum_{i=1}^n \sum_{t=1}^T (e_{it} - \bar{e}_{i\cdot})^2}{nT - n - K}. \quad (11-40)$$

This is the variance estimator in the fixed effects model in (11-35), appropriately corrected for degrees of freedom. It remains to estimate  $\sigma_u^2$ . Return to the original model specification in (11-35). In spite of the correlation across observations, this is a classical regression model in which the ordinary least squares slopes and variance estimators are both consistent and, in most cases, unbiased. Therefore, using the ordinary least squares residuals from the model with only a single overall constant, we have

$$\text{plim } s_{Pooled}^2 = \text{plim } \frac{\mathbf{e}'\mathbf{e}}{nT - K - 1} = \sigma_e^2 + \sigma_u^2. \quad (11-41)$$

This provides the two estimators needed for the variance components; the second would be  $\hat{\sigma}_u^2 = s_{Pooled}^2 - s_{LSDV}^2$ . A possible complication is that this second estimator could be negative. But, recall that for feasible generalized least squares, we do not need an unbiased estimator of the variance, only a consistent one. As such, we may drop the degrees of freedom corrections in (11-40) and (11-41). If so, then the two variance estimators must be nonnegative, since the sum of squares in the LSDV model cannot be larger than that in the simple regression with only one constant term. Alternative estimators have been proposed, all based on this principle of using two different sums of squared residuals.<sup>15</sup> This is a point on which modern software varies greatly. Generally, programs begin with (11-40) and (11-41) to estimate the variance components. What they do next when the estimate of  $\sigma_u^2$  is nonpositive is far from uniform. Dropping the degrees of freedom correction is a frequently used strategy, but at least one widely used program simply sets  $\sigma_u^2$  to zero, and others resort to different strategies based on, for example, the group means estimator. The unfortunate implication for the unwary is that different programs can systematically produce different results using the same

<sup>14</sup>A formal proof of this proposition may be found in Maddala (1971) or in Judge et al. (1985, p. 551).

<sup>15</sup>See, for example, Wallace and Hussain (1969), Maddala (1971), Fuller and Battese (1974), and Amemiya (1971).

## 376 PART II ♦ Generalized Regression Model and Equation Systems

model and the same data. The practitioner is strongly advised to consult the program documentation for resolution.

There is a remaining complication. If there are any regressors that do not vary within the groups, the LSDV estimator cannot be computed. For example, in a model of family income or labor supply, one of the regressors might be a dummy variable for location, family structure, or living arrangement. Any of these could be perfectly collinear with the fixed effect for that family, which would prevent computation of the LSDV estimator. In this case, it is still possible to estimate the random effects variance components. Let  $[\mathbf{b}, a]$  be any consistent estimator of  $[\boldsymbol{\beta}, \alpha]$  in (11-35), such as the ordinary least squares estimator. Then, (11-41) provides a consistent estimator of  $m_{ee} = \sigma_e^2 + \sigma_u^2$ . The mean squared residuals using a regression based only on the  $n$  group means in (11-35) provides a consistent estimator of  $m_{**} = \sigma_u^2 + (\sigma_e^2/T)$ , so we can use

$$\begin{aligned}\hat{\sigma}_e^2 &= \frac{T}{T-1}(m_{ee} - m_{**}) \\ \hat{\sigma}_u^2 &= \frac{T}{T-1}m_{**} - \frac{1}{T-1}m_{ee} = \omega m_{**} + (1-\omega)m_{ee},\end{aligned}$$

where  $\omega > 1$ . As before, this estimator can produce a negative estimate of  $\sigma_u^2$  that, once again, calls the specification of the model into question. [Note, finally, that the residuals in (11-40) and (11-41) could be based on the same coefficient vector.]

There is, perhaps surprisingly, a simpler way out of the dilemma posed by time-invariant regressors. In (11-36), we find that the group mean deviations estimator still provides a consistent estimator of  $\sigma_e^2$ . The time-invariant variables fall out of the model so it is not possible to estimate the full coefficient vector  $\boldsymbol{\beta}$ . But, recall, estimation of  $\boldsymbol{\beta}$  is not the objective at this step, estimation of  $\sigma_e^2$  is. Therefore, it follows that the residuals from the group mean deviations (LSDV) estimator can still be used to estimate  $\sigma_e^2$ . By the same logic, the first differences could also be used. (See Section 11.3.5.) The residual variance in the first difference regression would estimate  $2\sigma_e^2$ . These outcomes are irrespective of whether there are time-invariant regressors in the model.

### 11.5.4 TESTING FOR RANDOM EFFECTS

Breusch and Pagan (1980) have devised a **Lagrange multiplier test** for the random effects model based on the OLS residuals.<sup>16</sup> For

$$\begin{aligned}H_0: \sigma_u^2 &= 0 \quad (\text{or } \text{Corr}[\eta_{it}, \eta_{is}] = 0), \\ H_1: \sigma_u^2 &\neq 0,\end{aligned}$$

the test statistic is

$$\text{LM} = \frac{nT}{2(T-1)} \left[ \frac{\sum_{i=1}^n \left[ \sum_{t=1}^T e_{it} \right]^2}{\sum_{i=1}^n \sum_{t=1}^T e_{it}^2} - 1 \right]^2 = \frac{nT}{2(T-1)} \left[ \frac{\sum_{i=1}^n (T \bar{e}_{i.})^2}{\sum_{i=1}^n \sum_{t=1}^T e_{it}^2} - 1 \right]^2. \quad (11-42)$$

<sup>16</sup>We have focused thus far strictly on generalized least squares and moments based consistent estimation of the variance components. The LM test is based on maximum likelihood estimation, instead. See Maddala (1971) and Balestra and Nerlove (1966, 2003) for this approach to estimation.

## CHAPTER 11 ♦ Models for Panel Data 377

Under the null hypothesis, the limiting distribution of LM is chi-squared with one degree of freedom.

**Example 11.6 Testing for Random Effects**

We are interested in comparing the random and fixed effects estimators in the Cornwell and Rupert wage equation. As we saw earlier, there are three time-invariant variables in the equation: *Ed*, *Fem*, and *Blk*. As such, we cannot directly compare the two estimators. The **random effects model** can provide separate estimates of the parameters on the time-invariant variables while the fixed effects estimator cannot. For purposes of the illustration, then, we will for the present time conf<sup>?</sup> attention to the restricted common effects model,

$$\ln \text{Wage}_{it} = \beta_1 \text{Exp}_{it} + \beta_2 \text{Exp}_{it}^2 + \beta_3 \text{Wks}_{it} + \beta_4 \text{Occ}_{it} + \beta_5 \text{Ind}_{it} + \beta_6 \text{South}_{it} \\ + \beta_7 \text{SMSA}_{it} + \beta_8 \text{MS}_{it} + \beta_9 \text{Union}_{it} + c_i + \varepsilon_{it}.$$

The fixed and random effects models differ in the treatment of  $c_i$ .

Least squares estimates of the parameters including a constant term appear in Table 11.6. We then computed the group mean residuals for the seven observations for each individual. The sum of squares of the means is 53.824384. The total sum of squared residuals for the regression is 607.1265. With  $T$  and  $n$  equal to 7 and 595, respectively, (11-42) produces a chi-squared statistic of 3881.34. This far exceeds the 95 percent critical value for the chi-squared distribution with one degree of freedom, 3.84. At this point, we conclude that the classical regression model with a single constant term is inappropriate for these data. The result of the test is to reject the null hypothesis in favor of the random effects model. But, it is best to reserve judgment on that, because there is another competing specification that might induce these same results, the fixed effects model. We will examine this possibility in the subsequent examples.

**TABLE 11.6** Estimates of the Wage Equation

<i>Variable</i>	<i>Pooled Least Squares</i>		<i>Fixed Effects LSDV</i>		<i>Random Effects FGLS</i>		
	<i>Estimate</i>	<i>Std.Error<sup>a</sup></i>	<i>Estimate</i>	<i>Std.Error</i>	<i>Estimate</i>	<i>Std.Error</i>	<i>Robust</i>
<i>Exp</i>	0.0361	0.004533	0.1132	0.002471	0.08906	0.002280	0.01276
<i>Exp</i> <sup>2</sup>	-0.0006550	0.0001016	-0.0004184	0.0000546	-0.0007577	0.00005036	0.00031
<i>Wks</i>	0.004461	0.001728	0.0008359	0.0005997	0.001066	0.0005939	0.00331
<i>Occ</i>	-0.3176	0.02726	-0.02148	0.01378	-0.1067	0.01269	0.05424
<i>Ind</i>	0.03213	0.02526	0.01921	0.01545	-0.01637	0.01391	0.05303
<i>South</i>	-0.1137	0.02868	-0.001861	0.03430	-0.06899	0.02354	0.05984
<i>SMSA</i>	0.1586	0.02602	-0.04247	0.01943	-0.01530	0.01649	0.05421
<i>MS</i>	0.3203	0.03494	-0.02973	0.01898	-0.02398	0.01711	0.06989
<i>Union</i>	0.06975	0.02667	0.03278	0.01492	0.03597	0.01367	0.05653
<i>Constant</i>	5.8802	0.09673			5.3455	0.04361	0.19866
			<i>Mundlak: Group Means</i>		<i>Mundlak: Time Varying</i>		
<i>Exp</i>			-0.08574	0.005821	0.1132	0.002474	
<i>Exp</i> <sup>2</sup>			-0.0001168	0.0001281	-0.0004184	0.00005467	
<i>Wks</i>			0.008020	0.004006	0.0008359	0.0006004	
<i>Occ</i>			-0.3321	0.03363	-0.02148	0.01380	
<i>Ind</i>			0.02677	0.03203	0.01921	0.01547	
<i>South</i>			-0.1064	0.04444	-0.001861	0.03434	
<i>SMSA</i>			0.2239	0.03421	0.04247	0.01945	
<i>MS</i>			0.4134	0.03984	-0.02972	0.01901	
<i>Union</i>			0.05637	0.03549	0.03278	0.01494	
<i>Constant</i>					5.7222	0.1906	

<sup>a</sup>Robust standard errors

## 378 PART II ♦ Generalized Regression Model and Equation Systems

With the variance estimators in hand, FGLS can be used to estimate the parameters of the model. All of our earlier results for FGLS estimators apply here. In particular, all that is needed for efficient estimation of the model parameters are consistent estimators of the variance components, and there are several. [See Hsiao (2003), Baltagi (2005), Nerlove (2002), Berzeg (1979), and Maddala and Mount (1973).]

### **Example 11.7 Estimates of the Random Effects Model**

In the previous example, we found the total sum of squares for the least squares estimator was 607.1265. The fixed effects (LSDV) estimates for this model appear in Table 11.5 (and 17) where the sum of squares given is 82.26732. Therefore, the moment estimators of the parameters are

$$\hat{\sigma}_e^2 + \hat{\sigma}_u^2 = \frac{607.1265}{4165 - 10} = 0.1461195.$$

and

$$\hat{\sigma}_e^2 = \frac{82.26732}{4165 - 595 - 9} = 0.0231023.$$

The implied estimator of  $\sigma_u^2$  is 0.12301719. (No problem of negative variance components has emerged.) The estimate of  $\theta$  for FGLS is

$$\hat{\theta} = 1 - \sqrt{\frac{0.0231023}{0.0231023 + 7(0.12301719)}} = 0.8383608.$$

FGLS estimates are computed by regressing the partial differences of  $\ln Wage_{it}$  on the partial differences of the constant and the nine regressors, using this estimate of  $\theta$  in (11-39). Estimates of the parameters using the OLS, fixed effects and random effects estimators appear in Table 11.6.

None of the desirable properties of the estimators in the random effects model rely on  $T$  going to infinity.<sup>17</sup> Indeed,  $T$  is likely to be quite small. The estimator of  $\sigma_e^2$  is equal to an average of  $n$  estimators, each based on the  $T$  observations for unit  $i$ . [See (11-39).] Each component in this average is, in principle, consistent. That is, its variance is of order  $1/T$  or smaller. Because  $T$  is small, this variance may be relatively large. But, each term provides some information about the parameter. The average over the  $n$  cross-sectional units has a variance of order  $1/(nT)$ , which will go to zero if  $n$  increases, even if we regard  $T$  as fixed. The conclusion to draw is that nothing in this treatment relies on  $T$  growing large. Although it can be shown that some consistency results will follow for  $T$  increasing, the typical panel data set is based on data sets for which it does not make sense to assume that  $T$  increases without bound or, in some cases, at all.<sup>18</sup> As a general proposition, it is necessary to take some care in devising estimators whose properties hinge on whether  $T$  is large or not. The widely used conventional ones we have discussed here do not, but we have not exhausted the possibilities.

The random effects model was developed by Balestra and Nerlove (1966). Their formulation included a time-specific component,  $\kappa_t$ , as well as the individual effect:

$$y_{it} = \alpha + \beta' \mathbf{x}_{it} + \varepsilon_{it} + u_i + \kappa_t.$$

<sup>17</sup>See Nickell (1981).

<sup>18</sup>In this connection, Chamberlain (1984) provided some innovative treatments of panel data that, in fact, take  $T$  as given in the model and that base consistency results solely on  $n$  increasing. Some additional results for dynamic models are given by Bhargava and Sargan (1983).

The extended formulation is rather complicated analytically. In Balestra and Nerlove's study, it was made even more so by the presence of a lagged dependent variable. A full set of results for this extended model, including a method for handling the lagged dependent variable, has been developed.<sup>19</sup> We will turn to this in Section 11.8.

### 11.5.5 HAUSMAN'S SPECIFICATION TEST FOR THE RANDOM EFFECTS MODEL

At various points, we have made the distinction between fixed and random effects models. An inevitable question is, Which should be used? From a purely practical standpoint, the dummy variable approach is costly in terms of degrees of freedom lost. On the other hand, the fixed effects approach has one considerable virtue. There is little justification for treating the individual effects as uncorrelated with the other regressors, as is assumed in the random effects model. The random effects treatment, therefore, may suffer from the inconsistency due to this correlation between the included variables and the random effect.<sup>20</sup>

The **specification test** devised by Hausman (1978)<sup>21</sup> is used to test for orthogonality of the common effects and the regressors. The test is based on the idea that under the hypothesis of no correlation, both OLS in the LSDV model and GLS are consistent, but OLS is inefficient,<sup>22</sup> whereas under the alternative, OLS is consistent, but GLS is not. Therefore, under the null hypothesis, the two estimates should not differ systematically, and a test can be based on the difference. The other essential ingredient for the test is the covariance matrix of the difference vector,  $[\mathbf{b} - \hat{\beta}]$ :

$$\text{Var}[\mathbf{b} - \hat{\beta}] = \text{Var}[\mathbf{b}] + \text{Var}[\hat{\beta}] - \text{Cov}[\mathbf{b}, \hat{\beta}] - \text{Cov}[\hat{\beta}, \mathbf{b}]. \quad (11-43)$$

Hausman's essential result is that *the covariance of an efficient estimator with its difference from an inefficient estimator is zero*, which implies that

$$\text{Cov}[(\mathbf{b} - \hat{\beta}), \hat{\beta}] = \text{Cov}[\mathbf{b}, \hat{\beta}] - \text{Var}[\hat{\beta}] = \mathbf{0}$$

or that

$$\text{Cov}[\mathbf{b}, \hat{\beta}] = \text{Var}[\hat{\beta}].$$

Inserting this result in (11-43) produces the required covariance matrix for the test,

$$\text{Var}[\mathbf{b} - \hat{\beta}] = \text{Var}[\mathbf{b}] - \text{Var}[\hat{\beta}] = \Psi.$$

The chi-squared test is based on the Wald criterion:

$$W = \chi^2[K - 1] = [\mathbf{b} - \hat{\beta}]' \hat{\Psi}^{-1} [\mathbf{b} - \hat{\beta}]. \quad (11-44)$$

For  $\hat{\Psi}$ , we use the estimated covariance matrices of the slope estimator in the LSDV model and the estimated covariance matrix in the random effects model, excluding the constant term. Under the null hypothesis,  $W$  has a limiting chi-squared distribution with  $K - 1$  degrees of freedom.

<sup>19</sup>See Balestra and Nerlove (1966), Fomby, Hill, and Johnson (1980), Judge et al. (1985), Hsiao (1986), Anderson and Hsiao (1982), Nerlove (1971a, 2002), and Baltagi (2005).

<sup>20</sup>See Hausman and Taylor (1981) and Chamberlain (1978).

<sup>21</sup>Related results are given by Baltagi (1986).

<sup>22</sup>Referring to the GLS matrix weighted average given earlier, we see that the efficient weight uses  $\theta$ , whereas OLS sets  $\theta = 1$ .

## 380 PART II ♦ Generalized Regression Model and Equation Systems

The **Hausman test** is a useful device for determining the preferred specification of the common effects model. As developed here, it has one practical shortcoming. The construction in (11-43) conforms to the theory of the test. However, it does not guarantee that the difference of the two covariance matrices will be positive definite in a finite sample. The implication is that nothing prevents the statistic from being negative when it is computed according to (11-44). One can, in that event, conclude that the random effects model is not rejected, since the similarity of the covariance matrices is what is causing the problem, and under the alternative (fixed effects) hypothesis, they would be significantly different. There are, however, several alternative methods of computing the statistic for the Hausman test, some asymptotically equivalent and others actually numerically identical. Baltagi (2005, pp. 65–73) provides an extensive analysis. One particularly convenient form of the test fineshes the practical problem noted here. An asymptotically equivalent test statistic is given by

$$H' = (\hat{\beta}_{LSDV} - \hat{\beta}_{MEANS})' \left[ \text{Asy.Var}[\hat{\beta}_{LSDV}] + \text{Asy.Var}[\hat{\beta}_{MEANS}] \right]^{-1} (\hat{\beta}_{LSDV} - \hat{\beta}_{MEANS}) \quad (11-45)$$

where  $\hat{\beta}_{MEANS}$  is the group means estimator discussed in Section 11.3.4. As noted, this is one of several equivalent forms of the test. The advantage of this form is that the covariance matrix will always be nonnegative definite.

### **Example 11.8 Hausman Test for Fixed versus Random Effects**

Using the results of the preceding example, we retrieved the coefficient vector and estimated asymptotic covariance matrix,  $\mathbf{b}_{FE}$  and  $\mathbf{V}_{FE}$  from the fixed effects results and the first nine elements of  $\hat{\beta}_{RE}$  and  $\mathbf{V}_{RE}$  (excluding the constant term). The test statistic is

$$H = (\mathbf{b}_{FE} - \hat{\beta}_{RE})' [\mathbf{V}_{FE} - \mathbf{V}_{RE}]^{-1} (\mathbf{b}_{FE} - \hat{\beta}_{RE})$$

The value of the test statistic is 2,636.08. The critical value from the chi-squared table is 16.919 so the null hypothesis of the random effects model is rejected. We conclude that the fixed effects model is the preferred specification for these data. This is an unfortunate turn of events, as the main object of the study is the impact of education, which is a time-invariant variable in this sample. Using (11-42) instead, we obtain a test statistic of 3,177.58. Of course, this does not change the conclusion.

Imbens and Wooldridge (2007) have argued that in spite of the practical considerations about the Hausman test in (11-44) and (11-45), the test should be based on robust covariance matrices that do not depend on the assumption of the null hypothesis (the random effects model). (I.e., “It makes no sense to report a fully robust variance matrix for FE and RE but then to compute a Hausman test that maintains the full set of RE assumptions.”) Their suggested approach amounts to the variable addition test described in the next section, with a robust covariance matrix.

### **11.5.6 EXTENDING THE UNOBSERVED EFFECTS MODEL: MUNDLAK’S APPROACH**

Even with the Hausman test available, choosing between the fixed and random effects specifications presents a bit of a dilemma. Both specifications have unattractive shortcomings. The fixed effects approach is robust to correlation between the omitted heterogeneity and the regressors, but it proliferates parameters and cannot accommodate time-invariant regressors. The random effects model hinges on an unlikely assumption, that the omitted heterogeneity is uncorrelated with the regressors. Several authors have

suggested modifications of the random effects model that would at least partly overcome its deficit. The failure of the random effects approach is that the mean independence assumption,  $E[c_i | \mathbf{X}_i] = 0$ , is untenable. **Mundlak's (1978) approach** would suggest the specification

$$E[c_i | \mathbf{X}_i] = \bar{\mathbf{x}}_i' \boldsymbol{\gamma}^{23}$$

Substituting this in the random effects model, we obtain

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}' \boldsymbol{\beta} + c_i + \varepsilon_{it} \\ &= \mathbf{x}_{it}' \boldsymbol{\beta} + \bar{\mathbf{x}}_i' \boldsymbol{\gamma} + \varepsilon_{it} + (c_i - E[c_i | \mathbf{X}_i]) \\ &= \mathbf{x}_{it}' \boldsymbol{\beta} + \bar{\mathbf{x}}_i' \boldsymbol{\gamma} + \varepsilon_{it} + u_i. \end{aligned} \quad (11-46)$$

This preserves the specification of the random effects model, but (one hopes) deals directly with the problem of correlation of the effects and the regressors. Note that the additional terms in  $\bar{\mathbf{x}}_i' \boldsymbol{\gamma}$  will only include the time-varying variables—the time-invariant variables are already group means. This additional set of estimates is shown in the lower panel of Table 11.6 in Example 11.6.

Mundlak's approach is frequently used as a compromise between the fixed and random effects models. One side benefit of the specification is that it provides another convenient approach to the Hausman test. As the model is formulated above, the difference between the “fixed effects” model and the “random effects” model is the nonzero  $\boldsymbol{\gamma}$ . As such, a statistical test of the null hypothesis that  $\boldsymbol{\gamma}$  equals zero should provide an alternative approach to the two methods suggested earlier.

#### **Example 11.9 Variable Selection Test for Fixed versus Random Effects**

Using the results in Example 11.6, we recovered the subvector of the estimates in the lower half of Table 11.6 corresponding to  $\boldsymbol{\gamma}$ , and the corresponding submatrix of the full covariance matrix. The test statistic is

$$H' = \hat{\boldsymbol{\gamma}}' [\text{Est. Asy. Var}(\hat{\boldsymbol{\gamma}})]^{-1} \hat{\boldsymbol{\gamma}}$$

The value of the test statistic is 3193.69. The critical value from the chi-squared table for nine degrees of freedom is 16.919, so the null hypothesis of the random effects model is rejected. We conclude as before that the fixed effects estimator is the preferred specification for this model.

#### **11.5.7 EXTENDING THE RANDOM AND FIXED EFFECTS MODELS: CHAMBERLAIN'S APPROACH**

The linear unobserved effects model is

$$y_{it} = c_i + \mathbf{x}_{it}' \boldsymbol{\beta} + \varepsilon_{it}. \quad (11-47)$$

The **random effects** model assumes that  $E[c_i | \mathbf{X}_i] = \alpha$ , where the  $T$  rows of  $\mathbf{X}_i$  are  $\mathbf{x}_{it}'$ . As we saw in Section 11.5.1, this model can be estimated consistently by ordinary least squares. Regardless of how  $\varepsilon_{it}$  is modeled, there is autocorrelation induced by

<sup>23</sup>Other analyses, for example, Chamberlain (1982) and Wooldridge (2002a), interpret the linear function as the *projection* of  $c_i$  on the group means, rather than the conditional mean. The difference is that we need not make any particular assumptions about the conditional mean function while there always exists a linear projection. The conditional mean interpretation does impose an additional assumption on the model but brings considerable simplification. Several authors have analyzed the extension of the model to projection on the full set of individual observations rather than the means. The additional generality provides the bases of several other estimators including minimum distance [Chamberlain (1982)], GMM [Arellano and Bover (1995)], and constrained seemingly unrelated regressions and three-stage least squares [Wooldridge (2002a)].

## 382 PART II ♦ Generalized Regression Model and Equation Systems

the common, unobserved  $c_i$ , so the generalized regression model applies. The random effects formulation is based on the assumption  $E[\mathbf{w}_i \mathbf{w}'_i | \mathbf{X}_i] = \sigma_\varepsilon^2 \mathbf{I}_T + \sigma_u^2 \mathbf{ii}'$ , where  $w_{it} = (\varepsilon_{it} + u_i)$ . We developed the GLS and FGLS estimators for this formulation as well as a strategy for robust estimation of the OLS covariance matrix. Among the implications of the development of Section 11.5 is that this formulation of the disturbance covariance matrix is more restrictive than necessary, given the information contained in the data. The assumption that  $E[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}'_i | \mathbf{X}_i] = \sigma_\varepsilon^2 \mathbf{I}_T$  assumes that the correlation across periods is equal for all pairs of observations, and arises solely through the persistent  $c_i$ . In Section 10.2.6, we estimated the equivalent model with an unrestricted covariance matrix,  $E[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}'_i | \mathbf{X}_i] = \Sigma$ . The implication is that the random effects treatment includes two restrictive assumptions, mean independence,  $E[c_i | \mathbf{X}_i] = \alpha$ , and homoscedasticity,  $E[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}'_i | \mathbf{X}_i] = \sigma_\varepsilon^2 \mathbf{I}_T$ . [We do note, dropping the second assumption will cost us the identification of  $\sigma_u^2$  as an estimable parameter. This makes sense—if the correlation across periods  $t$  and  $s$  can arise from either their common  $u_i$  or from correlation of  $(\varepsilon_{it}, \varepsilon_{is})$  then there is no way for us separately to estimate a variance for  $u_i$  apart from the covariances of  $\varepsilon_{it}$  and  $\varepsilon_{is}$ .] It is useful to note, however, that the panel data model can be viewed and formulated as a seemingly unrelated regressions model with common coefficients in which each period constitutes an equation. Indeed, it is possible, albeit unnecessary, to impose the restriction  $E[\mathbf{w}_i \mathbf{w}'_i | \mathbf{X}_i] = \sigma_\varepsilon^2 \mathbf{I}_T + \sigma_u^2 \mathbf{ii}'$ .

The mean independence assumption is the major shortcoming of the random effects model. The central feature of the fixed effects model in Section 11.4 is the possibility that  $E[c_i | \mathbf{X}_i]$  is a nonconstant  $g(\mathbf{X}_i)$ . As such, least squares regression of  $y_{it}$  on  $\mathbf{x}_{it}$  produces an inconsistent estimator of  $\beta$ . The dummy variable model considered in Section 11.4 is the natural alternative. The **fixed effects** approach has the advantage of dispensing with the unlikely assumption that  $c_i$  and  $\mathbf{x}_{it}$  are uncorrelated. However, it has the shortcoming of requiring estimation of the  $n$  “parameters,”  $\alpha_i$ .

Chamberlain (1982, 1984) and Mundlak (1978) suggested alternative approaches that lie between these two. Their modifications of the fixed effects model augment it with the **projections** of  $c_i$  on all the rows of  $\mathbf{X}_i$  (Chamberlain) or the group means (Mundlak). (See Section 11.5.5.) Consider the first of these, and assume (as it requires) a balanced panel of  $T$  observations per group. For purposes of this development, we will assume  $T = 3$ . The generalization will be obvious at the conclusion. Then, the projection suggested by Chamberlain is

$$c_i = \alpha + \mathbf{x}'_{i1} \boldsymbol{\gamma}_1 + \mathbf{x}'_{i2} \boldsymbol{\gamma}_2 + \mathbf{x}'_{i3} \boldsymbol{\gamma}_3 + r_i \quad (11-48)$$

where now, by construction,  $r_i$  is orthogonal to  $\mathbf{x}_{it}$ .<sup>24</sup> Insert (11-48) into (11-44) to obtain

$$y_{it} = \alpha + \mathbf{x}'_{i1} \boldsymbol{\gamma}_1 + \mathbf{x}'_{i2} \boldsymbol{\gamma}_2 + \mathbf{x}'_{i3} \boldsymbol{\gamma}_3 + \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it} + r_i.$$

<sup>24</sup>There are some fine points here that can only be resolved theoretically. If the projection in (11-48) is not the conditional mean, then we have  $E[r_i \times \mathbf{x}_{it}] = 0$ ,  $t = 1, \dots, T$  but not  $E[r_i | \mathbf{X}_i] = 0$ . This does not affect the asymptotic properties of the FGLS estimator to be developed here, although it does have implications, for example, for unbiasedness. Consistency will hold regardless. The assumptions behind (11-48) do not include that  $\text{Var}[r_i | \mathbf{X}_i]$  is homoscedastic. It might not be. This could be investigated empirically. The implication here concerns efficiency, not consistency. The FGLS estimator to be developed here would remain consistent, but a GMM estimator would be more efficient—see Chapter 13. Moreover, without homoscedasticity, it is not certain that the FGLS estimator suggested here is more efficient than OLS (with a robust covariance matrix estimator). Our intent is to begin the investigation here. Further details can be found in Chamberlain (1984) and, e.g., Im, Ahn, Schmidt, and Wooldridge (1999).

CHAPTER 11 ♦ Models for Panel Data **383**

Estimation of the  $1 + 3K + K$  parameters of this model presents a number of complications. [We do note, this approach has the potential to (wildly) proliferate parameters. For our quite small regional productivity model in Example 11.19, the original model with six main coefficients plus the treatment of the constants becomes a model with  $1 + 6 + 17(6) = 109$  parameters to be estimated.]

If only the  $n$  observations for period 1 are used, then the parameter vector,

$$\boldsymbol{\theta}_1 = \alpha, (\boldsymbol{\beta} + \boldsymbol{\gamma}_1), \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3 = \alpha, \boldsymbol{\pi}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \quad (11-49)$$

can be estimated consistently, albeit inefficiently, by ordinary least squares. The “model” is

$$y_{i1} = \mathbf{z}'_{i1} \boldsymbol{\theta}_1 + w_{i1}, i = 1, \dots, n.$$

Collecting the  $n$  observations, we have

$$\mathbf{y}_1 = \mathbf{Z}_1 \boldsymbol{\theta}_1 + \mathbf{w}_1.$$

If, instead, only the  $n$  observations from period 2 or period 3 are used, then OLS estimates, in turn,

$$\boldsymbol{\theta}_2 = \alpha, \boldsymbol{\gamma}_1, (\boldsymbol{\beta} + \boldsymbol{\gamma}_2), \boldsymbol{\gamma}_3 = \alpha, \boldsymbol{\gamma}_1, \boldsymbol{\pi}_2, \boldsymbol{\gamma}_3,$$

or

$$\boldsymbol{\theta}_3 = \alpha, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, (\boldsymbol{\beta} + \boldsymbol{\gamma}_3) = \alpha, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\pi}_3.$$

It remains to reconcile the multiple estimates of the same parameter vectors. In terms of the preceding layouts above, we have the following:

$$\begin{aligned} \text{OLS Estimates: } & a_1, \mathbf{p}_1, \mathbf{c}_{2,1}, \mathbf{c}_{3,1}, & a_2, \mathbf{c}_{1,2}, \mathbf{p}_2, \mathbf{c}_{3,2}, & a_3, \mathbf{c}_{1,3}, \mathbf{c}_{2,3}, \mathbf{p}_3; \\ \text{Estimated Parameters: } & \alpha, (\boldsymbol{\beta} + \boldsymbol{\gamma}_1), \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, & \alpha, \boldsymbol{\gamma}_1, (\boldsymbol{\beta} + \boldsymbol{\gamma}_2), \boldsymbol{\gamma}_3, & \alpha, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, (\boldsymbol{\beta} + \boldsymbol{\gamma}_3); \\ \text{Structural Parameters: } & \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3. \end{aligned} \quad (11-50)$$

Chamberlain suggested a minimum distance estimator (MDE). For this problem, the MDE is essentially a weighted average of the several estimators of each part of the parameter vector. We will examine the MDE for this application in more detail in Chapter 13. (For another simpler application of minimum distance estimation that shows the “weighting” procedure at work, see the reconciliation of four competing estimators of a single parameter at the end of Example 11.20.) Here is an alternative way to formulate the estimator that is a bit more transparent. For the first period,

$$\mathbf{y}_1 = \begin{pmatrix} y_{1,1} \\ y_{2,1} \\ \vdots \\ y_{n,1} \end{pmatrix} = \begin{bmatrix} 1 & \mathbf{x}_{1,1} & \mathbf{x}_{1,1} & \mathbf{x}_{1,2} & \mathbf{x}_{1,3} \\ 1 & \mathbf{x}_{2,2} & \mathbf{x}_{2,1} & \mathbf{x}_{2,2} & \mathbf{x}_{2,3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \mathbf{x}_{n,1} & \mathbf{x}_{n,1} & \mathbf{x}_{n,1} & \mathbf{x}_{n,1} \end{bmatrix} \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \\ \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \\ \boldsymbol{\gamma}_3 \end{pmatrix} + \begin{pmatrix} r_{1,1} \\ r_{2,1} \\ \vdots \\ r_{n,1} \end{pmatrix} = \tilde{\mathbf{X}}_1 \boldsymbol{\theta} + \mathbf{r}_1. \quad (11-51)$$

We treat this as the first equation in a  $T$  equation seemingly unrelated regressions model. The second equation, for period 2, is the same (same coefficients), with the data from the second period appearing in the blocks, then likewise for period 3 (and periods

### 384 PART II ♦ Generalized Regression Model and Equation Systems

4, ...,  $T$  in the general case). Stacking the data for the  $T$  equations (periods), we have

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{X}}_1 \\ \tilde{\mathbf{X}}_2 \\ \vdots \\ \tilde{\mathbf{X}}_T \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \\ \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_T \end{pmatrix} + \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_T \end{pmatrix} = \tilde{\mathbf{X}}\boldsymbol{\theta} + \mathbf{r}, \quad (11-52)$$

where  $E[\tilde{\mathbf{X}}'\mathbf{r}] = \mathbf{0}$  and (by assumption),  $E[\mathbf{r}\mathbf{r}'|\tilde{\mathbf{X}}] = \boldsymbol{\Sigma} \otimes \mathbf{I}_n$ . With the homoscedasticity assumption for  $r_{i,t}$ , this is precisely the application in Section 10.2.6. The parameters can be estimated by FGLS as shown in Section 10.2.6.

#### **Example 11.10 Hospital Costs**

Carey (1997) examined hospital costs for a sample of 1,733 hospitals observed in five years, 1987–1991. The model estimated is

$$\begin{aligned} \ln(TC/P)_{it} = & \alpha_i + \beta_D DIS_{it} + \beta_O OPV_{it} + \beta_3 ALS_{it} + \beta_4 CM_{it} \\ & + \beta_5 DIS_{it}^2 + \beta_6 DIS_{it}^3 + \beta_7 OPV_{it}^2 + \beta_8 OPV_{it}^3 \\ & + \beta_9 ALS_{it}^2 + \beta_{10} ALS_{it}^3 + \beta_{11} DIS_{it} \times OPV_{it} \\ & + \beta_{12} FA_{it} + \beta_{13} HI_{it} + \beta_{14} HT_i + \beta_{15} LT_i + \beta_{16} Large_i \\ & + \beta_{17} Small_i + \beta_{18} NonProfit_i + \beta_{19} Profit_i \\ & + \varepsilon_{it}, \end{aligned}$$

where

TC	= total cost,
P	= input price index,
DIS	= discharges,
OPV	= outpatient visits,
ALS	= average length of stay,
CM	= case mix index,
FA	= fixed assets,
HI	= Hirfindahl index of market concentration at county level,
HT	= dummy for high teaching load hospital,
LT	= dummy variable for low teaching load hospital,
Large	= dummy variable for large urban area,
Small	= dummy variable for small urban area,
Nonprofit	= dummy variable for nonprofit hospital,
Profit	= dummy variable for profit hospital.

We have used subscripts “D” and “O” for the coefficients on DIS and OPV as these will be isolated in the following discussion. The model employed in the study is that in (11-47) and (11-48). Initial OLS estimates are obtained for the full cost function in each year. SUR estimates are then obtained using a restricted version of the Chamberlain system. This second step involved a hybrid model that modified (11-49) so that in each period the coefficient vector was

$$\boldsymbol{\theta}_t = [\alpha_t, \beta_{Dt}(\boldsymbol{\gamma}), \beta_{Ot}(\boldsymbol{\gamma}), \beta_{3t}(\boldsymbol{\gamma}), \beta_{4t}(\boldsymbol{\gamma}), \beta_{5t}, \dots, \beta_{19t}]$$

where  $\beta_{Dt}(\boldsymbol{\gamma})$  indicates that all five years of the variable ( $DIS_{it}$ ) are included in the equation and, likewise for  $\beta_{Ot}(\boldsymbol{\gamma})$  ( $OPV$ ),  $\beta_{3t}(\boldsymbol{\gamma})$  ( $ALS$ ) and  $\beta_{4t}(\boldsymbol{\gamma})$  ( $CM$ ). This is equivalent to using

$$c_t = \alpha + \sum_{i=1987}^{1991} (DIS, OPV, ALS, CM)_{it} \boldsymbol{\gamma}_t + r_t$$

in (11-48).

**TABLE 11.7** Coefficient Estimates in SUR Model for Hospital Costs

<i>Equation</i>	<i>Coefficient on Variable in the Equation</i>				
	<i>DIS87</i>	<i>DIS88</i>	<i>DIS89</i>	<i>DIS90</i>	<i>DIS91</i>
SUR87	$\beta_{D,87} + \gamma_{D,87}$ 1.76	$\gamma_{D,88}$ 0.116	$\gamma_{D,89}$ -0.0881	$\gamma_{D,90}$ 0.0570	$\gamma_{D,91}$ -0.0617
SUR88	$\gamma_{D,87}$ 0.254	$\beta_{D,88} + \gamma_{D,88}$ 1.61	$\gamma_{D,89}$ -0.0934	$\gamma_{D,90}$ 0.0610	$\gamma_{D,91}$ -0.0514
SUR89	$\gamma_{D,87}$ 0.217	$\gamma_{D,88}$ 0.0846	$\beta_{D,89} + \gamma_{D,89}$ 1.51	$\gamma_{D,90}$ 0.0454	$\gamma_{D,91}$ -0.0253
SUR90	$\gamma_{D,87}$ 0.179	$\gamma_{D,88}$ 0.0822 <sup>a</sup>	$\gamma_{D,89}$ 0.0295	$\beta_{D,90} + \gamma_{D,90}$ 1.57	$\gamma_{D,91}$ 0.0244
SUR91	$\gamma_{D,87}$ 0.153	$\gamma_{D,88}$ 0.0363	$\gamma_{D,89}$ -0.0422	$\gamma_{D,90}$ 0.0813	$\beta_{D,91} + \gamma_{D,91}$ 1.70

<sup>a</sup>The value reported in the published paper is 8.22. The correct value is 0.0822. (Personal communication from the author.)

The unrestricted SUR system estimated at the second step provides multiple estimates of the various model parameters. For example, each of the five equations provides an estimate of  $(\beta_5, \dots, \beta_{19})$ . The author added one more layer to the model in allowing the coefficients on  $DIS_{it}$  and  $OPV_{it}$  to vary over time. Therefore, the structural parameters of interest are  $(\beta_{D1}, \dots, \beta_{D5})$ ,  $(\gamma_{D1}, \dots, \gamma_{D5})$  (the coefficients on DIS) and  $(\beta_{O1}, \dots, \beta_{O5})$ ,  $(\gamma_{O1}, \dots, \gamma_{O5})$  (the coefficients on OPV). There are, altogether, 20 parameters of interest. The SUR estimates produce, in each year (equation), parameters on DIS for the five years and on OPV for the five years, so there is a total of 50 estimates. Reconciling all of them means imposing a total of 30 restrictions. Table 11.7 shows the relationships for the time varying parameter on  $DIS_{it}$  in the five-equation model. The numerical values reported by the author are shown following the theoretical results. A similar table would apply for the coefficients on OPV, ALS, and CM. (In the latter two, the  $\beta$  coefficient was not assumed to be time varying.) It can be seen in the table, for example, that there are directly four different estimates of  $\gamma_{D,87}$  in the second to fifth equations, and likewise for each of the other parameters. Combining the entries in Table 11.7 with the counterpart for the coefficients on OPV, we see 50 SUR/FGLS estimates to be used to estimate 20 underlying parameters. The author used a minimum distance approach to reconcile the different estimates. We will return to this example in Example 13.6, where we will develop the MDE in more detail.

## 11.6 NONSPHERICAL DISTURBANCES AND ROBUST COVARIANCE ESTIMATION

Because the models considered here are extensions of the classical regression model, we can treat heteroscedasticity in the same way that we did in Chapter 9. That is, we can compute the ordinary or feasible generalized least squares estimators and obtain an appropriate robust covariance matrix estimator, or we can impose some structure on the disturbance variances and use generalized least squares. In the panel data settings, there is greater flexibility for the second of these without making strong assumptions about the nature of the heteroscedasticity.

### 11.6.1 ROBUST ESTIMATION OF THE FIXED EFFECTS MODEL

As noted in Section 11.3.2, in a panel data set, the correlation across observations within a group is likely to be a more substantial influence on the estimated covariance matrix of

## 386 PART II ♦ Generalized Regression Model and Equation Systems

the least squares estimator than is heteroscedasticity. This is evident in the estimates in Table 11.1. In the fixed (or random) effects model, the intent of explicitly including the common effect in the model is to account for the source of this correlation. However, accounting for the common effect in the model does not remove heteroscedasticity—it centers the conditional mean properly. Here, we consider the straightforward extension of White's estimator to the fixed and random effects models.

In the fixed effects model, the full regressor matrix is  $\mathbf{Z} = [\mathbf{X}, \mathbf{D}]$ . The White heteroscedasticity consistent covariance matrix for OLS—that is, for the fixed effects estimator—is the lower right block of the partitioned matrix

$$\text{Est. Asy. Var}[\mathbf{b}, \mathbf{a}] = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{E}^2\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1},$$

where  $\mathbf{E}$  is a diagonal matrix of least squares (fixed effects estimator) residuals. This computation promises to be formidable, but fortunately, it works out very simply. The White estimator for the slopes is obtained just by using the data in group mean deviation form [see (11-15) and (11-18)] in the familiar computation of  $\mathbf{S}_0$  [see (9-26) and (9-27)]. Also, the disturbance variance estimator in (11-18) is the counterpart to the one in (9-20), which we showed that after the appropriate scaling of  $\Omega$  was a consistent estimator of  $\sigma^2 = \text{plim}[1/(nT)] \sum_{i=1}^n \sum_{t=1}^T \sigma_{it}^2$ . The implication is that we may still use (11-18) to estimate the variances of the fixed effects.

A somewhat less general but useful simplification of this result can be obtained if we assume that the disturbance variance is constant within the  $i$ th group. If  $E[\varepsilon_{it}^2 | \mathbf{Z}_i] = \sigma_i^2$ , then, with a panel of data,  $\sigma_i^2$  is estimable by  $\mathbf{e}_i'\mathbf{e}_i/T$  using the least squares residuals. The center matrix in Est. Asy. Var $[\mathbf{b}, \mathbf{a}]$  may be replaced with  $\sum_i (\mathbf{e}_i'\mathbf{e}_i/T)\mathbf{Z}_i'\mathbf{Z}_i$ . Whether this estimator is preferable is unclear. If the groupwise model is correct, then it and the White estimator will estimate the same matrix. On the other hand, if the disturbance variances do vary within the groups, then this revised computation may be inappropriate.

Arellano (1987) and Arellano and Bover (1995) have taken this analysis a step further. If one takes the  $i$ th group as a whole, then we can treat the observations in

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \alpha_i\mathbf{i}_T + \boldsymbol{\varepsilon}_i$$

as a generalized regression model with disturbance covariance matrix  $\Omega_i$ . We saw in Section 11.3.2 that a model this general, with no structure on  $\Omega$ , offered little hope for estimation, robust or otherwise. But the problem is more manageable with a panel data set where correlation across units can be assumed to be zero. As before, let  $\mathbf{X}_{i*}$  denote the data in group mean deviation form. The counterpart to  $\mathbf{X}'\Omega\mathbf{X}$  here is

$$\mathbf{X}'_*\Omega\mathbf{X}_* = \sum_{i=1}^n (\mathbf{X}'_{i*}\Omega_i\mathbf{X}_{i*}).$$

By the same reasoning that we used to construct the White estimator in Chapter 9, we can consider estimating  $\Omega_i$  with the sample of one,  $\mathbf{e}_i\mathbf{e}_i'$ . As before, it is not consistent estimation of the individual  $\Omega_i$ 's that is at issue, but estimation of the sum. If  $n$  is large

enough, then we could argue that

$$\begin{aligned}
 \text{plim } \frac{1}{nT} \mathbf{X}'_* \boldsymbol{\Omega} \mathbf{X}_* &= \text{plim } \frac{1}{nT} \sum_{i=1}^n \mathbf{X}'_{i*} \boldsymbol{\Omega}_i \mathbf{X}_{*i} \\
 &= \text{plim } \frac{1}{n} \sum_{i=1}^n \frac{1}{T} \mathbf{X}'_{*i} \mathbf{e}_i \mathbf{e}'_i \mathbf{X}_{*i} \\
 &= \text{plim } \frac{1}{n} \sum_{i=1}^n \left( \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T e_{it} e_{is} \mathbf{x}_{*it} \mathbf{x}'_{*is} \right). \tag{11-53}
 \end{aligned}$$

This is the extension of (11-3) to the fixed effects case.

### 11.6.2 HETEROSCEDASTICITY IN THE RANDOM EFFECTS MODEL

Because the random effects model is a generalized regression model with a known structure, OLS with a robust estimator of the asymptotic covariance matrix is not the best use of the data. The GLS estimator is efficient whereas the OLS estimator is not. If a perfectly general covariance structure is assumed, then one might simply use Arellano's estimator described in the preceding section with a single overall constant term rather than a set of fixed effects. But, within the setting of the random effects model,  $\eta_{it} = \varepsilon_{it} + u_i$ , allowing the disturbance variance to vary across groups would seem to be a useful extension.

A series of papers, notably Mazodier and Trognon (1978), Baltagi and Griffin (1988), and the recent monograph by Baltagi (2005, pp. 77–79) suggest how one might allow the group-specific component  $u_i$  to be heteroscedastic. But, empirically, there is an insurmountable problem with this approach. In the final analysis, all estimators of the variance components must be based on sums of squared residuals, and, in particular, an estimator of  $\sigma_{ui}^2$  would be estimated using a set of residuals from the distribution of  $u_i$ . However, the data contain only a single observation on  $u_i$  repeated in each observation in group  $i$ . So, the estimators presented, for example, in Baltagi (2001), use, in effect, one residual in each case to estimate  $\sigma_{ui}^2$ . What appears to be a mean squared residual is only  $(1/T) \sum_{t=1}^T \hat{u}_i^2 = \hat{u}_i^2$ . The properties of this estimator are ambiguous, but efficiency seems unlikely. The estimators do not converge to any population figure as the sample size, even  $T$ , increases. [The counterpoint is made in Hsiao (2003, p. 56).] Heteroscedasticity in the unique component,  $\varepsilon_{it}$  represents a more tractable modeling possibility.

In Section 11.5.2, we introduced heteroscedasticity into estimation of the random effects model by allowing the group sizes to vary. But the estimator there (and its feasible counterpart in the next section) would be the same if, instead of  $\theta_i = 1 - \sigma_\varepsilon / (T_i \sigma_u^2 + \sigma_\varepsilon^2)^{1/2}$ , we were faced with

$$\theta_i = 1 - \frac{\sigma_{\varepsilon i}}{\sqrt{\sigma_{\varepsilon i}^2 + T_i \sigma_u^2}}.$$

Therefore, for computing the appropriate feasible generalized least squares estimator, once again we need only devise consistent estimators for the variance components and

## 388 PART II ♦ Generalized Regression Model and Equation Systems

then apply the GLS transformation shown earlier. One possible way to proceed is as follows: Because pooled OLS is still consistent, OLS provides a usable set of residuals. Using the OLS residuals for the specific groups, we would have, for each group,

$$\widehat{\sigma_{\varepsilon i}^2 + u_i^2} = \frac{\mathbf{e}'_i \mathbf{e}_i}{T}.$$

The residuals from the dummy variable model are purged of the individual specific effect,  $u_i$ , so  $\widehat{\sigma_{\varepsilon i}^2}$  may be consistently (in  $T$ ) estimated with

$$\widehat{\sigma_{\varepsilon i}^2} = \frac{\mathbf{e}_i^{lsdv} \mathbf{e}_i^{lsdv}}{T}$$

where  $e_i^{lsdv} = y_{it} - \mathbf{x}'_i \mathbf{b}^{lsdv} - a_i$ . Combining terms, then,

$$\hat{\sigma}_u^2 = \frac{1}{n} \sum_{i=1}^n \left[ \left( \frac{\mathbf{e}_i^{ols} \mathbf{e}_i^{ols}}{T} \right) - \left( \frac{\mathbf{e}_i^{lsdv} \mathbf{e}_i^{lsdv}}{T} \right) \right] = \frac{1}{n} \sum_{i=1}^n \widehat{(u_i^2)}.$$

We can now compute the FGLS estimator as before.

### 11.6.3 AUTOCORRELATION IN PANEL DATA MODELS

Serial correlation of regression disturbances  will be considered in detail in Section 20.10. Rather than defer the topic in connection to panel data to Chapter 20, we will briefly note it here. As we saw in Section 11.3.2 and Example 11.1, “autocorrelation”—that is, correlation across the observations in the groups in a panel—is likely to be a substantive feature of the model. Our treatment of the effect there, however, was meant to accommodate autocorrelation in its broadest sense, that is, nonzero covariances across observations in a group. The results there would apply equally to clustered observations, as observed in Section 11.3.3. An important element of that specification was that with clustered data, there might be no obvious structure to the autocorrelation. When the panel data set consists explicitly of groups of time series, and especially if the time series are relatively long as in Example 11.11, one might want to begin to invoke the more detailed, structured time series models which are discussed in Chapter 20.

### 11.6.4 CLUSTER (AND PANEL) ROBUST COVARIANCE MATRICES FOR FIXED AND RANDOM EFFECTS ESTIMATORS

As suggested earlier, in situations in which cluster corrections are appropriate, there might be a residual correlation within groups that is not fully accounted for by a generalized least squares estimator or a fixed effects model. A counterpart to (11-4) for the fixed and random effects estimators is straightforward to construct based on results we have already obtained.

For the fixed effects estimator, based on (11-14) and (11-20), we have

$$\mathbf{b}_{LSDV} = \left[ \sum_{g=1}^G \sum_{i=1}^{n_g} (\Delta(1)\mathbf{x}_{ig})(\Delta(1)\mathbf{x}_{ig})' \right]^{-1} \left[ \sum_{g=1}^G \sum_{i=1}^{n_g} (\Delta(1)\mathbf{x}_{ig}) (\Delta(1)y_{ig}) \right] \quad (11-54)$$

where  $\Delta(1)\mathbf{x}_{it} = \mathbf{x}_{it} - (1)\bar{\mathbf{x}}_i$  is the deviation of  $\mathbf{x}_{it}$  from one times the group mean vector. The motivation for the “(1)” will be evident shortly. In the same fashion as (11-3), we

will construct a robust covariance matrix estimator using

$$\text{Est.Asy.Var}[\mathbf{b}_{LSDV}] = \left[ \sum_{g=1}^G \left\{ \sum_{i=1}^{n_g} (\Delta(1)\mathbf{x}_{ig}) e_{ig} \right\} \left\{ \sum_{i=1}^{n_g} (\Delta(1)\mathbf{x}_{ig}) e_{ig} \right\}' \right]^{-1} \times \\ \left[ \sum_{g=1}^G \left\{ \sum_{i=1}^{n_g} (\Delta(1)\mathbf{x}_{ig}) (\Delta(1)\mathbf{x}_{ig})' \right\} \right]^{-1} \quad (11-55)$$

This estimator is equivalent to (11-3) based on the data in deviations from their cluster means. (With a slight change in notation, it becomes a robust estimator for the covariance matrix of the fixed effects estimator.) From (11-32) and (11-33), the GLS estimator of  $\beta$  for the random effects model is

$$\hat{\beta}_{GLS} = \left[ \sum_{g=1}^G \mathbf{X}_g' \Sigma_g^{-1} \mathbf{X}_g \right]^{-1} \left[ \sum_{g=1}^G \mathbf{X}_g' \Sigma_g^{-1} y_g \right] \\ \left[ \sum_{g=1}^G \left\{ \sum_{i=1}^{n_g} (\Delta(\theta_g)\mathbf{x}_{ig}) (\Delta(\theta_g)\mathbf{x}_{ig})' \right\} \right]^{-1} \left[ \sum_{g=1}^G \left\{ \sum_{i=1}^{n_g} (\Delta(\theta_g)\mathbf{x}_{ig}) (\Delta(\theta_g)y_{ig}) \right\} \right], \quad (11-56)$$

where  $\theta_g = 1 - (\sigma_\varepsilon / \sqrt{\sigma_\varepsilon^2 + n_g \sigma_u^2})$ . It follows that the estimator of the asymptotic covariance matrix would be

$$\text{Est.Asy.Var}[\hat{\beta}_{GLS}] = \left[ \sum_{g=1}^G \left\{ \sum_{i=1}^{n_g} (\Delta(\theta_g)\mathbf{x}_{ig}) e_{ig} \right\} \left\{ \sum_{i=1}^{n_g} (\Delta(\theta_g)\mathbf{x}_{ig}) e_{ig} \right\}' \right]^{-1} \times \\ \left[ \sum_{g=1}^G \left\{ \sum_{i=1}^{n_g} (\Delta(\theta_g)\mathbf{x}_{ig}) (\Delta(\theta_g)\mathbf{x}_{ig})' \right\} \right]^{-1} \quad (11-57)$$

See, also, Cameron and Trivedi (2005, pp. 838–839).

#### **Example 11.11 Robust Standard Errors for Fixed and Random Effects Estimators**

Table 11.8 presents the estimates of the fixed random effects models that appear in Tables 11.5 and 11.6. The correction of the standard errors results in a fairly substantial change in the estimates. The effect is especially pronounced in the random effects case, where the estimated standard errors increase by a factor of five or more.

### 11.7 SPATIAL AUTOCORRELATION

The nested random effects structure in Example 11.12 was motivated by an expectation that effects of neighboring states would spill over into each other, creating a sort of correlation across space, rather than across time as we have focused on thus far. The

**390 PART II ♦ Generalized Regression Model and Equation Systems**
**TABLE 11.8** Cluster Corrections for Fixed and Random Effects Estimators

<i>Variable</i>	<i>Fixed Effects</i>			<i>Random Effects</i>		
	<i>Estimate</i>	<i>Std.Error</i>	<i>Robust</i>	<i>Estimate</i>	<i>Std.Error</i>	<i>Robust</i>
Constant				5.3455	0.04361	0.19866
Exp	0.1132	0.002471	0.00437	0.08906	0.002280	0.01276
Exp <sup>2</sup>	-0.00042	0.000055	0.000089	-0.0007577	0.00005036	0.00031
Wks	0.00084	0.000600	0.00094	0.001066	0.0005939	0.00331
Occ	-0.02148	0.01378	0.02052	-0.1067	0.01269	0.05424
Ind	0.01921	0.01545	0.02450	-0.01637	0.01391	0.053003
South	-0.00186	0.03430	0.09646	-0.06899	0.02354	0.05984
SMSA	-0.04247	0.01942	0.03185	-0.01530	0.01649	0.05421
MS	-0.02973	0.01898	0.02902	-0.02398	0.01711	0.06984
Union	0.03278	0.01492	0.02708	0.03597	0.01367	0.05653

effect should be common in cross-region studies, such as in agriculture, urban economics, and regional science. Recent studies of the phenomenon include Case's (1991) study of expenditure patterns, Bell and Bockstael's (2000) study of real estate prices, and Baltagi and Li's (2001) analysis of R&D spillovers. Models of **spatial autocorrelation** [see Anselin (1988, 2001) for the canonical reference and Le Sage and Pace (2009) for a recent survey], are constructed to formalize this notion.

A model with spatial autocorrelation can be formulated as follows: The regression model takes the familiar panel structure,

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it} + u_i, i = 1, \dots, n; t = 1, \dots, T.$$

The common  $u_i$  is the usual unit (e.g., country) effect. The correlation across space is implied by the spatial autocorrelation structure

$$\varepsilon_{it} = \lambda \sum_{j=1}^n W_{ij} \varepsilon_{jt} + v_t.$$

The scalar  $\lambda$  is the **spatial autoregression coefficient**. The elements  $W_{ij}$  are spatial (or **contiguity**) weights that are assumed known. The elements that appear in the sum above are a row of the spatial weight or **contiguity matrix**,  $\mathbf{W}$ , so that for the  $n$  units, we have

$$\boldsymbol{\varepsilon}_t = \lambda \mathbf{W} \boldsymbol{\varepsilon}_t + \mathbf{v}_t, \mathbf{v}_t = v_t \mathbf{i}.$$

The structure of the model is embodied in the symmetric weight matrix,  $\mathbf{W}$ . Consider for an example counties or states arranged geographically on a grid or some linear scale such as a line from one coast of the country to another. Typically  $W_{ij}$  will equal one for  $i, j$  pairs that are neighbors and zero otherwise. Alternatively,  $W_{ij}$  may reflect distances across space, so that  $W_{ij}$  decreases with increases in  $|i - j|$ . This would be similar to a temporal autocorrelation matrix. Assuming that  $|\lambda|$  is less than one, and that the elements of  $\mathbf{W}$  are such that  $(\mathbf{I} - \lambda \mathbf{W})$  is nonsingular, we may write

$$\boldsymbol{\varepsilon}_t = (\mathbf{I}_n - \lambda \mathbf{W})^{-1} \mathbf{v}_t,$$

so for the  $n$  observations at time  $t$ ,

$$\mathbf{y}_t = \mathbf{X}_t \boldsymbol{\beta} + (\mathbf{I}_n - \lambda \mathbf{W})^{-1} \mathbf{v}_t + \mathbf{u}.$$

We further assume that  $u_i$  and  $v_i$  have zero means, variances  $\sigma_u^2$  and  $\sigma_v^2$  and are independent across countries and of each other. It follows that a generalized regression model

applies to the  $n$  observations at time  $t$ :

$$E[\mathbf{y}_t | \mathbf{X}_t] = \mathbf{X}_t \boldsymbol{\beta},$$

$$\text{Var}[\mathbf{y}_t | \mathbf{X}_t] = (\mathbf{I}_n - \lambda \mathbf{W})^{-1} [\sigma_v^2 \mathbf{i} \mathbf{i}'] (\mathbf{I}_n - \lambda \mathbf{W})^{-1} + \sigma_u^2 \mathbf{I}_n.$$

At this point, estimation could proceed along the lines of Chapter 9, save for the need to estimate  $\lambda$ . There is no natural residual based estimator of  $\lambda$ . Recent treatments of this model have added a normality assumption and employed maximum likelihood methods. [The log likelihood function for this model and numerous references appear in Baltagi (2005, p. 196). Extensive analysis of the estimation problem is given in Bell and Bockstael (2000).]

A natural first step in the analysis is a test for spatial effects. The standard procedure for a cross section is Moran's (1950)  $I$  statistic, which would be computed for each set of residuals,  $\mathbf{e}_t$ , using

$$I_t = \frac{n \sum_{i=1}^n \sum_{j=1}^n W_{ij}(e_{it} - \bar{e}_t)(e_{jt} - \bar{e}_t)}{\left( \sum_{i=1}^n \sum_{j=1}^n W_{i,j} \right) \sum_{i=1}^n (e_{it} - \bar{e}_t)^2}. \quad (11-58)$$

For a panel of  $T$  independent sets of observations,  $\bar{I} = \frac{1}{T} \sum_{t=1}^T I_t$  would use the full set of information. A large sample approximation to the variance of the statistic under the null hypothesis of no spatial autocorrelation is

$$V^2 = \frac{1}{T} \frac{n^2 \sum_{i=1}^n \sum_{j=1}^n W_{ij}^2 + 3 \left( \sum_{i=1}^n \sum_{j=1}^n W_{ij} \right)^2 - n \sum_{i=1}^n \left( \sum_{j=1}^n W_{ij} \right)^2}{(n^2 - 1) \left( \sum_{i=1}^n \sum_{j=1}^n W_{ij} \right)^2}. \quad (11-59)$$

The statistic  $\bar{I}/V$  will converge to standard normality under the null hypothesis and can form the basis of the test. (The assumption of independence across time is likely to be dubious at best, however.) Baltagi, Song, and Koh (2003) identify a variety of LM tests based on the assumption of normality. Two that apply to cross section analysis [See Bell and Bockstael (2000, p. 78)] are

$$LM(1) = \frac{(\mathbf{e}' \mathbf{W} \mathbf{e} / s^2)^2}{\text{tr}(\mathbf{W}' \mathbf{W} + \mathbf{W}^2)}$$

for spatial autocorrelation and

$$LM(2) = \frac{(\mathbf{e}' \mathbf{W} \mathbf{y} / s^2)^2}{\mathbf{b}' \mathbf{X}' \mathbf{W} \mathbf{M} \mathbf{W} \mathbf{X} \mathbf{b} / s^2 + \text{tr}(\mathbf{W}' \mathbf{W} + \mathbf{W}^2)}$$

for spatially lagged dependent variables, where  $\mathbf{e}$  is the vector of OLS residuals,  $s^2 = \mathbf{e}' \mathbf{e} / n$ , and  $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'$ . [See Anselin and Hudak (1992).]

Anselin (1988) identifies several possible extensions of the spatial model to dynamic regressions. A “pure space-recursive model” specifies that the autocorrelation pertains to neighbors in the previous period:

$$y_{it} = \gamma [\mathbf{W} \mathbf{y}_{t-1}]_i + \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it}.$$

A “time-space recursive model” specifies dependence that is purely autoregressive with respect to neighbors in the previous period:

$$y_{it} = \rho y_{i,t-1} + \gamma [\mathbf{W} \mathbf{y}_{t-1}]_i + \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it}.$$

## 392 PART II ♦ Generalized Regression Model and Equation Systems

A “time-space simultaneous” model specifies that the spatial dependence is with respect to neighbors in the current period:

$$y_{it} = \rho y_{i,t-1} + [\lambda \mathbf{W} \mathbf{y}_t]_i + \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it}.$$

Finally, a “time-space dynamic model” specifies that autoregression depends on neighbors in both the current and last period:

$$y_{it} = \rho y_{i,t-1} + [\lambda \mathbf{W} \mathbf{y}_t]_i + \gamma [\mathbf{W} \mathbf{y}_{t-1}]_i + \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it}.$$

### **Example 11.12 Spatial Autocorrelation in Real Estate Sales**

Bell and Bockstaal analyzed the problem of modeling spatial autocorrelation in large samples. This is likely to become an increasingly common problem with GIS (geographic information system) data sets. The central problem is maximization of a likelihood function that involves a sparse matrix,  $(\mathbf{I} - \lambda \mathbf{W})$ . Direct approaches to the problem can encounter severe inaccuracies in evaluation of the inverse and determinant. Kelejian and Prucha (1999) have developed a moment-based estimator for  $\lambda$  that helps to alleviate the problem. Once the estimate of  $\lambda$  is in hand, estimation of the spatial autocorrelation model is done by FGLS. The authors applied the method to analysis of a cross section of 1,000 residential sales in Anne Arundel County, Maryland, from 1993 to 1996. The parcels sold all involved houses built within one year prior to the sale. GIS software was used to measure attributes of interest.

The model is

$$\begin{aligned} \ln Price &= \alpha + \beta_1 \ln \text{Assessed value (LIV)} \\ &\quad + \beta_2 \ln \text{Lot size (LLT)} \\ &\quad + \beta_3 \ln \text{Distance in km to Washington, DC (LDC)} \\ &\quad + \beta_4 \ln \text{Distance in km to Baltimore (LBA)} \\ &\quad + \beta_5 \% \text{ land surrounding parcel in publicly owned space (POPN)} \\ &\quad + \beta_6 \% \text{ land surrounding parcel in natural privately owned space (PNAT)} \\ &\quad + \beta_7 \% \text{ land surrounding parcel in intensively developed use (PDEV)} \\ &\quad + \beta_8 \% \text{ land surrounding parcel in low density residential use (PLOW)} \\ &\quad + \beta_9 \text{ Public sewer service (1 if existing or planned, 0 if not) (PSEW)} \\ &\quad + \varepsilon. \end{aligned}$$

(Land surrounding the parcel is all parcels in the GIS data whose centroids are within 500 meters of the transacted parcel.) For the full model, the specification is

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \varepsilon,$$

$$\varepsilon = \lambda \mathbf{W} \varepsilon + \mathbf{v}.$$

The authors defined four contiguity matrices:

- W1:  $W_{ij} = 1/\text{distance between } i \text{ and } j \text{ if distance} < 600 \text{ meters, 0 otherwise,}$
- W2:  $W_{ij} = 1 \text{ if distance between } i \text{ and } j < 200 \text{ meters, 0 otherwise,}$
- W3:  $W_{ij} = 1 \text{ if distance between } i \text{ and } j < 400 \text{ meters, 0 otherwise,}$
- W4:  $W_{ij} = 1 \text{ if distance between } i \text{ and } j < 600 \text{ meters, 0 otherwise.}$

All contiguity matrices were row-standardized. That is, elements in each row are scaled so that the row sums to one. One of the objectives of the study was to examine the impact of row standardization on the estimation. It is done to improve the numerical stability of the optimization process. Because the estimates depend numerically on the normalization, it is not completely innocent.

Test statistics for spatial autocorrelation based on the OLS residuals are shown in Table 11.9. (These are taken from the authors' Table 3.) The Moran statistics are distributed as standard normal while the LM statistics are distributed as chi-squared with one degree

**TABLE 11.9** Test Statistics for Spatial Autocorrelation

	<b>W1</b>	<b>W2</b>	<b>W3</b>	<b>W4</b>
Moran's <i>I</i>	7.89	9.67	13.66	6.88
LM(1)	49.95	84.93	156.48	36.46
LM(2)	7.40	17.22	2.33	7.42

**TABLE 11.10** Estimated Spatial Regression Models

<b>Parameter</b>	<b>OLS</b>		<b>FGLS<sup>a</sup></b>		<i>Spatial based on W1 ML</i>		<i>Spatial based on W1 Gen. Moments</i>	
	<b>Estimate</b>	<b>Std.Err.</b>	<b>Estimate</b>	<b>Std.Err.</b>	<b>Estimate</b>	<b>Std.Err.</b>	<b>Estimate</b>	<b>Std.Err.</b>
$\alpha$	4.7332	0.2047	4.7380	0.2048	5.1277	0.2204	5.0648	0.2169
$\beta_1$	0.6926	0.0124	0.6924	0.0214	0.6537	0.0135	0.6638	0.0132
$\beta_2$	0.0079	0.0052	0.0078	0.0052	0.0002	0.0052	0.0020	0.0053
$\beta_3$	-0.1494	0.0195	-0.1501	0.0195	-0.1774	0.0245	-0.1691	0.0230
$\beta_4$	-0.0453	0.0114	-0.0455	0.0114	-0.0169	0.0156	-0.0278	0.0143
$\beta_5$	-0.0493	0.0408	-0.0484	0.0408	-0.0149	0.0414	-0.0269	0.0413
$\beta_6$	0.0799	0.0177	0.0800	0.0177	0.0586	0.0213	0.0644	0.0204
$\beta_7$	0.0677	0.0180	0.0680	0.0180	0.0253	0.0221	0.0394	0.0211
$\beta_8$	-0.0166	0.0194	-0.0168	0.0194	-0.0374	0.0224	-0.0313	0.0215
$\beta_9$	-0.1187	0.0173	-0.1192	0.0174	-0.0828	0.0180	-0.0939	0.0179
$\lambda$	—	—	—	—	0.4582	0.0454	0.3517	—

<sup>a</sup>The author reports using a heteroscedasticity model  $\sigma_i^2 \times f(LIV_i, LIV_i^2)$ . The function  $f(\cdot)$  is not identified.

of freedom. All but the LM(2) statistic for W3 are larger than the 99% critical value from the respective table, so we would conclude that there is evidence of spatial autocorrelation. Estimates from some of the regressions are shown in Table 11.10. In the remaining results in the study, the authors find that the outcomes are somewhat sensitive to the specification of the spatial weight matrix, but not particularly so to the method of estimating  $\lambda$ .

#### **Example 11.13 Spatial Lags in Health Expenditures**

Moscone, Knapp, and Tosetti (2007) investigated the determinants of mental health expenditure over six years in 148 British local authorities using two forms of the spatial correlation model to incorporate possible interaction among authorities as well as unobserved spatial heterogeneity. The models estimated, in addition to pooled regression and a random effects model, were as follows. The first is a model with **spatial lags**:

$$\mathbf{y}_t = \gamma_t \mathbf{i} + \rho \mathbf{W} \mathbf{y}_t + \mathbf{X}_t \boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\varepsilon}_t,$$

where  $\mathbf{u}$  is a  $148 \times 1$  vector of random effects and  $\mathbf{i}$  is a  $148 \times 1$  column of ones. For each local authority,

$$y_{it} = \gamma_t + \rho (\mathbf{w}'_i \mathbf{y}_t) + \mathbf{x}'_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it},$$

where  $\mathbf{w}'_i$  is the  $i$ th row of the contiguity matrix,  $\mathbf{W}$ . Contiguities were defined in  $\mathbf{W}$  as one if the locality shared a border or vertex and zero otherwise. (The authors also experimented with other contiguity matrices based on “sociodemographic” differences.) The second model estimated is of **spatial error correlation**

$$\mathbf{y}_t = \gamma_t \mathbf{i} + \mathbf{X}_t \boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\varepsilon}_t,$$

$$\boldsymbol{\varepsilon}_t = \lambda \mathbf{W} \boldsymbol{\varepsilon}_t + \mathbf{v}_t.$$

## 394 PART II ♦ Generalized Regression Model and Equation Systems

For each local authority, this model implies

$$y_{it} = \gamma_t + \mathbf{x}'_{it}\beta + u_i + \lambda \sum_j w_{ij} \varepsilon_{jt} + v_{it}.$$

The authors use maximum likelihood to estimate the parameters of the model. To simplify the computations, they note that the maximization can be done using a two-step procedure. As we have seen in other applications, when  $\Omega$  in a generalized regression model is known, the appropriate estimator is GLS. For both of these models, with known spatial autocorrelation parameter, a GLS transformation of the data produces a classical regression model. [See (9-11).] The method used is to iterate back and forth between simple OLS estimation of  $\gamma_t$ ,  $\beta$  and  $\sigma^2_\varepsilon$  and maximization of the “concentrated log likelihood” function which, given the other estimates, is a function of the spatial autocorrelation parameter,  $\rho$  or  $\lambda$ , and the variance of the heterogeneity,  $\sigma^2_u$ .

The dependent variable in the models is the log of per capita mental health expenditures. The covariates are the percentage of males and of people under 20 in the area, average mortgage rates, numbers of unemployment claims, employment, average house price, median weekly wage, percent of single parent households, dummy variables for Labour party or Liberal Democrat party authorities, and the density of population (“to control for supply-side factors”). The estimated spatial autocorrelation coefficients for the two models are 0.1579 and 0.1220, both more than twice as large as the estimated standard error. Based on the simple Wald tests, the hypothesis of no spatial correlation would be rejected. The log likelihood values for the two spatial models were +206.3 and +202.8, compared to -211.1 for the model with no spatial effects or region effects, so the results seem to favor the spatial models based on a chi-squared test statistic (with one degree of freedom) of twice the difference. However, there is an ambiguity in this result as the improved “fit” could be due to the region effects rather than the spatial effects. A simple random effects model shows a log likelihood value of +202.3, which bears this out. Measured against this value, the spatial lag model seems the preferred specification, whereas the spatial autocorrelation model does not add significantly to the log likelihood function compared to the basic random effects model.

### 11.8 ENDOGENEITY

Recent **panel data** applications have relied heavily on the methods of instrumental variables. We will develop this methodology in detail in Chapter 13 where we consider generalized method of moments (GMM) estimation. At this point, we can examine two major building blocks in this set of methods, Hausman and Taylor’s (1981) estimator for the random effects model and Bhargava and Sargan’s (1983) proposals for estimating a dynamic panel data model. These two tools play a significant role in the GMM estimators of dynamic panel models in Chapter 13.

#### 11.8.1 HAUSMAN AND TAYLOR’S INSTRUMENTAL VARIABLES ESTIMATOR

Recall the original specification of the linear model for panel data in (11-1):

$$y_{it} = \mathbf{x}'_{it}\beta + \mathbf{z}'_i\alpha + \varepsilon_{it}. \quad (11-60)$$

The random effects model is based on the assumption that the unobserved person-specific effects,  $\mathbf{z}_i$ , are uncorrelated with the included variables,  $\mathbf{x}_{it}$ . This assumption is a major shortcoming of the model. However, the random effects treatment does allow the model to contain observed time-invariant characteristics, such as demographic characteristics, while the fixed effects model does not—if present, they are simply absorbed into the fixed effects. **Hausman and Taylor’s (1981) estimator** for the random effects model suggests a way to overcome the first of these while accommodating the second.

Their model is of the form:

$$y_{it} = \mathbf{x}'_{1it}\boldsymbol{\beta}_1 + \mathbf{x}'_{2it}\boldsymbol{\beta}_2 + \mathbf{z}'_{1i}\boldsymbol{\alpha}_1 + \mathbf{z}'_{2i}\boldsymbol{\alpha}_2 + \varepsilon_{it} + u_i$$

where  $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)'$  and  $\boldsymbol{\alpha} = (\boldsymbol{\alpha}'_1, \boldsymbol{\alpha}'_2)'$ . In this formulation, all individual effects denoted  $\mathbf{z}_i$  are observed. As before, unobserved individual effects that are contained in  $\mathbf{z}'_i\boldsymbol{\alpha}$  in (11-60) are contained in the person specific random term,  $u_i$ . Hausman and Taylor define four sets of *observed* variables in the model:

- $\mathbf{x}_{1it}$  is  $K_1$  variables that are time varying and uncorrelated with  $u_i$ ,
- $\mathbf{z}_{1i}$  is  $L_1$  variables that are time-invariant and uncorrelated with  $u_i$ ,
- $\mathbf{x}_{2it}$  is  $K_2$  variables that are time varying and are correlated with  $u_i$ ,
- $\mathbf{z}_{2i}$  is  $L_2$  variables that are time-invariant and are correlated with  $u_i$ .

The assumptions about the random terms in the model are

$$\begin{aligned} E[u_i | \mathbf{x}_{1it}, \mathbf{z}_{1i}] &= 0 \text{ though } E[u_i | \mathbf{x}_{2it}, \mathbf{z}_{2i}] \neq 0, \\ \text{Var}[u_i | \mathbf{x}_{1it}, \mathbf{z}_{1i}, \mathbf{x}_{2it}, \mathbf{z}_{2i}] &= \sigma_u^2, \\ \text{Cov}[\varepsilon_{it}, u_i | \mathbf{x}_{1it}, \mathbf{z}_{1i}, \mathbf{x}_{2it}, \mathbf{z}_{2i}] &= 0, \\ \text{Var}[\varepsilon_{it} + u_i | \mathbf{x}_{1it}, \mathbf{z}_{1i}, \mathbf{x}_{2it}, \mathbf{z}_{2i}] &= \sigma^2 = \sigma_\varepsilon^2 + \sigma_u^2, \\ \text{Corr}[\varepsilon_{it} + u_i, \varepsilon_{is} + u_i | \mathbf{x}_{1it}, \mathbf{z}_{1i}, \mathbf{x}_{2it}, \mathbf{z}_{2i}] &= \rho = \sigma_u^2 / \sigma^2. \end{aligned}$$

Note the crucial assumption that one can distinguish sets of variables  $\mathbf{x}_1$  and  $\mathbf{z}_1$  that are uncorrelated with  $u_i$  from  $\mathbf{x}_2$  and  $\mathbf{z}_2$  which are not. The likely presence of  $\mathbf{x}_2$  and  $\mathbf{z}_2$  is what complicates specification and estimation of the random effects model in the first place.

We note in passing that we can contrast the four assumptions with those made in Plümper and Troeger's (2007) FEVD formulation in Section 11.4.5 which, in the notation of this formulation, would be that  $\mathbf{x}_{1it}$  and  $\mathbf{x}_{2it}$  are time varying and both freely correlated with  $u_i$  while  $\mathbf{z}_{1i}$  and  $\mathbf{z}_{2i}$  are time invariant and are both uncorrelated with  $u_i$ . For both formulations, (11-61) applies. The two approaches differ in the additional moment conditions,  $E[\text{variable} \times (u_i + \varepsilon_{it})] = 0$ , that are used to identify the parameters  $\boldsymbol{\alpha}_1$  and  $\boldsymbol{\alpha}_2$ .

By construction, any OLS or GLS estimators of this model are inconsistent when the model contains variables that are correlated with the random effects. Hausman and Taylor have proposed an instrumental variables estimator that uses only the information within the model (i.e., as already stated). The strategy for estimation is based on the following logic: First, by taking deviations from group means, we find that

$$y_{it} - \bar{y}_{i.} = (\mathbf{x}_{1it} - \bar{\mathbf{x}}_{1i.})'\boldsymbol{\beta}_1 + (\mathbf{x}_{2it} - \bar{\mathbf{x}}_{2i.})'\boldsymbol{\beta}_2 + \varepsilon_{it} - \bar{\varepsilon}_{i.}, \quad (11-61)$$

which implies that both parts of  $\boldsymbol{\beta}$  can be consistently estimated by least squares, *in spite of the correlation between  $\mathbf{x}_2$  and  $u$* . This is the familiar, fixed effects, least squares dummy variable estimator—the transformation to deviations from group means  moves from the model the part of the disturbance that is correlated with  $\mathbf{x}_{2it}$ . Now,  the original model, Hausman and Taylor show that the group mean deviations can be used as  $(K_1 + K_2)$  instrumental variables for estimation of  $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ . That is the implication of (11-61). Because  $\mathbf{z}_1$  is uncorrelated with the disturbances, it can likewise serve as a set of  $L_1$  instrumental variables. That leaves a necessity for  $L_2$  instrumental variables. The authors show that the group means for  $\mathbf{x}_1$  can serve as these remaining instruments, and the model will be identified so long as  $K_1$  is greater than or equal to  $L_2$ . *For identification purposes, then,  $K_1$  must be at least as large as  $L_2$* . As usual,

## 396 PART II ♦ Generalized Regression Model and Equation Systems

**feasible GLS** is better than OLS, and available. Likewise, FGLS is an improvement over simple instrumental variable estimation of the model, which is consistent but inefficient.

The authors propose the following set of steps for consistent and efficient estimation:

**Step 1.** Obtain the LSDV (fixed effects) estimator of  $\beta = (\beta'_1, \beta'_2)'$  based on  $\mathbf{x}_1$  and  $\mathbf{x}_2$ . The residual variance estimator from this step is a consistent estimator of  $\sigma_\varepsilon^2$ .

**Step 2.** Form the within-groups residuals,  $e_{it}$ , from the LSDV regression at step 1. Stack the group means of these residuals in a full-sample-length data vector. Thus,  $e_{it}^* = \bar{e}_i = \frac{1}{T} \sum_{t=1}^T (y_{it} - \mathbf{x}'_{it} \mathbf{b}_w)$ ,  $t = 1, \dots, T, i = 1, \dots, n$ . (The individual constant term,  $a_i$ , is not included in  $e_{it}^*$ .) These group means are used as the dependent variable in an instrumental variable regression on  $\mathbf{z}_1$  and  $\mathbf{z}_2$  with instrumental variables  $\mathbf{z}_1$  and  $\mathbf{x}_1$ . (Note the identification requirement that  $K_1$ , the number of variables in  $\mathbf{x}_1$  be at least as large as  $L_2$ , the number of variables in  $\mathbf{z}_2$ .) The time-invariant variables are each repeated  $T$  times in the data matrices in this regression. This provides a consistent estimator of  $\alpha$ .

**Step 3.** The residual variance in the regression in step 2 is a consistent estimator of  $\sigma^{*2} = \sigma_u^2 + \sigma_\varepsilon^2/T$ . From this estimator and the estimator of  $\sigma_\varepsilon^2$  in step 1, we deduce an estimator of  $\sigma_u^2 = \sigma^{*2} - \sigma_\varepsilon^2/T$ . We then form the weight for feasible GLS in this model by forming the estimate of

$$\theta = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_u^2}}.$$

**Step 4.** The final step is a weighted instrumental variable estimator. Let the full set of variables in the model be

$$\mathbf{w}'_{it} = (\mathbf{x}'_{1it}, \mathbf{x}'_{2it}, \mathbf{z}'_{1i}, \mathbf{z}'_{2i}).$$

Collect these  $nT$  observations in the rows of data matrix  $\mathbf{W}$ . The transformed variables for GLS are, as before when we first fit the random effects model,

$$\mathbf{w}_{it}^{*'} = \mathbf{w}'_{it} - \hat{\theta} \bar{\mathbf{w}}'_i \quad \text{and} \quad y_{it}^* = y_{it} - \hat{\theta} \bar{y}_i.$$

where  $\hat{\theta}$  denotes the sample estimate of  $\theta$ . The transformed data are collected in the rows data matrix  $\mathbf{W}^*$  and in column vector  $\mathbf{y}^*$ . Note in the case of the time-invariant variables in  $\mathbf{w}_{it}$ , the group mean is the original variable, and the transformation just multiplies the variable by  $1 - \hat{\theta}$ . The instrumental variables are

$$\mathbf{v}'_{it} = [(\mathbf{x}_{1it} - \bar{\mathbf{x}}_{1i})', (\mathbf{x}_{2it} - \bar{\mathbf{x}}_{2i})', \mathbf{z}'_{1i}, \bar{\mathbf{x}}'_{1i}].$$

These are stacked in the rows of the  $nT \times (K_1 + K_2 + L_1 + K_1)$  matrix  $\mathbf{V}$ . Note for the third and fourth sets of instruments, the time-invariant variables and group means are repeated for each member of the group. The instrumental variable estimator would be

$$(\hat{\beta}', \hat{\alpha}')_{IV}' = [(\mathbf{W}^{*'} \mathbf{V})(\mathbf{V}' \mathbf{V})^{-1} (\mathbf{V}' \mathbf{W}^*)]^{-1} [(\mathbf{W}^{*'} \mathbf{V})(\mathbf{V}' \mathbf{V})^{-1} (\mathbf{V}' \mathbf{y}^*)].^{25} \quad (11-62)$$

<sup>25</sup>Note that the FGLS random effects estimator would be  $(\hat{\beta}', \hat{\alpha}')_{RE}' = [\mathbf{W}^{*'} \mathbf{W}^*]^{-1} \mathbf{W}^{*'} \mathbf{y}^*$ .

CHAPTER 11 ♦ Models for Panel Data **397**

The instrumental variable estimator is consistent if the data are not weighted, that is, if  $\mathbf{W}$  rather than  $\mathbf{W}^*$  is used in the computation. But, this is inefficient, in the same way that OLS is consistent but inefficient in estimation of the simpler random effects model.

**Example 11.14 The Returns to Schooling**

The economic returns to schooling have been a frequent topic of study by econometricians. The PSID and NLS data sets have provided a rich source of panel data for this effort. In wage (or log wage) equations, it is clear that the economic benefits of schooling are correlated with latent, unmeasured characteristics of the individual such as innate ability, intelligence, drive, or perseverance. As such, there is little question that simple random effects models based on panel data will suffer from the effects noted earlier. The fixed effects model is the obvious alternative, but these rich data sets contain many useful variables, such as race, union membership, and marital status, which are generally time invariant. Worse yet, the variable most of interest, years of schooling, is also time invariant. Hausman and Taylor (1981) proposed the estimator described here as a solution to these problems. The authors studied the effect of schooling on (the log of) wages using a random sample from the PSID of 750 men aged 25–55, observed in two years, 1968 and 1972. The two years were chosen so as to minimize the effect of serial correlation apart from the persistent unmeasured individual effects. The variables used in their model were as follows:

Experience = age—years of schooling—5,  
Years of schooling,  
Bad Health = a dummy variable indicating general health,  
Race = a dummy variable indicating nonwhite (70 of 750 observations),  
Union = a dummy variable indicating union membership,  
Unemployed = a dummy variable indicating previous year's unemployment.

 The model also included a constant term and a period indicator. [The coding of the latter is not given, but any two distinct values, including 0 for 1968 and 1 for 1972, would produce identical results. (Why?)]

The primary focus of the study is the coefficient on schooling in the log wage equation. Because schooling and, probably, Experience and Unemployed are correlated with the latent effect, there is likely to be serious bias in conventional estimates of this equation. Table 11.11 reports some of their reported results. The OLS and random effects GLS results in the first two columns provide the benchmark for the rest of the study. The schooling coefficient is estimated at 0.0669, a value which the authors suspected was far too small. As we saw earlier, even in the presence of correlation between measured and latent effects, in this model, the LSDV estimator provides a consistent estimator of the coefficients on the time varying variables. Therefore, we can use it in the **Hausman specification test** for correlation between the included variables and the latent heterogeneity. The calculations are shown in Section 11.5.4, result (11-42). Because there are three variables remaining in the LSDV equation, the chi-squared statistic has three degrees of freedom. The reported value of 20.2 is far larger than the 95 percent critical value of 7.81, so the results suggest that the random effects model is misspecified.

Hausman and Taylor proceeded to reestimate the log wage equation using their proposed estimator. The fourth and fifth sets of results in Table 11.11 present the instrumental variable estimates. The specification test given with the fourth set of results suggests that the procedure has produced the desired result. The hypothesis of the modified random effects model is now not rejected; the chi-squared value of 2.24 is much smaller than the critical value. The schooling variable is treated as endogenous (correlated with  $u_i$ ) in both cases. The difference between the two is the treatment of Unemployed and Experience. In the preferred equation, they are included in   the end result of the exercise is, again, the coefficient on schooling, which has risen from 0.0669 in the worst specification (OLS) to 0.2169 in the last one, a difference of over 200 percent. As the authors note, at the same time, the measured effect of race nearly vanishes.

**398 PART II ♦ Generalized Regression Model and Equation Systems**
**TABLE 11.11** Estimated Log Wage Equations

	<i>Variables</i>	<i>OLS</i>	<i>GLS/RE</i>	<i>LSDV</i>	<i>HT/IV-GLS</i>	<i>HT/IV-GLS</i>
$\mathbf{x}_1$	Experience	0.0132 (0.0011) <sup>a</sup>	0.0133 (0.0017)	0.0241 (0.0042)	0.0217 (0.0031)	
	Bad health	-0.0843 (0.0412)	-0.0300 (0.0363)	-0.0388 (0.0460)	-0.0278 (0.0307)	-0.0388 (0.0348)
	Unemployed	-0.0015 (0.0267)	-0.0402 (0.0207)	-0.0560 (0.0295)	-0.0559 (0.0246)	
	Last Year					
	Time	NR <sup>b</sup>	NR	NR	NR	NR 0.0241 (0.0045)
$\mathbf{x}_2$	Experience					-0.0560 (0.0279)
	Unemployed					
$\mathbf{z}_1$	Race	-0.0853 (0.0328)	-0.0878 (0.0518)		-0.0278 (0.0752)	-0.0175 (0.0764)
	Union	0.0450 (0.0191)	0.0374 (0.0296)		0.1227 (0.0473)	0.2240 (0.2863)
	Schooling	0.0669 (0.0033)	0.0676 (0.0052)			
	Constant	NR	NR	NR	NR	NR
$\mathbf{z}_2$	Schooling				0.1246 (0.0434)	0.2169 (0.0979)
	$\sigma_\epsilon$	0.321	0.192	0.160	0.190	0.629
$\rho = \sqrt{\sigma_u^2 / (\sigma_u^2 + \sigma_\epsilon^2)}$			0.632		0.661	0.817
Spec. Test [3]			20.2		2.24	0.00

<sup>a</sup>Estimated asymptotic standard errors are given in parentheses.

<sup>b</sup>NR indicates that the coefficient estimate was not reported in the study.

**11.8.2 CONSISTENT ESTIMATION OF DYNAMIC PANEL DATA MODELS: ANDERSON AND HSIAO'S IV ESTIMATOR**

Consider a homogeneous dynamic panel data model,

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}'_{it} \boldsymbol{\beta} + c_i + \varepsilon_{it}, \quad (11-63)$$

where  $c_i$  is, as in the preceding sections of this chapter, individual unmeasured heterogeneity, that may or may not be correlated with  $\mathbf{x}_{it}$ . We consider methods of estimation for this model when  $T$  is fixed and relatively small, and  $n$  may be large and increasing.

Pooled OLS is obviously inconsistent. Rewrite (11-63) as

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}'_{it} \boldsymbol{\beta} + w_{it}.$$

The disturbance in this pooled regression may be correlated with  $\mathbf{x}_{it}$ , but either way, it is surely correlated with  $y_{i,t-1}$ . By substitution,

$$\text{Cov}[y_{i,t-1}, (c_i + \varepsilon_{it})] = \sigma_c^2 + \gamma \text{Cov}[y_{i,t-2}, (c_i + \varepsilon_{it})],$$

and so on. By repeated substitution, it can be seen that for  $|\gamma| < 1$  and moderately large  $T$ ,

$$\text{Cov}[y_{i,t-1}, (c_i + \varepsilon_{it})] \approx \sigma_c^2 / (1 - \gamma). \quad (11-64)$$

[It is useful to obtain this result from a different direction. If the stochastic process that is generating  $(y_{it}, c_i)$  is *stationary*, then  $\text{Cov}[y_{i,t-1}, c_i] = \text{Cov}[y_{i,t-2}, c_i]$ , from which we would obtain (11-64) directly. The assumption  $|\gamma| < 1$  would be required for stationarity. We will return to this subject in Chapters 21 and 22.] Consequently, OLS and GLS are inconsistent. The fixed effects approach does not solve the problem either. Taking deviations from individual means, we have

$$y_{it} - \bar{y}_{i\cdot} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i\cdot})'\boldsymbol{\beta} + \gamma(y_{i,t-1} - \bar{y}_{i\cdot}) + (\varepsilon_{it} - \bar{\varepsilon}_{i\cdot}).$$

Anderson and Hsiao (1981, 1982) show that

$$\begin{aligned}\text{Cov}[(y_{it} - \bar{y}_{i\cdot}), (\varepsilon_{it} - \bar{\varepsilon}_{i\cdot})] &\approx \frac{-\sigma_\varepsilon^2}{T(1-\gamma)^2} \left[ \frac{(T-1)-T\gamma+\gamma^T}{T} \right] \\ &= \frac{-\sigma_\varepsilon^2}{T(1-\gamma)^2} \left[ (1-\gamma) - \frac{1-\gamma^T}{T} \right].\end{aligned}$$

This does converge to zero as  $T$  increases, but, again, we are considering cases in which  $T$  is small or moderate, say 5 to 15, in which case, the bias in the OLS estimator could be 15 percent to 60 percent. The implication is that the “within” transformation does not produce a consistent estimator.

It is easy to see that taking first differences is likewise ineffective. The first differences of the observations are

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \gamma(y_{i,t-1} - y_{i,t-2}) + (\varepsilon_{it} - \varepsilon_{i,t-1}). \quad (11-65)$$

As before, the correlation between the last regressor and the disturbance persists, so OLS or GLS based on first differences would also be inconsistent. There is another approach. Write the regression in differenced form as

$$\Delta y_{it} = \Delta \mathbf{x}'_{it} \boldsymbol{\beta} + \gamma \Delta y_{i,t-1} + \Delta \varepsilon_{it}$$

or, defining  $\mathbf{x}^*_{it} = [\Delta \mathbf{x}_{it}, \Delta y_{i,t-1}]$ ,  $\varepsilon^*_{it} = \Delta \varepsilon_{it}$  and  $\boldsymbol{\theta} = [\boldsymbol{\beta}', \gamma]'$

$$y^*_{it} = \mathbf{x}^*_{it}' \boldsymbol{\theta} + \varepsilon^*_{it}.$$

For the pooled sample, beginning with  $t = 3$ , write this as

$$\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\theta} + \boldsymbol{\varepsilon}^*.$$

The least squares estimator based on the first differenced data is

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \left[ \frac{1}{n(T-3)} \mathbf{X}^{*\prime} \mathbf{X}^* \right]^{-1} \left( \frac{1}{n(T-3)} \mathbf{X}^{*\prime} \mathbf{y}^* \right) \\ &= \boldsymbol{\theta} + \left[ \frac{1}{n(T-3)} \mathbf{X}^{*\prime} \mathbf{X}^* \right]^{-1} \left( \frac{1}{n(T-3)} \mathbf{X}^{*\prime} \boldsymbol{\varepsilon}^* \right).\end{aligned}$$

Assuming that the inverse matrix in brackets converges to a positive definite matrix—that remains to be shown—the inconsistency in this estimator arises because the vector in parentheses does not converge to zero. The last element is  $\text{plim}_{n \rightarrow \infty} [1/(n(T-3))] \sum_{t=3}^n (y_{i,t-1} - y_{i,t-2})(\varepsilon_{it} - \varepsilon_{i,t-1})$  which is not zero.

Suppose there were a variable  $\mathbf{z}^*$  such that  $\text{plim}[1/(n(T-3))] \mathbf{z}^{*\prime} \boldsymbol{\varepsilon}^* = 0$  and  $\text{plim}[1/(n(T-3))] \mathbf{z}^{*\prime} \mathbf{X}^* \neq \mathbf{0}$ . Let  $\mathbf{Z} = [\Delta \mathbf{X}, \mathbf{z}^*]$ ;  $z^*_{it}$  replaces  $\Delta y_{i,t-1}$  in  $\mathbf{x}^*_{it}$ . By this

## 400 PART II ♦ Generalized Regression Model and Equation Systems

construction, it appears we have a consistent estimator. Consider

$$\begin{aligned}\hat{\theta}_{IV} &= (\mathbf{Z}'\mathbf{X}^*)^{-1}\mathbf{Z}'\mathbf{y}^*. \\ &= (\mathbf{Z}'\mathbf{X}^*)^{-1}\mathbf{Z}'(\mathbf{X}^*\boldsymbol{\theta} + \boldsymbol{\varepsilon}^*) \\ &= \boldsymbol{\theta} + (\mathbf{Z}'\mathbf{X}^*)^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}^*.\end{aligned}$$

Then, after multiplying throughout by  $1/(n(T - 3))$  as before, we find

$$\text{Plim } \hat{\theta}_{IV} = \boldsymbol{\theta} + \text{plim}\{[1/(n(T - 3))](\mathbf{Z}'\mathbf{X}^*)\}^{-1} \times \mathbf{0},$$

which seems to solve the problem of consistent estimation.

The variable  $z^*$  is an **instrumental variable**, and the estimator is an **instrumental variable estimator** (hence the subscript on the preceding estimator). Finding suitable, valid instruments, that is, variables that satisfy the necessary assumptions, for models in which the right-hand variables are correlated with omitted factors is often challenging. In this setting, there is a natural candidate—in fact, there are several. From (11-65), we have at period  $t = 3$

$$y_{i3} - y_{i2} = (\mathbf{x}_{i3} - \mathbf{x}_{i2})'\boldsymbol{\beta} + \gamma(y_{i2} - y_{i1}) + (\varepsilon_{i3} - \varepsilon_{i2}).$$

We could use  $y_{i1}$  as the needed variable, because it is not correlated  $\varepsilon_{i3} - \varepsilon_{i2}$ . Continuing in this fashion, we see that for  $t = 3, 4, \dots, T$ ,  $y_{i,t-2}$  appears to satisfy our requirements. Alternatively, beginning from period  $t = 4$ , we can see that  $z_{it} = (y_{i,t-2} - y_{i,t-3})$  once again satisfies our requirements. This is Anderson and Hsiao's (1981) result for instrumental variable estimation of the dynamic panel data model. It now becomes a question of which approach, levels  $(y_{i,t-2}, t = 3, \dots, T)$ , or differences  $(y_{i,t-2} - y_{i,t-3}, t = 4, \dots, T)$  is a preferable approach. Arellano (1989) and Kiviet (1995) obtain results that suggest that the estimator based on levels is more efficient.

### 11.8.3 EFFICIENT ESTIMATION OF DYNAMIC PANEL DATA MODELS—THE ARELLANO/BOND ESTIMATORS

A leading contemporary application of the methods of this chapter is the **dynamic panel data model**, which we now write

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \delta y_{i,t-1} + c_i + \varepsilon_{it}.$$

Several applications are described in Example 11.21. The basic assumptions of the model are

1. Strict exogeneity:  $E[\varepsilon_{it} | \mathbf{X}_i, c_i] = 0,$
2. Homoscedasticity:  $E[\varepsilon_{it}^2 | \mathbf{X}_i, c_i] = \sigma_\varepsilon^2,$
3. Nonautocorrelation:  $E[\varepsilon_{it}\varepsilon_{is} | \mathbf{X}_i, c_i] = 0 \text{ if } t \neq s,$
4. Uncorrelated observations:  $E[\varepsilon_{it}\varepsilon_{js} | \mathbf{X}_i, c_i, \mathbf{X}_j, c_j] = 0 \text{ for } i \neq j \text{ and for all } t \text{ and } s,$

where the rows of the  $T \times K$  data matrix  $\mathbf{X}_i$  are  $\mathbf{x}'_{it}$ . We will not assume mean independence. The “effects” may be fixed or random, so we allow

$$E[c_i | \mathbf{X}_i] = g(\mathbf{X}_i).$$

(See Section 11.2.1.) We will also assume a fixed number of periods,  $T$ , for convenience. The treatment here (and in the literature) can be modified to accommodate unbalanced

CHAPTER 11 ♦ Models for Panel Data **401**

panels, but it is a bit inconvenient. (It involves the placement of zeros at various places in the data matrices defined below and, of course, changing the terminal indexes in summations from 1 to  $T$ .)

The presence of the lagged dependent variable in this model presents a considerable obstacle to estimation. Consider, first, the straightforward application of Assumption A.I3 in Section 8.2. The compound disturbance in the model is  $(c_i + \varepsilon_{it})$ . The correlation between  $y_{i,t-1}$  and  $(c_i + \varepsilon_{i,t})$  is obviously nonzero because  $y_{i,t-1} = \mathbf{x}'_{i,t-1}\boldsymbol{\beta} + \delta y_{i,t-2} + c_i + \varepsilon_{i,t-1}$ :

$$\text{Cov}[y_{i,t-1}, (c_i + \varepsilon_{it})] = \sigma_c^2 + \delta \text{Cov}[y_{i,t-2}, (c_i + \varepsilon_{it})].$$

If  $T$  is large and  $-\delta < 1$ , then this covariance will be approximately  $\sigma_c^2/(1 - \delta)$ . The large  $T$  assumption is not going to be met in most cases. But, because  $\delta$  will generally be positive, we can expect that this covariance will be at least larger than  $\sigma_c^2$ . The implication is that both (pooled) OLS and GLS in this model will be inconsistent. Unlike the case for the static model ( $\delta = 0$ ), the fixed effects treatment does not solve the problem. Taking group mean differences, we obtain

$$y_{i,t} - \bar{y}_{i\cdot} = (\mathbf{x}_{i,t} - \bar{\mathbf{x}}_{i\cdot})'\boldsymbol{\beta} + \delta(y_{i,t-1} - \bar{y}_{i\cdot}) + (\varepsilon_{i,t} - \bar{\varepsilon}_{i\cdot}).$$

As shown in Anderson and Hsiao (1981, 1982),

$$\text{Cov}[(y_{i,t-1} - \bar{y}_{i\cdot}), (\varepsilon_{i,t} - \bar{\varepsilon}_{i\cdot})] \approx \frac{-\sigma_\varepsilon^2}{T^2} \frac{(T-1) - T\delta + \delta^T}{(1-\delta)^2}.$$

This result is  $O(1/T)$ , which would generally be no problem if the asymptotics in our model were with respect to increasing  $T$ . But, in this panel data model,  $T$  is assumed to be fixed and relatively small. For conventional values of  $T$ , say 5 to 15, the proportional bias in estimation of  $\delta$  could be on the order of, say, 15 to 60 percent.

Neither OLS nor GLS are useful as estimators. There are, however, instrumental variables available within the structure of the model. Anderson and Hsiao (1981, 1982) proposed an approach based on first differences rather than differences from group means,

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \delta(y_{i,t-1} - y_{i,t-2}) + \varepsilon_{it} - \varepsilon_{i,t-1}.$$

For the first full observation,

$$y_{i3} - y_{i2} = (\mathbf{x}_{i3} - \mathbf{x}_{i2})'\boldsymbol{\beta} + \delta(y_{i2} - y_{i1}) + \varepsilon_{i3} - \varepsilon_{i2}, \quad (11-66)$$

the variable  $y_{i1}$  (assuming initial point  $t = 0$  is where our data generating process begins) satisfies the requirements, because  $\varepsilon_{i1}$  is predetermined with respect to  $(\varepsilon_{i3} - \varepsilon_{i2})$ . [That is, if we used only the data from periods 1 to 3 constructed as in (11-66), then the instrumental variables for  $(y_{i2} - y_{i1})$  would be  $\mathbf{z}_{i(3)}$  where  $\mathbf{z}_{i(3)} = (y_{1,1}, y_{2,1}, \dots, y_{n,1})$  for the  $n$  observations.] For the next observation,

$$y_{i4} - y_{i3} = (\mathbf{x}_{i4} - \mathbf{x}_{i3})'\boldsymbol{\beta} + \delta(y_{i3} - y_{i2}) + \varepsilon_{i4} - \varepsilon_{i3},$$

variables  $y_{i2}$  and  $(y_{i2} - y_{i1})$  are both available.

Based on the preceding paragraph, one might begin to suspect that there is, in fact, rather than a paucity of instruments, a large surplus. In this limited development, we have a choice between differences and levels. Indeed, we could use both and, moreover, in any period after the fourth, not only is  $y_{i2}$  available as an instrument, but so also is  $y_{i1}$ , and so

## 402 PART II ♦ Generalized Regression Model and Equation Systems

on. This is the essential observation behind the Arellano, Bover, and Bond (1991, 1995) estimators, which are based on the very large number of candidates for instrumental variables in this panel data model. To begin, with the model in first differences form, for  $y_{i3} - y_{i2}$ , variable  $y_{i1}$  is available. For  $y_{i4} - y_{i3}$ ,  $y_{i1}$  and  $y_{i2}$  are both available; for  $y_{i5} - y_{i4}$ , we have  $y_{i1}$ ,  $y_{i2}$ , and  $y_{i3}$ , and so on. Consider, as well, that we have not used the exogenous variables. With strictly exogenous regressors, not only are all lagged values of  $y_{is}$  for  $s$  previous to  $t - 1$ , but all values of  $\mathbf{x}_{it}$  are also available as instruments. For example, for  $y_{i4} - y_{i3}$ , the candidates are  $y_{i1}, y_{i2}$  and  $(\mathbf{x}'_{i1}, \mathbf{x}'_{i2}, \dots, \mathbf{x}'_{iT})$  for all  $T$  periods. The number of candidates for instruments is, in fact, potentially huge. [See Ahn and Schmidt (1995) for a very detailed analysis.] If the exogenous variables are only predetermined, rather than strictly exogenous, then only  $E[\varepsilon_{it} | \mathbf{x}_{i,t}, \mathbf{x}_{i,t-1}, \dots, \mathbf{x}_{i1}] = 0$ , and only vectors  $\mathbf{x}_{is}$  from 1 to  $t - 1$  will be valid instruments in the differenced equation that contains  $\varepsilon_{it} - \varepsilon_{i,t-1}$ . [See Baltagi and Levin (1986) for an application.] This is hardly a limitation, given that in the end, for a moderate sized model, we may be considering potentially hundreds or thousands of instrumental variables for estimation of what is usually a small handful of parameters.

We now formulate the model in a more familiar form, so we can apply the instrumental variable estimator. In terms of the differenced data, the basic equation is

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \delta(y_{i,t-1} - y_{i,t-2}) + \varepsilon_{it} - \varepsilon_{i,t-1},$$

or

$$\Delta y_{it} = (\Delta \mathbf{x}_{it})'\boldsymbol{\beta} + \delta(\Delta y_{i,t-1}) + \Delta \varepsilon_{it}, \quad (11-67)$$

where  $\Delta$  is the first difference operator,  $\Delta a_t = a_t - a_{t-1}$  for any time-series variable (or vector)  $a_t$ . (It should be noted that a constant term and any time-invariant variables in  $\mathbf{x}_{it}$  will fall out of the first differences. We will recover these below after we develop the estimator for  $\boldsymbol{\beta}$ .) The parameters of the model to be estimated are  $\boldsymbol{\theta} = (\boldsymbol{\beta}', \delta)'$  and  $\sigma_\varepsilon^2$ . For convenience, write the model as

$$\tilde{y}_{it} = \tilde{\mathbf{x}}'_{it}\boldsymbol{\theta} + \tilde{\varepsilon}_{it}$$

We are going to define an instrumental variable estimator along the lines of (8-9) and (8-10). Because our data set is a panel, the counterpart to

$$\mathbf{Z}'\tilde{\mathbf{X}} = \sum_{i=1}^n \mathbf{z}_i \tilde{\mathbf{x}}'_i \quad (11-68)$$

in the cross-section case would seem to be

$$\begin{aligned} \mathbf{Z}'\tilde{\mathbf{X}} &= \sum_{i=1}^n \sum_{t=1}^T \mathbf{z}_{it} \tilde{\mathbf{x}}'_{it} = \sum_{i=1}^n \mathbf{Z}'_i \tilde{\mathbf{X}}'_i \\ \tilde{\mathbf{y}}_i &= \begin{bmatrix} \Delta y_{i3} \\ \Delta y_{i4} \\ \vdots \\ \Delta y_{iT_i} \end{bmatrix}, \quad \tilde{\mathbf{X}}_i = \begin{bmatrix} \Delta \mathbf{x}'_{i3} & \Delta y_{i2} \\ \Delta \mathbf{x}'_{i4} & \Delta y_{i3} \\ \vdots & \vdots \\ \Delta \mathbf{x}'_{iT_i} & \Delta y_{i,T-1} \end{bmatrix}, \end{aligned} \quad (11-69)$$

where there are  $(T - 2)$  observations (rows) and  $K + 1$  columns in  $\tilde{\mathbf{X}}_i$ . There is a complication, however, in that the number of instruments we have defined may vary by period, so the matrix computation in (11-69) appears to sum matrices of different sizes.

CHAPTER 11 ♦ Models for Panel Data **403**

Consider an alternative approach. If we used only the first full observations defined in (11-67), then the cross-section version would apply, and the set of instruments  $\mathbf{Z}$  in (11-68) with strictly exogenous variables would be the  $n \times (1 + KT)$  matrix

$$\mathbf{Z}_{(3)} = \begin{bmatrix} y_{1,1}, \mathbf{x}'_{1,1}, \mathbf{x}'_{1,2}, \dots, \mathbf{x}'_{1,T} \\ y_{2,1}, \mathbf{x}'_{2,1}, \mathbf{x}'_{2,2}, \dots, \mathbf{x}'_{2,T} \\ \vdots \\ y_{n,1}, \mathbf{x}'_{n,1}, \mathbf{x}'_{n,2}, \dots, \mathbf{x}'_{n,T} \end{bmatrix},$$

and the instrumental variable estimator of (8-9) would be based on

$$\tilde{\mathbf{X}}_{(3)} = \begin{bmatrix} \mathbf{x}'_{1,3} - \mathbf{x}'_{1,2} & y_{1,4} - y_{1,3} \\ \mathbf{x}'_{2,3} - \mathbf{x}'_{2,2} & y_{2,4} - y_{2,3} \\ \vdots & \vdots \\ \mathbf{x}'_{n,3} - \mathbf{x}'_{n,2} & y_{n,4} - y_{n,3} \end{bmatrix} \text{ and } \tilde{\mathbf{y}}_{(3)} = \begin{bmatrix} y_{1,3} - y_{1,2} \\ y_{2,3} - y_{2,2} \\ \vdots \\ y_{n,3} - y_{n,2} \end{bmatrix}.$$

The subscript “(3)” indicates the first observation used for the left-hand side of the equation. Neglecting the other observations, then, we could use these data to form the IV estimator in (8-9), which we label for the moment  $\hat{\theta}_{IV(3)}$ . Now, repeat the construction using the next (fourth) observation as the first, and, again, using only a single year of the panel. The data matrices are now

$$\tilde{\mathbf{X}}_{(4)} = \begin{bmatrix} \mathbf{x}'_{1,4} - \mathbf{x}'_{1,3} & y_{1,3} - y_{1,2} \\ \mathbf{x}'_{2,4} - \mathbf{x}'_{2,3} & y_{2,3} - y_{2,2} \\ \vdots & \vdots \\ \mathbf{x}'_{n,4} - \mathbf{x}'_{n,3} & y_{n,3} - y_{n,2} \end{bmatrix}, \tilde{\mathbf{y}}_{(4)} = \begin{bmatrix} y_{1,4} - y_{1,3} \\ y_{2,4} - y_{2,3} \\ \vdots \\ y_{n,4} - y_{n,3} \end{bmatrix}, \text{ and} \\ \mathbf{Z}_{(4)} = \begin{bmatrix} y_{1,1}, y_{1,2}, \mathbf{x}'_{1,1}, \mathbf{x}'_{1,2}, \dots, \mathbf{x}'_{1,T} \\ y_{2,1}, y_{2,2}, \mathbf{x}'_{2,1}, \mathbf{x}'_{2,2}, \dots, \mathbf{x}'_{2,T} \\ \vdots \\ y_{n,1}, y_{n,2}, \mathbf{x}'_{n,1}, \mathbf{x}'_{n,2}, \dots, \mathbf{x}'_{n,T} \end{bmatrix} \quad (11-70)$$

and we have a second IV estimator,  $\hat{\theta}_{IV(4)}$ , also based on  $n$  observations, but, now,  $2 + KT$  instruments. And so on.

We now need to reconcile the  $T - 2$  estimators of  $\theta$  that we have constructed,  $\hat{\theta}_{IV(3)}, \hat{\theta}_{IV(4)}, \dots, \hat{\theta}_{IV(T)}$ . We faced this problem in Section 11.5.8 where we examined Chamberlain's formulation of the fixed effects model. The minimum distance estimator suggested there and used in Carey's (1997) study of hospital costs in Example 11.10 provides a means of efficiently “averaging” the multiple estimators of the parameter vector. We will (as promised) return to the MDE in Chapter 13. For the present, we consider, instead, **Arellano and Bond's** (1991) [and Arellano and Bover's (1995)] **approach** to this problem. We will collect the full set of estimators in a counterpart to (11-56) and (11-57). First, combine the sets of instruments in a single matrix,  $\mathbf{Z}$ , where for each individual, we obtain the  $(T - 2) \times L$  matrix  $\mathbf{Z}_i$ . The definition of the rows of  $\mathbf{Z}_i$  depend on whether the regressors are assumed to be strictly exogenous or predetermined. For

## 404 PART II ♦ Generalized Regression Model and Equation Systems

strictly exogenous variables,

$$\mathbf{Z}_i = \begin{bmatrix} y_{i,1}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \dots, \mathbf{x}'_{i,T} & 0 & \dots & 0 \\ 0 & y_{i,1}, y_{i,2}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \dots, \mathbf{x}'_{i,T} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & y_{i,1}, y_{i,2}, \dots, y_{i,T-2}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \dots, \mathbf{x}'_{i,T} \end{bmatrix}, \quad (11-71a)$$

and  $L = \sum_{i=1}^{T-2}(i + TK) = (T - 2)(T - 1)/2 + (T - 2)TK$ . For only predetermined variables, the matrix of instrumental variables is

$$\mathbf{Z}_i = \begin{bmatrix} y_{i,1}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2} & 0 & \dots & 0 \\ 0 & y_{i,1}, y_{i,2}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \mathbf{x}'_{i,3} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & y_{i,1}, y_{i,2}, \dots, y_{i,T-2}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \dots, \mathbf{x}'_{i,T-1} \end{bmatrix}, \quad (11-71b)$$

and  $L = \sum_{i=1}^{T-2}(i(K+1) + K) = [(T-2)(T-1)/2](1+K) + (T-2)K$ . This construction does proliferate instruments (moment conditions, as we will see in Chapter 13). In the application in Example 11.15, we have a small panel with only  $T = 7$  periods, and we fit a model with only  $K = 4$  regressors in  $\mathbf{x}_{it}$ , plus the lagged dependent variable. The strict exogeneity assumption produces a  $\mathbf{Z}_i$  matrix that is  $(5 \times 135)$  for this case. With only the assumption of predetermined  $\mathbf{x}_{it}$ ,  $\mathbf{Z}_i$  collapses slightly to  $(5 \times 95)$ . For purposes of the illustration, we have used only the two previous observations on  $\mathbf{x}_{it}$ . This further reduces the matrix to

$$\mathbf{Z}_i = \begin{bmatrix} y_{i,1}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2} & 0 & \dots & 0 \\ 0 & y_{i,1}, y_{i,2}, \mathbf{x}'_{i,2}, \mathbf{x}'_{i,3} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & y_{i,1}, y_{i,2}, \dots, y_{i,T-2}, \mathbf{x}'_{i,T-2}, \mathbf{x}'_{i,T-1} \end{bmatrix}, \quad (11-71c)$$

which, with  $T = 7$  and  $K = 4$ , will be  $(5 \times 55)$ . [Baltagi (2005, Chapter 8) presents some alternative configurations of  $\mathbf{Z}_i$  that allow for mixtures of strictly exogenous and predetermined variables.]

Now, we can compute the two-stage least squares estimator in (11-10) using our definitions of the data matrices  $\mathbf{Z}_i$ ,  $\tilde{\mathbf{X}}_i$ , and  $\tilde{\mathbf{y}}_i$  and (11-69). This will be

$$\begin{aligned} \hat{\theta}_{IV} &= \left[ \left( \sum_{i=1}^n \tilde{\mathbf{X}}_i' \mathbf{Z}_i \right) \left( \sum_{i=1}^n \mathbf{Z}_i' \mathbf{Z}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{Z}_i' \tilde{\mathbf{X}}_i \right) \right]^{-1} \\ &\quad \times \left[ \left( \sum_{i=1}^n \tilde{\mathbf{X}}_i' \mathbf{Z}_i \right) \left( \sum_{i=1}^n \mathbf{Z}_i' \mathbf{Z}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{Z}_i' \tilde{\mathbf{y}}_i \right) \right]. \end{aligned} \quad (11-72)$$

The natural estimator of the asymptotic covariance matrix for the estimator would be

$$\text{Est. Asy. Var} [\hat{\theta}_{IV}] = \hat{\sigma}_{\Delta\varepsilon}^2 \left[ \left( \sum_{i=1}^n \tilde{\mathbf{X}}_i' \mathbf{Z}_i \right) \left( \sum_{i=1}^n \mathbf{Z}_i' \mathbf{Z}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{Z}_i' \mathbf{X}_i \right) \right]^{-1}, \quad (11-73)$$

CHAPTER 11 ♦ Models for Panel Data **405**

where

$$\hat{\sigma}_{\Delta\varepsilon}^2 = \frac{\sum_{i=1}^n \sum_{t=3}^T [(y_{it} - y_{i,t-1}) - (\mathbf{x}'_{it} - \mathbf{x}'_{i,t-1})' \hat{\beta} - \hat{\delta}(y_{i,t-1} - y_{i,t-2})]^2}{n(T-2)}. \quad (11-74)$$

However, this variance estimator is likely to underestimate the true asymptotic variance because the observations are autocorrelated for one period. Because  $(y_{it} - y_{i,t-1}) = \tilde{\mathbf{x}}'_it \theta + (\varepsilon_{it} - \varepsilon_{i,t-1}) = \tilde{\mathbf{x}}'_it \theta + v_{it}$ ,

$$\text{Cov}[v_{it}, v_{i,t-1}] = \text{Cov}[v_{it}, v_{i,t+1}] = -\sigma_v^2.$$

Covariances at longer lags or leads are zero. In the differenced model, though the disturbance covariance matrix is not  $\sigma_v^2 \mathbf{I}$ , it does take a particularly simple form.

$$\text{Cov} \begin{pmatrix} \varepsilon_{i,3} - \varepsilon_{i,2} \\ \varepsilon_{i,4} - \varepsilon_{i,3} \\ \varepsilon_{i,5} - \varepsilon_{i,4} \\ \dots \\ \varepsilon_{i,T} - \varepsilon_{i,T-1} \end{pmatrix} = \sigma_\varepsilon^2 \begin{bmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ 0 & -1 & 2 & \dots & 0 \\ \dots & \dots & -1 & \dots & -1 \\ 0 & 0 & \dots & -1 & 2 \end{bmatrix} = \sigma_\varepsilon^2 \boldsymbol{\Omega}_i. \quad (11-75)$$

The implication is that the estimator in (11-74) estimates not  $\sigma_\varepsilon^2$  but  $2\sigma_\varepsilon^2$ . However, simply dividing the estimator by two does not produce the correct asymptotic covariance matrix because the observations themselves are autocorrelated. As such, the matrix in (11-73) is inappropriate. (We encountered this issue in Theorem 9.1 and in Sections 9.2.3, 9.4.3, and 11.3.2.) An appropriate correction can be based on the counterpart to the White estimator that we developed in (11-3). For simplicity, let

$$\hat{\mathbf{A}} = \left[ \left( \sum_{i=1}^n \tilde{\mathbf{X}}'_i \mathbf{Z}_i \right) \left( \sum_{i=1}^n \mathbf{Z}'_i \mathbf{Z}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{Z}'_i \tilde{\mathbf{X}}_i \right) \right]^{-1}.$$

Then, a robust covariance matrix that accounts for the autocorrelation would be

$$\hat{\mathbf{A}} \left[ \left( \sum_{i=1}^n \tilde{\mathbf{X}}'_i \mathbf{Z}_i \right) \left( \sum_{i=1}^n \mathbf{Z}'_i \mathbf{Z}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{Z}'_i \hat{\mathbf{v}}_i \hat{\mathbf{v}}'_i \mathbf{Z}_i \right) \left( \sum_{i=1}^n \mathbf{Z}'_i \mathbf{Z}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{Z}'_i \tilde{\mathbf{X}}_i \right) \right] \hat{\mathbf{A}}. \quad (11-76)$$

[One could also replace the  $\hat{\mathbf{v}}_i \hat{\mathbf{v}}'_i$  in (11-76) with  $\hat{\sigma}_\varepsilon^2 \boldsymbol{\Omega}_i$  in (11-72) because this is the known expectation.]

It will be useful to digress briefly and examine the estimator in (11-72). The computations are less formidable than it might appear. Note that the rows of  $\mathbf{Z}_i$  in (11-71a,b,c) are orthogonal. It follows that the matrix

$$\mathbf{F} = \sum_{i=1}^n \mathbf{Z}'_i \mathbf{Z}_i$$

in (11-72) is block-diagonal with  $T-2$  blocks. The specific blocks in  $\mathbf{F}$  are

$$\begin{aligned} \mathbf{F}_t &= \sum_{i=1}^n \mathbf{z}'_{it} \mathbf{z}'_{it} \\ &= \mathbf{Z}'_{(t)} \mathbf{Z}_{(t)}, \end{aligned}$$

## 406 PART II ♦ Generalized Regression Model and Equation Systems

for  $t = 3, \dots, T$ . Because the number of instruments is different in each period—see (11-71)—these blocks are of different sizes, say,  $(L_t \times L_t)$ . The same construction shows that the matrix  $\sum_{i=1}^n \tilde{\mathbf{X}}_i' \mathbf{Z}_i$  is actually a partitioned matrix of the form

$$\sum_{i=1}^n \tilde{\mathbf{X}}_i' \mathbf{Z}_i = [\tilde{\mathbf{X}}_{(3)}' \mathbf{Z}_{(3)} \quad \tilde{\mathbf{X}}_{(4)}' \mathbf{Z}_{(4)} \quad \dots \quad \tilde{\mathbf{X}}_{(T)}' \mathbf{Z}_{(T)}],$$

where, again, the matrices are of different sizes; there are  $T - 2$  rows in each but the number of columns differs. It follows that the inverse matrix,  $(\sum_{i=1}^n \mathbf{Z}_i' \mathbf{Z}_i)^{-1}$ , is also block-diagonal, and that the matrix quadratic form in (11-72) can be written

$$\begin{aligned} \left( \sum_{i=1}^n \tilde{\mathbf{X}}_i' \mathbf{Z}_i \right) \left( \sum_{i=1}^n \tilde{\mathbf{Z}}_i' \mathbf{Z}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{Z}_i' \tilde{\mathbf{X}}_i \right) &= \sum_{t=3}^T (\tilde{\mathbf{X}}_{(t)}' \mathbf{Z}_{(t)}) (\mathbf{Z}_{(t)}' \mathbf{Z}_{(t)})^{-1} (\mathbf{Z}_{(t)}' \tilde{\mathbf{X}}_{(t)}) \\ &= \sum_{t=3}^T (\hat{\tilde{\mathbf{X}}}_{(t)}' \hat{\tilde{\mathbf{X}}}_{(t)}) \\ &= \sum_{t=3}^T \mathbf{W}_{(t)}, \end{aligned}$$

[see (8-9) and the preceding result]. Continuing in this fashion, we find

$$\left( \sum_{i=1}^n \tilde{\mathbf{X}}_i' \mathbf{Z}_i \right) \left( \sum_{i=1}^n \tilde{\mathbf{Z}}_i' \mathbf{Z}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{Z}_i' \tilde{\mathbf{y}}_i \right) = \sum_{t=3}^T \hat{\tilde{\mathbf{X}}}_{(t)}' \mathbf{y}_{(t)}.$$

From (8-10), we can see that

$$\begin{aligned} \hat{\tilde{\mathbf{X}}}_{(t)}' \mathbf{y}_{(t)} &= (\hat{\tilde{\mathbf{X}}}_{(t)}' \hat{\tilde{\mathbf{X}}}_{(t)}) \hat{\theta}_{IV}(t) \\ &= \mathbf{W}_{(t)} \hat{\theta}_{IV}(t). \end{aligned}$$

Combining the terms constructed thus far, we find that the estimator in (11-72) can be written in the form

$$\begin{aligned} \hat{\theta}_{IV} &= \left( \sum_{t=3}^T \mathbf{W}_{(t)} \right)^{-1} \left( \sum_{t=3}^T \mathbf{W}_{(t)} \hat{\theta}_{IV}(t) \right) \\ &= \sum_{t=3}^T \mathbf{R}_{(t)} \hat{\theta}_{IV}(t), \end{aligned}$$

where

$$\mathbf{R}_{(t)} = \left( \sum_{t=3}^T \mathbf{W}_{(t)} \right)^{-1} \mathbf{W}_{(t)} \text{ and } \sum_{t=3}^T \mathbf{R}_{(t)} = \mathbf{I}.$$

In words, we find that, as might be expected, the Arellano and Bond estimator of the parameter vector is a matrix weighted average of the  $T - 2$  period specific two-stage least squares estimators, where the instruments used in each period may differ. Because the estimator is an average of estimators, a question arises, is it an efficient average— are the weights chosen to produce an efficient estimator? Perhaps not surprisingly, the

CHAPTER 11 ♦ Models for Panel Data **407**

answer for this  $\hat{\theta}$  is no; there is a more efficient set of weights that can be constructed for this model. We will assemble them when we examine the generalized method of moments estimator in Chapter 13



There remains a loose end in the preceding. After (11-64), it was noted that this treatment discards a constant term and any time-invariant variables that appear in the model. The Hausman and Taylor (1981) approach developed in the preceding section suggests a means by which the model could be completed to accommodate this possibility. Expand the basic formulation to include the time-invariant effects, as

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \delta y_{i,t-1} + \alpha + \mathbf{f}'_i\boldsymbol{\gamma} + c_i + \varepsilon_{it},$$

where  $\mathbf{f}_i$  is the set of time-invariant variables and  $\boldsymbol{\gamma}$  is the parameter vector yet to be estimated. This model is consistent with the entire preceding development, as the component  $\alpha + \mathbf{f}'_i\boldsymbol{\gamma}$  would have fallen out of the differenced equation along with  $c_i$  at the first step at (11-63). Having developed a consistent estimator for  $\boldsymbol{\theta} = (\boldsymbol{\beta}', \delta)',$  we now turn to estimation of  $(\alpha, \boldsymbol{\gamma})'.$  The residuals from the IV regression (11-72),

$$w_{it} = \mathbf{x}'_{it}\hat{\boldsymbol{\beta}}_{IV} - \hat{\delta}_{IV}y_{i,t-1}$$

are pointwise consistent estimators of

$$\omega_{it} = \alpha + \mathbf{f}'_i\boldsymbol{\gamma} + c_i + \varepsilon_{it}.$$

Thus, the group means of the residuals can form the basis of a second-step regression;

$$\bar{w}_i = \alpha + \mathbf{f}'_i\boldsymbol{\gamma} + c_i + \bar{\varepsilon}_i + \eta_i \quad (11-76)$$

where  $\eta_i = (\bar{w}_i - \bar{\omega}_i)$  is the estimation error that converges to zero as  $\hat{\theta}$  converges to  $\boldsymbol{\theta}.$  The implication would seem to be that we can now linearly regress these group mean residuals on a constant and the time-invariant variables  $\mathbf{f}_i$  to estimate  $\alpha$  and  $\boldsymbol{\gamma}.$  The flaw in the strategy, however, is that the initial assumptions of the model do not state that  $c_i$  is uncorrelated with the other variables in the model, including the implicit time invariant terms,  $\mathbf{f}_i.$  Therefore, least squares is not a usable estimator here unless the random effects model is assumed, which we specifically sought to avoid at the outset. As in Hausman and Taylor's treatment, there is a workable strategy if it can be assumed that there are some variables in the model, including possibly some among the  $\mathbf{f}_i$  as well as others among  $\mathbf{x}_{it}$  that are uncorrelated with  $c_i$  and  $\varepsilon_{it}.$  These are the  $\mathbf{z}_1$  and  $\mathbf{x}_1$  in the Hausman and Taylor estimator (see step 2 in the development of the preceding section). Assuming that these variables are available—this is an identification assumption that must be added to the model—then we do have a usable instrumental variable estimator, using as instruments the constant term (1), any variables in  $\mathbf{f}_i$  that are uncorrelated with the latent effects or the disturbances (call this  $\mathbf{f}_{i1}$ ), and the group means of any variables in  $\mathbf{x}_{it}$  that are also exogenous. There must be enough of these to provide a sufficiently large set of instruments to fit all the parameters in (11-76). This is, once again, the same identification we saw in step 2 of the Hausman and Taylor estimator,  $K_1,$  the number of exogenous variables in  $\mathbf{x}_{it}$  must be at least as large as  $L_2,$  which is the number of endogenous variables in  $\mathbf{f}_i.$  With all this in place, we then have the instrumental variable estimator in which the dependent variable is  $\bar{w}_i,$  the right-hand-side variables are  $(1, \mathbf{f}_i),$  and the instrumental variables are  $(1, \mathbf{f}_{i1}, \bar{\mathbf{x}}_{i1}).$

## 408 PART II ♦ Generalized Regression Model and Equation Systems

There is yet another direction that we might extend this estimation method. In (11-73), we have implicitly allowed a more general covariance matrix to govern the generation of the disturbances  $\varepsilon_{it}$  and computed a robust covariance matrix for the simple IV estimator. We could take this a step further and look for a more efficient estimator. As a library of recent studies has shown, panel data sets are rich in information that allows the analyst to specify highly general models and to exploit the implied relationships among the variables to construct much more efficient generalized method of moments (GMM) estimators. [See, in particular, Arellano and Bover (1995) and Blundell and Bond (1998).] We will return to this development in Chapter 13.

### **Example 11.15 Dynamic Labor Supply Equation**

In Example 8.5, we used instrumental variables fit a labor supply equation,

$$Wks_{it} = \gamma_1 + \gamma_2 \ln Wage_{it} + \gamma_3 Ed_i + \gamma_4 Union_{it} + \gamma_5 Fem_i + u_{it}.$$

To illustrate the computations of this section, we will extend this model as follows:

$$\begin{aligned} Wks_{it} = & \beta_1 \ln Wage_{it} + \beta_2 Union_{it} + \beta_3 Occ_{it} + \beta_4 Exp_{it} + \delta Wks_{i,t-1} \\ & + \alpha + \gamma_1 Ed_i + \gamma_2 Fem_i + c_i + \varepsilon_{it}. \end{aligned}$$

(We have rearranged the variables and parameter names to conform to the notation in this section.) We note, in theoretical terms, as suggested in the earlier example, it may not be appropriate to treat  $\ln Wage_{it}$  as uncorrelated with  $\varepsilon_{it}$  or  $c_i$ . However, we will be analyzing the model in first differences. It may well be appropriate to treat changes in wages as exogenous. That would depend on the theoretical underpinnings of the model. We will treat the variable as predetermined here, and proceed. There are two time-invariant variables in the model,  $Fem_i$ , which is clearly exogenous, and  $Ed_i$ , which might be endogenous. The identification requirement for estimation of  $(\alpha, \gamma_1, \gamma_2)$  is met by the presence of three exogenous variables,  $Union_{it}$ ,  $Occ_{it}$ , and  $Exp_{it}$  ( $K_1 = 3$  and  $L_2 = 1$ ).

The differenced equation analyzed at the first step is

$$\Delta Wks_{it} = \beta_1 \Delta \ln Wage_{it} + \beta_2 \Delta Union_{it} + \beta_3 \Delta Occ_{it} + \beta_4 \Delta Exp_{it} + \delta \Delta Wks_{i,t-1} + \varepsilon_{it}$$

We estimated the parameters and the asymptotic covariance matrix according to (11-72) and (11-76). For specification of the instrumental variables, we used the one previous observation on  $x_{it}$ , as shown in the text.<sup>26</sup> Table 11.12 presents the computations with several other inconsistent estimators.

The various estimates are quite far apart. In the absence of the common effects (and autocorrelation of the disturbances), all five estimators shown would be consistent. Given the very wide disparities, one might suspect that common effects are an important feature of the data. The second standard errors given with the IV estimates are based on the uncorrected matrix in (11-73) with  $\hat{\sigma}_{\Delta e}^2$  in (11-74) divided by two. We found the estimator to be quite volatile, as can be seen in the table. The estimator is also very sensitive to the choice of instruments that comprise  $Z_i$ . Using (11-71a) instead of (11-71b) produces wild swings in the estimates and, in fact, produces implausible results. One possible explanation in this particular example is that the instrumental variables we are using are dummy variables that have relatively little variation over time.

<sup>26</sup>This estimator and the GMM estimators in Chapter 13 are built into some contemporary computer programs, including *NLOGIT* and *Stata*. Many researchers use Gauss programs that are distributed by M. Arellano, <http://www.cemfi.es/%7Earellano/#dpd>, or program the calculations themselves using *MatLab* or *R*. We have programmed the matrix computations directly for this application using the matrix package in *NLOGIT*.

**TABLE 11.12** Estimated Dynamic Panel Data Model Using Arellano and Bond's Estimator

Variable	<i>OLS, Full Eqn.</i>		<i>OLS, Differenced</i>		<i>IV, Differenced</i>		<i>Random Effects</i>		<i>Fixed Effects</i>	
	<i>Estimate</i>	<i>Std. Err.</i>	<i>Estimate</i>	<i>Std. Err.</i>	<i>Estimate</i>	<i>Std. Err.</i>	<i>Estimate</i>	<i>Std. Err.</i>	<i>Estimate</i>	<i>Std. Err.</i>
In Wage	0.2966	0.2052	-0.1100	0.4565	-1.1402	0.2639	0.2281	0.2405	0.5886	0.4790
Union	-1.2945	0.1713	1.1640	0.4222	2.7089	0.3684	-1.4104	0.2199	0.1444	0.4369
Occ	0.4163	0.2005	0.8142	0.3924	2.2808	1.3105	0.5191	0.2484	1.0064	0.4030
Exp	-0.0295	0.00728	-0.0742	0.0975	-0.0208	0.1126	-0.0353	0.01021	-0.1683	0.05954
Wks <sub>t-1</sub>	0.3804	0.01477	-0.3527	0.01609	0.1304	0.04760	0.2100	0.01511	0.0148	0.01705
Constant	28.918	1.4490			-0.4110	0.3364	37.461	1.6778		
Ed	-0.0690	0.03703			0.0321	0.02587	-0.0657	0.04988		
Fem	-0.8607	0.2544			-0.0122	0.1554	-1.1463	0.3513		
Sample	<i>t = 2 - 7</i>		<i>t = 3 - 7</i>		<i>t = 3 - 7; n = 595</i>		<i>t = 2 - 7</i>		<i>t = 2 - 7</i>	
	<i>n = 595</i>		<i>n = 595</i>		Means used <i>t = 7</i>		<i>n = 595</i>		<i>n = 595</i>	

## 410 PART II ♦ Generalized Regression Model and Equation Systems

### 11.8.4 NONSTATIONARY DATA AND PANEL DATA MODELS

Some of the discussion thus far (and to follow) focuses on “small  $T$ ” statistical results. Panels are taken to contain a fixed and small  $T$  observations on a large  $n$  individual units. Recent research using cross-country data sets such as the Penn World Tables ([http://pwt.econ.upenn.edu/php\\_site/pwt\\_index.php](http://pwt.econ.upenn.edu/php_site/pwt_index.php)), which now include data on nearly 200 countries for well over 50 years, have begun to analyze panels with  $T$  sufficiently large that the time-series properties of the data become an important consideration. In particular, the recognition and accommodation of nonstationarity that is now a standard part of single time-series analyses (as in Chapter 23) are now seen to be appropriate for large scale cross-country studies, such as income growth studies based on the Penn World Tables, cross-country studies of health care expenditure, and analyses of purchasing power parity.

The analysis of long panels, such as in the growth and convergence literature, typically involves dynamic models, such as

$$y_{it} = \alpha_i + \gamma_i y_{i,t-1} + \mathbf{x}'_{it} \boldsymbol{\beta}_i + \varepsilon_{it}. \quad (11-77)$$

In single time-series analysis involving low-frequency macroeconomic flow data such as income, consumption, investment, the current account deficit, and so on, it has long been recognized that estimated regression relations can be distorted by nonstationarity in the data. What appear to be persistent and strong regression relationships can be entirely spurious and due to underlying characteristics of the time-series processes rather than actual connections among the variables. Hypothesis tests about long-run effects will be considerably distorted by unit roots in the data. It has become evident that the same influences, with the same deleterious effects, will be found in long panel data sets. The panel data application is further complicated by the possible heterogeneity of the parameters. The coefficients of interest in many cross-country studies are the lagged effects, such as  $\gamma_i$  in (11-77), and it is precisely here that the received results on nonstationary data have revealed the problems of estimation and inference. Valid tests for unit roots in panel data have been proposed in many studies. Three that are frequently cited are Levin and Lin (1992), Im, Pesaran, and Shin (2003) and Maddala and Wu (1999).

There have been numerous empirical applications of time series methods for non-stationary data in panel data settings, including Frankel and Rose's (1996) and Pedroni's (2001) studies of purchasing power parity, Fleissig and Strauss (1997) on real wage stationarity, Culver and Papell (1997) on inflation, Wu (2000) on the current account balance, McCoskey and Selden (1998) on health care expenditure, Sala-i-Martin (1996) on growth and convergence, McCoskey and Kao (1999) on urbanization and production, and Coakley et al. (1996) on savings and investment. An extensive enumeration appears in Baltagi (2005, Chapter 12).

A subtle problem arises in obtaining results useful for characterizing the properties of estimators of the model in (11-77). The asymptotic results based on large  $n$  and large  $T$  are not necessarily obtainable simultaneously, and great care is needed in deriving the asymptotic behavior of useful statistics. Phillips and Moon (1999, 2000) are standard references on the subject.

We will return to the topic of nonstationary data in Chapter 23. This is an emerging literature, most of which is well beyond the level of this text. We will rely on the several

detailed received surveys, such as Bannerjee (1999), Smith (2000), and Baltagi and Kao (2000) to fill in the details.

## 11.9 NONLINEAR REGRESSION WITH PANEL DATA

The extension of the panel data models to the nonlinear regression case is, perhaps surprisingly, not at all straightforward. Thus far, to accommodate the nonlinear model, we have generally applied familiar results to the linearized regression. This approach will carry forward to the case of clustered data. (See Section 11.3.3.) Unfortunately, this will not work with the standard panel data methods. The nonlinear regression will be the first of numerous panel data applications that we will consider in which the wisdom of the linear regression model cannot be extended to the more general framework.

### 11.9.1 A ROBUST COVARIANCE MATRIX FOR NONLINEAR LEAST SQUARES

The counterpart to (11-3) or (11-4) would simply replace  $\mathbf{X}_i$  with  $\hat{\mathbf{X}}_i^0$  where the rows are the pseudoregressors for cluster  $i$  as defined in (7-12) and “ $^0$ ” indicates that it is computed using the nonlinear least squares estimates of the parameters.

#### *Example 11.16 Health Care Utilization*

The recent literature in health economics includes many studies of health care utilization. A common measure of the dependent variable of interest is a count of the number of encounters with the health care system, either through visits to a physician or to a hospital. These counts of occurrences are usually studied with the Poisson regression model described in Section 19.1. The nonlinear regression model is

$$E[y_i | \mathbf{x}_i] = \exp(\mathbf{x}_i'\beta).$$

A recent study in this genre is “Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation” by Riphahn, Wambach, and Million (2003). The authors were interested in counts of physician visits and hospital visits. In this application, they were particularly interested in the impact that the presence of private insurance has on the utilization counts of interest, that is, whether the data contain evidence of moral hazard.

The raw data are published on the *Journal of Applied Econometrics* data archive web site. The URL for the data file is <http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/>. The variables in the data file are listed in Appendix Table F7.1. The sample is an unbalanced panel of 7,293 households, the German Socioeconomic Panel data set. The number of observations varies from one to seven (1,525; 1,079; 825; 926; 1,311; 1,000; 887) with a total number of observations of 27,326. We will use these data in several examples here and later in the book.

The following model uses a simple specification for the count of number of visits to the physician in the observation year,

$$\mathbf{x}_{it} = (1, age_{it}, educ_{it}, income_{it}, kids_{it})$$

Table 11.13 details the nonlinear least squares iterations and the results. The convergence criterion for the iterations is  $\mathbf{e}^0\mathbf{X}^0(\mathbf{X}^0\mathbf{X}^0)^{-1}\mathbf{e}^0 < 10^{-10}$ . Although this requires 11 iterations, the function actually reaches the minimum in 7. The estimates of the asymptotic standard errors are computed using the conventional method,  $s^2(\hat{\mathbf{X}}^0\hat{\mathbf{X}}^0)^{-1}$  and then by the cluster correction in (11-4). The corrected standard errors are considerably larger, as might be expected given that these are a panel data set.

## 412 PART II ♦ Generalized Regression Model and Equation Systems

**TABLE 11.13** Nonlinear Least Squares Estimates of a Utilization Equation

Begin NLSQ iterations. Linearized regression.  
 Iteration = 1; Sum of squares = 1014865.00; Gradient = 156281.794  
 Iteration = 2; Sum of squares = 8995221.17; Gradient = 8131951.67  
 Iteration = 3; Sum of squares = 1757006.18; Gradient = 897066.012  
 Iteration = 4; Sum of squares = 930876.806; Gradient = 73036.2457  
 Iteration = 5; Sum of squares = 860068.332; Gradient = 2430.80472  
 Iteration = 6; Sum of squares = 857614.333; Gradient = 12.8270683  
 Iteration = 7; Sum of squares = 857600.927; Gradient = 0.411851239E-01  
 Iteration = 8; Sum of squares = 857600.883; Gradient = 0.190628165E-03  
 Iteration = 9; Sum of squares = 857600.883; Gradient = 0.904650588E-06  
 Iteration = 10; Sum of squares = 857600.883; Gradient = 0.430441193E-08  
 Iteration = 11; Sum of squares = 857600.883; Gradient = 0.204875467E-10

Convergence achieved

Variable	Estimate	Standard Error	Robust Standard Error
Constant	0.9801	0.08927	0.12522
Age	0.01873	0.001053	0.00142
Education	-0.03613	0.005732	0.00780
Income	-0.5911	0.07173	0.09702
Kids	-0.1692	0.02642	0.03330

### 11.9.2 FIXED EFFECTS

The nonlinear panel data regression model would appear

$$y_{it} = h(\mathbf{x}_{it}, \boldsymbol{\beta}) + \varepsilon_{it}, t = 1, \dots, T_i, i = 1, \dots, n.$$

Consider a model with latent heterogeneity,  $c_i$ . An ambiguity immediately emerges; how should heterogeneity enter the model. Building on the linear model, an additive term might seem natural, as in

$$y_{it} = h(\mathbf{x}_{it}, \boldsymbol{\beta}) + c_i + \varepsilon_{it}, t = 1, \dots, T_i, i = 1, \dots, n. \quad (11-78)$$

But we can see in the previous application that this is likely to be inappropriate. The loglinear model of the previous section is constrained to ensure that  $E[y_{it} | \mathbf{x}_{it}]$  is positive. But an additive random term  $c_i$  as in (11-78) could subvert this; unless the range of  $c_i$  is restricted, the conditional mean could be negative. The most common application of nonlinear models is the **index function model**,

$$y_{it} = h(\mathbf{x}'_{it}\boldsymbol{\beta} + c_i) + \varepsilon_{it}.$$

This is the natural extension of the linear model, but only in the appearance of the conditional mean. Neither the fixed effects nor the random effects model can be estimated as they were in the linear case.

Consider the fixed effects model first. We would write this as

$$y_{it} = h(\mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i) + \varepsilon_{it},$$



where the parameters to be estimated are  $\boldsymbol{\beta}$  and  $\alpha_i, i = 1, \dots, n$ . Transforming the data to deviations from group means does not remove the fixed effects from the model.

For example,

$$y_{it} - \bar{y}_{it} = h(\mathbf{x}'_{it}\beta + \alpha_i) - \frac{1}{T_i} \sum_{s=1}^{T_i} h(\mathbf{x}'_{is}\beta + \alpha_i), \quad (11-79)$$

which does not simplify things at all. Transforming the regressors to deviations is likewise pointless. To estimate the parameters, it is necessary to minimize the sum of squares with respect to all  $n + K$  parameters simultaneously. Because the number of dummy variable coefficients can be huge—the preceding example is based on a data set with 7,293 groups—this can be a difficult or impractical computation. A method of maximizing a function (such as the negative of the sum of squares) that contains an unlimited number of dummy variable coefficients is shown in Chapter 17. As we will examine later in the book, the difficulty with nonlinear models that contain large numbers of dummy variable coefficients is not necessarily the practical one of computing the estimates. That is generally a solvable problem. The difficulty with such models is an intriguing phenomenon known as the **incidental parameters problem**. In most (not all, as we shall find) nonlinear panel data models that contain  $n$  dummy variable coefficients, such as the one in (11-79), as a consequence of the fact that the number of parameters increases with the number of individuals in the sample, the estimator of  $\beta$  is biased and inconsistent, to a degree that is  $O(1/T)$ . Because  $T$  is only 7 or less in our application, this would seem to be a case in point.

#### **Example 11.17 Exponential Model with Fixed Effects**

The exponential model of the preceding example is actually one of a small handful of known special cases in which it is possible to “condition” out the dummy variables. Consider the sum of squared residuals,

$$S_n = \frac{1}{2} \sum_{i=1}^n \sum_{t=1}^{T_i} [y_{it} - \exp(\mathbf{x}'_{it}\beta + \alpha_i)]^2.$$

The first order condition for minimizing  $S_n$  with respect to  $\alpha_i$  is

$$\frac{\partial S_n}{\partial \alpha_i} = \sum_{t=1}^{T_i} -[y_{it} - \exp(\mathbf{x}'_{it}\beta + \alpha_i)] \exp(\mathbf{x}'_{it}\beta + \alpha_i) = 0. \quad (11-80)$$

Let  $\gamma_i = \exp(\alpha_i)$ . Then, an equivalent necessary condition would be

$$\frac{\partial S_n}{\partial \gamma_i} = \sum_{t=1}^{T_i} -[y_{it} - \gamma_i \exp(\mathbf{x}'_{it}\beta)] [\gamma_i \exp(\mathbf{x}'_{it}\beta)] = 0,$$

or

$$\gamma_i \sum_{t=1}^{T_i} [y_{it} \exp(\mathbf{x}'_{it}\beta)] = \gamma_i^2 \sum_{t=1}^{T_i} [\exp(\mathbf{x}'_{it}\beta)]^2.$$

Obviously, if we can solve the equation for  $\gamma_i$ , we can obtain  $\alpha_i = \ln \gamma_i$ . The preceding equation can, indeed, be solved for  $\gamma_i$ , at least conditionally. At the minimum of the sum of squares, it will be true that

$$\hat{\gamma}_i = \frac{\sum_{t=1}^{T_i} y_{it} \exp(\mathbf{x}'_{it}\hat{\beta})}{\sum_{t=1}^{T_i} [\exp(\mathbf{x}'_{it}\hat{\beta})]^2}. \quad (11-81)$$

We can now insert (11-81) into (11-80) to eliminate  $\alpha_i$ . (This is a counterpart to taking deviations from means in the linear case. As noted, this is possible only for a very few special

## 414 PART II ♦ Generalized Regression Model and Equation Systems

models—this happens to be one of them. The process is also known as “concentrating out” the parameters  $\gamma_i$ . Note that the solution,  $\hat{\gamma}_i$ , is obtained as the slope in a regression without a constant term of  $y_{it}$  on  $\exp(\mathbf{x}'_{it}\beta)$  using  $T_i$  observations.) The result in (11-81) must hold at the solution. Thus, (11-81) inserted in (11-80) restricts the search for  $\beta$  to those values that satisfy the restrictions in (11-81). The resulting sum of squares function is now a function only of the data and  $\beta$ , and can be minimized with respect to this vector of  $K$  parameters. With the estimate of  $\beta$  in hand,  $\alpha_i$  can be estimated using the log of the result in (11-81) (which is positive by construction).

The preceding example presents a mixed picture for the fixed effects model. In nonlinear cases, two problems emerge that were not present earlier, the practical one of actually computing the dummy variable parameters and the theoretical incidental parameters problem that we have yet to investigate, but which promises to be a significant shortcoming of the fixed effects model. We also note we have focused on a particular form of the model, the “single index” function, in which the conditional mean is a nonlinear function of a linear function. In more general cases, it may be unclear how the unobserved heterogeneity should enter the regression function.

### 11.9.3 RANDOM EFFECTS

The random effects nonlinear model also presents complications both for specification and for estimation. We might begin with a general model

$$y_{it} = h(\mathbf{x}'_{it}\beta, u_i) + \varepsilon_{it}. \quad (11-82)$$

The “random effects” assumption would be, as usual, mean independence,

$$E[u_i | \mathbf{X}_i] = 0.$$

Unlike the linear model, the nonlinear regression cannot be consistently estimated by (nonlinear) least squares. In practical terms, we can see why in (7-28)–(7-30). In the linearized regression, the conditional mean at the expansion point  $\beta^0$  [see (7-28)] as well as the pseudoregressors are both functions of the unobserved  $u_i$ . This is true in the general case (11-81) as well as the simpler case of a single index model,

$$y_{it} = h(\mathbf{x}'_{it}\beta + u_i) + \varepsilon_{it}. \quad (11-83)$$

Thus, it is not possible to compute the iterations for nonlinear least squares. As in the fixed effects case, neither deviations from group means nor first differences solves the problem. Ignoring the problem—that is, simply computing the nonlinear least squares estimator without accounting for heterogeneity—does not produce a consistent estimator, for the same reasons. In general, the benign effect of latent heterogeneity (random effects) that we observe in the linear model only carries over to a very few nonlinear models and, unfortunately, this is not one of them.

The problem of computing partial effects in a random effects model such as (11-83) is that when  $E[y_{it} | \mathbf{x}_{it}, u_i]$  is given by (11-83),

$$\frac{\partial E[y_{it} | \mathbf{x}'_{it}\beta + u_i]}{\partial \mathbf{x}'_{it}} = [h'(\mathbf{x}'_{it}\beta + u_i)]\beta$$

is a function of the unobservable  $u_i$ . Two ways to proceed from here are the fixed effects approach of the previous section and a random effects approach. The fixed effects approach is feasible but may be hindered by the incidental parameters problem

CHAPTER 11 ♦ Models for Panel Data **415**

noted earlier. A random effects approach might be preferable, but comes at the price of assuming that  $\mathbf{x}_{it}$  and  $u_i$  are uncorrelated, which may be unreasonable. Papke and Wooldridge (2008) examined several cases and proposed the Mundlak approach of projecting  $u_i$  on the group means of  $\mathbf{x}_{it}$ . The working specification of the model is then

$$E^*[y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i, v_i] = h(\mathbf{x}'_{it}\boldsymbol{\beta} + \alpha + \bar{\mathbf{x}}'_i\boldsymbol{\theta} + v_i).$$

This leaves the practical problem of how to compute the estimates of the parameters and how to compute the partial effects. Papke and Wooldridge (2008) suggest a useful result if it can be assumed that  $v_i$  is normally distributed with mean zero and variance  $\sigma_v^2$ . In that case,

$$E[y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}] = E_{v_i} E[y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}, v_i] = h\left(\frac{\mathbf{x}'_{it}\boldsymbol{\beta} + \alpha + \bar{\mathbf{x}}'_i\boldsymbol{\theta}}{\sqrt{1 + \sigma_v^2}}\right) = h(\mathbf{x}'_{it}\boldsymbol{\beta}_v + \alpha_v + \bar{\mathbf{x}}'_i\boldsymbol{\theta}_v).$$

The implication is that nonlinear least squares regression will estimate the scaled coefficients, after which the average partial effect can be estimated for a particular value of the covariates,  $\mathbf{x}_0$ , with

$$\hat{\Delta}(\mathbf{x}_0) = \frac{1}{n} \sum_{i=1}^n h'(\mathbf{x}'_0 \hat{\boldsymbol{\beta}}_v + \hat{\alpha}_v + \bar{\mathbf{x}}'_i \hat{\boldsymbol{\theta}}_v) \hat{\boldsymbol{\beta}}_v.$$

They applied the technique to a case of test pass rates, which are a fraction bounded by zero and one. Loudermilk (2007) is another application with an extension to a dynamic model.

### 11.10 SYSTEMS OF EQUATIONS

Extensions of the SUR model to panel data applications have been made in two directions. Several studies have layered the familiar random effects treatment of Section 11.5 on top of the generalized regression. An alternative treatment of the fixed and random effects models as a form of seemingly unrelated regressions model suggested by Chamberlain (1982, 1984) has provided some of the foundation of recent treatments of dynamic panel data models, as in Sections 11.8.2 and 11.8.3.

Avery (1977) suggested a natural extension of the random effects model to multiple equations,

$$y_{it,j} = \mathbf{x}'_{it,j}\boldsymbol{\beta}_j + \varepsilon_{it,j} + u_{i,j},$$

where  $j$  indexes the equation,  $i$  indexes individuals, and  $t$  is the time index as before. Each equation can be treated as a random effects model. In this instance, however, the efficient estimator when the equations are *actually* unrelated (that is,  $\text{Cov}[\varepsilon_{it,m}, \varepsilon_{it,l} | \mathbf{X}] = 0$  and  $\text{Cov}[u_{i,m}, u_{i,l} | \mathbf{X}] = 0$ ) is equation by equation GLS as developed in Section 11.5, not OLS. That is, without the cross-equation correlation, each equation constitutes a random effects model. The cross-equation correlation takes the form

$$E[\varepsilon_{it,j}\varepsilon_{it,l} | \mathbf{X}] = \sigma_{jl}$$

and

$$E[u_{i,j}u_{i,l} | \mathbf{X}] = \theta_{jl}.$$

## 416 PART II ♦ Generalized Regression Model and Equation Systems

Observations remain uncorrelated across individuals,  $(\varepsilon_{it,j}, \varepsilon_{rs,l})$  and  $(u_{ij}, u_{r,l})$  when  $i \neq r$ . The “noise” terms,  $\varepsilon_{it,j}$  are also uncorrelated across time for all individuals and across individuals. Correlation over time arises through the influence of the common effect, which produces persistent random effects for the given individual, both within the equation and across equations through  $\theta_{jl}$ . Avery developed a two-step estimator for the model. At the first step, as usual, estimates of the variance components are based on OLS residuals. The second step is FGLS. Subsequent studies have added features to the model. Magnus (1982) derived the log likelihood function for normally distributed disturbances, the likelihood equations for the MLE, and a method of estimation. Verbon (1980) added heteroscedasticity to the model.

There have also been a handful of applications, including Howrey and Varian's (1984) analysis of electricity pricing and the impact of time of day rates, Brown et al.'s (1983) treatment of a form of the capital asset pricing model (CAPM), Sickles's (1985) analysis of airline costs, and Wan et al.'s (1992) development of a nonlinear panel data SUR model for agricultural output.

### **Example 11.18 Demand for Electricity and Gas**

Beierlein, Dunn, and McConnon (1981) proposed a dynamic panel data SUR model for demand for electricity and natural gas in the northeastern United States. The central equation of the model is

$$\begin{aligned} \ln Q_{it,j} = & \beta_0 + \beta_1 \ln P_{\text{natural gas}}_{it,j} + \beta_2 \ln P_{\text{electricity}}_{it,j} + \beta_3 \ln P_{\text{fuel oil}}_{it,j} \\ & + \beta_4 \ln \text{per capita income}_{it,j} + \beta_5 \ln Q_{i,t-1,j} + w_{it,j} \\ w_{it,j} = & \varepsilon_{it,j} + u_{i,j} + v_{t,j} \end{aligned}$$

where

$j$  = consuming sectors (natural gas, electricity)  $\times$  (residential, commercial, industrial)

$i$  = state (New England plus New York, New Jersey, Pennsylvania)

$t$  = year, 1957, ..., 1977.

Note that this model has both time and state random effects and a lagged dependent variable in each equation.

### 11.11 PARAMETER HETEROGENEITY

The treatment so far has essentially treated the slope parameters of the model as fixed constants, and the intercept as varying randomly from group to group. An equivalent formulation of the pooled, fixed, and random effects models is

$$y_{it} = (\alpha + u_i) + \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it},$$

where  $u_i$  is a person-specific random variable with conditional variance zero in the pooled model, positive in the others, and conditional mean dependent on  $\mathbf{X}_i$  in the fixed effects model and constant in the random effects model. By any of these, the heterogeneity in the model shows up as variation in the constant terms in the regression model. There is ample evidence in many studies—we will examine two later—that suggests that the other parameters in the model also vary across individuals. In the

## CHAPTER 11 ♦ Models for Panel Data 417

dynamic model we consider in Section 11.11.3, cross-country variation in the slope parameter in a production function is the central focus of the analysis. This section will consider several approaches to analyzing parameter heterogeneity in panel data models.

### 11.11.1 THE RANDOM COEFFICIENTS MODEL

Parameter heterogeneity across individuals or groups can be modeled as stochastic variation.<sup>27</sup> Suppose that we write

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \\ E[\boldsymbol{\varepsilon}_i | \mathbf{X}_i] &= \mathbf{0}, \\ E[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}'_i | \mathbf{X}_i] &= \sigma^2 \mathbf{I}_T, \end{aligned} \tag{11-84}$$

where

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{u}_i \tag{11-85}$$

and

$$\begin{aligned} E[\mathbf{u}_i | \mathbf{X}_i] &= \mathbf{0}, \\ E[\mathbf{u}_i \mathbf{u}'_i | \mathbf{X}_i] &= \boldsymbol{\Gamma}. \end{aligned} \tag{11-86}$$

(Note that if only the constant term in  $\boldsymbol{\beta}$  is random in this fashion and the other parameters are fixed as before, then this reproduces the random effects model we studied in Section 11.5.) Assume for now that there is no autocorrelation or cross-section correlation in  $\boldsymbol{\varepsilon}_i$ . We also assume for now that  $T > K$ , so that when desired, it is possible to compute the linear regression of  $\mathbf{y}_i$  on  $\mathbf{X}_i$  for each group. Thus, the  $\boldsymbol{\beta}_i$  that applies to a particular cross-sectional unit is the outcome of a random process with mean vector  $\boldsymbol{\beta}$  and covariance matrix  $\boldsymbol{\Gamma}$ .<sup>28</sup> By inserting (11-85) into (11-84) and expanding the result, we obtain a generalized regression model for each block of observations:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + (\boldsymbol{\varepsilon}_i + \mathbf{X}_i \mathbf{u}_i),$$

so

$$\boldsymbol{\Omega}_{ii} = E[(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})' | \mathbf{X}_i] = \sigma^2 \mathbf{I}_T + \mathbf{X}_i \boldsymbol{\Gamma} \mathbf{X}'_i.$$

For the system as a whole, the disturbance covariance matrix is block diagonal, with  $T \times T$  diagonal block  $\boldsymbol{\Omega}_{ii}$ . We can write the GLS estimator as a matrix weighted average of the group specific OLS estimators:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{y} = \sum_{i=1}^n \mathbf{W}_i \mathbf{b}_i, \tag{11-87}$$

<sup>27</sup>The most widely cited studies are Hildreth and Houck (1968), Swamy (1970, 1971, 1974), Hsiao (1975), and Chow (1984). See also Breusch and Pagan (1979). Some recent discussions are Swamy and Tavlas (1995, 2001) and Hsiao (2003). The model bears some resemblance to the Bayesian approach of Chapter 10, but, the similarity is only superficial. We are maintaining the classical approach to estimation throughout.

<sup>28</sup>Swamy and Tavlas (2001) label this the “first-generation random coefficients model” (RCM). We will examine the “second generation” (the current generation) of random coefficients models in the next section.

## 418 PART II ♦ Generalized Regression Model and Equation Systems

where

$$\mathbf{W}_i = \left[ \sum_{i=1}^n \left( \boldsymbol{\Gamma} + \sigma_e^2 (\mathbf{X}'_i \mathbf{X}_i)^{-1} \right)^{-1} \right]^{-1} (\boldsymbol{\Gamma} + \sigma_e^2 (\mathbf{X}'_i \mathbf{X}_i)^{-1})^{-1}.$$

Empirical implementation of this model requires an estimator of  $\boldsymbol{\Gamma}$ . One approach [see, e.g., Swamy (1971)] is to use the empirical variance of the set of  $n$  least squares estimates,  $\mathbf{b}_i$  minus the average value of  $s_i^2 (\mathbf{X}'_i \mathbf{X}_i)^{-1}$ :

$$\mathbf{G} = [1/(n-1)] [\Sigma_i \mathbf{b}_i \mathbf{b}'_i - n \bar{\mathbf{b}} \bar{\mathbf{b}}'] - (1/N) \Sigma_i \mathbf{V}_i, \quad (11-88)$$

where



$$\bar{\mathbf{b}} = (1/n) \Sigma_i \mathbf{b}_i$$

and

$$\mathbf{V}_i = s_i^2 (\mathbf{X}'_i \mathbf{X}_i)^{-1}.$$

This matrix may not be positive definite, however, in which case [as Baltagi (2005) suggests], one might drop the second term.

A chi-squared test of the random coefficients model against the alternative of the classical regression (no randomness of the coefficients) can be based on

$$C = \Sigma_i (\mathbf{b}_i - \mathbf{b}_*)' \mathbf{V}_i^{-1} (\mathbf{b}_i - \mathbf{b}_*),$$

where

$$\mathbf{b}_* = \left[ \Sigma_i \mathbf{V}_i^{-1} \right]^{-1} \Sigma_i \mathbf{V}_i^{-1} \mathbf{b}_i.$$

Under the null hypothesis of homogeneity,  $C$  has a limiting chi-squared distribution with  $(n-1)K$  degrees of freedom. The best linear unbiased individual predictors of the group-specific coefficient vectors are matrix weighted averages of the GLS estimator,  $\hat{\beta}$ , and the group specific OLS estimates,  $\mathbf{b}_i$ ,<sup>29</sup>

$$\hat{\beta}_i = \mathbf{Q}_i \hat{\beta} + [\mathbf{I} - \mathbf{Q}_i] \mathbf{b}_i, \quad (11-89)$$

where

$$\mathbf{Q}_i = [(1/s_i^2) \mathbf{X}'_i \mathbf{X}_i + \mathbf{G}^{-1}]^{-1} \mathbf{G}^{-1}.$$

### Example 11.19 Random Coefficients Model

In Example 10.1, we examined Munnell's production model for gross state product,

$$\begin{aligned} \ln gsp_{it} &= \beta_1 + \beta_2 \ln pc_{it} + \beta_3 \ln hwy_{it} + \beta_4 \ln water_{it} \\ &\quad + \beta_5 \ln util_{it} + \beta_6 \ln emp_{it} + \beta_7 \ln unemp_{it} + \varepsilon_{it}, \quad i = 1, \dots, 48; t = 1, \dots, 17. \end{aligned}$$

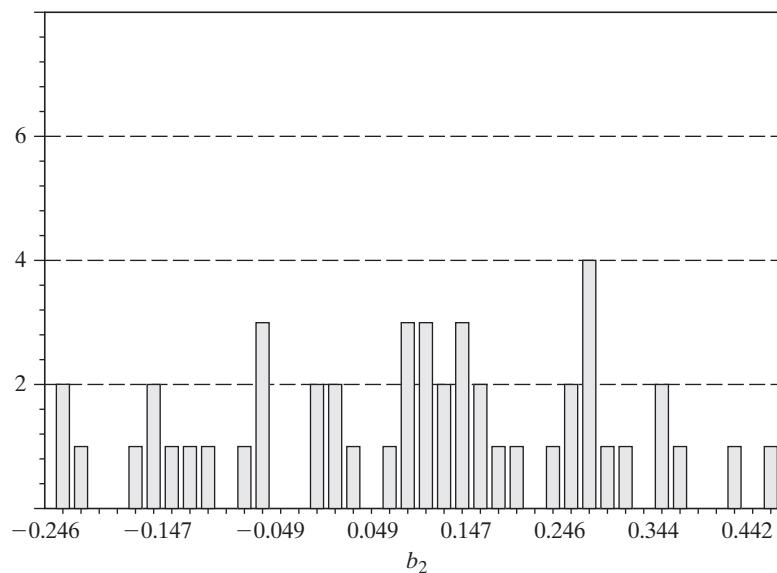
The panel consists of state level data for 17 years. The model in Example 10.1 (and Munnell's) provide no means for parameter heterogeneity save for the constant term. We have reestimated the model using the Hildreth and Houck approach. The OLS and Feasible GLS results given in Table 11.14. The chi-squared statistic for testing the null hypothesis of parameter homogeneity is 25,556.26, with  $7(47) = 329$  degrees of freedom. The critical value from the table is 372.299, so the hypothesis would be rejected.

Unlike the other cases we have examined in this chapter, the FGLS estimates are very different from OLS in these estimates, in spite of the fact that both estimators are consistent and the sample is fairly large. The underlying standard deviations are computed using  $\mathbf{G}$  as

<sup>29</sup>See Hsiao (2003, pp. 144–149).

CHAPTER 11 ♦ Models for Panel Data **419****TABLE 11.14** Estimated Random Coefficients Models

Variable	Least Squares		Feasible GLS		
	Estimate	Standard Error	Estimate	Standard Error	Popn. Std. Deviation
Constant	1.9260	0.05250	1.6533	1.08331	7.0782
$\ln p_V$	0.3120	0.01109	0.09409	0.05152	0.3036
$\ln hwy$	0.05888	0.01541	0.1050	0.1736	1.1112
$\ln water$	0.1186	0.01236	0.07672	0.06743	0.4340
$\ln util$	0.00856	0.01235	-0.01489	0.09886	0.6322
$\ln emp$	0.5497	0.01554	0.9190	0.1044	0.6595
$unemp$	-0.00727	0.001384	-0.004706	0.002067	0.01266
$\sigma_e$	0.08542		0.2129		
$\ln L$	853.1372				

**FIGURE 11.1** Estimates of Coefficient on Private Capital.

the covariance matrix. [For these data, subtracting the second matrix rendered  $\mathbf{G}$  not positive definite, so in the table, the standard deviations are based on the estimates using only the first term in (11-88).] The increase in the standard errors is striking. This suggests that there is considerable variation in the parameters across states. We have used (11-89) to compute the estimates of the state specific coefficients. Figure 11.1 shows a histogram for the coefficient on private capital. As suggested, there is a wide variation in the estimates.

## 420 PART II ♦ Generalized Regression Model and Equation Systems

### 11.11.2 A HIERARCHICAL LINEAR MODEL

Many researchers have employed a two-step approach to estimate two-level models. In a common form of the application, a panel data set is employed to estimate the model,

$$\mathbf{y}_{it} = \mathbf{x}'_{it}\boldsymbol{\beta}_i + \varepsilon_{it}, i = 1, \dots, n, t = 1, \dots, T,$$

$$\boldsymbol{\beta}_{t,k} = \mathbf{z}'_i\boldsymbol{\alpha}_k + u_{i,k}, i = 1, \dots, n.$$

Assuming the panel is long enough, the first equation is estimated  $n$  times, once for each individual  $i$ , and then the estimated coefficient on  $x_{itk}$  in each regression forms an observation for the second-step regression.<sup>30</sup> (This is the approach we took in the previous section; each  $a_i$  is computed by a linear regression of  $\mathbf{X}_i\mathbf{b}_{LSDV}$  on a column of ones.)

#### *Example 11.20 Fannie Mae's Pass Through*

Fannie Mae is the popular name for the Federal National Mortgage Corporation. Fannie Mae is the secondary provider for mortgage money for nearly all the small- and moderate-sized home mortgages in the United States. Loans in the study described here are termed "small" if they are for less than \$100,000. A loan is termed a "conforming" in the language of the literature on this market if (as of 2004), it was for no more than \$333,700. A larger than conforming loan is called a "jumbo" mortgage. Fannie Mae provides the capital for nearly all conforming loans and no nonconforming loans. The question pursued in the study described here was whether the clearly observable spread between the rates on jumbo loans and conforming loans reflects the cost of raising the capital in the market. Fannie Mae is a "government sponsored enterprise" (GSE). It was created by the U.S. Congress, but it is not an arm of the government; it is a private corporation. In spite of, or perhaps because of this ambiguous relationship to the government, apparently, capital markets believe that there is some benefit to Fannie Mae in raising capital. Purchasers of the GSE's debt securities seem to believe that the debt is implicitly backed by the government—this in spite of the fact that Fannie Mae explicitly states otherwise in its publications. This emerges as a "funding advantage" (GFA) estimated by the authors of the study of about 17 basis points (hundredths of one percent). In a study of the residential mortgage market, Passmore (2005) and Passmore, Sherlund, and Burgess (2005) sought to determine whether this implicit subsidy to the GSE was passed on to the mortgagees or was, instead, passed on to the stockholders. Their approach utilized a very large data set and a two-level, two-step estimation procedure. The first step equation estimated was a mortgage rate equation using a sample of roughly 1 million closed mortgages. All were conventional 30-year fixed-rate loans closed between April 1997 and May 2003. The dependent variable of interest is the rate on the mortgage,  $RM_{it}$ . The first level equation is

$$RM_{it} = \beta_{1i} + \beta_{2,i}J_{it} + \text{terms for "loan to value ratio," "new home dummy variable," "small mortgage"} \\ + \text{terms for "fees charged" and whether the mortgage was originated by a mortgage company} + \varepsilon_{it}.$$

The main variable of interest in this model is  $J_{it}$ , which is a dummy variable for whether the loan is a jumbo mortgage. The " $i$ " in this setting is a (state, time) pair for California, New Jersey, Maryland, Virginia, and all other states, and months from April 1997 to May 2003. There were 370 groups in total. The regression model was estimated for each group. At the second step, the coefficient of interest is  $\beta_{2,i}$ . On overall average, the spread between jumbo

<sup>30</sup>An extension of the model in which " $ui$ " is heteroscedastic is developed at length in Saxonhouse (1976) and revisited by Achen (2005).

CHAPTER 11 ♦ Models for Panel Data **421**

and conforming loans at the time was roughly 16 basis points. The second-level equation is

$$\begin{aligned}\beta_{2,i} = & \alpha_1 + \alpha_2 \text{GFA}_i \\ & + \alpha_3 \text{one-year treasury rate} \\ & + \alpha_4 \text{10-year treasury rate} \\ & + \alpha_5 \text{credit risk} \\ & + \alpha_6 \text{prepayment risk} \\ & + \text{measures of maturity mismatch risk} \\ & + \text{quarter and state fixed effects} \\ & + \text{mortgage market capacity} \\ & + \text{mortgage market development} \\ & + u_i.\end{aligned}$$

The result ultimately of interest is the coefficient on GFA,  $\alpha_2$ , which is interpreted as the fraction of the GSE funding advantage that is passed through to the mortgage holders. Four different estimates of  $\alpha_2$  were obtained, based on four different measures of corporate debt liquidity; the estimated values were  $(\hat{\alpha}_2^1, \hat{\alpha}_2^2, \hat{\alpha}_2^3, \hat{\alpha}_2^4) = (0.07, 0.31, 0.17, 0.10)$ . The four estimates were averaged using a **minimum distance estimator** (MDE). Let  $\hat{\Omega}$  denote the estimated  $4 \times 4$  asymptotic covariance matrix for the estimators. Denote the distance vector

$$\mathbf{d} = (\hat{\alpha}_2^1 - \alpha_2, \hat{\alpha}_2^2 - \alpha_2, \hat{\alpha}_2^3 - \alpha_2, \hat{\alpha}_2^4 - \alpha_2)'$$

The minimum distance estimator is the value for  $\alpha_2$  that minimizes  $\mathbf{d}' \hat{\Omega}^{-1} \mathbf{d}$ . For this study,  $\hat{\Omega}$  is a diagonal matrix. It is straightforward to show that in this case, the MDE is

$$\hat{\alpha}_2 = \sum_{j=1}^4 \hat{\alpha}_2^j \left( \frac{1/\hat{\omega}_j}{\sum_{m=1}^4 1/\hat{\omega}_m} \right).$$

 The final answer is roughly 16 percent. By implication, then, the authors estimated that about 16 percent of the GSE funding advantage was kept within the company or passed through to stockholders.

### 11.11.3 PARAMETER HETEROGENEITY AND DYNAMIC PANEL DATA MODELS

The analysis in this section has involved static models and relatively straightforward estimation problems. We have seen as this section has progressed that parameter heterogeneity introduces a fair degree of complexity to the treatment. Dynamic effects in the model, with or without heterogeneity, also raise complex new issues in estimation and inference. There are numerous cases in which dynamic effects and parameter heterogeneity coincide in panel data models. This section will explore a few of the specifications and some applications. The familiar estimation techniques (OLS, FGTS, etc.) are not effective in these cases. The proposed solutions are developed in Chapter 8 where we present the technique of instrumental variables and in Chapter 13 where we present the GMM estimator and its application to **dynamic panel data models**.

#### **Example 11.21 Dynamic Panel Data Models**

The antecedent of much of the current research on panel data is Balestra and Nerlove's (1966) study of the natural gas market. [See, also, Nerlove (2002, Chapter 2).] The model is a

## 422 PART II ♦ Generalized Regression Model and Equation Systems

stock-flow description of the derived demand for fuel for gas using appliances. The central equation is a model for total demand,

$$G_{it} = G_{it}^* + (1 - r)G_{i,t-1},$$

where  $G_{it}$  is current total demand. Current demand consists of new demand,  $G_{it}^*$ , that is created by additions to the stock of appliances plus old demand which is a proportion of the previous period's demand,  $r$  being the depreciation rate for gas using appliances. New demand is due to net increases in the stock of gas using appliances, which is modeled as

$$G_{it}^* = \beta_0 + \beta_1 Price_{it} + \beta_2 \Delta Pop_{it} + \beta_3 Pop_{it} + \beta_4 \Delta Income_{it} + \beta_5 Income_{it} + \varepsilon_{it},$$

where  $\Delta$  is the first difference (change) operator,  $\Delta X_t = X_t - X_{t-1}$ . The reduced form of the model is a dynamic equation,

$$G_{it} = \beta_0 + \beta_1 Price_{it} + \beta_2 \Delta Pop_{it} + \beta_3 Pop_{it} + \beta_4 \Delta Income_{it} + \beta_5 Income_{it} + \gamma G_{i,t-1} + \varepsilon_{it}.$$

The authors analyzed a panel of 36 states over a six-year period (1957–1962). Both fixed effects and random effects approaches were considered.

An equilibrium model for steady state growth has been used by numerous authors [e.g., Robertson and Symons (1992), Pesaran and Smith (1995), Lee, Pesaran, and Smith (1997), Pesaran, Shin, and Smith (1999), Nerlove (2002) and Hsiao, Pesaran, and Tahmisioglu (2002)] for cross industry or country comparisons. Robertson and Symons modeled real wages in 13 OECD countries over the period 1958 to 1986 with a wage equation

$$W_{it} = \alpha_i + \beta_{1i} k_{it} + \beta_{2i} \Delta wedge_{it} + \gamma_i W_{i,t-1} + \varepsilon_{it},$$

where  $W_{it}$  is the real product wage for country  $i$  in year  $t$ ,  $k_{it}$  is the capital-labor ratio, and  $wedge$  is the “tax and import price wedge.”

Lee, Pesaran, and Smith (1997) compared income growth across countries with a steady-state income growth model of the form

$$\ln y_{it} = \alpha_i + \theta_i t + \lambda_i \ln y_{i,t-1} + \varepsilon_{it},$$

where  $\theta_i = (1 - \lambda_i)\delta_i$ ,  $\delta_i$  is the technological growth rate for country  $i$  and  $\lambda_i$  is the convergence parameter. The rate of convergence to a steady state is  $1 - \lambda_i$ .

Pesaran and Smith (1995) analyzed employment in a panel of 38 UK industries observed over 29 years, 1956–1984. The main estimating equation was

$$\begin{aligned} \ln e_{it} = & \alpha_i + \beta_{1i} t + \beta_{2i} \ln y_{it} + \beta_{3i} \ln y_{i,t-1} + \beta_{4i} \ln \bar{y}_t + \beta_{5i} \ln \bar{y}_{t-1} \\ & + \beta_{6i} \ln w_{it} + \beta_{7i} \ln w_{i,t-1} + \gamma_{1i} \ln e_{i,t-1} + \gamma_{2i} \ln e_{i,t-2} + \varepsilon_{it}, \end{aligned}$$

where  $y_{it}$  is industry output,  $\bar{y}_t$  is total (not average) output, and  $w_{it}$  is real wages.

In the growth models, a quantity of interest is the **long-run multiplier** or **long-run elasticity**. Long-run effects are derived through the following conceptual experiment. The essential feature of the models above is a dynamic equation of the form

$$y_t = \alpha + \beta x_t + \gamma y_{t-1}.$$

Suppose at time  $t$ ,  $x_t$  is fixed from that point forward at  $\bar{x}$ . The value of  $y_t$  at that time will then be  $\alpha + \beta \bar{x} + \gamma y_{t-1}$ , given the previous value. If this process continues, and if  $|\gamma| < 1$ , then eventually  $y_s$  will reach an equilibrium at a value such that  $y_s = y_{s-1} = \bar{y}$ . If so, then  $\bar{y} = \alpha + \beta \bar{x} + \gamma \bar{y}$ , from which we can deduce that  $\bar{y} = (\alpha + \bar{x})/(1 - \gamma)$ . The path to this equilibrium from time  $t$  into the future is governed by the **adjustment equation**

$$y_s - \bar{y} = (y_t - \bar{y})\gamma^{s-t}, s \geq t.$$

CHAPTER 11 ♦ Models for Panel Data **423**

The experiment, then, is to ask: What is the impact on the equilibrium of a change in the input,  $\bar{x}$ ? The result is  $\partial \bar{y} / \partial \bar{x} = \beta / (1 - \gamma)$ . This is the long-run multiplier, or **equilibrium multiplier** in the model. In the preceding Pesaran and Smith model, the inputs are in logarithms, so the multipliers are long-run elasticities. For example, with two lags of  $\ln e_{it}$  in Pesaran and Smith's model, the long-run effects for wages are

$$\phi_i = (\beta_{6i} + \beta_{7i}) / (1 - \gamma_{1i} - \gamma_{2i}).$$

In this setting, in contrast to the preceding treatments, the number of units,  $n$ , is generally taken to be fixed, though often it will be fairly large. The Penn World Tables ([http://pwt.econ.upenn.edu/php\\_site/pwt\\_index.php](http://pwt.econ.upenn.edu/php_site/pwt_index.php)) that provide the database for many of these analyses now contain information on almost 200 countries for well over 50 years. Asymptotic results for the estimators are with respect to increasing  $T$ , though we will consider in general, cases in which  $T$  is small. Surprisingly, increasing  $T$  and  $n$  at the same time need not simplify the derivations. ~~We will revisit this issue in the next section.~~ 

The parameter of interest in many studies is the average long-run effect, say  $\bar{\phi} = (1/n) \sum_i \phi_i$ , in the Pesaran and Smith example. Because  $n$  is taken to be fixed, the “parameter”  $\bar{\phi}$  is a definable object of estimation—that is, with  $n$  fixed, we can speak of  $\bar{\phi}$  as a parameter rather than as an estimator of a parameter. There are numerous approaches one might take. For estimation purposes, pooling, fixed effects, random effects, group means, or separate regressions are all possibilities. (Unfortunately, nearly all are inconsistent.) In addition, there is a choice to be made whether to compute the average of long-run effects or compute the long-run effect from averages of the parameters. The choice of the average of functions,  $\bar{\phi}$  versus the function of averages,

$$\bar{\phi}^* = \frac{\frac{1}{n} \sum_{i=1}^n (\hat{\beta}_{6i} + \hat{\beta}_{7i})}{1 - \frac{1}{n} \sum_{i=1}^n (\hat{\gamma}_{1i} + \hat{\gamma}_{2i})}$$

turns out to be of substance. For their UK industry study, Pesaran and Smith report estimates of  $-0.33$  for  $\bar{\phi}$  and  $-0.45$  for  $\bar{\phi}^*$ . (The authors do not express a preference for one over the other.)

The development to this point is implicitly based on estimation of separate models for each unit (country, industry, etc.). There are also a variety of other estimation strategies one might consider. We will assume for the moment that the data series are stationary in the dimension of  $T$ . (See Chapter 23.) This is a transparently false assumption, as revealed by a simple look at the trends in macroeconomic data, but maintaining it for the moment allows us to proceed. We will reconsider it later.

We consider the generic, dynamic panel data model,

$$y_{it} = \alpha_i + \beta_i x_{it} + \gamma_i y_{i,t-1} + \varepsilon_{it}. \quad (11-90)$$

Assume that  $T$  is large enough that the individual regressions can be computed. In the absence of autocorrelation in  $\varepsilon_{it}$ , it has been shown [e.g., Griliches (1961), Maddala and Rao (1973)] that the OLS estimator of  $\gamma_i$  is biased downward, but consistent in  $T$ . Thus,  $E[\hat{\gamma}_i - \gamma_i] = \theta_i / T$  for some  $\theta_i$ . The implication for the individual estimator of the long-run multiplier,  $\phi_i = \beta_i / (1 - \gamma_i)$ , is unclear in this case, however. The denominator is overestimated. But it is not clear whether the estimator of  $\beta_i$  is overestimated or underestimated. It is true that whatever bias there is  $O(1/T)$ . For this application,  $T$  is fixed

## 424 PART II ♦ Generalized Regression Model and Equation Systems

and possibly quite small. The end result is that it is unlikely that the individual estimator of  $\phi_i$  is unbiased, and by construction, it is inconsistent, because  $T$  cannot be assumed to be increasing. If that is the case, then  $\hat{\phi}$  is likewise inconsistent for  $\bar{\phi}$ . We are averaging  $n$  estimators, each of which has bias and variance that are  $O(1/T)$ . The variance of the mean is, therefore,  $O(1/nT)$  which goes to zero, but the bias remains  $O(1/T)$ . It follows that the average of the  $n$  means is not converging to  $\bar{\phi}$ ; it is converging to the average of whatever these biased estimators are estimating. The problem vanishes with large  $T$ , but that is not relevant to the current context. However, in the Pesaran and Smith study,  $T$  was 29, which is large enough that these effects are probably moderate. For macroeconomic cross-country studies such as those based on the Penn World Tables, the data series might be yet longer than this.

One might consider aggregating the data to improve the results. Smith and Pesaran (1995) suggest an average based on country means. Averaging the observations over  $T$  in (11-90) produces

$$\bar{y}_{i\cdot} = \alpha_i + \beta_i \bar{x}_{i\cdot} + \gamma_i \bar{y}_{-1,i} + \bar{\varepsilon}_{i\cdot} \quad (11-91)$$

A linear regression using the  $n$  observations would be inconsistent for two reasons: First,  $\bar{\varepsilon}_{i\cdot}$  and  $\bar{y}_{-1,i}$  must be correlated. Second, because of the parameter heterogeneity, it is not clear without further assumptions what the OLS slopes estimate under the false assumption that all coefficients are equal. But  $\bar{y}_{i\cdot}$  and  $\bar{y}_{-1,i}$  differ by only the first and last observations;  $\bar{y}_{-1,i} = \bar{y}_{i\cdot} - (y_{iT} - y_{i0})/T = \bar{y}_{i\cdot} - [\Delta_T(y)/T]$ . Inserting this in (11-89) produces

$$\begin{aligned} \bar{y}_{i\cdot} &= \alpha_i + \beta_i \bar{x}_{i\cdot} + \gamma_i \bar{y}_{i\cdot} - \gamma_i [\Delta_T(y)/T] + \bar{\varepsilon}_{i\cdot} \\ &= \frac{\alpha_i}{1 - \gamma_i} + \frac{\beta_i}{1 - \gamma_i} \bar{x}_{i\cdot} - \frac{\gamma_i}{1 - \gamma_i} [\Delta_T(y)/T] + \bar{\varepsilon}_{i\cdot} \\ &= \delta_i + \phi_i \bar{x}_{i\cdot} + \tau_i [\Delta_T(y)/T] + \bar{\varepsilon}_{i\cdot}. \end{aligned} \quad (11-92)$$

We still seek to estimate  $\bar{\phi}$ . The form in (11-92) does not solve the estimation problem, since the regression suggested using the group means is still heterogeneous. If it could be assumed that the individual long-run coefficients differ randomly from the averages in the fashion of the random parameters model of the previous section, so  $\delta_i = \bar{\delta} + u_{\delta,i}$  and likewise for the other parameters, then the model could be written

$$\begin{aligned} \bar{y}_{i\cdot} &= \bar{\delta} + \bar{\phi} \bar{x}_{i\cdot} + \bar{\tau} [\Delta_T(y)/T]_i + \bar{\varepsilon}_{i\cdot} + \{u_{\delta,i} + u_{\phi,i} \bar{x}_{i\cdot} + u_{\tau,i} [\Delta_T(y)/T]_i\} \\ &= \bar{\delta} + \bar{\phi} \bar{x}_{i\cdot} + \bar{\tau} [\Delta_T(y)/T]_i + \bar{\varepsilon}_{i\cdot} + w_i. \end{aligned}$$

At this point, the equation appears to be a heteroscedastic regression amenable to least squares estimation, but for one loose end. Consistency follows if the terms  $[\Delta_T(y)/T]_i$  and  $\bar{\varepsilon}_{i\cdot}$  are uncorrelated. Because the first is a rate of change and the second is in levels, this should generally be the case. Another interpretation that serves the same purpose is that the rates of change in  $[\Delta_T(y)/T]_i$  should be uncorrelated with the levels in  $\bar{x}_{i\cdot}$ , in which case, the regression can be partitioned, and simple linear regression of the country means of  $y_{it}$  on the country means of  $x_{it}$  and a constant produces consistent estimates of  $\bar{\phi}$  and  $\bar{\delta}$ .

## CHAPTER 11 ♦ Models for Panel Data 425

Alternatively, consider a time-series approach. We average the observation in (11-90) across countries at each time period rather than across time within countries. In this case, we have

$$\bar{y}_{it} = \bar{\alpha} + \frac{1}{n} \sum_{i=1}^n \beta_i x_{it} + \frac{1}{n} \sum_{i=1}^n \gamma_i y_{i,t-1} + \frac{1}{n} \sum_{i=1}^n \varepsilon_{it}.$$

Let  $\bar{\gamma} = \frac{1}{n} \sum_{i=1}^n \gamma_i$  so that  $\gamma_i = \bar{\gamma} + (\gamma_i - \bar{\gamma})$  and  $\beta_i = \bar{\beta} + (\beta_i - \bar{\beta})$ . Then,

$$\begin{aligned}\bar{y}_{it} &= \bar{\alpha} + \bar{\beta} \bar{x}_{it} + \bar{\gamma} \bar{y}_{i,t-1} + [\bar{\varepsilon}_{it} + (\beta_i - \bar{\beta}) \bar{x}_{it} + (\gamma_i - \bar{\gamma}) \bar{y}_{i,t-1}] \\ &= \bar{\alpha} + \bar{\beta} \bar{x}_{it} + \bar{\gamma} \bar{y}_{i,t-1} + \bar{\varepsilon}_{it} + w_{it}.\end{aligned}$$

Unfortunately, the regressor,  $\bar{\gamma} \bar{y}_{i,t-1}$  is surely correlated with  $w_{it}$ , so neither OLS or GLS will provide a consistent estimator for this model. (One might consider an instrumental variable estimator, however, there is no natural instrument available in the model as constructed.) Another possibility is to pool the entire data set, possibly with random or fixed effects for the constant terms. Because pooling, even with country-specific constant terms, imposes homogeneity on the other parameters, the same problems we have just observed persist.

Finally, returning to (11-90), one might treat it as a formal random parameters model,

$$\begin{aligned}y_{it} &= \alpha_i + \beta_i x_{it} + \gamma_i y_{i,t-1} + \varepsilon_{it}, \\ \alpha_i &= \alpha + u_{\alpha,i}, \\ \beta_i &= \beta + u_{\beta,i}, \\ \gamma_i &= \gamma + u_{\gamma,i}.\end{aligned}$$

(11-90)

Equation no.  
here?

The assumptions needed to formulate the model in this fashion are those of the previous section. As Pesaran and Smith (1995) observe, this model can be estimated using the “Swamy (1971)” estimator, which is the matrix weighted average of the least squares estimators discussed in Section 11.11.1. The estimator requires that  $T$  be large enough to fit each country regression by least squares. That has been the case for the received applications. Indeed, for the applications we have examined, both  $n$  and  $T$  are relatively large. If not, then one could still use the mixed models approach developed in Chapter 17, a compromise that appears to work well for panels with moderate sized  $n$  and  $T$ . The “mixed-fixed” model suggested in Hsiao (1986, 2003) and Weinhold (1999). The dynamic model in (11-90) is formulated as a partial fixed effects model,

$$\begin{aligned}y_{it} &= \alpha_i d_{it} + \beta_i x_{it} + \gamma_i d_{it} y_{i,t-1} + \varepsilon_{it}, \\ \beta_i &= \beta + u_{\beta,i},\end{aligned}$$

where  $d_{it}$  is a dummy variable that equals one for country  $i$  in every period and zero otherwise (i.e., the usual fixed effects approach). Note that  $d_{it}$  also appears with  $y_{i,t-1}$ . As stated, the model has “fixed effects,” one random coefficient, and a total of  $2n+1$  coefficients to estimate, in addition to the two variance components,  $\sigma_\varepsilon^2$  and  $\sigma_u^2$ . The model could be estimated inefficiently by using ordinary least squares—the random coefficient induces heteroscedasticity (see Section 11.11.1)—by using the Hildreth–Hoover–Swamy approach, or with the mixed linear model approach developed in Chapter 17.

## 426 PART II ♦ Generalized Regression Model and Equation Systems

### **Example 11.22 A Mixed Fixed Growth Model for Developing Countries**

Weinhold (1996) and Nair-Reichert and Weinhold (2001) analyzed growth and development in a panel of 24 developing countries observed for 25 years, 1971–1995. The model they employed was a variant of the mixed-fixed model proposed by Hsiao (1986, 2003). In their specification,

$$\begin{aligned} GGDP_{i,t} = & \alpha_i d_t + \gamma_i d_t GGDP_{i,t-1} \\ & + \beta_{1i} GGDI_{i,t-1} + \beta_{2i} GFDI_{i,t-1} + \beta_{3i} GEXP_{i,t-1} + \beta_{4i} INF_{i,t-1} + \varepsilon_{it}, \end{aligned}$$

where

*GGDP* = Growth rate of gross domestic product,

*GGDI* = Growth rate of gross domestic investment,

*GFDI* = Growth rate of foreign direct investment (inflows),

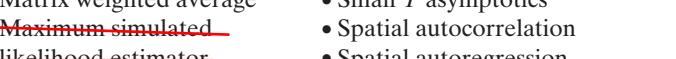
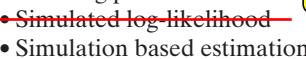
*GEXP* = Growth rate of exports of goods and services,

*INF* = Inflation rate.

## 11.12 SUMMARY AND CONCLUSIONS

This chapter has shown a few of the extensions of the classical model that can be obtained when panel data are available. In principle, any of the models we have examined before this chapter and all those we will consider later, including the multiple equation models, can be extended in the same way. The main advantage, as we noted at the outset, is that with panel data, one can formally model dynamic effects and the heterogeneity across groups that are typical in microeconomic data.

### Key Terms and Concepts

- Adjustment equation
- Autocorrelation
- Arellano and Bond's 
- Balanced panel
- Between groups
- Cluster estimator
- Contiguity
- Contiguity matrix
- Contrasts
- Dynamic panel data model
- Equilibrium multiplier
- Error components model
- Estimator
- Feasible GLS
- First difference
- Fixed effects
- Fixed effects vector decomposition
- Fixed panel
- Group means
- Group means estimator
- Hausman specification test
- Heterogeneity
- Hierarchical linear model
- Hierarchical model
- Hausman and Taylor's 
- Incidental parameters problem
- Index function model
- Individual effect
- Instrumental variable
- Instrumental variable estimator
- Lagrange multiplier test
- Least squares dummy variable estimator
- Long run elasticity
- Long run multiplier 
- Longitudinal data 
- Matrix weighted average 
- Maximum simulated likelihood estimator 
- Mean independence
- Measurement error
- Minimum distance estimator
- Mixed model
- Mundlak's approach
- Nested random effects
- Panel data
- Parameter heterogeneity
- Partial effects
- Pooled model
- Pooled regression
- Population averaged model
- Projections
- Random coefficients model
- Random effects
- Random parameters
- Robust covariance matrix
- Rotating panel 
- Simulated log likelihood 
- Simulation based estimation
- Small *T* asymptotics
- Spatial autocorrelation
- Spatial autoregression coefficient
- Spatial error correlation
- Spatial lags
- Specification test

CHAPTER 11 ♦ Models for Panel Data **427**

- Strict exogeneity
- Time-invariant

- Two-step estimation
- Unbalanced panel

- Variable addition test
- Within groups

**Exercises**

1. The following is a panel of data on investment ( $y$ ) and profit ( $x$ ) for  $n = 3$  firms over  $T = 10$  periods.

t	<i>i</i> = 1		<i>i</i> = 2		<i>i</i> = 3	
	y	x	y	x	y	x
1	13.32	12.85	20.30	22.93	8.85	8.65
2	26.30	25.69	17.47	17.96	19.60	16.55
3	2.62	5.48	9.31	9.16	3.87	1.47
4	14.94	13.79	18.01	18.73	24.19	24.91
5	15.80	15.41	7.63	11.31	3.99	5.01
6	12.20	12.59	19.84	21.15	5.73	8.34
7	14.93	16.64	13.76	16.13	26.68	22.70
8	29.82	26.45	10.00	11.61	11.49	8.36
9	20.32	19.64	19.51	19.55	18.49	15.44
10	4.77	5.43	18.32	17.06	20.84	17.87

- Pool the data and compute the least squares regression coefficients of the model  $y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}$ .
- Estimate the fixed effects model of (11-13), and then test the hypothesis that the constant term is the same for all three firms.
- Estimate the random effects model of (11-2) and then carry out the Lagrange multiplier test of the hypothesis that the classical model without the common effect applies.
- Carry out Hausman's specification test for the random versus the fixed effect model.
- Suppose that the fixed effects model is formulated with an overall constant term and  $n - 1$  dummy variables (dropping, say, the last one). Investigate the effect that this supposition has on the set of dummy variable coefficients and on the least squares estimates of the slopes.
- Unbalanced design for random effects.* Suppose that the random effects model of Section 9.7 is to be estimated with a panel in which the groups have different numbers of observations. Let  $T_i$  be the number of observations in group  $i$ .
  - Show that the pooled least squares estimator is unbiased and consistent despite this complication.
  - Show that the estimator in (11-40) based on the pooled least squares estimator of  $\beta$  (or, for that matter, *any* consistent estimator of  $\beta$ ) is a consistent estimator of  $\sigma_\varepsilon^2$ .
- What are the probability limits of  $(1/n)\text{LM}$ , where LM is defined in (11-42) under the null hypothesis that  $\sigma_u^2 = 0$  and under the alternative that  $\sigma_u^2 \neq 0$ ?
- A two-way fixed effects model.* Suppose that the fixed effects model is modified to include a time-specific dummy variable as well as an individual-specific variable. Then  $y_{it} = \alpha_i + \gamma_t + \mathbf{x}'_{it}\beta + \varepsilon_{it}$ . At every observation, the individual- and

## 428 PART II ♦ Generalized Regression Model and Equation Systems

time-specific dummy variables sum to 1, so there are some redundant coefficients. The discussion in Section 11.4.4 shows that one way to remove the redundancy is to include an overall constant and drop one of the time specific *and* one of the time-dummy variables. The model is, thus,

$$y_{it} = \mu + (\alpha_i - \alpha_1) + (\gamma_t - \gamma_1) + \mathbf{x}'_{it}\beta + \varepsilon_{it}.$$

(Note that the respective time- or individual-specific variable is zero when  $t$  or  $i$  equals one.) Ordinary least squares estimates of  $\beta$  are then obtained by regression of  $y_{it} - \bar{y}_i - \bar{y}_t + \bar{y}$  on  $\mathbf{x}_{it} - \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_t + \bar{\mathbf{x}}$ . Then  $(\alpha_i - \alpha_1)$  and  $(\gamma_t - \gamma_1)$  are estimated using the expressions in (9.2), while  $m = \bar{y} - \bar{\mathbf{x}}'\bar{\beta}$ . Using the following data, estimate the full set of coefficients for the least squares dummy variable model:

	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$	$t = 9$	$t = 10$
$i = 1$										
$y$	21.7	10.9	33.5	22.0	17.6	16.1	19.0	18.1	14.9	23.2
$x_1$	26.4	17.3	23.8	17.6	26.2	21.1	17.5	22.9	22.9	14.9
$x_2$	5.79	2.60	8.36	5.50	5.26	1.03	3.11	4.87	3.79	7.24
$i = 2$										
$y$	21.8	21.0	33.8	18.0	12.2	30.0	21.7	24.9	21.9	23.6
$x_1$	19.6	22.8	27.8	14.0	11.4	16.0	28.8	16.8	11.8	18.6
$x_2$	3.36	1.59	6.19	3.75	1.59	9.87	1.31	5.42	6.32	5.35
$i = 3$										
$y$	25.2	41.9	31.3	27.8	13.2	27.9	33.3	20.5	16.7	20.7
$x_1$	13.4	29.7	21.6	25.1	14.1	24.1	10.5	22.1	17.0	20.5
$x_2$	9.57	9.62	6.61	7.24	1.64	5.99	9.00	1.75	1.74	1.82
$i = 4$										
$y$	15.3	25.9	21.9	15.5	16.7	26.1	34.8	22.6	29.0	37.1
$x_1$	14.2	18.0	29.9	14.1	18.4	20.1	27.6	27.4	28.5	28.6
$x_2$	4.09	9.56	2.18	5.43	6.33	8.27	9.16	5.24	7.92	9.63

Test the hypotheses that (1) the “period” effects are all zero, (2) the “group” effects are all zero, and (3) both period and group effects are zero. Use an  $F$  test in each case.

6. *Two-way random effects model.* We modify the random effects model by the addition of a time-specific disturbance. Thus,

$$y_{it} = \alpha + \mathbf{x}'_{it}\beta + \varepsilon_{it} + u_i + v_t,$$

where

$$E[\varepsilon_{it} | \mathbf{X}] = E[u_i | \mathbf{X}] = E[v_t | \mathbf{X}] = 0,$$

$$E[\varepsilon_{it}u_j | \mathbf{X}] = E[\varepsilon_{it}v_s | \mathbf{X}] = E[u_i v_t | \mathbf{X}] = 0 \quad \text{for all } i, j, t, s$$

$$\text{Var}[\varepsilon_{it} | \mathbf{X}] = \sigma_\varepsilon^2, \quad \text{Cov}[\varepsilon_{it}, \varepsilon_{js} | \mathbf{X}] = 0 \quad \text{for all } i, j, t, s$$

$$\text{Var}[u_i | \mathbf{X}] = \sigma_u^2, \quad \text{Cov}[u_i, u_j | \mathbf{X}] = 0 \quad \text{for all } i, j$$

$$\text{Var}[v_t | \mathbf{X}] = \sigma_v^2, \quad \text{Cov}[v_t, v_s | \mathbf{X}] = 0 \quad \text{for all } t, s.$$

Write out the full disturbance covariance matrix for a data set with  $n = 2$  and  $T = 2$ .

CHAPTER 11 ♦ Models for Panel Data **429**

## 7. The model

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}$$

satisfies the groupwise heteroscedastic regression model of Section 9.7.2 All variables have zero means. The following sample second-moment matrix is obtained from a sample of 20 observations:

$$\begin{array}{cccc} & y_1 & y_2 & x_1 & x_2 \\ y_1 & \left[ \begin{array}{cccc} 20 & 6 & 4 & 3 \end{array} \right] \\ y_2 & \left[ \begin{array}{cccc} 6 & 10 & 3 & 6 \end{array} \right] \\ x_1 & \left[ \begin{array}{cccc} 4 & 3 & 5 & 2 \end{array} \right] \\ x_2 & \left[ \begin{array}{cccc} 3 & 6 & 2 & 10 \end{array} \right] \end{array}.$$

- a. Compute the two separate OLS estimates of  $\boldsymbol{\beta}$ , their sampling variances, the estimates of  $\sigma_1^2$  and  $\sigma_2^2$ , and the  $R^2$ 's in the two regressions.
- b. Carry out the Lagrange multiplier test of the hypothesis that  $\sigma_1^2 = \sigma_2^2$ .
- c. Compute the two-step FGLS estimate of  $\boldsymbol{\beta}$  and an estimate of its sampling variance. Test the hypothesis that  $\boldsymbol{\beta}$  equals 1.
- d. Carry out the Wald test of equal disturbance variances.
- e. Compute the maximum likelihood estimates of  $\boldsymbol{\beta}$ ,  $\sigma_1^2$ , and  $\sigma_2^2$  by iterating the FGLS estimates to convergence.
- f. Carry out a likelihood ratio test of equal disturbance variances.
- 8. Suppose that in the groupwise heteroscedasticity model of Section 9.7.2,  $\mathbf{X}_i$  is the same for all  $i$ . What is the generalized least squares estimator of  $\boldsymbol{\beta}$ ? How would you compute the estimator if it were necessary to estimate  $\sigma_i^2$ ?
- 9. The following table presents a hypothetical panel of data:

t	<i>i</i> = 1		<i>i</i> = 2		<i>i</i> = 3	
	y	x	y	x	y	x
1	30.27	24.31	38.71	28.35	37.03	21.16
2	35.59	28.47	29.74	27.38	43.82	26.76
3	17.90	23.74	11.29	12.74	37.12	22.21
4	44.90	25.44	26.17	21.08	24.34	19.02
5	37.58	20.80	5.85	14.02	26.15	18.64
6	23.15	10.55	29.01	20.43	26.01	18.97
7	30.53	18.40	30.38	28.13	29.64	21.35
8	39.90	25.40	36.03	21.78	30.25	21.34
9	20.44	13.57	37.90	25.65	25.41	15.86
10	36.85	25.60	33.90	11.66	26.04	13.28

- a. Estimate the groupwise heteroscedastic model of Section 9.7.2. Include an estimate of the asymptotic variance of the slope estimator. Use a two-step procedure, basing the FGLS estimator at the second step on residuals from the pooled least squares regression.
- b. Carry out the Wald and Lagrange multiplier tests of the hypothesis that the variances are all equal.

## 430 PART II ♦ Generalized Regression Model and Equation Systems

### Applications

As usual, the following applications below require econometric software. The computations can be done with any modern software package, so no specific program is recommended.

1. The data in Appendix Table F10.4 were used by Grunfeld (1958) and dozens of researchers since, including Zellner (1962, 1963) and Zellner and Huang (1962), to study rent estimators for panel data and linear regression systems. [See Kleiber (2010) and Zeileis.] The model is an investment equation

$$I_{it} = \beta_1 + \beta_2 F_{it} + \beta_3 C_{it} + \varepsilon_{it}, \quad t = 1, \dots, 20, \quad i = 1, \dots, 10,$$

where

$I_{it}$  = real gross investment for firm  $i$  in year  $t$ ,

$F_{it}$  = real value of the firm—shares outstanding,

$C_{it}$  = real value of the capital stock.

For present purposes, this is a balanced panel data set.

- a. Fit the pooled regression model.
- b. Referring to the results in part a, is there evidence of within groups correlation? Compute the robust standard errors for your pooled OLS estimator and compare them to the conventional ones.
- c. Compute the fixed effects estimator for these data, then, using an  $F$  test, test the hypothesis that the constants for the 10 firms are all the same.
- d. Use a Lagrange multiplier statistic to test for the presence of common effects in the data.
- e. Compute the one-way random effects estimator and report all estimation results. Explain the difference between this specification and the one in part c.
- f. Use a Hausman test to determine whether a fixed or random effects specification is preferred for these data.
2. The data in Appendix Table F6.1 are an unbalanced panel on 25 U.S. airlines in the pre-deregulation days of the 1970s and 1980s. The group sizes range from 2 to 15. Data in the file are the following variables. (Variable names contained in the data file are constructed to indicate the variable contents.)

Total cost,

Expenditures on Capital, Labor, Fuel, Materials, Property, and Equipment,

Price measures for the six inputs,

Quantity measures for the six inputs,

Output measured in revenue passenger miles, converted to an index number for the airline,

Load factor = the average percentage capacity utilization of the airline's fleet,

Stage = the average flight (stage) length in miles,

Points = the number of points served by the airline,

Year = the calendar year,

T = Year - 1969,

TI = the number of observations for the airline, repeated for each year.

**CHAPTER 11 ♦ Models for Panel Data 431**

Use these data to build a cost model for airline service. Allow for cross-airline heterogeneity in the constants in the model. Use both random and fixed effects specifications, and use available statistical tests to determine which is the preferred model. An appropriate cost model to begin the analysis would be

$$\ln \text{cost}_{it} = \alpha_i + \sum_{k=1}^6 \beta_k \ln \text{Price}_{k,it} + \gamma \ln \text{Output}_{it} + \varepsilon_{it}.$$

It is necessary to impose linear homogeneity in the input prices on the cost function, which you would do by dividing five of the six prices and the total cost by the sixth price (choose any one), then using  $\ln(\text{cost}/P_6)$  and  $\ln(P_k/P_6)$  in the regression. You might also generalize the cost function by including a quadratic term in the log of output in the function. A translog model would include the unique squares and cross products of the input prices and products of log output with the logs of the prices. The data include three additional factors that may influence costs, stage length, load factor and number of points served. Include them in your model, and use the appropriate test statistic to test whether they are, indeed, relevant to the determination of (log) total cost.

# 12

## ESTIMATION FRAMEWORKS IN ECONOMETRICS

---

### 12.1 INTRODUCTION

This chapter begins our treatment of methods of estimation. Contemporary econometrics offers the practitioner a remarkable variety of estimation methods, ranging from tightly parameterized likelihood-based techniques at one end to thinly stated nonparametric methods that assume little more than mere association between variables at the other, and a rich variety in between. Even the experienced researcher could be forgiven for wondering how they should choose from this long menu. It is certainly beyond our scope to answer this question here, but a few principles can be suggested. Recent research has leaned when possible toward methods that require few (or fewer) possibly unwarranted or improper assumptions. This explains the ascendance of the GMM estimator in situations where strong likelihood-based parameterizations can be avoided and robust estimation can be done in the presence of heteroscedasticity and serial correlation. (It is intriguing to observe that this is occurring at a time when advances in computation have helped bring about *increased* acceptance of very heavily parameterized Bayesian methods.)

As a general proposition, the progression from full to semi- to non-parametric estimation relaxes strong assumptions, but at the cost of weakening the conclusions that can be drawn from the data. As much as anywhere else, this is clear in the analysis of discrete choice models, which provide one of the most active literatures in the field. (A sampler appears in Chapter 17.) A formal probit or logit model allows estimation of probabilities, marginal effects, and a host of ancillary results, but at the cost of imposing the normal or logistic distribution on the data. Semiparametric and nonparametric estimators allow one to relax the restriction but often provide, in return, only ranges of probabilities, if that, and in many cases, preclude estimation of probabilities or useful marginal effects. One does have the virtue of robustness in the conclusions, however. [See, e.g., the symposium in Angrist (2001) for a spirited discussion on these points.]

Estimation properties is another arena in which the different approaches can be compared. Within a class of estimators, one can define “the best” (most efficient) means of using the data. (See Example 12.2 for an application.) Sometimes comparisons can be made across classes as well. For example, when they are estimating the same parameters—this remains to be established—the best parametric estimator will generally outperform the best semiparametric estimator. That is the value of the information, of course. The other side of the comparison, however, is that the semiparametric estimator will carry the day if the parametric model is misspecified in a fashion to which the semiparametric estimator is robust (and the parametric model is not).

CHAPTER 12 ♦ Estimation Frameworks in Econometrics **433**

Schools of thought have entered this conversation for a long time. Proponents of **Bayesian estimation** often took an almost theological viewpoint in their criticism of their classical colleagues. [See, for example, Poirier (1995).] Contemporary practitioners are usually more pragmatic than this. Bayesian estimation has gained currency as a set of techniques that can, in very many cases, provide both elegant and tractable solutions to problems that have heretofore been out of reach. Thus, for example, the **simulation-based estimation** advocated in the many papers of Chib and Greenberg (e.g., 1996) have provided solutions to a variety of computationally challenging problems.<sup>1</sup> Arguments as to the methodological virtue of one approach or the other have received much less attention than before.

Chapters 2 through 7 of this book have focused on the classical regression model and a particular estimator, least squares (linear and nonlinear). In this and the next four chapters, we will examine several general estimation strategies that are used in a wide variety of situations. This chapter will survey a few methods in the three broad areas we have listed. Chapter 13 discusses the **generalized method of moments**, which has emerged as the centerpiece of semiparametric estimation. Chapter 14 presents the method of **maximum likelihood**, the broad platform for parametric, classical estimation in econometrics. Chapter 15 discusses simulation-based estimation and bootstrapping. This is a recently developed body of techniques that have been made feasible by advances in estimation technology and which has made quite straightforward many estimators which were previously only scarcely used because of the sheer difficulty of the computations. Finally, Chapter 16 introduces the methods of Bayesian econometrics.

The list of techniques presented here is far from complete. We have chosen a set that constitutes the mainstream of econometrics. Certainly there are others that might be considered. [See, for example, Mittelhammer, Judge, and Miller (2000) for a lengthy catalog.] Virtually all of them are the subject of excellent monographs on the subject. In this chapter we will present several applications, some from the literature, some home grown, to demonstrate the range of techniques that are current in econometric practice. We begin in Section 12.2 with parametric approaches, primarily maximum likelihood. Because this is the subject of much of the remainder of this book, this section is brief. Section 12.2 also introduces Bayesian estimation, which in its traditional form, is as heavily parameterized as maximum likelihood estimation. Section 12.3 is on semiparametric estimation. GMM estimation is the subject of all of Chapter 13, so it is only introduced here. The technique of least absolute deviations is presented here as well. A range of applications from the recent literature is also surveyed. Section 12.4 describes nonparametric estimation. The fundamental tool, the kernel density estimator is developed, then applied to a problem in regression analysis. Two applications are presented here as well. Being focused on application, this chapter will say very little about the statistical theory for these techniques—such as their asymptotic properties.

<sup>1</sup>The penetration of Bayesian econometrics could be overstated. It is fairly well represented in current journals such as the *Journal of Econometrics*, *Journal of Applied Econometrics*, *Journal of Business and Economic Statistics*, and so on. On the other hand, in the six major general treatments of econometrics published in 2000, four (Hayashi, Ruud, Patterson, Davidson) do not mention Bayesian methods at all, a buffet of 32 essays (Baltagi) devotes only one to the subject, and the one that displays any preference (Mittelhammer  devotes nearly 10 percent (70) of its pages to Bayesian estimation, but all to the broad metatheory of linear regression model and none to the more elaborate applications that form the received applications in the many journals in the field.

## 434 PART III ♦ Instrumental Variables and Simultaneous Equations Models

(The results are developed at length in the literature, of course.) We will turn to the subject of the properties of estimators briefly at the end of the chapter, in Section 12.5, then in greater detail in Chapters 13 through 16.

### 12.2 PARAMETRIC ESTIMATION AND INFERENCE

Parametric estimation departs from a full statement of the **density** or probability model that provides the **data generating mechanism** for a random variable of interest. For the sorts of applications we have considered thus far, we might say that the joint density of a scalar random variable, “ $y$ ” and a random vector, “ $\mathbf{x}$ ” of interest can be specified by

$$f(y, \mathbf{x}) = g(y | \mathbf{x}, \boldsymbol{\beta}) \times h(\mathbf{x} | \boldsymbol{\theta}), \quad (12-1)$$

with unknown parameters  $\boldsymbol{\beta}$  and  $\boldsymbol{\theta}$ . To continue the application that has occupied us since Chapter 2, consider the linear regression model with normally distributed disturbances. The assumption produces a full statement of the **conditional density** that is the population from which an observation is drawn;

$$y_i | \mathbf{x}_i \sim N[\mathbf{x}'_i \boldsymbol{\beta}, \sigma^2].$$

All that remains for a full definition of the population is knowledge of the specific values taken by the *unknown*, but *fixed* parameters. With those in hand, the conditional probability distribution for  $y_i$  is completely defined—mean, variance, probabilities of certain events, and so on. (The marginal density for the conditioning variables is usually not of particular interest.) Thus, the signature features of this modeling platform are specifications of both the density and the features (parameters) of that density.

The **parameter space** for the parametric model is the set of allowable values of the parameters that satisfy some prior specification of the model. For example, in the regression model specified previously, the  $K$  regression slopes may take any real value, but the variance must be a positive number. Therefore, the parameter space for that model is  $[\boldsymbol{\beta}, \sigma^2] \in \mathbb{R}^K \times \mathbb{R}_+$ . “Estimation” in this context consists of specifying a criterion for ranking the points in the parameter space, then choosing that point (a point estimate) or a set of points (an interval estimate) that optimizes that criterion, that is, has the best ranking. Thus, for example, we chose linear least squares as one **estimation criterion** for the linear model. “Inference” in this setting is a process by which some regions of the (already specified) parameter space are deemed not to contain the unknown parameters, though, in more practical terms, we typically define a criterion and then, state that by that criterion, certain regions are *unlikely* to contain the true parameters.

#### 12.2.1 CLASSICAL LIKELIHOOD-BASED ESTIMATION

The most common (by far) class of parametric estimators used in econometrics is the maximum likelihood estimators. The underlying philosophy of this class of estimators is the idea of “sample information.” When the density of a sample of observations is completely specified, apart from the unknown parameters, then the joint density of those observations (assuming they are independent), is the **likelihood function**

$$f(y_1, y_2, \dots, \mathbf{x}_1, \mathbf{x}_2, \dots) = \prod_{i=1}^n f(y_i, \mathbf{x}_i | \boldsymbol{\beta}, \boldsymbol{\theta}). \quad (12-2)$$

## CHAPTER 12 ♦ Estimation Frameworks in Econometrics 435

This function contains all the information available in the sample about the population from which those observations were drawn. The strategy by which that information is used in estimation constitutes the estimator.

The **maximum likelihood estimator** [Fisher (1925)] is the function of the data that (as its name implies) maximizes the likelihood function (or, because it is usually more convenient, the log of the likelihood function). The motivation for this approach is most easily visualized in the setting of a discrete random variable. In this case, the likelihood function gives the joint probability for the observed sample observations, and the maximum likelihood estimator is the function of the sample information that makes the observed data most probable (at least by that criterion). Though the analogy is most intuitively appealing for a discrete variable, it carries over to continuous variables as well. Since this estimator is the subject of Chapter 14, which is quite lengthy, we will defer any formal discussion until then and consider instead two applications to illustrate the techniques and underpinnings.

### **Example 12.1 The Linear Regression Model**

Least squares weighs negative and positive deviations equally and gives disproportionate weight to large deviations in the calculation. This property can be an advantage or a disadvantage, depending on the data generating process. For normally distributed disturbances, this method is precisely the one needed to use the data most efficiently. If the data are generated by a normal distribution, then the log of the likelihood function is

$$\ln L = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

You can easily show that least squares is the estimator of choice for this model. Maximizing the function means minimizing the exponent, which is done by least squares for  $\boldsymbol{\beta}$ , then  $\mathbf{e}'\mathbf{e}/n$  follows as the estimator for  $\sigma^2$ .

If the appropriate distribution is deemed to be something other than normal—perhaps on the basis of an observation that the tails of the disturbance distribution are too thick—see Example 4.7 and Section 14.9.5.a—then there are three ways one might proceed. First, as we have observed, the consistency of least squares is robust to this failure of the specification, so long as the conditional mean of the disturbances is still zero. Some correction to the standard errors is necessary for proper inferences. Second, one might want to proceed to an estimator with better finite sample properties. The least absolute deviations estimator discussed in Section 12.3.2 is a candidate. Finally, one might consider some other distribution which accommodates the observed discrepancy. For example, Ruud (2000) examines in some detail a linear regression model with disturbances distributed according to the  $t$  distribution with  $v$  degrees of freedom. As long as  $v$  is finite, this random variable will have a larger variance than the normal. Which way should one proceed? The third approach is the least appealing. Surely if the normal distribution is inappropriate, then it would be difficult to come up with a plausible mechanism whereby the  $t$  distribution would not be. The LAD estimator might well be preferable if the sample were small. If not, then least squares would probably remain the estimator of choice, with some allowance for the fact that standard inference tools would probably be misleading. Current practice is generally to adopt the first strategy.

### **Example 12.2 The Stochastic Frontier Model**

The **stochastic frontier model**, discussed in detail in Chapter 19, is a regression-like model with a disturbance distribution that is asymmetric and distinctly nonnormal. The conditional density for the dependent variable in this model is

$$f(y | \mathbf{x}, \boldsymbol{\beta}, \sigma, \lambda) = \frac{\sqrt{2}}{\sigma \sqrt{\pi}} \exp \left[ \frac{-(y - \alpha - \mathbf{x}'\boldsymbol{\beta})^2}{2\sigma^2} \right] \Phi \left( \frac{-\lambda(y - \alpha - \mathbf{x}'\boldsymbol{\beta})}{\sigma} \right).$$

## 436 PART III ♦ Instrumental Variables and Simultaneous Equations Models

This produces a log-likelihood function for the model,

$$\ln L = -n \ln \sigma - \frac{n}{2} \ln \frac{2}{\pi} - \frac{1}{2} \sum_{i=1}^n \left( \frac{\varepsilon_i}{\sigma} \right)^2 + \sum_{i=1}^n \ln \Phi \left( \frac{-\lambda \varepsilon_i}{\sigma} \right).$$

There are at least two fully parametric estimators for this model. The maximum likelihood estimator is discussed in Section 18.2. Greene (2007) presents the following **method of moments** estimator: For the regression slopes, excluding the constant term, use least squares. For the parameters  $\sigma$ , and  $\lambda$ , based on the second and third moments of the least squares residuals and least squares constant, solve

$$\begin{aligned} m_2 &= \sigma_v^2 + [1 - 2/\pi]\sigma_u^2, \\ m_3 &= (2/\pi)^{1/2}[1 - 4/\pi]\sigma_u^3, \\ a &= \alpha + (2/\pi)^2\sigma_u, \end{aligned}$$

where  $\lambda = \sigma_u/\sigma_v$  and  $\sigma^2 = \sigma_u^2 + \sigma_v^2$ .

Both estimators are fully parametric. The maximum likelihood estimator is for the reasons discussed earlier. The method of moments estimators (see Section 13.1) are appropriate only for this distribution. Which is preferable? As we will see in Chapter 18, both estimators are consistent and asymptotically normally distributed. By virtue of the Cramér-Rao theorem, the maximum likelihood estimator has a smaller asymptotic variance. Neither has any small sample optimality properties. Thus, the only virtue of the method of moments estimator is that one can compute it with any standard regression/statistics computer package and a hand calculator whereas the maximum likelihood estimator requires specialized software (only somewhat—it is reasonably common).

### 12.2.2 MODELING JOINT DISTRIBUTIONS WITH COPULA FUNCTIONS

Specifying the likelihood function commits the analyst to a possibly strong assumption about the distribution of the random variable of interest. The payoff, of course, is the stronger inferences that this permits. However, when there is more than one random variable of interest, such as in a joint household decision on health care usage in the example to follow, formulating the full likelihood involves specifying the marginal distributions, which might be comfortable, and a full specification of the joint distribution, which is likely to be less so. In the typical situation, the model might involve two similar random variables and an ill-formed specification of correlation between them. Implicitly, this case involves specification of the marginal distributions. The joint distribution is an empirical necessity to allow the correlation to be nonzero. The **copula function** approach provides a mechanism that the researcher can use to steer around this situation.

Trivedi and Zimmer (2007) suggest a variety of applications that fit this description:

- Financial institutions are often concerned with the prices of different, related (dependent) assets. The typical multivariate normality assumption is problematic because of GARCH effects (see Section 20.13) and thick tails in the distributions. While specifying appropriate marginal distributions may be reasonably straightforward, specifying the joint distribution is anything but that. Klugman and Parsa (2000) is an application.
- There are many microeconometric applications in which straightforward marginal distributions cannot be readily combined into a natural joint distribution. The

CHAPTER 12 ♦ Estimation Frameworks in Econometrics **437**

bivariate event count model analyzed in Munkin and Trivedi (1999) and in the next example is an application.

- In the linear self-selection model of Chapter 19, the necessary joint distribution is part of a larger model. The likelihood function for the observed outcome involves the joint distribution of a variable of interest, hours, wages, income, and so on, and the probability of observation. The typical application is based on a joint normal distribution. Smith (2003, 2005) suggests some applications in which a flexible copula representation is more appropriate. [In an intriguing early application of copula modeling that was not labeled as such, since it greatly predates the econometric literature, Lee (1983) modeled the outcome variable in a selectivity model as normal, the observation probability as logistic, and the connection between them using what amounted to the “Gaussian” copula function shown next.]

Although the antecedents in the statistics literature date to Sklar's (1973) derivations, the applications in econometrics and finance are quite recent, with most applications appearing since 2000. [See the excellent survey by Trivedi and Zimmer (2007) for an extensive description.]

Consider a modeling problem in which the marginal cdfs of two random variables can be fully specified as  $F_1(y_1 | \bullet)$  and  $F_2(y_2 | \bullet)$ , where we condition on sample information (data) and parameters denoted “ $\bullet$ .“ For the moment, assume these are continuous random variables that obey all the axioms of probability. The bivariate cdf is  $F_{12}(y_1, y_2 | \bullet)$ . A (bivariate) copula function (the results also extend to multivariate functions) is a function  $C(u_1, u_2)$  defined over the unit square  $[(0 \leq u_1 \leq 1) \times (0 \leq u_2 \leq 1)]$  that satisfies

- (1)  $C(1, u_2) = u_2$  and  $C(u_1, 1) = u_1$ ,
- (2)  $C(0, u_2) = C(u_1, 0) = 0$ ,
- (3)  $\partial C(u_1, u_2)/\partial u_1 \geq 0$  and  $\partial C(u_1, u_2)/\partial u_2 \geq 0$ .

These are properties of bivariate cdfs for random variables  $u_1$  and  $u_2$  that are bounded in the unit square. It follows that the copula function is a two-dimensional cdf defined over the unit square that has one-dimensional marginal distributions that are standard uniform in the unit interval [that is, property (1)]. To make profitable use of this relationship, we note that the cdf of a random variable,  $F_1(y_1 | \bullet)$ , is, itself, a uniformly distributed random variable. This is the **fundamental probability transform** that we use for generating random numbers. (See Section 15.2.) In Sklar's (1973) **theorem**, the marginal cdfs play the roles of  $u_1$  and  $u_2$ . The theorem states that there exists a copula function,  $C(\cdot, \cdot)$  such that

$$F_{12}(y_1, y_2 | \bullet) = C[F_1(y_1 | \bullet), F_2(y_2 | \bullet)].$$

If  $F_{12}(y_1, y_2 | \bullet) = C[F_1(y_1 | \bullet), F_2(y_2 | \bullet)]$  is continuous and if the marginal cdfs have quantile (inverse) functions  $F_j^{-1}(u_j)$  where  $0 \leq u_j \leq 1$ , then the copula function can be expressed as

$$\begin{aligned} F_{12}(y_1, y_2 | \bullet) &= F_{12}[F_1^{-1}(u_1 | \bullet), F_2^{-1}(u_2 | \bullet)] \\ &= \text{Prob}[U_1 \leq u_1, U_2 \leq u_2] \\ &= C(u_1, u_2). \end{aligned}$$

### 438 PART III ♦ Instrumental Variables and Simultaneous Equations Models

In words, the theorem implies that the joint density can be written as the copula function evaluated at the two cumulative probability functions.

Copula functions allow the analyst to assemble joint distributions when only the marginal distributions can be specified. To fill in the desired element of correlation between the random variables, the copula function is written

$$F_{12}(y_1, y_2 | \bullet) = C[F_1(y_1 | \bullet), F_2(y_2 | \bullet), \theta],$$

where  $\theta$  is a “dependence parameter.” For continuous random variables, the joint pdf is then the mixed partial derivative,

$$\begin{aligned} f_{12}(y_1, y_2 | \bullet) &= c_{12}[F_1(y_1 | \bullet), F_2(y_2 | \bullet), \theta] \\ &= \partial^2 C[F_1(y_1 | \bullet), F_2(y_2 | \bullet), \theta] / \partial y_1 \partial y_2 \\ &= [\partial^2 C(., ., \theta) / \partial F_1 \partial F_2] f_1(y_1 | \bullet) f_2(y_2 | \bullet). \end{aligned} \quad (12-3)$$

A log-likelihood function can now be constructed using the logs of the right-hand sides of (12-3). Taking logs of (12-3) reveals the utility of the copula approach. The contribution of the joint observation to the log likelihood is

$$\ln f_{12}(y_1, y_2 | \bullet) = \ln[\partial^2 C(., ., \theta) / \partial F_1 \partial F_2] + \ln f_1(y_1 | \bullet) + \ln f_2(y_2 | \bullet).$$

Some of the common copula functions that have been used in applications are as follows:

Product:  $C[u_1, u_2, \theta] = u_1 \times u_2,$

FGM:  $C[u_1, u_2, \theta] = u_1 u_2 [1 + \theta(1 - u_1)(1 - u_2)],$

Gaussian:  $C[u_1, u_2, \theta] = \Phi_2[\Phi^{-1}(u_1), \Phi^{-1}(u_2), \theta],$

Clayton:  $C[u_1, u_2, \theta] = [u_1^{-\theta} + u_2^{-\theta} - 1]^{-1/\theta},$

Frank:  $C[u_1, u_2, \theta] = \frac{1}{\theta} \ln \left[ 1 + \frac{\exp(\theta u_1 - 1)\exp(\theta u_2 - 1)}{\exp(\theta) - 1} \right].$

The product copula implies that the random variables are independent, because it implies that the joint cdf is the product of the marginals. In the FGM (Fairlie, Gumbel, Morgenstern) copula, it can be seen that  $\theta = 0$  implies the product copula, or independence. The same result can be shown for the Clayton copula. In the Gaussian function, the copula is the bivariate normal cdf if the marginals happen to be normal to begin with. The essential point is that the marginals need not be normal to construct the copula function, so long as the marginal cdfs can be specified. (The dependence parameter is not the correlation between the variables. Trivedi and Zimmer provide transformations of  $\theta$  that are closely related to correlations for each copula function listed.)

The essence of the copula technique is that the researcher can specify and analyze the marginals and the copula functions separately. The likelihood function is obtained by formulating the cdfs [or the densities, because the differentiation in (12-3) will reduce the joint density to a convenient function of the marginal densities] and the copula.

#### **Example 12.3 Joint Modeling of a Pair of Event Counts**

The standard regression modeling approach for a random variable,  $y$ , that is a count of events is the Poisson regression model,

$$\text{Prob}[Y = y | \mathbf{x}] = \exp(-\lambda) \lambda^y / y!, \text{ where } \lambda = \exp(\mathbf{x}' \boldsymbol{\beta}), y = 0, 1, \dots$$

CHAPTER 12 ♦ Estimation Frameworks in Econometrics **439**

More intricate specifications use the negative binomial model (version 2, NB2),

$$\text{Prob}[Y = y | \mathbf{x}] = \frac{\Gamma(y + \alpha)}{\Gamma(\alpha) \Gamma(y + 1)} \left( \frac{\alpha}{\lambda + \alpha} \right)^{\alpha} \left( \frac{\lambda}{\lambda + \alpha} \right)^y, y = 0, 1, \dots,$$

where  $\alpha$  is an overdispersion parameter. (See Chapter 19.) A satisfactory, appropriate specification for bivariate outcomes has been an ongoing topic of research. Early suggestions were based on a latent mixture model,

$$y_1 = z + w_1,$$

$$y_2 = z + w_2,$$

where  $w_1$  and  $w_2$  have the Poisson or NB2 distributions specified earlier with conditional means  $\lambda_1$  and  $\lambda_2$  and  $z$  is taken to be an unobserved Poisson or NB variable. This formulation induces correlation between the variables but is unsatisfactory because that correlation must be positive. In a natural application,  $y_1$  is doctor visits and  $y_2$  is hospital visits. These could be negatively correlated. Munkin and Trivedi (1999) specified the jointness in the conditional mean functions, in the form of latent, common heterogeneity;

$$\lambda_j = \exp(\mathbf{x}'_j \boldsymbol{\beta}_j + \varepsilon)$$

where  $\varepsilon$  is common to the two functions. Cameron et al. (2004) used a bivariate copula approach to analyze Australian data on self-reported and actual physician visits (the latter maintained by the Health Insurance Commission). They made two adjustments to the preceding model we developed above. First, they adapted the basic copula formulation to these discrete random variables. Second, the variable of interest to them was not the actual or self-reported count, but the difference. Both of these are straightforward modifications of the basic copula model.

## 12.3 SEMIPARAMETRIC ESTIMATION

Semiparametric estimation is based on fewer assumptions than parametric estimation. In general, the distributional assumption is removed, and an estimator is devised from certain more general characteristics of the population. Intuition suggests two (correct) conclusions. First, the semiparametric estimator will be more robust than the parametric estimator—it will retain its properties, notably consistency across a greater range of specifications. Consider our most familiar example. The least squares slope estimator is consistent whenever the data are well behaved and the disturbances and the regressors are uncorrelated. This is even true for the frontier function in Example 12.2, which has an asymmetric, nonnormal disturbance. But, second, this robustness comes at a cost. The distributional assumption usually makes the preferred estimator more efficient than a robust one. The best robust estimator in its class will usually be inferior to the parametric estimator when the assumption of the distribution is correct. Once again, in the frontier function setting, least squares may be robust for the slopes, and it is the most efficient estimator that uses only the orthogonality of the disturbances and the regressors, but it will be inferior to the maximum likelihood estimator when the two-part normal distribution is the correct assumption.

### 12.3.1 GMM ESTIMATION IN ECONOMETRICS

Recent applications in economics include many that base estimation on the **method of moments**. The generalized method of moments departs from a set of model based

## 440 PART III ♦ Instrumental Variables and Simultaneous Equations Models

moment equations,  $E[\mathbf{m}(y_i, \mathbf{x}_i, \boldsymbol{\beta})] = \mathbf{0}$ , where the set of equations specifies a relationship known to hold in the population. We used one of these in the preceding paragraph. The least squares estimator can be motivated by noting that the essential assumption is that  $E[\mathbf{x}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta})] = \mathbf{0}$ . The estimator is obtained by seeking a parameter estimator,  $\hat{\boldsymbol{\beta}}$ , which mimics the population result;  $(1/n)\sum_i[\mathbf{x}_i(y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}})] = \mathbf{0}$ . These are, of course, the normal equations for least squares. Note that the estimator is specified without benefit of any  distributional assumption. Method of moments estimation is the subject of Chapter 15, so we will defer further analysis until then.

### 12.3.2 MAXIMUM EMPIRICAL LIKELIHOOD ESTIMATION

Empirical likelihood methods are suggested as a semiparametric alternative to maximum likelihood. As we shall see shortly, the estimator is closely related to the GMM estimator. Let  $\pi_i$  denote generically the probability that  $y_i|\mathbf{x}_i$  takes the realized value in the sample. Intuition suggests (correctly) that with no further information,  $\pi_i$  will equal  $1/n$ . The **empirical likelihood function** is

$$EL = \prod_{i=1}^n \pi_i^{1/n}.$$

The **maximum empirical likelihood estimator** maximizes  $EL$ . Equivalently, we maximize the log of the empirical likelihood,

$$ELL = \frac{1}{n} \sum_{i=1}^n \ln \pi_i.$$

As a maximization problem, this program lacks sufficient structure to admit a solution—the solutions for  $\pi_i$  are unbounded. If we impose the restrictions that  $\pi_i$  are probabilities that sum to one, we can use a Langragean formulation to solve the optimization problem,

$$ELL = \left[ \frac{1}{n} \sum_{i=1}^n \ln \pi_i \right] + \lambda \left[ 1 - \sum_{i=1}^n \pi_i \right].$$

This slightly restricts the problem since with  $0 < \pi_i < 1$  and  $\sum_i \pi_i = 1$ , the solution suggested earlier becomes obvious. (There is nothing in the problem that differentiates the  $\pi_i$ 's, so they must all be equal to each other.) Inserting this result in the derivative with respect to any specific  $\pi_i$  produces the remaining result,  $\lambda = 1$ .

The maximization problem becomes meaningful when we impose a structure on the data. To develop an example, we'll recall Example 7.5, a nonlinear regression equation for *Income* for the German Socioeconomic Panel data, where we specified

$$E[Income|Age, Sex, Education] = \exp(\mathbf{x}' \boldsymbol{\beta}) = h(\mathbf{x}, \boldsymbol{\beta}).$$

For purpose of an example, assume that *Education* may be endogenous in this equation, but we have available a set of instruments,  $\mathbf{z}$ , say (*Age*, *Health*, *Sex*, *MarketCondition*). We have assumed that there are more instruments (4) than included variables (3), so that the parameters will be overidentified (and the example will be complicated enough to be interesting). (See Sections 8.3.4 and 8.6.) The orthogonality conditions for nonlinear instrumental variable estimation are that the disturbances be uncorrelated with the instrumental variables, so

$$E\{\mathbf{z}_i[Income_i - h(\mathbf{x}_i, \boldsymbol{\beta})]\} = E[\mathbf{m}_i(\boldsymbol{\beta})] = \mathbf{0}.$$

## CHAPTER 12 ♦ Estimation Frameworks in Econometrics 441

The nonlinear least squares solution to this problem was developed in Example 8.10. A GMM estimator will minimize with respect to  $\beta$  the criterion function

$$q = \bar{\mathbf{m}}'(\beta) \mathbf{A} \bar{\mathbf{m}}(\beta)$$

where  $\mathbf{A}$  is the chosen weighting matrix. Note that for our example, including the constant term, there are four elements in  $\beta$  and five moment equations, so the parameters are overidentified.

If we impose the restrictions implied by our moment equations on the empirical likelihood function, instead, we obtain the population moment condition

$$\left[ \sum_{i=1}^n \pi_i \mathbf{z}_i (Income_i - h(\mathbf{x}_i, \beta)) \right] = 0.$$

(The probabilities are population quantities, so this is the expected value.) This produces the constrained empirical log likelihood

$$ELL = \left[ \frac{1}{n} \sum_{i=1}^n \ln \pi_i \right] + \lambda \left[ 1 - \sum_{i=1}^n \pi_i \right] + \gamma' \left[ \sum_{i=1}^n \pi_i \mathbf{z}_i (Income_i - h(\mathbf{x}_i, \beta)) \right].$$

The function is now maximized with respect to  $\pi_i$ ,  $\lambda$ ,  $\beta$  ( $K$  elements) and  $\gamma$  ( $L$  elements, the number of instrumental variables.) At the solution, the values of  $\pi_i$  provide, essentially, a set of weights. Cameron and Trivedi (2005, p. 205) provide a solution for  $\hat{\pi}_i$  in terms of  $(\beta, \gamma)$  and show, once again, that  $\lambda = 1$ . The concentrated  $ELL$  function with these inserted provides a function of  $\gamma$  and  $\beta$  that remains to be maximized.

The empirical likelihood estimator has the same asymptotic properties as the GMM estimator. (This makes sense, given the resemblance of the estimation criteria—ultimately, both are focused on the moment equations.) There is evidence that at least in some cases, the finite sample properties of the empirical likelihood estimator might be better than GMM. A survey appears in Imbens (2002). One suggested modification of the procedure is to replace the core function in  $(1/n) \sum_i \ln \pi_i$  with the **entropy** measure,

$$Entropy = (1/n) \sum_i \pi_i \ln \pi_i.$$

The **maximum entropy** estimator is developed in Golan, Judge, and Miller (1996) and Golan (2009).

### 12.3.3 LEAST ABSOLUTE DEVIATIONS ESTIMATION AND QUANTILE REGRESSION

Least squares can be severely distorted by outlying observations in a small sample. Recent applications in microeconomics and financial economics involving thick-tailed disturbance distributions, for example, are particularly likely to be affected by precisely these sorts of observations. (Of course, in those applications in finance involving hundreds of thousands of observations, which are becoming commonplace, this discussion is moot.) These applications have led to the proposal of “robust” estimators that are unaffected by outlying observations. One of these, the least absolute deviations, or LAD estimator discussed in Section 7.3.1, is also useful in its own right as an estimator of the conditional median function in the modified model

$$\text{Med}[y|\mathbf{x}] = \mathbf{x}'\beta_{.50}.$$

## 442 PART III ♦ Estimation Methodology

That is, rather than providing a robust alternative to least squares as an estimator of the slopes of  $E[y|\mathbf{x}]$ , LAD is an estimator of a different feature of the population. This is essentially a semiparametric specification in that it specifies only a particular feature of the distribution, its median, but not the distribution itself. It also specifies that the conditional median be a linear function of  $\mathbf{x}$ .

The median, in turn, is only one possible quantile of interest. If the model is extended to other quantiles of the conditional distribution, we obtain

$$Q[y|\mathbf{x}, q] = \mathbf{x}'\boldsymbol{\beta}_q \text{ such that } \text{Prob}[y \leq \mathbf{x}'\boldsymbol{\beta}_q | \mathbf{x}] = q, 0 < q < 1.$$

This is essentially a nonparametric specification. No assumption is made about the distribution of  $y|\mathbf{x}$  or about its conditional variance. The fact that  $q$  can vary continuously (strictly) between zero and one means that there is an infinite number of possible “parameter vectors.” It seems reasonable to view the coefficients, which we might write  $\boldsymbol{\beta}(q)$  less as fixed “parameters,” as we do in the linear regression model, than loosely as *features* of the distribution of  $y|\mathbf{x}$ . For example, it is not likely to be meaningful to view  $\boldsymbol{\beta}(.49)$  to be discretely different from  $\boldsymbol{\beta}(.50)$  or to compute precisely a particular difference such as  $\boldsymbol{\beta}(.5) - \boldsymbol{\beta}(.3)$ . On the other hand, the qualitative difference, or possibly the lack of a difference, between  $\boldsymbol{\beta}(.3)$  and  $\boldsymbol{\beta}(.5)$  may well be an interesting characteristic of the population. The quantile regression model is examined in Section 7.3.2.

### 12.3.4 KERNEL DENSITY METHODS

The kernel density estimator is an inherently nonparametric method, so it fits more appropriately into the next section. But some models that use kernel methods are not completely nonparametric. The partially linear model in the preceding example is a case in point. Many models retain an index function formulation, that is, build the specification around a linear function,  $\mathbf{x}'\boldsymbol{\beta}$ , which makes them at least semiparametric, but nonetheless still avoid distributional assumptions by using kernel methods. Lewbel's (2000) estimator for the binary choice model is another example.

#### **Example 12.4 Semiparametric Estimation of Binary Choice Models**

The core binary choice model analyzed in Section 17.9, the probit model, is a fully parametric specification. Under the assumptions of the model, maximum likelihood is the efficient (and appropriate) estimator. However, as documented in a voluminous literature, the estimator of  $\boldsymbol{\beta}$  is fragile with respect to failures of the distributional assumption. We will examine a few semiparametric and nonparametric estimators in Section 17.9. To illustrate the nature of the modeling process, we consider an estimator recently suggested by Lewbel (2000). The probit model is based on the normal distribution, with  $\text{Prob}[y_i = 1 | \mathbf{x}_i] = \text{Prob}[\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0]$  where  $\varepsilon_i \sim \mathbf{N}[0, 1]$ . The estimator of  $\boldsymbol{\beta}$  under this specification will be inconsistent if the distribution is not normal or if  $\varepsilon_i$  is heteroscedastic. Lewbel suggests the following: If (a) it can be assumed that  $\mathbf{x}_i$  contains a “special” variable,  $v_i$ , whose coefficient has a known sign—a method is developed for determining the sign and (b) the density of  $\varepsilon_i$  is independent of this variable, then a consistent estimator of  $\boldsymbol{\beta}$  can be obtained by regression of  $[y_i - s(v_i)] / f(v_i | \mathbf{x}_i)$  on  $\mathbf{x}_i$  where  $s(v_i) = 1$  if  $v_i > 0$  and 0 otherwise and  $f(v_i | \mathbf{x}_i)$  is a kernel density estimator of the density of  $v_i | \mathbf{x}_i$ . Lewbel's estimator is robust to heteroscedasticity and distribution. A method is also suggested for estimating the distribution of  $\varepsilon_i$ . Note that Lewbel's estimator is semiparametric. His underlying model is a function of the parameters  $\boldsymbol{\beta}$ , but the distribution is unspecified.

## CHAPTER 12 ♦ Estimation Frameworks in Econometrics 443

## 12.3.5 COMPARING PARAMETRIC AND SEMIPARAMETRIC ANALYSES

It is often of interest to compare the outcomes of parametric and semiparametric models. As we have noted earlier, the strong assumptions of the fully parametric model come at a cost; the inferences from the model are only as robust as the underlying assumptions. Of course, the other side of that equation is that when the assumptions are met, parametric models represent efficient strategies for analyzing the data. The alternative, semiparametric approaches relax assumptions such as normality and homoscedasticity. It is important to note that the model extensions to which semiparametric estimators are typically robust render the more heavily parameterized estimators inconsistent. The comparison is not just one of efficiency. As a consequence, comparison of parameter estimates can be misleading—the parametric and semiparametric estimators are often estimating very different quantities.

**Example 12.5 A Model of Vacation Expenditures**

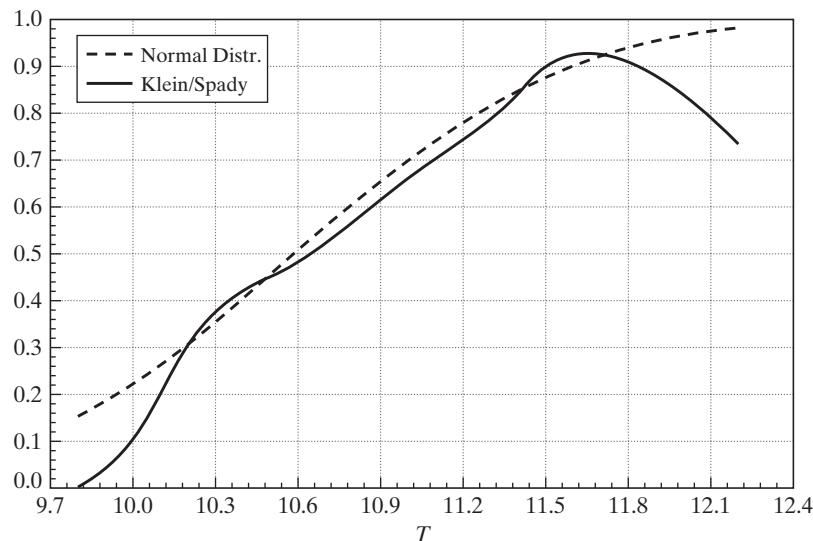
Melenberg and van Soest (1996) analyzed the 1981 vacation expenditures of a sample of 1,143 Dutch families. The important feature of the data that complicated the analysis was that 37 percent (423) of the families reported zero expenditures. A linear regression that ignores this feature of the data would be heavily skewed toward underestimating the response of expenditures to the covariates such as total family expenditures (budget), family size, age, or education. (See Section 18.3.) The standard parametric approach to analyzing data of this sort is the “Tobit,” or censored, regression model:

$$y_i^* = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \mathbf{N}[0, \sigma^2], \\ y_i = \max(0, y_i^*).$$



(Maximum likelihood estimation of this model is examined in detail in Section 18.3.) The model rests on two strong assumptions, normality and homoscedasticity. Both assumptions can be relaxed in a more elaborate parametric framework, but the authors found that test statistics persistently rejected one or both of the assumptions even with the extended specifications. An alternative approach that is robust to both is Powell's (1984, 1986a, b) censored least absolute deviations estimator, which is a more technically demanding computation based on the LAD estimator in Section 7.3.1. Not surprisingly, the parameter estimates produced by the two approaches vary widely. The authors computed a variety of estimators of  $\boldsymbol{\beta}$ . A useful exercise that they did not undertake would be to compare the partial effects from the different models. This is a benchmark on which the differences between the different estimators cannot sometimes be reconciled. In the Tobit model,  $\partial E[y_i | \mathbf{x}_i] / \partial \mathbf{x}_i = \Phi(\mathbf{x}'_i \boldsymbol{\beta} / \sigma) \boldsymbol{\beta}$  (see Section 18.3). It is unclear how to compute the counterpart in the semiparametric model, since the underlying specification holds only that  $\text{Med}[\varepsilon_i | \mathbf{x}_i] = 0$ . (The authors report on the *Journal of Applied Econometrics* data archive site that these data are proprietary. As such, we were unable to extend the analysis to obtain estimates of partial effects.) This highlights a significant difficulty with the semiparametric approach to estimation. In a nonlinear model such as this one, it is often the partial effects that are of interest, not the coefficients. But, one of the byproducts of the more “robust” specification is that the partial effects are undefined.

In a second stage of the analysis, the authors decomposed their expenditure equation into a “participation” equation that modeled probabilities for the binary outcome “expenditure = 0 or 1” and a conditional expenditure equation for those with positive expenditure. [In Chapter 25, we will label this a “hurdle” model. See Mullahy (1986).] For this step, the authors once again used a parametric model based on the normal distribution (the probit model—see Section 17.3) and a semiparametric model that is robust to distribution and heteroscedasticity developed by Klein and Spady (1993). As before, the coefficient estimates differ substantially.

**444 PART III ♦ Instrumental Variables and Simultaneous Equations Models**


**FIGURE 12.1** Predicted Probabilities of Positive Expenditure.

However, in this instance, the specification tests are considerably more sympathetic to the parametric model. Figure 12.1, which reproduces their Figure 2, compares the predicted probabilities from the two models. The dashed curve is the probit model. Within the range of most of the data, the models give quite similar predictions. Once again, however, it is not possible to compare partial effects. The interesting outcome from this part of the analysis seems to be that the failure of the parametric specification resides more in the modeling of the continuous expenditure variable than with the model that separates the two subsamples based on zero or positive expenditures.

## 12.4 NONPARAMETRIC ESTIMATION

Researchers have long held reservations about the strong assumptions made in parametric models fit by maximum likelihood. The linear regression model with normal disturbances is a leading example. Splines, translog models, and polynomials all represent attempts to generalize the functional form. Nonetheless, questions remain about how much generality can be obtained with such approximations. The techniques of nonparametric estimation discard essentially all fixed assumptions about functional form and distribution. Given their very limited structure, it follows that nonparametric specifications rarely provide very precise inferences. The benefit is that what information is provided is extremely robust. The centerpiece of this set of techniques is the kernel density estimator that we have used in the preceding examples. We will examine some examples, then examine an application to a bivariate regression.<sup>2</sup>

<sup>2</sup>The set of literature in this area of econometrics is large and rapidly growing. Major references which provide an applied and theoretical foundation are Härdle (1990), Pagan and Ullah (1999), and Li and Racine (2007).

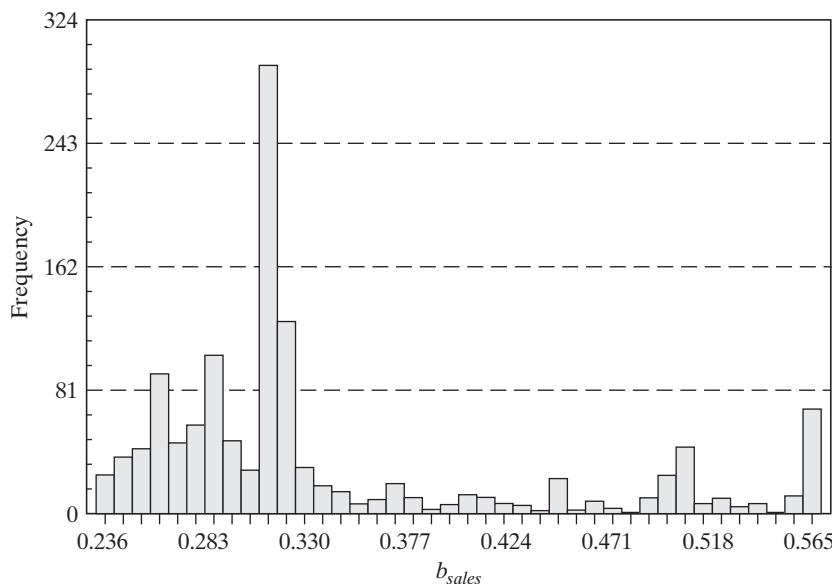
### 12.4.1 KERNEL DENSITY ESTIMATION

Sample statistics such as a mean, variance, and range give summary information about the values that a random variable may take. But, they do not suffice to show the distribution of values that the random variable takes, and these may be of interest as well. The density of the variable is used for this purpose. A fully parametric approach to density estimation begins with an assumption about the form of a distribution. Estimation of the density is accomplished by estimation of the parameters of the distribution. To take the canonical example, if we decide that a variable is generated by a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , then the density is fully characterized by these parameters. It follows that

$$\hat{f}(x) = f(x | \hat{\mu}, \hat{\sigma}^2) = \frac{1}{\hat{\sigma}} \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{x - \hat{\mu}}{\hat{\sigma}} \right)^2 \right].$$

One may be unwilling to make a narrow distributional assumption about the density. The usual approach in this case is to begin with a **histogram** as a descriptive device. Consider an example. In Examples 15.6 and in Greene (2004a), we estimate a model that produces a conditional estimator of a slope vector for each of the 1,270 firms in our sample. We might be interested in the distribution of these estimators across firms. In particular, the conditional estimates of the estimated slope on  $\ln \text{sales}$  for the 1,270 firms have a sample mean of 0.3428, a standard deviation of 0.08919, a minimum of 0.2361, and a maximum of 0.5664. This tells us little about the distribution of values, though the fact that the mean is well below the midrange of .4013 might suggest some skewness. The histogram in Figure 12.2 is much more revealing. Based on what we see

**FIGURE 12.2** Histogram for Estimated  $b_{\text{sales}}$  Coefficients.



## 446 PART III ♦ Instrumental Variables and Simultaneous Equations Models

thus far, an assumption of normality might not be appropriate. The distribution seems to be bimodal, but certainly no particular functional form seems natural.

The histogram is a crude density estimator. The rectangles in the figure are called bins. By construction, they are of equal width. (The parameters of the histogram are the number of bins, the bin width, and the leftmost starting point. Each is important in the shape of the end result.) Because the frequency count in the bins sums to the sample size, by dividing each by  $n$ , we have a density estimator that satisfies an obvious requirement for a density; it sums (integrates) to one. We can formalize this by laying out the method by which the frequencies are obtained. Let  $x_k$  be the midpoint of the  $k$ th bin and let  $h$  be the width of the bin—we will shortly rename  $h$  to be the bandwidth for the density estimator. The distances to the left and right boundaries of the bins are  $h/2$ . The frequency count in each bin is the number of observations in the sample which fall in the range  $x_k \pm h/2$ . Collecting terms, we have our “estimator”

$$\hat{f}(x) = \frac{1}{n} \frac{\text{frequency in bin}_x}{\text{width of bin}_x} = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} \mathbf{1}\left(x - \frac{h}{2} < x_i < x + \frac{h}{2}\right),$$

where  $\mathbf{1}(\text{statement})$  denotes an indicator function which equals 1 if the statement is true and 0 if it is false and  $\text{bin}_x$  denotes the bin which has  $x$  as its midpoint. We see, then, that the histogram is an estimator, at least in some respects, like other estimators we have encountered. The event in the indicator can be rearranged to produce an equivalent form

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} \mathbf{1}\left(-\frac{1}{2} < \frac{x_i - x}{h} < \frac{1}{2}\right).$$

This form of the estimator simply counts the number of points that are within one half-bin width of  $x_k$ .

Albeit rather crude, this “naive” (its formal name in the literature) estimator is in the form of **kernel density estimators** that we have met at various points;

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left[\frac{x_i - x}{h}\right], \quad \text{where } K[z] = \mathbf{1}[-1/2 < z < 1/2].$$

The naive estimator has several shortcomings. It is neither smooth nor continuous. Its shape is partly determined by where the leftmost and rightmost terminals of the histogram are set. (In constructing a histogram, one often chooses the bin width to be a specified fraction of the sample range. If so, then the terminals of the lowest and highest bins will equal the minimum and maximum values in the sample, and this will partly determine the shape of the histogram. If, instead, the bin width is set irrespective of the sample values, then this problem is resolved.) More importantly, the shape of the histogram will be crucially dependent on the bandwidth itself. (Unfortunately, this problem remains even with more sophisticated specifications.)

The crudeness of the weighting function in the estimator is easy to remedy. Rosenblatt's (1956) suggestion was to substitute for the naive estimator some other weighting function which is continuous and which also integrates to one. A number of candidates have been suggested, including the (long) list in Table 12.1. Each of these is smooth, continuous, symmetric, and equally attractive. The logit and normal kernels are defined so that the weight only asymptotically falls to zero whereas the others fall to zero at

CHAPTER 12 ♦ Estimation Frameworks in Econometrics **447****TABLE 12.1** Kernels for Density Estimation

<i>Kernel</i>	<i>Formula K[z]</i>
Epanechnikov	$0.75(1 - 0.2z^2)/2.236$ if $ z  \leq 5$ , 0 else
Normal	$\phi(z)$ (normal density),
Logit	$\Lambda(z)[1 - \Lambda(z)]$ (logistic density)
Uniform	0.5 if $ z  \leq 1$ , 0 else
Beta	$0.75(1 - z)(1 + z)$ if $ z  \leq 1$ , 0 else
Cosine	$1 + \cos(2\pi z)$ if $ z  \leq 0.5$ , 0 else
Triangle	$1 -  z $ , if $ z  \leq 1$ , 0 else
Parzen	$4/3 - 8z^2 + 8 z ^3$ if $ z  \leq 0.5$ , $8(1 -  z )^3/3$ if $0.5 <  z  \leq 1$ , 0 else.

specific points. It has been observed that in constructing a density estimator, the choice of kernel function is rarely crucial, and is usually minor in importance compared to the more difficult problem of choosing the bandwidth. (The logit and normal kernels appear to be the default choice in many applications.)

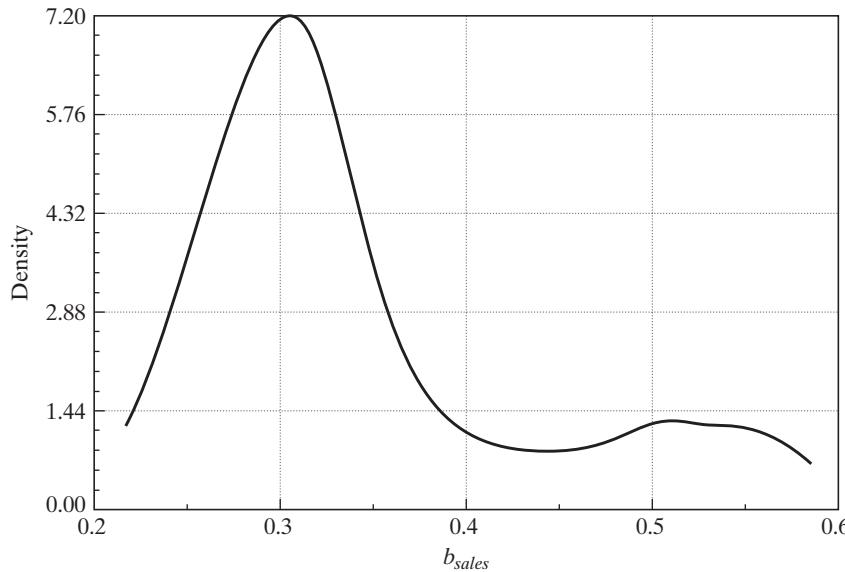
The kernel density function is an estimator. For any specific  $x$ ,  $\hat{f}(x)$  is a sample statistic,

$$\hat{f}(z) = \frac{1}{n} \sum_{i=1}^n g(x_i | z, h).$$

Because  $g(x_i | z, h)$  is nonlinear, we should expect a bias in a finite sample. It is tempting to apply our usual results for sample moments, but the analysis is more complicated because the bandwidth is a function of  $n$ . Pagan and Ullah (1999) have examined the properties of kernel estimators in detail and found that under certain assumptions, the estimator is consistent and asymptotically normally distributed but biased in finite samples. The bias is a function of the bandwidth, but for an appropriate choice of  $h$ , the bias does vanish asymptotically. As intuition might suggest, the larger is the bandwidth, the greater is the bias, but at the same time, the smaller is the variance. This might suggest a search for an optimal bandwidth. After a lengthy analysis of the subject, however, the authors' conclusion provides little guidance for finding one. One consideration does seem useful. For the proportion of observations captured in the bin to converge to the corresponding area under the density, the width itself must shrink more slowly than  $1/n$ . Common applications typically use a **bandwidth** equal to some multiple of  $n^{-1/5}$  for this reason. Thus, the one we used earlier is  $h = 0.9 \times s/n^{1/5}$ . To conclude the illustration begun earlier, Figure 12.3 is a logit-based kernel density estimator for the distribution of slope estimates for the model estimated earlier. The resemblance to the histogram in Figure 12.2 is to be expected.

## 12.5 PROPERTIES OF ESTIMATORS

The preceding has been concerned with methods of estimation. We have surveyed a variety of techniques that have appeared in the applied literature. We have not yet examined the statistical properties of these estimators. Although, as noted earlier, we will leave extensive analysis of the asymptotic theory for more advanced treatments, it is appropriate to spend at least some time on the fundamental theoretical platform which underlies these techniques.

**448 PART III ♦ Instrumental Variables and Simultaneous Equations Models**

**FIGURE 12.3** Kernel Density for  $b_{sales}$  Coefficients.

**12.5.1 STATISTICAL PROPERTIES OF ESTIMATORS**

Properties that we have considered are as follows:

- **Unbiasedness:** This is a finite sample property that can be established in only a very small number of cases. Strict unbiasedness is rarely of central importance outside the linear regression model. However, “asymptotic unbiasedness” (whereby the expectation of an estimator converges to the true parameter as the sample size grows), might be of interest. [See, e.g., Pagan and Ullah (1999, Section 2.5.1 on the subject of the kernel density estimator).] In most cases, however, discussions of asymptotic unbiasedness are actually directed toward consistency, which is a more desirable property.
- **Consistency:** This is a much more important property. Econometricians are rarely willing to place much credence in an estimator for which consistency cannot be established.
- **Asymptotic normality:** This property forms the platform for most of the statistical inference that is done with common estimators. When asymptotic normality cannot be established, it sometimes becomes difficult to find a method of progressing beyond simple presentation of the numerical values of estimates (with caveats). However, most of the contemporary literature in macroeconomics and time-series analysis is strongly focused on estimators that are decidedly not asymptotically normally distributed. The implication is that this property takes its importance only in context, not as an absolute virtue.
- **Asymptotic efficiency:** Efficiency can rarely be established in absolute terms. Efficiency within a class often can, however. Thus, for example, a great deal can be said about the relative efficiency of maximum likelihood and GMM estimators

## CHAPTER 12 ♦ Estimation Frameworks in Econometrics 449

in the class of consistent and asymptotically normally distributed (CAN) estimators. There are two important practical considerations in this setting. First, the researcher will want to know that he or she has not made demonstrably suboptimal use of the data. (The literature contains discussions of GMM estimation of fully specified parametric probit models—GMM estimation in this context is unambiguously inferior to maximum likelihood.) Thus, when possible, one would want to avoid obviously inefficient estimators. On the other hand, it will usually be the case that the researcher is not choosing from a list of available estimators; he or she has one at hand, and questions of relative efficiency are moot.

### 12.5.2 EXTREMUM ESTIMATORS

An **extremum estimator** is one that is obtained as the optimizer of a **criterion function**  $q(\theta | \text{data})$ . Three that have occupied much of our effort thus far are

- Least squares:  $\hat{\theta}_{LS} = \text{Argmax} [-(1/n) \sum_{i=1}^n (y_i - h(\mathbf{x}_i, \theta_{LS}))^2]$ ,
- Maximum likelihood:  $\hat{\theta}_{ML} = \text{Argmax} [(1/n) \sum_{i=1}^n \ln f(y_i | \mathbf{x}_i, \theta_{ML})]$ , and
- GMM:  $\hat{\theta}_{GMM} = \text{Argmax} [-\bar{\mathbf{m}}(\text{data}, \theta_{GMM})' \mathbf{W} \bar{\mathbf{m}}(\text{data}, \theta_{GMM})]$ .

(We have changed the signs of the first and third only for convenience so that all three may be cast as the same type of optimization problem.) The least squares and maximum likelihood estimators are examples of **M estimators**, which are defined by optimizing over a sum of terms. Most of the familiar theoretical results developed here and in other treatises concern the behavior of extremum estimators. Several of the estimators considered in this chapter are extremum estimators, but a few—including the Bayesian estimators, some of the semiparametric estimators, and all of the nonparametric estimators—are not. Nonetheless, we are interested in establishing the properties of estimators in all these cases, whenever possible. The end result for the practitioner will be the set of statistical properties that will allow him or her to draw with confidence conclusions about the data generating process(es) that have motivated the analysis in the first place.

Derivations of the behavior of extremum estimators are pursued at various levels in the literature. (See, for example, any of the sources mentioned in Footnote 1 of this chapter.) Amemiya (1985) and Davidson and MacKinnon (2004) are very accessible treatments. Newey and McFadden (1994) is a rigorous analysis that provides a current, standard source. Our discussion at this point will only suggest the elements of the analysis. The reader is referred to one of these sources for detailed proofs and derivations.

### 12.5.3 ASSUMPTIONS FOR ASYMPTOTIC PROPERTIES OF EXTREMUM ESTIMATORS

Some broad results are needed in order to establish the asymptotic properties of the classical (not Bayesian) conventional extremum estimators noted above.

1. **The parameter space** (see Section 12.2) must be convex and the parameter vector that is the object of estimation must be a point in its interior. The first requirement rules out ill-defined estimation problems such as estimating a parameter which can only take one of a finite discrete set of values. Thus, searching for the date of a structural break in a time-series model as if it were a conventional parameter

## 450 PART III ♦ Instrumental Variables and Simultaneous Equations Models

leads to a nonconvexity. Some proofs in this context are simplified by assuming that the parameter space is compact. (A compact set is closed and bounded.) However, assuming compactness is usually restrictive, so we will opt for the weaker requirement.

2. **The criterion function** must be concave in the parameters. (See Section A.8.2.) This assumption implies that with a given data set, the objective function has an interior optimum and that we can locate it. Criterion functions need not be “globally concave”; they may have multiple optima. But, if they are not at least “locally concave,” then we cannot speak meaningfully about optimization. One would normally only encounter this problem in a badly structured model, but it is possible to formulate a model in which the estimation criterion is monotonically increasing or decreasing in a parameter. Such a model would produce a nonconcave criterion function.<sup>3</sup> The distinction between compactness and concavity in the preceding condition is relevant at this point. If the criterion function is strictly continuous in a compact parameter space, then it has a maximum in that set and assuming concavity is not necessary. The problem for estimation, however, is that this does not rule out having that maximum occur on the (assumed) boundary of the parameter space. This case interferes with proofs of consistency and asymptotic normality. The overall problem is solved by assuming that the criterion function is concave in the neighborhood of the true parameter vector.
3. **Identifiability of the parameters.** Any statement that begins with “the true parameters of the model,  $\theta_0$  are identified if . . .” is problematic because if the parameters are “not identified,” then arguably, they are not *the* parameters of the (any) model. (For example, there is no “true” parameter vector in the unidentified model of Example 2.5.) A useful way to approach this question that avoids the ambiguity of trying to define *the* true parameter vector first and then asking if it is identified (estimable) is as follows, where we borrow from Davidson and MacKinnon (1993, p. 591): Consider the parameterized model,  $M$ , and the set of allowable data generating processes for the model,  $\mu$ . Under a particular parameterization  $\mu$ , let there be an assumed “true” parameter vector,  $\theta(\mu)$ . Consider any parameter vector  $\theta$  in the parameter space,  $\Theta$ . Define

$$q_\mu(\mu, \theta) = \text{plim}_\mu q_n(\theta | \text{data}).$$

This function is the probability limit of the objective function under the assumed parameterization  $\mu$ . If this probability limit exists (is a finite constant) and moreover, if

$$q_\mu[\mu, \theta(\mu)] > q_\mu(\mu, \theta) \text{ if } \theta \neq \theta(\mu),$$

then, if the parameter space is compact, the parameter vector is identified by the criterion function. We have not assumed compactness. For a convex parameter

---

<sup>3</sup>In their Exercise 23.6, Griffiths, Hill, and Judge (1993), based (alas) on the first edition of this text, suggest a probit model for statewide voting outcomes that includes dummy variables for region: Northeast, Southeast, West, and Mountain. One would normally include three of the four dummy variables in the model, but Griffiths et al. carefully dropped two of them because in addition to the dummy variable trap, the Southeast variable is always zero when the dependent variable is zero. Inclusion of this variable produces a nonconcave likelihood function—the parameter on this variable diverges. Analysis of a closely related case appears as a caveat on page 272 of Amemiya (1985).

## CHAPTER 12 ♦ Estimation Frameworks in Econometrics 451

space, we would require the additional condition that there exist no sequences without limit points  $\theta^m$  such that  $q(\mu, \theta^m)$  converges to  $q[\mu, \theta(\mu)]$ .

The approach taken here is to assume first that the model has *some* set of parameters. The identifiability criterion states that assuming this is the case, the probability limit of the criterion is maximized at these parameters. This result rests on convergence of the criterion function to a finite value at any point in the interior of the parameter space. Because the criterion function is a function of the data, this convergence requires a statement of the properties of the data—for example, well behaved in some sense. Leaving that aside for the moment, interestingly, the results to this point already establish the consistency of the M estimator. In what might seem to be an extremely terse fashion, Amemiya (1985) defined identifiability simply as “existence of a consistent estimator.” We see that identification and the conditions for consistency of the M estimator are substantively the same.

This form of identification is necessary, in theory, to establish the consistency arguments. In any but the simplest cases, however, it will be extremely difficult to verify in practice. Fortunately, there are simpler ways to secure identification that will appeal more to the intuition:

- For the least squares estimator, a sufficient condition for identification is that any two different parameter vectors,  $\theta$  and  $\theta_0$ , must be able to produce different values of the conditional mean function. This means that for any two different parameter vectors, there must be an  $\mathbf{x}_i$  that produces different values of the conditional mean function. You should verify that for the linear model, this is the full rank assumption A.2. For the model in Example 2.5, we have a regression in which  $x_2 = x_3 + x_4$ . In this case, any parameter vector of the form  $(\beta_1, \beta_2 - a, \beta_3 + a, \beta_4)$  produces the same conditional mean as  $(\beta_1, \beta_2, \beta_3, \beta_4)$  regardless of  $\mathbf{x}_i$ ; so this model is not identified. The full rank assumption is needed to preclude this problem. For nonlinear regressions, the situation is much more complicated, and there is no simple generality. Example 11.2 shows a nonlinear regression model that is not identified and how the lack of identification is remedied.
  - For the maximum likelihood estimator, a condition similar to that for the regression model is needed. For any two parameter vectors,  $\theta \neq \theta_0$ , it must be possible to produce different values of the density  $f(y_i | \mathbf{x}_i, \theta)$  for some data vector  $(y_i, \mathbf{x}_i)$ . Many econometric models that are fit by maximum likelihood are “index function” models that involve densities of the form  $f(y_i | \mathbf{x}_i, \theta) = f(y_i | \mathbf{x}'_i \theta)$ . When this is the case, the same full rank assumption that applies to the regression model may be sufficient. (If there are no other parameters in the model, then it will be sufficient.)
  - For the GMM estimator, not much simplicity can be gained. A sufficient condition for identification is that  $E[\bar{\mathbf{m}}(\mathbf{data}, \theta)] \neq \mathbf{0}$  if  $\theta \neq \theta_0$ .
4. **Behavior of the data** has been discussed at various points in the preceding text. The estimators are based on means of functions of observations. (You can see this in all three of the preceding definitions. Derivatives of these criterion functions will likewise be means of functions of observations.) Analysis of their large sample behaviors will turn on determining conditions under which certain sample means of functions of observations will be subject to laws of large numbers such as the Khinchine (D.5) or Chebychev (D.6) theorems, and what must be assumed in

## 452 PART III ♦ Instrumental Variables and Simultaneous Equations Models

order to assert that “root- $n$ ” times sample means of functions will obey central limit theorems such as the Lindeberg–Feller (D.19) or Lyapounov (D.20) theorems for cross sections or the Martingale Difference Central Limit theorem for dependent observations (Theorem 19.3). Ultimately, this is the issue in establishing the statistical properties. The convergence property claimed above must occur in the context of the data. These conditions have been discussed in Sections 4.4.1 and 4.4.2 under the heading of “well-behaved data.” At this point, we will assume that the data are well behaved.

### 12.5.4 ASYMPTOTIC PROPERTIES OF ESTIMATORS

With all this apparatus in place, the following are the standard results on asymptotic properties of M estimators:

#### THEOREM 12.1 Consistency of M Estimators

*If (a) the parameter space is convex and the true parameter vector is a point in its interior, (b) the criterion function is concave, (c) the parameters are identified by the criterion function, and (d) the data are well behaved, then the M estimator converges in probability to the true parameter vector.*

Proofs of consistency of M estimators rely on a fundamental convergence result that, itself, rests on assumptions (a) through (d) in Theorem 12.1. We have assumed identification. The fundamental device is the following: Because of its dependence on the data,  $q(\theta | \text{data})$  is a random variable. We assumed in (c) that  $\text{plim } q(\theta | \text{data}) = q_0(\theta)$  for any point in the parameter space. Assumption (c) states that the maximum of  $q_0(\theta)$  occurs at  $q_0(\theta_0)$ , so  $\theta_0$  is the maximizer of the probability limit. By its definition, the estimator  $\hat{\theta}$  is the maximizer of  $q(\theta | \text{data})$ . Therefore, consistency requires the limit of the maximizer,  $\hat{\theta}$  be equal to the maximizer of the limit,  $\theta_0$ . Our identification condition establishes this. We will use this approach in somewhat greater detail in Section 14.4.5.a where we establish consistency of the maximum likelihood estimator.

#### THEOREM 12.2 Asymptotic Normality of M Estimators

*If*

- (i)  $\hat{\theta}$  is a consistent estimator of  $\theta_0$  where  $\theta_0$  is a point in the interior of the parameter space;
- (ii)  $q(\theta | \text{data})$  is concave and twice continuously differentiable in  $\theta$  in a neighborhood of  $\theta_0$ ;
- (iii)  $\sqrt{n}[\partial q(\theta_0 | \text{data})/\partial\theta_0] \xrightarrow{d} N[\mathbf{0}, \Phi]$ ;
- (iv) for any  $\theta$  in  $\Theta$ ,  $\lim_{n \rightarrow \infty} \Pr[|(\partial^2 q(\theta | \text{data})/\partial\theta_k \partial\theta_m) - h_{km}(\theta)| > \varepsilon] = 0 \forall \varepsilon > 0$  where  $h_{km}(\theta)$  is a continuous finite valued function of  $\theta$ ;
- (v) the matrix of elements  $\mathbf{H}(\theta)$  is nonsingular at  $\theta_0$ , then  $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\{\mathbf{0}, [\mathbf{H}^{-1}(\theta_0)\Phi\mathbf{H}^{-1}(\theta_0)]\}$ .

**CHAPTER 12 ♦ Estimation Frameworks in Econometrics 453**

The proof of asymptotic normality is based on the mean value theorem from calculus and a Taylor series expansion of the derivatives of the maximized criterion function around the true parameter vector;

$$\sqrt{n} \frac{\partial q(\hat{\theta} | \text{data})}{\partial \hat{\theta}} = \mathbf{0} = \sqrt{n} \frac{\partial q(\theta_0 | \text{data})}{\partial \theta_0} + \frac{\partial^2 q(\bar{\theta} | \text{data})}{\partial \bar{\theta} \partial \bar{\theta}'} \sqrt{n} (\hat{\theta} - \theta_0).$$

The second derivative is evaluated at a point  $\bar{\theta}$  that is between  $\hat{\theta}$  and  $\theta_0$ , that is,  $\bar{\theta} = w\hat{\theta} + (1-w)\theta_0$  for some  $0 < w < 1$ . Because we have assumed  $\text{plim } \hat{\theta} = \theta_0$ , we see that the matrix in the second term on the right must be converging to  $\mathbf{H}(\theta_0)$ . The assumptions in the theorem can be combined to produce the claimed normal distribution. Formal proof of this set of results appears in Newey and McFadden (1994). A somewhat more detailed analysis based on this theorem appears in Section 14.4.5.b, where we establish the asymptotic normality of the maximum likelihood estimator.

The preceding was restricted to M estimators, so it remains to establish counterparts for the important GMM estimator. Consistency follows along the same lines used earlier, but asymptotic normality is a bit more difficult to establish. We will return to this issue in Chapter 13, where, once again, we will sketch the formal results and refer the reader to a source such as Newey and McFadden (1994) for rigorous derivation.

The preceding results are not straightforward in all estimation problems. For example, the least absolute deviations (LAD) is not among the estimators noted earlier, but it is an M estimator and it shares the results given here. The analysis is complicated because the criterion function is not continuously differentiable. Nonetheless, consistency and asymptotic normality have been established. [See Koenker and Bassett (1982) and Amemiya (1985, pp. 152–154).] Some of the semiparametric and all of the nonparametric estimators noted require somewhat more intricate treatments. For example, Pagan and Ullah (Sections 2.5 and 2.6) are able to establish the familiar desirable properties for the kernel density estimator  $f(x^*)$ , but it requires a somewhat more involved analysis of the function and the data than is necessary, say, for the linear regression or binomial logit model. The interested reader can find many lengthy and detailed analyses of asymptotic properties of estimators in, for example, Amemiya (1985), Newey and McFadden (1994), Davidson and MacKinnon (2004) and Hayashi (2000). In practical terms, it is rarely possible to verify the conditions for an estimation problem at hand, and they are usually simply assumed. However, finding violations of the conditions is sometimes more straightforward, and this is worth pursuing. For example, lack of parametric identification can often be detected by analyzing the model itself.

#### **12.5.5 TESTING HYPOTHESES**

The preceding describes a set of results that (more or less) unifies the theoretical underpinnings of three of the major classes of estimators in econometrics, least squares, maximum likelihood, and GMM. A similar body of theory has been produced for the familiar test statistics, Wald, likelihood ratio (LR), and Lagrange multiplier (LM). [See Newey and McFadden (1994).] All of these have been laid out in practical terms elsewhere in this text, so in the interest of brevity, we will refer the interested reader to the background sources listed for the technical details.

**454 PART III ♦ Instrumental Variables and Simultaneous Equations Models****12.6 SUMMARY AND CONCLUSIONS**

This chapter has presented a short overview of estimation in econometrics. There are various ways to approach such a survey. The current literature can be broadly grouped by three major types of estimators—parametric, semiparametric, and nonparametric. It has been suggested that the overall drift in the literature is from the first toward the third of these, but on a closer look, we see that this is probably not the case. Maximum likelihood is still the estimator of choice in many settings. New applications have been found for the GMM estimator, but at the same time, new Bayesian and simulation estimators, all fully parametric, are emerging at a rapid pace. Certainly, the range of tools that can be applied in any setting is growing steadily.

**Key Terms and Concepts**

- Bandwidth
- Bayesian estimation
- Bootstrap
- Conditional density
- Copula function
- Criterion function
- Data generating mechanism
- Density
- Empirical likelihood function
- Entropy
- Estimation criterion
- Extremum estimator
- Fundamental probability transform
- Generalized method of moments
- Histogram
- Identifiability
- Kernel density estimator
- Least absolute deviations (LAD)
- Likelihood function
- M estimator
- Maximum empirical likelihood estimator
- Maximum entropy
- Maximum likelihood estimator
- Method of moments
- Nearest neighbor
- Nonparametric estimators
- Parameter space
- Parametric estimation
- Partially linear model
- Quantile regression
- Semiparametric estimation
- Simulation-based estimation
- Sklar's theorem
- Smoothing function
- Stochastic frontier model

**Exercise and Question**

1. Compare the fully parametric and semiparametric approaches to estimation of a discrete choice model such as the multinomial logit model discussed in Chapter 17. What are the benefits and costs of the semiparametric approach?

## 13

# MINIMUM DISTANCE ESTIMATION AND THE GENERALIZED METHOD OF MOMENTS

---

## 13.1 INTRODUCTION

The **maximum likelihood estimator** presented in Chapter 14 is fully efficient among consistent and asymptotically normally distributed estimators, *in the context of the specified parametric model*. The possible shortcoming in this result is that to attain that efficiency, it is necessary to make possibly strong, restrictive assumptions about the distribution, or data generating process. The generalized method of moments (GMM) estimators discussed in this chapter move away from parametric assumptions, toward estimators that are robust to some variations in the underlying data generating process.

This chapter will present a number of fairly general results on parameter estimation. We begin with perhaps the oldest formalized theory of estimation, the classical theory of the method of moments. This body of results dates to the pioneering work of Fisher (1925). The use of sample moments as the building blocks of estimating equations is fundamental in econometrics. GMM is an extension of this technique that, as will be clear shortly, encompasses nearly all the familiar estimators discussed in this book. Section 13.2 will introduce the estimation framework with the method of moments. The technique of minimum distance estimation is developed in Section 13.3. Formalities of the GMM estimator are related in Section 13.4. Section 13.5 discusses hypothesis testing based on moment equations. Major applications, including dynamic panel data models, are described in Section 13.6.

### **Example 13.1 Euler Equations and Life Cycle Consumption**

One of the most often cited applications of the GMM principle for estimating econometric models is Hall's (1978) permanent income model of consumption. The original form of the model (with some small changes in notation) posits a hypothesis about the optimizing behavior of a consumer over the life cycle. Consumers are hypothesized to act according to the model:

$$\text{Maximize } E_t \left[ \sum_{\tau=0}^{T-t} \left( \frac{1}{1+\delta} \right)^{\tau} U(c_{t+\tau}) \mid \Omega_t \right] \text{ subject to } \sum_{\tau=0}^{T-t} \left( \frac{1}{1+r} \right)^{\tau} (c_{t+\tau} - w_{t+\tau}) = A_t.$$

The information available at time  $t$  is denoted  $\Omega_t$  so that  $E_t$  denotes the expectation formed at time  $t$  based on the information set  $\Omega_t$ . The maximand is the expected discounted stream of future utility from consumption from time  $t$  until the end of life at time  $T$ . The individual's subjective rate of time preference is  $\beta = 1/(1+\delta)$ . The real rate of interest,  $r \geq \delta$  is assumed to be constant. The utility function  $U(c_t)$  is assumed to be strictly concave and time separable (as shown in the model). One period's consumption is  $c_t$ . The intertemporal budget constraint states that the present discounted excess of  $c_t$  over earnings,  $w_t$ , over the lifetime equals

## 456 PART III ♦ Estimation Methodology

total assets  $A_t$  not including human capital. In this model, it is claimed that the only source of uncertainty is  $w_t$ . No assumption is made about the stochastic properties of  $w_t$  except that there exists an expected future earnings,  $E_t[w_{t+\tau} | \Omega_t]$ . Successive values are not assumed to be independent and  $w_t$  is not assumed to be stationary.

Hall's major "theorem" in the paper is the solution to the optimization problem, which states

$$E_t[U'(c_{t+1}) | \Omega_t] = \frac{1 + \delta}{1 + r} U'(c_t).$$

For our purposes, the major conclusion of the paper is "Corollary 1" which states "No information available in time  $t$  apart from the level of consumption,  $c_t$ , helps predict future consumption,  $c_{t+1}$ , in the sense of affecting the expected value of marginal utility. In particular, income or wealth in periods  $t$  or earlier are irrelevant once  $c_t$  is known." We can use this as the basis of a model that can be placed in the GMM framework. To proceed, it is necessary to assume a form of the utility function. A common (convenient) form of the utility function is  $U(c_t) = c_t^{1-\alpha}/(1 - \alpha)$ , which is monotonic,  $U' = c_t^{-\alpha} > 0$  and concave,  $U''/U' = -\alpha/c_t < 0$ . Inserting this form into the solution, rearranging the terms, and reparameterizing it for convenience, we have

$$E_t \left[ (1 + r) \left( \frac{1}{1 + \delta} \right) \left( \frac{c_{t+1}}{c_t} \right)^{-\alpha} - 1 | \Omega_t \right] = E_t [\beta(1 + r) R_{t+1}^\lambda - 1 | \Omega_t] = 0,$$

where  $R_{t+1} = c_{t+1}/c_t$  and  $\lambda = -\alpha$ .

Hall assumed that  $r$  was constant over time. Other applications of this modeling framework [for example, Hansen and Singleton (1982)] have modified the framework so as to involve a forecasted interest rate,  $r_{t+1}$ . How one proceeds from here depends on what is in the information set. The unconditional mean does not identify the two parameters. The corollary states that the only relevant information in the information set is  $c_t$ . Given the form of the model, the more natural instrument might be  $R_t$ . This assumption exactly identifies the two parameters in the model:

$$E_t \left[ (\beta(1 + r_{t+1}) R_{t+1}^\lambda - 1) \begin{pmatrix} 1 \\ R_t \end{pmatrix} \right] = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

As stated, the model has no testable implications. These two moment equations would exactly identify the two unknown parameters. Hall hypothesized several models involving income and consumption which would overidentify and thus place restrictions on the model.

### 13.2 CONSISTENT ESTIMATION: THE METHOD OF MOMENTS

Sample statistics such as the mean and variance can be treated as simple descriptive measures. In our discussion of estimation in Appendix C, however, we argue that, in general, sample statistics each have a counterpart in the population, for example, the correspondence between the sample mean and the population expected value. The natural (perhaps obvious) next step in the analysis is to use this analogy to justify using the sample "moments" as estimators of these population parameters. What remains to establish is whether this approach is the best, or even a good way to use the sample data to infer the characteristics of the population.

The basis of the **method of moments** is as follows: In random sampling, under generally benign assumptions, a sample statistic will converge in probability to some constant. For example, with i.i.d. random sampling,  $\bar{m}_2' = (1/n) \sum_{i=1}^n y_i^2$  will converge in

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 457

mean square to the variance plus the square of the mean of the random variable,  $y_i$ . This constant will, in turn, be a function of the unknown parameters of the distribution. To estimate  $K$  parameters,  $\theta_1, \dots, \theta_K$ , we can compute  $K$  such statistics,  $\bar{m}_1, \dots, \bar{m}_K$ , whose **probability limits** are known functions of the parameters. These  $K$  moments are equated to the  $K$  functions, and the functions are inverted to express the parameters as functions of the moments. The moments will be consistent by virtue of a law of large numbers (Theorems D.4–D.9). They will be asymptotically normally distributed by virtue of the Lindeberg–Levy **Central Limit theorem** (D.18). The derived parameter estimators will inherit consistency by virtue of the Slutsky theorem (D.12) and asymptotic normality by virtue of the delta method (Theorem D.21).

This section will develop this technique in some detail, partly to present it in its own right and partly as a prelude to the discussion of the generalized method of moments, or GMM, estimation technique, which is treated in Section 13.4.

### 13.2.1 RANDOM SAMPLING AND ESTIMATING THE PARAMETERS OF DISTRIBUTIONS

Consider independent, identically distributed random sampling from a distribution  $f(y | \theta_1, \dots, \theta_K)$  with finite moments up to  $E[y^{2K}]$ . The **random sample** consists of  $n$  observations,  $y_1, \dots, y_n$ . The  $k$ th “raw” or **uncentered moment** is

$$\bar{m}'_k = \frac{1}{n} \sum_{i=1}^n y_i^k.$$

By Theorem D.4,

$$E[\bar{m}'_k] = \mu'_k = E[y_i^k],$$

and

$$\text{Var}[\bar{m}'_k] = \frac{1}{n} \text{Var}[y_i^k] = \frac{1}{n} (\mu'_{2k} - \mu'^2_k).$$

By convention,  $\mu'_1 = E[y_i] = \mu$ . By the Khinchine theorem, D.5,

$$\text{plim } \bar{m}'_k = \mu'_k = E[y_i^k].$$

Finally, by the Lindeberg–Levy central limit theorem,

$$\sqrt{n}(\bar{m}'_k - \mu'_k) \xrightarrow{d} N[0, \mu'_{2k} - \mu'^2_k].$$

In general,  $\mu'_k$  will be a function of the underlying parameters. By computing  $K$  raw moments and equating them to these functions, we obtain  $K$  equations that can (in principle) be solved to provide estimates of the  $K$  unknown parameters.

#### **Example 13.2 Method of Moments Estimator for $N[\mu, \sigma^2]$**

In random sampling from  $N[\mu, \sigma^2]$ ,

$$\text{plim } \frac{1}{n} \sum_{i=1}^n y_i = \text{plim } \bar{m}'_1 = E[y_i] = \mu,$$

and

$$\text{plim } \frac{1}{n} \sum_{i=1}^n y_i^2 = \text{plim } \bar{m}'_2 = \text{Var}[y_i] + \mu^2 = \sigma^2 + \mu^2.$$

## 458 PART III ♦ Estimation Methodology

Equating the right- and left-hand sides of the probability limits gives moment estimators

$$\hat{\mu} = \bar{m}'_1 = \bar{y},$$

and

$$\hat{\sigma}^2 = \bar{m}'_2 - \bar{m}'_1^2 = \left( \frac{1}{n} \sum_{i=1}^n y_i^2 \right) - \left( \frac{1}{n} \sum_{i=1}^n y_i \right)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2.$$

Note that  $\hat{\sigma}^2$  is biased, although both estimators are consistent.

Although the moments based on powers of  $y$  provide a natural source of information about the parameters, other functions of the data may also be useful. Let  $m_k(\cdot)$  be a continuous and differentiable function not involving the sample size  $n$ , and let

$$\bar{m}_k = \frac{1}{n} \sum_{i=1}^n m_k(y_i), \quad k = 1, 2, \dots, K.$$

These are also “moments” of the data. It follows from Theorem D.4 and the corollary, (D-5), that

$$\text{plim } \bar{m}_k = E[m_k(y_i)] = \mu_k(\theta_1, \dots, \theta_K).$$

We assume that  $\mu_k(\cdot)$  involves some of or all the parameters of the distribution. With  $K$  parameters to be estimated, the  **$K$  moment equations**,

$$\bar{m}_1 - \mu_1(\theta_1, \dots, \theta_K) = 0,$$

$$\bar{m}_2 - \mu_2(\theta_1, \dots, \theta_K) = 0,$$

...

$$\bar{m}_K - \mu_K(\theta_1, \dots, \theta_K) = 0,$$

provide  $K$  equations in  $K$  unknowns,  $\theta_1, \dots, \theta_K$ . If the equations are continuous and functionally independent, then **method of moments estimators** can be obtained by solving the system of equations for

$$\hat{\theta}_k = \hat{\theta}_k[\bar{m}_1, \dots, \bar{m}_K].$$

As suggested, there may be more than one set of moments that one can use for estimating the parameters, or there may be more moment equations available than are necessary.

### Example 13.3 Inverse Gaussian (Wald) Distribution

The inverse Gaussian distribution is used to model survival times, or elapsed times from some beginning time until some kind of transition takes place. The standard form of the density for this random variable is

$$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left[-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right], \quad y > 0, \lambda > 0, \mu > 0.$$

The mean is  $\mu$  while the variance is  $\mu^3/\lambda$ . The efficient maximum likelihood estimators of the two parameters are based on  $(1/n) \sum_{i=1}^n y_i$  and  $(1/n) \sum_{i=1}^n (1/y_i)$ . Because the mean and variance are simple functions of the underlying parameters, we can also use the sample mean and sample variance as moment estimators of these functions. Thus, an alternative pair of method of moments estimators for the parameters of the Wald distribution can be based on  $(1/n) \sum_{i=1}^n y_i$  and  $(1/n) \sum_{i=1}^n y_i^2$ . The precise formulas for these two pairs of estimators is left as an exercise.

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 459

**Example 13.4 Mixtures of Normal Distributions**

Quandt and Ramsey (1978) analyzed the problem of estimating the parameters of a mixture of normal distributions. Suppose that each observation in a random sample is drawn from one of two different normal distributions. The probability that the observation is drawn from the first distribution,  $N[\mu_1, \sigma_1^2]$ , is  $\lambda$ , and the probability that it is drawn from the second is  $(1 - \lambda)$ . The density for the observed  $y$  is

$$\begin{aligned} f(y) &= \lambda N[\mu_1, \sigma_1^2] + (1 - \lambda) N[\mu_2, \sigma_2^2], \quad 0 \leq \lambda \leq 1 \\ &= \frac{\lambda}{(2\pi\sigma_1^2)^{1/2}} e^{-1/2[(y-\mu_1)/\sigma_1]^2} + \frac{1-\lambda}{(2\pi\sigma_2^2)^{1/2}} e^{-1/2[(y-\mu_2)/\sigma_2]^2}. \end{aligned}$$

Before proceeding, we note [ ] this density is precisely the same as the finite mixture model described in Section 14.7.d. Maximum likelihood estimation of the model using the method described there would be simpler than the method of moment generating functions developed here.

The sample mean and second through fifth **central moments**,

$$\bar{m}_k = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^k, \quad k = 2, 3, 4, 5,$$

provide five equations in five unknowns that can be solved (via a ninth-order polynomial) for consistent estimators of the five parameters. Because  $\bar{y}$  converges in probability to  $E[y_i] = \mu$ , the theorems given earlier for  $\bar{m}'_k$  as an estimator of  $\mu'_k$  apply as well to  $\bar{m}_k$  as an estimator of

$$\mu_k = E[(y_i - \mu)^k].$$

For the mixed normal distribution, the mean and variance are

$$\mu = E[y_i] = \lambda\mu_1 + (1 - \lambda)\mu_2,$$

and

$$\sigma^2 = \text{Var}[y_i] = \lambda\sigma_1^2 + (1 - \lambda)\sigma_2^2 + 2\lambda(1 - \lambda)(\mu_1 - \mu_2)^2,$$

which suggests how complicated the familiar method of moments is likely to become. An alternative method of estimation proposed by the authors is based on

$$E[e^{ty_i}] = \lambda e^{t\mu_1 + t^2\sigma_1^2/2} + (1 - \lambda)e^{t\mu_2 + t^2\sigma_2^2/2} = \Lambda_t,$$

where  $t$  is any value not necessarily an integer. Quandt and Ramsey (1978) suggest choosing five values of  $t$  that are not too close together and using the statistics

$$\bar{M}_t = \frac{1}{n} \sum_{i=1}^n e^{ty_i}$$

to estimate the parameters. The moment equations are  $\bar{M}_t - \Lambda_t(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda) = 0$ . They label this procedure the **method of moment generating functions**. (See Section B.6 for definition of the moment generating function.)

In most cases, method of moments estimators are not efficient. The exception is in random sampling from **exponential families** of distributions.

## 460 PART III ♦ Estimation Methodology

### DEFINITION 13.1 Exponential Family

An exponential (parametric) family of distributions is one whose log-likelihood is of the form

$$\ln L(\boldsymbol{\theta} | \text{data}) = a(\text{data}) + b(\boldsymbol{\theta}) + \sum_{k=1}^K c_k(\text{data})s_k(\boldsymbol{\theta}),$$

where  $a(\cdot)$ ,  $b(\cdot)$ ,  $c_k(\cdot)$ , and  $s_k(\cdot)$  are functions. The members of the “family” are distinguished by the different parameter values.

If the log-likelihood function is of this form, then the functions  $c_k(\cdot)$  are called **sufficient statistics**.<sup>1</sup> When sufficient statistics exist, method of moments estimator(s) can be functions of them. In this case, the method of moments estimators will also be the maximum likelihood estimators, so, of course, they will be efficient, at least asymptotically. We emphasize, in this case, the probability distribution is fully specified. Because the normal distribution is an exponential family with sufficient statistics  $\bar{m}'_1$  and  $\bar{m}'_2$ , the estimators described in Example 13.2 are fully efficient. (They are the maximum likelihood estimators.) The mixed normal distribution is not an exponential family. We leave it as an exercise to show that the Wald distribution in Example 13.3 is an exponential family. You should be able to show that the sufficient statistics are the ones that are suggested in Example 13.3 as the bases for the MLEs of  $\mu$  and  $\lambda$ .

#### Example 13.5 Gamma Distribution

The gamma distribution (see Section B.4.5) is

$$f(y) = \frac{\lambda^P}{\Gamma(P)} e^{-\lambda y} y^{P-1}, \quad y \geq 0, P > 0, \lambda > 0.$$

The log-likelihood function for this distribution is

$$\frac{1}{n} \ln L = [P \ln \lambda - \ln \Gamma(P)] - \lambda \frac{1}{n} \sum_{i=1}^n y_i + (P-1) \frac{1}{n} \sum_{i=1}^n \ln y_i.$$

This function is an exponential family with  $a(\text{data}) = 0$ ,  $b(\boldsymbol{\theta}) = n[P \ln \lambda - \ln \Gamma(P)]$  and two sufficient statistics,  $\frac{1}{n} \sum_{i=1}^n y_i$  and  $\frac{1}{n} \sum_{i=1}^n \ln y_i$ . The method of moments estimators based on  $\frac{1}{n} \sum_{i=1}^n y_i$  and  $\frac{1}{n} \sum_{i=1}^n \ln y_i$  would be the maximum likelihood estimators. But, we also have

$$\text{plim } \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} y_i \\ y_i^2 \\ \ln y_i \\ 1/y_i \end{bmatrix} = \begin{bmatrix} P/\lambda \\ P(P+1)/\lambda^2 \\ \Psi(P) - \ln \lambda \\ \lambda/(P-1) \end{bmatrix}.$$

(The functions  $\Gamma(P)$  and  $\Psi(P) = d \ln \Gamma(P) / dP$  are discussed in Section E.2.3.) Any two of these can be used to estimate  $\lambda$  and  $P$ .

<sup>1</sup>Stuart and Ord (1989, pp. 1–29) give a discussion of sufficient statistics and exponential families of distributions. A result that we will use in Chapter 17 is that if the statistics,  $c_k(\text{data})$  are sufficient statistics, then the conditional density  $f[y_1, \dots, y_n | c_k(\text{data})]$ ,  $k = 1, \dots, K$  is not a function of the parameters.

**CHAPTER 13 ♦ Minimum Distance Estimation and GMM 461**

For the income data in Example C.1, the four moments listed earlier are

$$(\bar{m}_1', \bar{m}_2', \bar{m}_*', \bar{m}_{-1}') = \frac{1}{n} \sum_{i=1}^n \left[ y_i, y_i^2, \ln y_i, \frac{1}{y_i} \right] = [31.278, 1453.96, 3.22139, 0.050014].$$

The method of moments estimators of  $\theta = (P, \lambda)$  based on the six possible pairs of these moments are as follows:

$$(\hat{P}, \hat{\lambda}) = \begin{bmatrix} \bar{m}_1' & \bar{m}_2' & \bar{m}_{-1}' \\ \bar{m}_2' & 2.05682, 0.065759 & \\ \bar{m}_{-1}' & 2.77198, 0.0886239 & 2.60905, 0.080475 \\ \bar{m}_*' & 2.4106, 0.0770702 & 2.26450, 0.071304 & 3.03580, 0.1018202 \end{bmatrix}.$$

The maximum likelihood estimates are  $\hat{\theta}(\bar{m}_1', \bar{m}_*') = (2.4106, 0.0770702)$ .

### 13.2.2 ASYMPTOTIC PROPERTIES OF THE METHOD OF MOMENTS ESTIMATOR

In a few cases, we can obtain the exact distribution of the method of moments estimator. For example, in sampling from the normal distribution,  $\hat{\mu}$  has mean  $\mu$  and variance  $\sigma^2/n$  and is normally distributed, while  $\hat{\sigma}^2$  has mean  $[(n-1)/n]\sigma^2$  and variance  $[(n-1)/n]^2 2\sigma^4/(n-1)$  and is exactly distributed as a multiple of a chi-squared variate with  $(n-1)$  degrees of freedom. If sampling is not from the normal distribution, the exact variance of the sample mean will still be  $\text{Var}[y]/n$ , whereas an asymptotic variance for the moment estimator of the population variance could be based on the leading term in (D-27), in Example D.10, but the precise distribution may be intractable.

There are cases in which no explicit expression is available for the variance of the underlying sample moment. For instance, in Example 13.4, the underlying sample statistic is

$$\bar{M}_t = \frac{1}{n} \sum_{i=1}^n e^{ty_i} = \frac{1}{n} \sum_{i=1}^n M_{it}.$$

The exact variance of  $\bar{M}_t$  is known only if  $t$  is an integer. But if sampling is random, and if  $\bar{M}_t$  is a sample mean: we can estimate its variance with  $1/n$  times the sample variance of the observations on  $M_{it}$ . We can also construct an estimator of the covariance of  $\bar{M}_t$  and  $\bar{M}_s$ :

$$\text{Est. Asy. Cov}[\bar{M}_t, \bar{M}_s] = \frac{1}{n} \left\{ \frac{1}{n} \sum_{i=1}^n [(e^{ty_i} - \bar{M}_t)(e^{sy_i} - \bar{M}_s)] \right\}.$$

In general, when the moments are computed as

$$\bar{m}_{n,k} = \frac{1}{n} \sum_{i=1}^n m_k(\mathbf{y}_i), \quad k = 1, \dots, K,$$

where  $\mathbf{y}_i$  is an observation on a vector of variables, an appropriate estimator of the asymptotic covariance matrix of  $\bar{\mathbf{m}}_n = [\bar{m}_{n,1}, \dots, \bar{m}_{n,K}]$  can be computed using

$$\frac{1}{n} \mathbf{F}_{jk} = \frac{1}{n} \left\{ \frac{1}{n} \sum_{i=1}^n [(m_j(\mathbf{y}_i) - \bar{m}_j)(m_k(\mathbf{y}_i) - \bar{m}_k)] \right\}, \quad j, k = 1, \dots, K.$$

(One might divide the inner sum by  $n-1$  rather than  $n$ . Asymptotically it is the same.) This estimator provides the asymptotic covariance matrix for the moments used in

## 462 PART III ♦ Estimation Methodology

computing the estimated parameters. Under the assumption of i.i.d. random sampling from a distribution with finite moments,  $n\mathbf{F}$  will converge in probability to the appropriate covariance matrix of the normalized vector of moments,  $\Phi = \text{Asy.Var}[\sqrt{n}\bar{\mathbf{m}}_n(\boldsymbol{\theta})]$ . Finally, under our assumptions of random sampling, although the precise distribution is likely to be unknown, we can appeal to the Lindeberg–Levy central limit theorem (D.18) to obtain an asymptotic approximation.

To formalize the remainder of this derivation, refer back to the moment equations, which we will now write

$$\bar{m}_{n,k}(\theta_1, \theta_2, \dots, \theta_K) = 0, \quad k = 1, \dots, K.$$

The subscript  $n$  indicates the dependence on a data set of  $n$  observations. We have also combined the sample statistic (sum) and function of parameters,  $\mu(\theta_1, \dots, \theta_K)$  in this general form of the moment equation. Let  $\bar{\mathbf{G}}_n(\boldsymbol{\theta})$  be the  $K \times K$  matrix whose  $k$ th row is the vector of partial derivatives

$$\bar{\mathbf{G}}'_{n,k} = \frac{\partial \bar{m}_{n,k}}{\partial \boldsymbol{\theta}'}.$$

Now, expand the set of solved moment equations around the true values of the parameters  $\boldsymbol{\theta}_0$  in a linear **Taylor series**. The linear approximation is

$$\mathbf{0} \approx [\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0)] + \bar{\mathbf{G}}'_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

Therefore,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx -[\bar{\mathbf{G}}_n(\boldsymbol{\theta}_0)]^{-1}\sqrt{n}[\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0)]. \quad (13-1)$$

(We have treated this as an approximation because we are not dealing formally with the higher order term in the Taylor series. We will make this explicit in the treatment of the GMM estimator in Section 13.4.) The argument needed to characterize the large sample behavior of the estimator,  $\hat{\boldsymbol{\theta}}$ , is discussed in Appendix D. We have from Theorem D.18 (the central limit theorem) that  $\sqrt{n}\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0)$  has a limiting normal distribution with mean vector  $\mathbf{0}$  and covariance matrix equal to  $\Phi$ . Assuming that the functions in the moment equation are continuous and functionally independent, we can expect  $\bar{\mathbf{G}}_n(\boldsymbol{\theta}_0)$  to converge to a nonsingular matrix of constants,  $\Gamma(\boldsymbol{\theta}_0)$ . Under general conditions, the limiting distribution of the right-hand side of (13-1) will be that of a linear function of a normally distributed vector. Jumping to the conclusion, we expect the asymptotic distribution of  $\hat{\boldsymbol{\theta}}$  to be normal with mean vector  $\boldsymbol{\theta}_0$  and covariance matrix  $(1/n) \times \{-[\Gamma(\boldsymbol{\theta}_0)]^{-1}\} \Phi \{-[\Gamma'(\boldsymbol{\theta}_0)]^{-1}\}$ . Thus, the asymptotic covariance matrix for the method of moments estimator may be estimated with

$$\text{Est. Asy. Var}[\hat{\boldsymbol{\theta}}] = \frac{1}{n}[\bar{\mathbf{G}}'_n(\hat{\boldsymbol{\theta}})\mathbf{F}^{-1}\bar{\mathbf{G}}_n(\hat{\boldsymbol{\theta}})]^{-1}.$$

### Example 13.5 (Continued)

Using the estimates  $\hat{\boldsymbol{\theta}}(m'_1, m'_*) = (2.4106, 0.0770702)$ ,

$$\hat{\bar{\mathbf{G}}} = \begin{bmatrix} -1/\hat{\lambda} & \hat{P}/\hat{\lambda}^2 \\ -\hat{\Psi}' & 1/\hat{\lambda} \end{bmatrix} = \begin{bmatrix} -12.97515 & 405.8353 \\ -0.51241 & 12.97515 \end{bmatrix}.$$

[The function  $\Psi'$  is  $d^2 \ln \Gamma(P)/dP^2 = (\Gamma \Gamma'' - \Gamma'^2)/\Gamma^2$ . With  $\hat{P} = 2.4106$ ,  $\hat{\Gamma} = 1.250832$ ,  $\hat{\Psi} = 0.658347$ , and  $\hat{\Psi}' = 0.512408$ .]<sup>2</sup> The matrix  $\mathbf{F}$  is the sample covariance matrix of  $y$

<sup>2</sup> $\Psi'$  is the trigamma function. Values for  $\Gamma(P)$ ,  $\Psi(P)$ , and  $\Psi'(P)$  are tabulated in Abramovitz and Stegun (1971). The values given were obtained using the IMSL computer program library.

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 463

and  $\ln y$  (using 19 as the divisor),

$$\mathbf{F} = \begin{bmatrix} 500.68 & 14.31 \\ 14.31 & 0.47746 \end{bmatrix}.$$

The product is

$$\frac{1}{n} [\hat{\mathbf{G}}' \mathbf{F}^{-1} \hat{\mathbf{G}}]^{-1} = \begin{bmatrix} 0.38978 & 0.014605 \\ 0.014605 & 0.00068747 \end{bmatrix}.$$

For the maximum likelihood estimator, the estimate of the asymptotic covariance matrix based on the expected (and actual) Hessian is

$$[-\mathbf{H}]^{-1} = \frac{1}{n} \begin{bmatrix} \Psi' & -1/\lambda \\ -1/\lambda & P/\lambda^2 \end{bmatrix}^{-1} = \begin{bmatrix} 0.51243 & 0.01638 \\ 0.01638 & 0.00064654 \end{bmatrix}.$$

The Hessian has the same elements as  $\mathbf{G}$  because we chose to use the sufficient statistics for the moment estimators, so the moment equations that we differentiated are, apart from a sign change, also the derivatives of the log-likelihood. The estimates of the two variances are 0.51203 and 0.00064654, respectively, which agrees reasonably well with the method of moments estimates. The difference would be due to sampling variability in a finite sample and the presence of  $\mathbf{F}$  in the first variance estimator.

### 13.2.3 SUMMARY—THE METHOD OF MOMENTS

In the simplest cases, the method of moments is robust to differences in the specification of the data generating process (DGP). A sample mean or variance estimates its population counterpart (assuming it exists), regardless of the underlying process. It is this freedom from unnecessary distributional assumptions that has made this method so popular in recent years. However, this comes at a cost. If more is known about the DGP, its specific distribution for example, then the method of moments may not make use of all of the available information. Thus, in Example 13.3, the natural estimators of the parameters of the distribution based on the sample mean and variance turn out to be inefficient. The method of maximum likelihood, which remains the foundation of much work in econometrics, is an alternative approach which utilizes this out of sample information and is, therefore, more efficient.

## 13.3 MINIMUM DISTANCE ESTIMATION

The preceding analysis has considered **exactly identified cases**. In each example, there were  $K$  parameters to estimate and we used  $K$  moments to estimate them. In Example 13.5, we examined the gamma distribution, a two-parameter family, and considered different pairs of moments that could be used to estimate the two parameters. (The most efficient estimator for the parameters of this distribution will be based on  $(1/n)\sum_i y_i$  and  $(1/n)\sum_i \ln y_i$ . This does raise a general question: How should we proceed if we have more moments than we need? It would seem counterproductive to simply discard the additional information. In this case, logically, the sample information provides more than one estimate of the model parameters, and it is now necessary to reconcile those competing estimators.

We have encountered this situation in several earlier examples: In Example 11.20, in Passmore's (2005) study of Fannie Mae, we have four independent estimators of a single

#### 464 PART III ♦ Estimation Methodology

parameter,  $\hat{\alpha}_j$ , with estimated asymptotic variance  $\hat{V}_j$ ,  $j = 1, \dots, 4$ . The estimators were combined using a **criterion function**:

$$\text{minimize with respect to } \alpha : q = \sum_{j=1}^4 \frac{(\hat{\alpha}_j - \alpha)^2}{\hat{V}_j}.$$

The solution to this minimization problem is

$$\hat{\alpha}_{\text{MDE}} = \sum_{j=1}^4 w_j \hat{\alpha}_j, \quad w_j = \frac{1/\hat{V}_j}{\sum_{s=1}^4 (1/\hat{V}_s)}, \quad j = 1, \dots, 4 \text{ and } \sum_{j=1}^4 w_j = 1.$$

In forming the two-stage least squares estimator of the parameters in a dynamic panel data model in Section 11.7.3, we obtained  $T - 2$  instrumental variable estimators of the parameter vector  $\theta$  by forming different instruments for each period for which we had sufficient data. The  $T - 2$  estimators of the same parameter vector are  $\hat{\theta}_{\text{IV}(t)}$ . The Arellano–Bond estimator of the single parameter vector in this setting is

$$\begin{aligned}\hat{\theta}_{\text{IV}} &= \left( \sum_{t=3}^T \mathbf{W}_{(t)} \right)^{-1} \left( \sum_{t=3}^T \mathbf{W}_{(t)} \hat{\theta}_{\text{IV}(t)} \right) \\ &= \sum_{t=3}^T \mathbf{R}_{(t)} \hat{\theta}_{\text{IV}(t)},\end{aligned}$$

where

$$\mathbf{W}_{(t)} = \left( \hat{\mathbf{X}}'_{(t)} \hat{\mathbf{X}}_{(t)} \right)$$

and

$$\mathbf{R}_{(t)} = \left( \sum_{t=3}^T \mathbf{W}_{(t)} \right)^{-1} \mathbf{W}_{(t)} \text{ and } \sum_{t=3}^T \mathbf{R}_{(t)} = \mathbf{I}.$$

Finally, Carey's (1997) analysis of hospital costs that we examined in Example 11.16 involved a seemingly unrelated regressions model that produced multiple estimates of several of the model parameters. We will revisit this application in Example 13.6.

**A minimum distance estimator (MDE)** is defined as follows: Let  $\bar{m}_{n,l}$  denote a sample statistic based on  $n$  observations such that

$$\text{plim } \bar{m}_{n,l} = g_l(\theta_0), \quad l = 1, \dots, L,$$

where  $\theta_0$  is a vector of  $K \leq L$  parameters to be estimated. Arrange these moments and functions in  $L \times 1$  vectors  $\bar{\mathbf{m}}_n$  and  $\mathbf{g}(\theta_0)$  and further assume that the statistics are jointly asymptotically normally distributed with  $\text{plim } \bar{\mathbf{m}}_n = \mathbf{g}(\theta)$  and  $\text{Asy. Var}[\bar{\mathbf{m}}_n] = (1/n)\Phi$ . Define the criterion function

$$q = [\bar{\mathbf{m}}_n - \mathbf{g}(\theta)]' \mathbf{W} [\bar{\mathbf{m}}_n - \mathbf{g}(\theta)]$$

for a positive definite **weighting matrix**,  $\mathbf{W}$ . The minimum distance estimator is the  $\hat{\theta}_{\text{MDE}}$  that minimizes  $q$ . Different choices of  $\mathbf{W}$  will produce different estimators, but the estimator has the following properties for any  $\mathbf{W}$ :

**THEOREM 13.1 Asymptotic Distribution of the Minimum Distance Estimator**

Under the assumption that  $\sqrt{n}[\bar{\mathbf{m}}_n - \mathbf{g}(\boldsymbol{\theta}_0)] \xrightarrow{d} N[\mathbf{0}, \Phi]$ , the asymptotic properties of the minimum distance estimator are as follows:

$$\text{plim } \hat{\boldsymbol{\theta}}_{\text{MDE}} = \boldsymbol{\theta}_0,$$

$$\begin{aligned}\text{Asy. Var} [\hat{\boldsymbol{\theta}}_{\text{MDE}}] &= \frac{1}{n} [\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)' \mathbf{W} (\boldsymbol{\Gamma} \boldsymbol{\theta}_0)]^{-1} [\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)' \mathbf{W} \boldsymbol{\Phi} \mathbf{W} \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)] [\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)' \mathbf{W} \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)]^{-1} \\ &= \frac{1}{n} \mathbf{V},\end{aligned}$$

where

$$\boldsymbol{\Gamma}(\boldsymbol{\theta}_0) = \text{plim } \mathbf{G}(\hat{\boldsymbol{\theta}}_{\text{MDE}}) = \text{plim} \frac{\partial \mathbf{g}(\hat{\boldsymbol{\theta}}_{\text{MDE}})}{\partial \hat{\boldsymbol{\theta}}'_{\text{MDE}}},$$

and

$$\hat{\boldsymbol{\theta}}_{\text{MDE}} \xrightarrow{a} N \left[ \boldsymbol{\theta}_0, \frac{1}{n} \mathbf{V} \right].$$

Proofs may be found in Malinvaud (1970) and Amemiya (1985). For our purposes, we can note that the MDE is an extension of the method of moments presented in the preceding section. One implication is that the estimator is consistent for any  $\mathbf{W}$ , but the asymptotic covariance matrix is a function of  $\mathbf{W}$ . This suggests that the choice of  $\mathbf{W}$  might be made with an eye toward the size of the covariance matrix and that there might be an optimal choice. That does indeed turn out to be the case. For minimum distance estimation, the weighting matrix that produces the smallest variance is

$$\begin{aligned}\text{optimal weighting matrix: } \mathbf{W}^* &= [\text{Asy. Var.} \sqrt{n} \{\bar{\mathbf{m}}_n - \mathbf{g}(\boldsymbol{\theta})\}]^{-1} \\ &= \boldsymbol{\Phi}^{-1}.\end{aligned}$$

[See Hansen (1982) for discussion.] With this choice of  $\mathbf{W}$ ,

$$\text{Asy. Var} [\hat{\boldsymbol{\theta}}_{\text{MDE}}] = \frac{1}{n} [\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)' \boldsymbol{\Phi}^{-1} \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)]^{-1},$$

which is the result we had earlier for the method of moments estimator.

The solution to the MDE estimation problem is found by locating the  $\hat{\boldsymbol{\theta}}_{\text{MDE}}$  such that

$$\frac{\partial q}{\partial \hat{\boldsymbol{\theta}}_{\text{MDE}}} = -\mathbf{G}(\hat{\boldsymbol{\theta}}_{\text{MDE}})' \mathbf{W} [\bar{\mathbf{m}}_n - \mathbf{g}(\hat{\boldsymbol{\theta}}_{\text{MDE}})] = \mathbf{0}.$$

An important aspect of the MDE arises in the exactly identified case. If  $K$  equals  $L$ , and if the functions  $g_l(\boldsymbol{\theta})$  are functionally independent, that is,  $\mathbf{G}(\boldsymbol{\theta})$  has full row rank,  $K$ , then it is possible to solve the moment equations exactly. That is, the minimization problem becomes one of simply solving the  $K$  moment equations,  $\bar{m}_{n,l} = g_l(\boldsymbol{\theta}_0)$  in the  $K$  unknowns,  $\hat{\boldsymbol{\theta}}_{\text{MDE}}$ . This is the method of moments estimator examined in the preceding

## 466 PART III ♦ Estimation Methodology

section. In this instance, the weighting matrix,  $\mathbf{W}$ , is irrelevant to the solution, because the MDE will now satisfy the moment equations

$$[\bar{\mathbf{m}}_n - \mathbf{g}(\hat{\theta}_{\text{MDE}})] = \mathbf{0}.$$

For the examples listed earlier, which are all for **overidentified cases**, the minimum distance estimators are defined by

$$q = ((\hat{\alpha}_1 - \alpha) \ (\hat{\alpha}_2 - \alpha) \ (\hat{\alpha}_3 - \alpha) \ (\hat{\alpha}_4 - \alpha)) \begin{bmatrix} \hat{V}_1 & 0 & 0 & 0 \\ 0 & \hat{V}_2 & 0 & 0 \\ 0 & 0 & \hat{V}_3 & 0 \\ 0 & 0 & 0 & \hat{V}_4 \end{bmatrix}^{-1} \begin{pmatrix} (\hat{\alpha}_1 - \alpha) \\ (\hat{\alpha}_2 - \alpha) \\ (\hat{\alpha}_3 - \alpha) \\ (\hat{\alpha}_4 - \alpha) \end{pmatrix}$$

for Passmore's analysis of Fannie Mae, and

$$q = ((\mathbf{b}_{\text{IV}(3)} - \boldsymbol{\theta}) \ \dots \ (\mathbf{b}_{\text{IV}(T)} - \boldsymbol{\theta}))' \begin{bmatrix} (\hat{\mathbf{X}}_{(3)}' \hat{\mathbf{X}}_{(3)}) & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & (\hat{\mathbf{X}}_{(T)}' \hat{\mathbf{X}}_{(T)}) \end{bmatrix}^{-1} \begin{pmatrix} (\mathbf{b}_{\text{IV}(3)} - \boldsymbol{\theta}) \\ \vdots \\ (\mathbf{b}_{\text{IV}(T)} - \boldsymbol{\theta}) \end{pmatrix}$$

for the Arellano–Bond estimator of the dynamic panel data model.

### Example 13.6 Minimum Distance Estimation of a Hospital Cost Function

In Carey's (1997) study of hospital costs in Example 11.16, Chamberlain's (1984) seemingly unrelated regressions approach to a panel data model produces five period-specific estimates of a parameter vector,  $\boldsymbol{\theta}_t$ . Some of the parameters are specific to the year while others (it is hypothesized) are common to all five years. There are two specific parameters of interest,  $\beta_D$  and  $\beta_O$ , that are allowed to vary by year, but are each estimated multiple times by the SUR model. We focus on just these parameters. The model states

$$y_{it} = \alpha_i + A_{it} + \beta_{D,t} DIS_{it} + \beta_{O,t} OUT_{it} + \varepsilon_{it},$$

where

$$\alpha_i = B_i + \sum_t \gamma_{D,t} DIS_{it} + \sum_t \gamma_{O,t} OUT_{it} + u_i, t = 1987, \dots, 1991,$$

$DIS_{it}$  is patient discharges, and  $OUT_{it}$  is outpatient visits. (We are changing Carey's notation slightly and suppressing parts of the model that are extraneous to the development here. The terms  $A_{it}$  and  $B_i$  contain those additional components.) The preceding model is estimated by inserting the expression for  $\alpha_i$  in the main equation, then fitting an unrestricted seemingly unrelated regressions model by FGLS. There are five years of data, hence five sets of estimates. Note, however, with respect to the discharge variable,  $DIS$ , although each equation provides separate estimates of  $(\gamma_{D,1}, \dots, (\beta_{D,t} + \gamma_{D,t}), \dots, \gamma_{D,5})$ , a total of five parameter estimates in each each equation (year), there are only 10, not 25 parameters to be estimated in total. The parameters on  $OUT_{it}$  are likewise overidentified. Table 13.1 reproduces the estimates in Table 10.2 for the discharge coefficients and adds the estimates for the outpatient variable.

Looking at the tables we see that the SUR model provides four direct estimates of  $\gamma_{D,87}$ , based on the 1988–1991 equations. It also implicitly provides four estimates of  $\beta_{D,87}$  since any of the four estimates of  $\gamma_{D,87}$  from the last four equations can be subtracted from the coefficient on  $DIS$  in the 1987 equation to estimate  $\beta_{D,87}$ . There are 50 parameter estimates of different functions of the 20 underlying parameters

$$\boldsymbol{\theta} = (\beta_{D,87}, \dots, \beta_{D,91}), (\gamma_{D,87}, \dots, \gamma_{D,91}), (\beta_{O,87}, \dots, \beta_{O,91}), (\gamma_{O,87}, \dots, \gamma_{O,91}),$$

and, therefore, 30 constraints to impose in finding a common, restricted estimator. An MDE was used to reconcile the competing estimators.

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 467

**TABLE 13.1a** Coefficient Estimates for DIS in SUR Model for Hospital Costs

<i>Equation</i>	<i>Coefficient on Variable in the Equation</i>				
	<i>DIS87</i>	<i>DIS88</i>	<i>DIS89</i>	<i>DIS90</i>	<i>DIS91</i>
<b>SUR87</b>	$\beta_{D,87} + \gamma_{D,87}$ 1.76	$\gamma_{D,88}$ 0.116	$\gamma_{D,89}$ -0.0881	$\gamma_{D,90}$ 0.0570	$\gamma_{D,91}$ -0.0617
<b>SUR88</b>	$\gamma_{D,87}$ 0.254	$\beta_{D,88} + \gamma_{D,88}$ 1.61	$\gamma_{D,89}$ -0.0934	$\gamma_{D,90}$ 0.0610	$\gamma_{D,91}$ -0.0514
<b>SUR89</b>	$\gamma_{D,87}$ 0.217	$\gamma_{D,88}$ 0.0846	$\beta_{D,89} + \gamma_{D,89}$ 1.51	$\gamma_{D,90}$ 0.0454	$\gamma_{D,91}$ -0.0253
<b>SUR90</b>	$\gamma_{D,87}$ 0.179	$\gamma_{D,88}$ 0.0822	$\gamma_{D,89}$ 0.0295	$\beta_{D,90} + \gamma_{D,90}$ 1.57	$\gamma_{D,91}$ 0.0244
<b>SUR91</b>	$\gamma_{D,87}$ 0.153	$\gamma_{D,88}$ 0.0363	$\gamma_{D,89}$ -0.0422	$\gamma_{D,90}$ 0.0813	$\beta_{D,91} + \gamma_{D,91}$ 1.70
<b>MDE</b>	$\beta = 1.50$ $\gamma = 0.219$	$\beta = 1.58$ $\gamma = 0.0666$	$\beta = 1.54$ $\gamma = -0.0539$	$\beta = 1.57$ $\gamma = 0.0690$	$\beta = 1.63$ $\gamma = -0.0213$

**TABLE 13.1b** Coefficient Estimates for OUT in SUR Model for Hospital Costs

<i>Equation</i>	<i>Coefficient on Variable in the Equation</i>				
	<i>OUT87</i>	<i>OUT88</i>	<i>OUT89</i>	<i>OUT90</i>	<i>OUT91</i>
<b>SUR87</b>	$\beta_{O,87} + \gamma_{D,87}$ 0.0139	$\gamma_{O,88}$ 0.00292	$\gamma_{O,89}$ 0.00157	$\gamma_{O,90}$ 0.000951	$\gamma_{O,91}$ 0.000678
<b>SUR88</b>	$\gamma_{O,87}$ 0.00347	$\beta_{O,88} + \gamma_{O,88}$ 0.0125	$\gamma_{O,89}$ 0.00501	$\gamma_{O,90}$ 0.00550	$\gamma_{O,91}$ 0.00503
<b>SUR89</b>	$\gamma_{O,87}$ 0.00118	$\gamma_{O,88}$ 0.00159	$\beta_{O,89} + \gamma_{O,89}$ 0.00832	$\gamma_{O,90}$ -0.00220	$\gamma_{O,91}$ -0.00156
<b>SUR90</b>	$\gamma_{O,87}$ -0.00226	$\gamma_{O,88}$ -0.00155	$\gamma_{O,89}$ 0.000401	$\beta_{O,90} + \gamma_{O,90}$ 0.00897	$\gamma_{O,91}$ 0.000450
<b>SUR91</b>	$\gamma_{O,87}$ 0.00278	$\gamma_{O,88}$ 0.00255	$\gamma_{O,89}$ 0.00233	$\gamma_{O,90}$ 0.00305	$\beta_{O,91} + \gamma_{O,91}$ 0.0105
<b>MDE</b>	$\beta = 0.0112$ $\gamma = 0.00177$	$\beta = 0.00999$ $\gamma = 0.00408$	$\beta = 0.0100$ $\gamma = -0.00011$	$\beta = 0.00915$ $\gamma = -0.00073$	$\beta = 0.00793$ $\gamma = 0.00267$

Let  $\hat{\beta}_t$  denote the  $10 \times 1$  period-specific estimator of the model parameters. Unlike the other cases we have examined, the individual estimates here are not uncorrelated. In the SUR model, the estimated asymptotic covariance matrix is the partitioned matrix given in (10-7). For the estimators of two equations,

$$\text{Est. Asy. Cov} [\hat{\beta}_t, \hat{\beta}_s] = \text{the } t, s \text{ block of } \begin{bmatrix} \hat{\sigma}^{11} \mathbf{X}'_1 \mathbf{X}_1 & \hat{\sigma}^{12} \mathbf{X}'_1 \mathbf{X}_2 & \dots & \hat{\sigma}^{15} \mathbf{X}'_1 \mathbf{X}_5 \\ \hat{\sigma}^{21} \mathbf{X}'_2 \mathbf{X}_1 & \hat{\sigma}^{22} \mathbf{X}'_2 \mathbf{X}_2 & \dots & \hat{\sigma}^{25} \mathbf{X}'_2 \mathbf{X}_5 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}^{51} \mathbf{X}'_5 \mathbf{X}_1 & \hat{\sigma}^{52} \mathbf{X}'_5 \mathbf{X}_2 & \dots & \hat{\sigma}^{55} \mathbf{X}'_5 \mathbf{X}_5 \end{bmatrix}^{-1} = \hat{\mathbf{V}}_{ts}$$

where  $\hat{\sigma}^{ts}$  is the  $t,s$  element of  $\hat{\Sigma}^{-1}$ . (We are extracting a submatrix of the relevant matrices here since Carey's SUR model contained 26 other variables in each equation in addition to

## 468 PART III ♦ Estimation Methodology

the five periods of DIS and OUT). The  $50 \times 50$  weighting matrix for the MDE is

$$\mathbf{W} = \begin{bmatrix} \hat{\mathbf{V}}_{87,87} & \hat{\mathbf{V}}_{87,88} & \hat{\mathbf{V}}_{87,89} & \hat{\mathbf{V}}_{87,90} & \hat{\mathbf{V}}_{87,91} \\ \hat{\mathbf{V}}_{88,87} & \hat{\mathbf{V}}_{88,88} & \hat{\mathbf{V}}_{88,89} & \hat{\mathbf{V}}_{88,90} & \hat{\mathbf{V}}_{88,91} \\ \hat{\mathbf{V}}_{89,87} & \hat{\mathbf{V}}_{89,88} & \hat{\mathbf{V}}_{89,89} & \hat{\mathbf{V}}_{89,90} & \hat{\mathbf{V}}_{89,91} \\ \hat{\mathbf{V}}_{90,87} & \hat{\mathbf{V}}_{90,88} & \hat{\mathbf{V}}_{90,89} & \hat{\mathbf{V}}_{90,90} & \hat{\mathbf{V}}_{90,91} \\ \hat{\mathbf{V}}_{91,87} & \hat{\mathbf{V}}_{91,88} & \hat{\mathbf{V}}_{91,89} & \hat{\mathbf{V}}_{91,90} & \hat{\mathbf{V}}_{91,91} \end{bmatrix}^{-1} = [\hat{\mathbf{V}}^{ts}] .$$

The vector of the quadratic form is a stack of five  $10 \times 1$  vectors; the first is

$$\bar{\mathbf{m}}_{n,87} - \mathbf{g}_{87}(\theta) = \left[ \begin{array}{l} \{\hat{\beta}_{D,87}^{87} - (\beta_{D,87} + \gamma_{D,87})\}, \{\hat{\beta}_{D,88}^{87} - \gamma_{D,88}\}, \{\hat{\beta}_{D,89}^{87} - \gamma_{D,89}\}, \{\hat{\beta}_{D,90}^{87} - \gamma_{D,90}\}, \{\hat{\beta}_{D,91}^{87} - \gamma_{D,91}\}, \\ \{\hat{\beta}_{O,87}^{87} - (\beta_{O,87} + \gamma_{O,87})\}, \{\hat{\beta}_{O,88}^{87} - \gamma_{O,88}\}, \{\hat{\beta}_{O,89}^{87} - \gamma_{O,89}\}, \{\hat{\beta}_{O,90}^{87} - \gamma_{O,90}\}, \{\hat{\beta}_{O,91}^{87} - \gamma_{O,91}\} \end{array} \right]^t$$

for the 1987 equation and likewise for the other four equations. The MDE criterion function for this model is

$$q = \sum_{t=1987}^{1991} \sum_{s=1997}^{1981} [\bar{\mathbf{m}}_t - \mathbf{g}_t(\theta)]' \hat{\mathbf{V}}^{ts} [\bar{\mathbf{m}}_s - \mathbf{g}_s(\theta)].$$

Note, there are 50 estimated parameters from the SUR equations (those are listed in Table 13.1) and 20 unknown parameters to be calibrated in the criterion function. The reported minimum distance estimates are shown in the last row of each table.

### 13.4 THE GENERALIZED METHOD OF MOMENTS (GMM) ESTIMATOR

A large proportion of the recent empirical work in econometrics, particularly in macroeconomics and finance, has employed GMM estimators. As we shall see, this broad class of estimators, in fact, includes most of the estimators discussed elsewhere in this book.

The GMM estimation technique is an extension of the minimum distance technique described in Section 13.3.<sup>3</sup> In the following, we will extend the generalized method of moments to other models beyond the generalized linear regression, and we will fill in some gaps in the derivation in Section 13.2.

#### 13.4.1 ESTIMATION BASED ON ORTHOGONALITY CONDITIONS

Consider the least squares estimator of the parameters in the classical linear regression model. An important assumption of the model is

$$E[\mathbf{x}_i \varepsilon_i] = E[\mathbf{x}_i (y_i - \mathbf{x}'_i \boldsymbol{\beta})] = \mathbf{0}.$$

<sup>3</sup>Formal presentation of the results required for this analysis are given by Hansen (1982); Hansen and Singleton (1988); Chamberlain (1987); Cumby, Huizinga, and Obstfeld (1983); Newey (1984, 1985a, 1985b); Davidson and MacKinnon (1993); and Newey and McFadden (1994). Useful summaries of GMM estimation and other developments in econometrics are provided by Pagan and Wickens (1989) and Matyas (1999). An application of some of these techniques that contains useful summaries is Pagan and Vella (1989). Some further discussion can be found in Davidson and MacKinnon (2004). Ruud (2000) provides many of the theoretical details. Hayashi (2000) is another extensive treatment of estimation centered on GMM estimators.

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 469

The sample analog is

$$\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \hat{\varepsilon}_i = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i (y_i - \mathbf{x}'_i \hat{\beta}) = \mathbf{0}.$$

The estimator of  $\beta$  is the one that satisfies these moment equations, which are just the normal equations for the least squares estimator. So, we see that the OLS estimator is a method of moments estimator.

For the instrumental variables estimator of Chapter 8, we relied on a large sample analog to the moment condition,

$$\text{plim}\left(\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \varepsilon_i\right) = \text{plim}\left(\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i (y_i - \mathbf{x}'_i \beta)\right) = \mathbf{0}.$$

We resolved the problem of having more instruments than parameters by solving the equations

$$\left(\frac{1}{n} \mathbf{X}' \mathbf{Z}\right) \left(\frac{1}{n} \mathbf{Z}' \mathbf{Z}\right)^{-1} \left(\frac{1}{n} \mathbf{Z}' \hat{\varepsilon}\right) = \frac{1}{n} \hat{\mathbf{X}}' \hat{\varepsilon} = \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{x}}_i \hat{\varepsilon}_i = \mathbf{0},$$

where the columns of  $\hat{\mathbf{X}}$  are the fitted values in regressions on all the columns of  $\mathbf{Z}$  (that is, the projections of these columns of  $\mathbf{X}$  into the column space of  $\mathbf{Z}$ ). (See Section 8.3.4 for further details.)

The nonlinear least squares estimator was defined similarly, although in this case, the normal equations are more complicated because the estimator is only implicit. The population **orthogonality condition** for the nonlinear regression model is  $E[\mathbf{x}_i^0 \varepsilon_i] = \mathbf{0}$ . The **empirical moment equation** is

$$\frac{1}{n} \sum_{i=1}^n \left( \frac{\partial E[y_i | \mathbf{x}_i, \beta]}{\partial \beta} \right) (y_i - E[y_i | \mathbf{x}_i, \beta]) = \mathbf{0}.$$

Maximum likelihood estimators are obtained by equating the derivatives of a log-likelihood to zero. The scaled log-likelihood function is

$$\frac{1}{n} \ln L = \frac{1}{n} \sum_{i=1}^n \ln f(y_i | \mathbf{x}_i, \theta),$$

where  $f(\cdot)$  is the density function and  $\theta$  is the parameter vector. For densities that satisfy the regularity conditions [see Chapter 16],

$$E\left[\frac{\partial \ln f(y_i | \mathbf{x}_i, \theta)}{\partial \theta}\right] = \mathbf{0}.$$

The maximum likelihood estimator is obtained by equating the sample analog to zero:

$$\frac{1}{n} \frac{\partial \ln L}{\partial \hat{\theta}} = \frac{1}{n} \sum_{i=1}^n \frac{\partial \ln f(y_i | \mathbf{x}_i, \hat{\theta})}{\partial \hat{\theta}} = \mathbf{0}.$$

(Dividing by  $n$  to make this result comparable to our earlier ones does not change the solution.) The upshot is that nearly all the estimators we have discussed and will encounter later can be construed as method of moments estimators. [Manski's (1992) treatment of **analog estimation** provides some interesting extensions and methodological discourse.]

## 470 PART III ♦ Estimation Methodology

As we extend this line of reasoning, it will emerge that most of the estimators defined in this book can be viewed as generalized method of moments estimators.

### 13.4.2 GENERALIZING THE METHOD OF MOMENTS

The preceding examples all have a common aspect. In each case listed, save for the general case of the instrumental variable estimator, there are exactly as many moment equations as there are parameters to be estimated. Thus, each of these are **exactly identified** cases. There will be a single solution to the moment equations, and at that solution, the equations will be exactly satisfied.<sup>4</sup> But there are cases in which there are more moment equations than parameters, so the system is overdetermined.

In Example 13.5, we defined four sample moments,

$$\bar{\mathbf{g}} = \frac{1}{n} \sum_{i=1}^n \left[ y_i, y_i^2, \frac{1}{y_i}, \ln y_i \right]$$

with probability limits  $P/\lambda$ ,  $P(P+1)/\lambda^2$ ,  $\lambda/(P-1)$ , and  $\psi(P) - \ln \lambda$ , respectively. Any pair could be used to estimate the two parameters, but as shown in the earlier example, the six pairs produce six somewhat different estimates of  $\boldsymbol{\theta} = (P, \lambda)$ .

In such a case, to use all the information in the sample it is necessary to devise a way to reconcile the conflicting estimates that may emerge from the overdetermined system. More generally, suppose that the model involves  $K$  parameters,  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_K)'$ , and that the theory provides a set of  $L > K$  moment conditions,

$$E[m_l(y_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta})] = E[m_{il}(\boldsymbol{\theta})] = 0,$$

where  $y_i$ ,  $\mathbf{x}_i$ , and  $\mathbf{z}_i$  are variables that appear in the model and the subscript  $i$  on  $m_{il}(\boldsymbol{\theta})$  indicates the dependence on  $(y_i, \mathbf{x}_i, \mathbf{z}_i)$ . Denote the corresponding sample means as

$$\bar{m}_l(\mathbf{y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n m_l(y_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n m_{il}(\boldsymbol{\theta}).$$

Unless the equations are functionally dependent, the system of  $L$  equations in  $K$  unknown parameters,

$$\bar{m}_l(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n m_l(y_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) = 0, \quad l = 1, \dots, L,$$

will not have a unique solution.<sup>5</sup> For convenience, the moment equations are defined implicitly here as opposed to equalities of moments to functions as in Section 13.3. It will be necessary to reconcile the  $\binom{L}{K}$  different sets of estimates that can be produced. One possibility is to minimize a criterion function, such as the sum of squares,<sup>6</sup>

$$q = \sum_{l=1}^L \bar{m}_l^2 = \bar{\mathbf{m}}(\boldsymbol{\theta})' \bar{\mathbf{m}}(\boldsymbol{\theta}). \tag{13-2}$$

<sup>4</sup>That is, of course if there is *any* solution. In the regression model with multicollinearity, there are  $K$  parameters but fewer than  $K$  independent moment equations.

<sup>5</sup>It may if  $L$  is greater than the sample size,  $n$ . We assume that  $L$  is strictly less than  $n$ .

<sup>6</sup>This approach is one that Quandt and Ramsey (1978) suggested for the problem in Example 13.4.

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 471

It can be shown [see, e.g., Hansen (1982)] that under the assumptions we have made so far, specifically that  $\text{plim } \bar{\mathbf{m}}(\boldsymbol{\theta}) = E[\bar{\mathbf{m}}(\boldsymbol{\theta})] = \mathbf{0}$ , the minimizer of  $q$  in (13-2) produces a consistent (albeit, as we shall see, possibly inefficient) estimator of  $\boldsymbol{\theta}$ . We can, in fact, use as the criterion a weighted sum of squares,

$$q = \bar{\mathbf{m}}(\boldsymbol{\theta})' \mathbf{W}_n \bar{\mathbf{m}}(\boldsymbol{\theta}),$$

where  $\mathbf{W}_n$  is *any* positive definite matrix that may depend on the data but is not a function of  $\boldsymbol{\theta}$ , such as  $\mathbf{I}$  in (13-2), to produce a consistent estimator of  $\boldsymbol{\theta}$ .<sup>7</sup> For example, we might use a diagonal matrix of weights if some information were available about the importance (by some measure) of the different moments. We do make the additional assumption that  $\text{plim } \mathbf{W}_n = \mathbf{W}$  a positive definite matrix,  $\mathbf{W}$ .

By the same logic that makes generalized least squares preferable to ordinary least squares, it should be beneficial to use a weighted criterion in which the weights are inversely proportional to the variances of the moments. Let  $\mathbf{W}$  be a diagonal matrix whose diagonal elements are the reciprocals of the variances of the individual moments,

$$w_{ll} = \frac{1}{\text{Asy. Var}[\sqrt{n} \bar{m}_l]} = \frac{1}{\phi_{ll}}.$$

(We have written it in this form to emphasize that the right-hand side involves the variance of a sample mean which is of order  $(1/n)$ .) Then, a **weighted least squares** estimator would minimize

$$q = \bar{\mathbf{m}}(\boldsymbol{\theta})' \Phi^{-1} \bar{\mathbf{m}}(\boldsymbol{\theta}). \quad (13-3)$$

In general, the  $L$  elements of  $\bar{\mathbf{m}}$  are freely correlated. In (13-3), we have used a diagonal  $\mathbf{W}$  that ignores this correlation. To use generalized least squares, we would define the full matrix,

$$\mathbf{W} = \{\text{Asy. Var}[\sqrt{n} \bar{\mathbf{m}}]\}^{-1} = \Phi^{-1}. \quad (13-4)$$

The estimators defined by choosing  $\boldsymbol{\theta}$  to minimize

$$q = \bar{\mathbf{m}}(\boldsymbol{\theta})' \mathbf{W}_n \bar{\mathbf{m}}(\boldsymbol{\theta})$$

are minimum distance estimators as defined in Section 13.3. The general result is that if  $\mathbf{W}_n$  is a positive definite matrix and if

$$\text{plim } \bar{\mathbf{m}}(\boldsymbol{\theta}) = \mathbf{0},$$

then the minimum distance (generalized method of moments, or GMM) estimator of  $\boldsymbol{\theta}$  is consistent.<sup>8</sup> Because the OLS criterion in (13-2) uses  $\mathbf{I}$ , this method produces a consistent estimator, as does the weighted least squares estimator and the full GLS estimator. What remains to be decided is the best  $\mathbf{W}$  to use. Intuition might suggest

<sup>7</sup>In principle, the weighting matrix can be a function of the parameters as well. See Hansen, Heaton, and Yaron (1996) for discussion. Whether this provides any benefit in terms of the asymptotic properties of the estimator seems unlikely. The one payoff the authors do note is that certain estimators become invariant to the sort of normalization that is discussed in Example 14.1. In practical terms, this is likely to be a consideration only in a fairly small class of cases.

<sup>8</sup>In the most general cases, a number of other subtle conditions must be met so as to assert consistency and the other properties we discuss. For our purposes, the conditions given will suffice. Minimum distance estimators are discussed in Malinvaud (1970), Hansen (1982), and Amemiya (1985).

## 472 PART III ♦ Estimation Methodology

(correctly) that the one defined in (13-4) would be optimal, once again based on the logic that motivates generalized least squares. This result is the now-celebrated one of Hansen (1982).

The asymptotic covariance matrix of this **generalized method of moments (GMM) estimator** is

$$\mathbf{V}_{GMM} = \frac{1}{n} [\boldsymbol{\Gamma}' \mathbf{W} \boldsymbol{\Gamma}]^{-1} = \frac{1}{n} [\boldsymbol{\Gamma}' \boldsymbol{\Phi}^{-1} \boldsymbol{\Gamma}]^{-1}, \quad (13-5)$$

where  $\boldsymbol{\Gamma}$  is the matrix of derivatives with  $j$ th row equal to

$$\boldsymbol{\Gamma}^j = \text{plim} \frac{\partial \bar{m}_j(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'},$$

and  $\boldsymbol{\Phi} = \text{Asy. Var}[\sqrt{n} \bar{\mathbf{m}}]$ . Finally, by virtue of the central limit theorem applied to the sample moments and the **Slutsky theorem** applied to this manipulation, we can expect the estimator to be asymptotically normally distributed. We will revisit the asymptotic properties of the estimator in Section 13.4.3.

### **Example 13.7 GMM Estimation of a Nonlinear Regression Model**

In Example 7.6, we examined a nonlinear regression model for income using the German Socioeconomic Panel Data set. The regression model was

$$\text{Income} = h(1, \text{Age}, \text{Education}, \text{Female}, \gamma) + \varepsilon,$$

where  $h(\cdot)$  is an exponential function of the variables. In the example, we used several interaction terms. In this application, we will simplify the conditional mean function somewhat, and use

$$\text{Income} = \exp(\gamma_1 + \gamma_2 \text{Age} + \gamma_3 \text{Education} + \gamma_4 \text{Female}) + \varepsilon,$$

which, for convenience, we will write

$$\begin{aligned} y_i &= \exp(\mathbf{x}'_i \boldsymbol{\gamma}) + \varepsilon_i \\ &= \mu_i + \varepsilon_i. \end{aligned}$$

The sample consists of the 1988 wave of the panel, less two observations for which *Income* equals zero. The resulting sample contains 4,481 observations. Descriptive statistics for the sample data are given in Table 7.2.

We will first consider nonlinear least squares estimation of the parameters. The normal equations for nonlinear least squares will be

$$(1/n) \sum_i [(y_i - \mu_i) \mu_i \mathbf{x}_i] = (1/n) \sum_i [\varepsilon_i \mu_i \mathbf{x}_i] = \mathbf{0}.$$

Note that the orthogonality condition involves the pseudoregressors,  $\partial \mu_i / \partial \boldsymbol{\gamma} = \mathbf{x}_i^0 = \mu_i \mathbf{x}_i$ . The implied population moment equation is

$$E[\varepsilon_i (\mu_i \mathbf{x}_i)] = \mathbf{0}.$$

Computation of the nonlinear least squares estimator is discussed in Section 7.2.6. The estimator of the asymptotic covariance matrix is

$$\text{Est. Asy. Var}[\hat{\boldsymbol{\gamma}}_{NLSQ}] = \frac{\sum_{i=1}^n (y_i - \hat{\mu}_i)^2}{(4,481 - 4)} \left[ \sum_{i=1}^{4,481} (\hat{\mu}_i \mathbf{x}_i) (\hat{\mu}_i \mathbf{x}_i)' \right]^{-1}, \quad \text{where } \hat{\mu}_i = \exp(\mathbf{x}'_i \hat{\boldsymbol{\gamma}}).$$

---

<sup>9</sup>We note that in this model, it is likely that *Education* is endogenous. It would be straightforward to accommodate that in the GMM estimator. However, for purposes of a straightforward numerical example, we will proceed assuming that *Education* is exogenous.

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 473

A simple method of moments estimator might be constructed from the hypothesis that  $\mathbf{x}_i$  (not  $\mathbf{x}_i^0$ ) is orthogonal to  $\varepsilon_i$ . Then,

$$E[\varepsilon_i \mathbf{x}_i] = E \left[ \varepsilon_i \begin{pmatrix} 1 \\ \text{Age}_i \\ \text{Education}_i \\ \text{Female}_i \end{pmatrix} \right] = \mathbf{0}$$

implies four moment equations. The sample counterparts will be

$$\bar{m}_k(\gamma) = \frac{1}{n} \sum_{i=1}^n (\gamma_i - \mu_i) x_{ik} = \frac{1}{n} \sum_{i=1}^n \varepsilon_i x_{ik}.$$

In order to compute the method of moments estimator, we will minimize the sum of squares,

$$\bar{\mathbf{m}}'(\gamma) \bar{\mathbf{m}}(\gamma) = \sum_{k=1}^4 \bar{m}_k^2(\gamma).$$

This is a nonlinear optimization problem that must be solved iteratively using the methods described in Section E.3.

With the first-step estimated parameters,  $\hat{\gamma}^0$ , in hand, the covariance matrix is estimated using (13-5).

$$\begin{aligned} \hat{\Phi} &= \left\{ \frac{1}{4,481} \sum_{i=1}^{4,481} \mathbf{m}_i(\hat{\gamma}^0) \mathbf{m}'_i(\hat{\gamma}^0) \right\} = \left\{ \frac{1}{4,481} \sum_{i=1}^{4,481} (\hat{\varepsilon}_i^0 \mathbf{x}_i) (\hat{\varepsilon}_i^0 \mathbf{x}_i)' \right\} \\ \bar{\mathbf{G}} &= \left\{ \frac{1}{4,481} \sum_{i=1}^n (\hat{\varepsilon}_i^0 \mathbf{x}_i) (-\hat{\mu}_i^0 \mathbf{x}_i)' \right\}. \end{aligned}$$

The asymptotic covariance matrix for the MOM estimator is computed using (13-5),

$$\text{Est. Asy. Var}[\hat{\gamma} \text{ MOM}] = \frac{1}{n} [\bar{\mathbf{G}} \hat{\Phi}^{-1} \bar{\mathbf{G}}']^{-1}.$$

Suppose we have in hand additional variables, *Health Satisfaction* and *Marital Status*, such that although the conditional mean function remains as given previously, we will use them to form a GMM estimator. This provides two additional moment equations,

$$E \left[ \varepsilon_i \begin{pmatrix} \text{Health Satisfaction}_i \\ \text{Marital Status}_i \end{pmatrix} \right]$$

for a total of six moment equations for estimating the four parameters. We construct the generalized method of moments estimator as follows: The initial step is the same as before, except the sum of squared moments,  $\bar{\mathbf{m}}'(\gamma) \bar{\mathbf{m}}(\gamma)$ , is summed over six rather than four terms. We then construct

$$\hat{\Phi} = \left\{ \frac{1}{4,481} \sum_{i=1}^{4,481} \mathbf{m}_i(\hat{\gamma}) \mathbf{m}'_i(\hat{\gamma}) \right\} = \left\{ \frac{1}{4,481} \sum_{i=1}^{4,481} (\hat{\varepsilon}_i \mathbf{z}_i) (\hat{\varepsilon}_i \mathbf{z}_i)' \right\},$$

where now,  $\mathbf{z}_i$  in the second term is a vector of exogenous variables, rather than the original four (including the constant term). Thus,  $\bar{\mathbf{m}}'$  is now a  $6 \times 6$  matrix. The optimal weighting matrix for estimation (developed in the next section) is  $\bar{\mathbf{G}}'$ . The GMM estimator is computed by minimizing with respect to  $\gamma$ ,

$$q = \bar{\mathbf{m}}'(\gamma) \hat{\Phi}^{-1} \bar{\mathbf{m}}(\gamma).$$

The asymptotic covariance matrix is computed using (13-5) as it was for the simple method of moments estimator.

## 474 PART III ♦ Estimation Methodology

**TABLE 13.2** Nonlinear Regression Estimates (Standard Errors in Parentheses)

Estimate	Nonlinear Least Squares	Method of Moments	First Step GMM	GMM
Constant	-1.69331 (0.04408)	-1.62969 (0.04214)	-1.45551 (0.10102)	-1.61192 (0.04163)
Age	0.00207 (0.00061)	0.00178 (0.00057)	-0.00028 (0.00100)	0.00092 (0.00056)
Education	0.04792 (0.00247)	0.04861 (0.00262)	0.03731 (0.00518)	0.04647 (0.00262)
Female	-0.00658 (0.01373)	0.00070 (0.01384)	-0.02205 (0.01445)	-0.01517 (0.01357)

Table 13.2 presents four sets of estimates, nonlinear least squares, method of moments, first-step GMM, and GMM using the optimal weighting matrix. Two comparisons are noted. The method of moments produces slightly different results from the nonlinear least squares estimator. This is to be expected, since they are different criteria. Judging by the standard errors, the GMM estimator seems to provide a very slight improvement over the nonlinear least squares and method of moments estimators. The conclusion, though, would seem to be that the two additional moments (variables) do not provide very much additional information for estimation of the parameters.

### 13.4.3 PROPERTIES OF THE GMM ESTIMATOR

We will now examine the properties of the GMM estimator in some detail. Because the GMM estimator includes other familiar estimators that we have already encountered, including least squares (linear and nonlinear), and instrumental variables, these results will extend to those cases. The discussion given here will only sketch the elements of the formal proofs. The assumptions we make here are somewhat narrower than a fully general treatment might allow, but they are broad enough to include the situations likely to arise in practice. More detailed and rigorous treatments may be found in, for example, Newey and McFadden (1994), White (2001), Hayashi (2000), Mittelhammer et al. (2000), or Davidson (2000).

The GMM estimator is based on the set of population orthogonality conditions,

$$E[\mathbf{m}_i(\boldsymbol{\theta}_0)] = \mathbf{0},$$

where we denote the true parameter vector by  $\boldsymbol{\theta}_0$ . The subscript  $i$  on the term on the left-hand side indicates dependence on the observed data,  $(\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i)$ . Averaging this over the sample observations produces the sample moment equation

$$E[\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0)] = \mathbf{0},$$

where

$$\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0) = \frac{1}{n} \sum_{i=1}^n \mathbf{m}_i(\boldsymbol{\theta}_0).$$

This moment is a set of  $L$  equations involving the  $K$  parameters. We will assume that this expectation exists and that the sample counterpart converges to it. The definitions are cast in terms of the population parameters and are indexed by the sample size. To fix the ideas, consider, once again, the empirical moment equations that define the instrumental variable estimator for a linear or nonlinear regression model.

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 475

**Example 13.8 Empirical Moment Equation for Instrumental Variables**

For the IV estimator in the linear or nonlinear regression model, we assume

$$E[\bar{\mathbf{m}}_n(\boldsymbol{\beta})] = E\left[\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i [y_i - h(\mathbf{x}_i, \boldsymbol{\beta})]\right] = \mathbf{0}.$$

There are  $L$  instrumental variables in  $\mathbf{z}_i$  and  $K$  parameters in  $\boldsymbol{\beta}$ . This statement defines  $L$  moment equations, one for each instrumental variable.

We make the following assumptions about the model and these empirical moments:

**ASSUMPTION 13.1. Convergence of the Empirical Moments:** *The data generating process is assumed to meet the conditions for a law of large numbers to apply, so that we may assume that the empirical moments converge in probability to their expectation. Appendix D lists several different laws of large numbers that increase in generality. What is required for this assumption is that*

$$\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0) = \frac{1}{n} \sum_{i=1}^n \mathbf{m}_i(\boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{0}.$$

The laws of large numbers that we examined in Appendix D accommodate cases of independent observations. Cases of dependent or correlated observations can be gathered under the **Ergodic theorem** (19.1). For this more general case, then, we would assume that the sequence of observations  $\mathbf{m}(\boldsymbol{\theta})$  constitutes a jointly ( $L \times 1$ ) stationary and ergodic process.

The empirical moments are assumed to be continuous and continuously differentiable functions of the parameters. For our earlier example, this would mean that the conditional mean function,  $h(\mathbf{x}_i, \boldsymbol{\beta})$  is a continuous function of  $\boldsymbol{\beta}$  (although not necessarily of  $\mathbf{x}_i$ ). With continuity and differentiability, we will also be able to assume that the derivatives of the moments,

$$\bar{\mathbf{G}}_n(\boldsymbol{\theta}_0) = \frac{\partial \bar{\mathbf{m}}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'_0} = \frac{1}{n} \sum_{i=1}^n \frac{\partial \mathbf{m}_{i,n}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'_0},$$

converge to a probability limit, say,  $\text{plim } \bar{\mathbf{G}}_n(\boldsymbol{\theta}_0) = \bar{\mathbf{G}}(\boldsymbol{\theta}_0)$ . [See (13-1), (13-5), and Theorem 13.1.] For sets of *independent* observations, the continuity of the functions and the derivatives will allow us to invoke the Slutsky theorem to obtain this result. For the more general case of sequences of *dependent* observations, Theorem 19.2, Ergodicity of Functions, will provide a counterpart to the Slutsky theorem for time-series data. In sum, if the moments themselves obey a law of large numbers, then it is reasonable to assume that the derivatives do as well.

**ASSUMPTION 13.2. Identification:** *For any  $n \geq K$ , if  $\boldsymbol{\theta}_1$  and  $\boldsymbol{\theta}_2$  are two different parameter vectors, then there exist data sets such that  $\bar{\mathbf{m}}_n(\boldsymbol{\theta}_1) \neq \bar{\mathbf{m}}_n(\boldsymbol{\theta}_2)$ . Formally, in Section 14.3, identification is defined to imply that the probability limit of the GMM criterion function is uniquely minimized at the true parameters,  $\boldsymbol{\theta}_0$ .*

## 476 PART III ♦ Estimation Methodology

Assumption 13.2 is a practical prescription for identification. More formal conditions are discussed in Section 12.5.3. We have examined two violations of this crucial assumption. In the linear regression model, one of the assumptions is full rank of the matrix of exogenous variables—the absence of multicollinearity in  $\mathbf{X}$ . In our discussion of the maximum likelihood estimator, we will encounter a case (Example 14.1) in which a normalization is needed to identify the vector of parameters. [See Hansen et al. (1996) for discussion of this case.] Both of these cases are included in this assumption. The identification condition has three important implications:

1. **Order condition.** The number of moment conditions is at least as large as the number of parameters;  $L \geq K$ . This is necessary, but not sufficient for identification.
2. **Rank condition.** The  $L \times K$  matrix of derivatives,  $\bar{\mathbf{G}}_n(\boldsymbol{\theta}_0)$  will have row rank equal to  $K$ . (Again, note that the number of rows must equal or exceed the number of columns.)
3. **Uniqueness.** With the continuity assumption, the identification assumption implies that the parameter vector that satisfies the population moment condition is unique. We know that at the true parameter vector,  $\text{plim } \bar{\mathbf{m}}_n(\boldsymbol{\theta}_0) = \mathbf{0}$ . If  $\boldsymbol{\theta}_1$  is any parameter vector that satisfies this condition, then  $\boldsymbol{\theta}_1$  must equal  $\boldsymbol{\theta}_0$ .

Assumptions 13.1 and 13.2 characterize the parameterization of the model. Together they establish that the parameter vector will be estimable. We now make the statistical assumption that will allow us to establish the properties of the GMM estimator.

**ASSUMPTION 13.3. Asymptotic Distribution of Empirical Moments:** *We assume that the empirical moments obey a central limit theorem. This assumes that the moments have a finite asymptotic covariance matrix,  $(1/n)\Phi$ , so that*

$$\sqrt{n} \bar{\mathbf{m}}_n(\boldsymbol{\theta}_0) \xrightarrow{d} N[\mathbf{0}, \Phi].$$

The underlying requirements on the data for this assumption to hold will vary and will be complicated if the observations comprising the empirical moment are not independent. For samples of independent observations, we assume the conditions underlying the Lindeberg–Feller (D.19) or Liapounov central limit theorem (D.20) will suffice. For the more general case, it is once again necessary to make some assumptions about the data. We have assumed that

$$E[\mathbf{m}_i(\boldsymbol{\theta}_0)] = \mathbf{0}.$$

If we can go a step further and assume that the functions  $\mathbf{m}_i(\boldsymbol{\theta}_0)$  are an ergodic, stationary **martingale difference series**,

$$E[\mathbf{m}_i(\boldsymbol{\theta}_0) | \mathbf{m}_{i-1}(\boldsymbol{\theta}_0), \mathbf{m}_{i-2}(\boldsymbol{\theta}_0), \dots] = \mathbf{0},$$

then we can invoke Theorem 20.3, the central limit theorem for the Martingale difference series. It will generally be fairly complicated to verify this assumption for nonlinear models, so it will usually be assumed outright. On the other hand, the assumptions are

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 477

likely to be fairly benign in a typical application. For regression models, the assumption takes the form

$$E[\mathbf{z}_i \varepsilon_i | \mathbf{z}_{i-1} \varepsilon_{i-1}, \dots] = \mathbf{0},$$

which will often be part of the central structure of the model.

With the assumptions in place, we have

**THEOREM 13.2 Asymptotic Distribution of the GMM Estimator**

*Under the preceding assumptions,*

$$\begin{aligned}\hat{\theta}_{GMM} &\xrightarrow{P} \theta_0, \\ \hat{\theta}_{GMM} &\xrightarrow{a} N[\theta_0, \mathbf{V}_{GMM}],\end{aligned}\tag{13-6}$$

where  $\mathbf{V}_{GMM}$  is defined in (13-5).

We will now sketch a proof of Theorem 13.2. The GMM estimator is obtained by minimizing the criterion function

$$q_n(\theta) = \bar{\mathbf{m}}_n(\theta)' \mathbf{W}_n \bar{\mathbf{m}}_n(\theta),$$

where  $\mathbf{W}_n$  is the weighting matrix used. Consistency of the estimator that minimizes this criterion can be established by the same logic that will be used for the maximum likelihood estimator. It must first be established that  $q_n(\theta)$  converges to a value  $q_0(\theta)$ . By our assumptions of strict continuity and Assumption 13.1,  $q_n(\theta_0)$  converges to 0. (We could apply the Slutsky theorem to obtain this result.) We will assume that  $q_n(\theta)$  converges to  $q_0(\theta)$  for other points in the parameter space as well. Because  $\mathbf{W}_n$  is positive definite, for any finite  $n$ , we know that

$$0 \leq q_n(\hat{\theta}_{GMM}) \leq q_n(\theta_0).\tag{13-7}$$

That is, in the finite sample,  $\hat{\theta}_{GMM}$  actually minimizes the function, so the sample value of the criterion is not larger at  $\hat{\theta}_{GMM}$  than at any other value, including the true parameters. But, at the true parameter values,  $q_n(\theta_0) \xrightarrow{P} 0$ . So, if (13-7) is true, then it must follow that  $q_n(\hat{\theta}_{GMM}) \xrightarrow{P} 0$  as well because of the identification assumption, 13.2. As  $n \rightarrow \infty$ ,  $q_n(\hat{\theta}_{GMM})$  and  $q_n(\theta)$  converge to the same limit. It must be the case, then, that as  $n \rightarrow \infty$ ,  $\bar{\mathbf{m}}_n(\hat{\theta}_{GMM}) \rightarrow \bar{\mathbf{m}}_n(\theta_0)$ , because the function is quadratic and  $\mathbf{W}$  is positive definite. The identification condition that we assumed earlier now assures that as  $n \rightarrow \infty$ ,  $\hat{\theta}_{GMM}$  must equal  $\theta_0$ . This establishes consistency of the estimator.

We will now sketch a proof of the asymptotic normality of the estimator: The first-order conditions for the GMM estimator are

$$\frac{\partial q_n(\hat{\theta}_{GMM})}{\partial \hat{\theta}_{GMM}} = 2\bar{\mathbf{G}}_n(\hat{\theta}_{GMM})' \mathbf{W}_n \bar{\mathbf{m}}_n(\hat{\theta}_{GMM}) = \mathbf{0}.\tag{13-8}$$

(The leading 2 is irrelevant to the solution, so it will be dropped at this point.) The orthogonality equations are assumed to be continuous and continuously differentiable. This allows us to employ the **mean value theorem** as we expand the empirical moments

### 478 PART III ♦ Estimation Methodology

in a linear Taylor series around the true value,  $\theta_0$

$$\bar{\mathbf{m}}_n(\hat{\theta}_{GMM}) = \bar{\mathbf{m}}_n(\theta_0) + \bar{\mathbf{G}}_n(\bar{\theta})(\hat{\theta}_{GMM} - \theta_0), \quad (13-9)$$

where  $\bar{\theta}$  is a point between  $\hat{\theta}_{GMM}$  and the true parameters,  $\theta_0$ . Thus, for each element  $\bar{\theta}_k = w_k \hat{\theta}_{k,GMM} + (1 - w_k)\theta_{0,k}$  for some  $w_k$  such that  $0 < w_k < 1$ . Insert (13-9) in (13-8) to obtain

$$\bar{\mathbf{G}}_n(\hat{\theta}_{GMM})' \mathbf{W}_n \bar{\mathbf{m}}_n(\theta_0) + \bar{\mathbf{G}}_n(\hat{\theta}_{GMM})' \mathbf{W}_n \bar{\mathbf{G}}_n(\bar{\theta})(\hat{\theta}_{GMM} - \theta_0) = \mathbf{0}.$$

Solve this equation for the estimation error and multiply by  $\sqrt{n}$ . This produces

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) = -[\bar{\mathbf{G}}_n(\hat{\theta}_{GMM})' \mathbf{W}_n \bar{\mathbf{G}}_n(\bar{\theta})]^{-1} \bar{\mathbf{G}}_n(\hat{\theta}_{GMM})' \mathbf{W}_n \sqrt{n} \bar{\mathbf{m}}_n(\theta_0).$$

Assuming that they have them, the quantities on the left- and right-hand sides have the same limiting distributions. By the consistency of  $\hat{\theta}_{GMM}$ , we know that  $\hat{\theta}_{GMM}$  and  $\bar{\theta}$  both converge to  $\theta_0$ . By the strict continuity assumed, it must also be the case that

$$\bar{\mathbf{G}}_n(\bar{\theta}) \xrightarrow{p} \bar{\mathbf{G}}(\theta_0) \text{ and } \bar{\mathbf{G}}_n(\hat{\theta}_{GMM}) \xrightarrow{p} \bar{\mathbf{G}}(\theta_0).$$

We have also assumed that the weighting matrix,  $\mathbf{W}_n$ , converges to a matrix of constants,  $\mathbf{W}$ . Collecting terms, we find that the limiting distribution of the vector on the left-hand side must be the same as that on the right-hand side in (13-10),

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) \xrightarrow{p} \left\{ [\bar{\mathbf{G}}(\theta_0)' \mathbf{W} \bar{\mathbf{G}}(\theta_0)]^{-1} \bar{\mathbf{G}}(\theta_0)' \mathbf{W} \right\} \sqrt{n} \bar{\mathbf{m}}_n(\theta_0). \quad (13-10)$$

We now invoke Assumption 13.3. The matrix in curled brackets is a set of constants. The last term has the normal limiting distribution given in Assumption 13.3. The mean and variance of this limiting distribution are zero and  $\Phi$ , respectively. Collecting terms, we have the result in Theorem 13.2, where

$$\mathbf{V}_{GMM} = \frac{1}{n} [\bar{\mathbf{G}}(\theta_0)' \mathbf{W} \bar{\mathbf{G}}(\theta_0)]^{-1} \bar{\mathbf{G}}(\theta_0)' \mathbf{W} \Phi \mathbf{W} \bar{\mathbf{G}}(\theta_0) [\bar{\mathbf{G}}(\theta_0)' \mathbf{W} \bar{\mathbf{G}}(\theta_0)]^{-1}. \quad (13-11)$$

The final result is a function of the choice of weighting matrix,  $\mathbf{W}$ . If the optimal weighting matrix,  $\mathbf{W} = \Phi^{-1}$ , is used, then the expression collapses to

$$\mathbf{V}_{GMM,optimal} = \frac{1}{n} [\bar{\mathbf{G}}(\theta_0)' \Phi^{-1} \bar{\mathbf{G}}(\theta_0)]^{-1}. \quad (13-12)$$

Returning to (13-11), there is a special case of interest. If we use least squares or instrumental variables with  $\mathbf{W} = \mathbf{I}$ , then

$$\mathbf{V}_{GMM} = \frac{1}{n} (\bar{\mathbf{G}}' \bar{\mathbf{G}})^{-1} \bar{\mathbf{G}}' \Phi \bar{\mathbf{G}} (\bar{\mathbf{G}}' \bar{\mathbf{G}})^{-1}.$$

This equation prescribes essentially the White or **Newey-West estimator**, which returns us to our departure point and provides a neat symmetry to the GMM principle. We will formalize this in Section 13.6.1.

## 13.5 TESTING HYPOTHESES IN THE GMM FRAMEWORK

The estimation framework developed in the previous section provides the basis for a convenient set of statistics for testing hypotheses. We will consider three groups of tests. The first is a pair of statistics that is used for testing the validity of the restrictions that produce the moment equations. The second is a trio of tests that correspond to the familiar Wald, LM, and LR tests. The third is a class of tests based on the theoretical underpinnings of the conditional moments that we used earlier to devise the GMM estimator.

### 13.5.1 TESTING THE VALIDITY OF THE MOMENT RESTRICTIONS

In the exactly identified cases we examined earlier (least squares, instrumental variables, maximum likelihood), the criterion for GMM estimation 

$$q = \bar{\mathbf{m}}(\boldsymbol{\theta})' \mathbf{W} \bar{\mathbf{m}}(\boldsymbol{\theta})$$


would be exactly zero because we can find a set of estimates for which  $\bar{\mathbf{m}}(\boldsymbol{\theta})$  is exactly zero. Thus in the exactly identified case when there are the same number of moment equations as there are parameters to estimate, the weighting matrix  $\mathbf{W}$  is irrelevant to the solution. But if the parameters are overidentified by the moment equations, then these equations imply substantive restrictions. As such, if the hypothesis of the model that led to the moment equations in the first place is incorrect, at least some of the sample moment restrictions will be systematically violated. This conclusion provides the basis for a test of the **overidentifying restrictions**. By construction, when the optimal weighting matrix is used,

$$nq = [\sqrt{n} \bar{\mathbf{m}}(\hat{\boldsymbol{\theta}})'] \{ \text{Est. Asy. Var}[\sqrt{n} \bar{\mathbf{m}}(\hat{\boldsymbol{\theta}})] \}^{-1} [\sqrt{n} \bar{\mathbf{m}}(\hat{\boldsymbol{\theta}})],$$

so  $nq$  is a Wald statistic. Therefore, under the hypothesis of the model,

$$nq \xrightarrow{d} \chi^2[L - K].$$

(For the exactly identified case, there are zero degrees of freedom and  $q = 0$ .)

#### **Example 13.9 Overidentifying Restrictions**

In Hall's consumption model, two orthogonality conditions noted in Example 13.1 exactly identify the two parameters. But his analysis of the model suggests a way to test the specification. The conclusion, "No information available in time  $t$  apart from the level of consumption,  $c_t$ , helps predict future consumption,  $c_{t+1}$ , in the sense of affecting the expected value of marginal utility. In particular, income or wealth in periods  $t$  or earlier are irrelevant once  $c_t$  is known" suggests how one might test the model. If lagged values of income ( $Y_t$  might equal the ratio of current income to the previous period's income) are added to the set of instruments, then the model is now overidentified by the orthogonality conditions;

$$E_t \left[ (\beta(1 + r_{t+1}) R_{t+1}^\lambda - 1) \times \begin{pmatrix} 1 \\ R_t \\ Y_{t-1} \\ Y_{t-2} \end{pmatrix} \right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

A simple test of the overidentifying restrictions would be suggestive of the validity of the corollary. Rejecting the restrictions casts doubt on the original model. Hall's proposed tests

## 480 PART III ♦ Estimation Methodology

to distinguish the life cycle–permanent income model from other theories of consumption involved adding two lags of income to the information set. Hansen and Singleton (1982) operated directly on this form of the model. Other studies, for example, Campbell and Mankiw's (1989) as well as Hall's, used the model's implications to formulate more conventional instrumental variable regression models.

The preceding is a **specification test**, not a test of parametric restrictions. However, there is a symmetry between the moment restrictions and restrictions on the parameter vector. Suppose  $\theta$  is subjected to  $J$  restrictions (linear or nonlinear) which restrict the number of free parameters from  $K$  to  $K - J$ . (That is, reduce the dimensionality of the parameter space from  $K$  to  $K - J$ .) The nature of the GMM estimation problem we have posed is not changed at all by the restrictions. The constrained problem may be stated in terms of

$$q_R = \bar{\mathbf{m}}(\theta_R)' \mathbf{W} \bar{\mathbf{m}}(\theta_R).$$

Note that the weighting matrix,  $\mathbf{W}$ , is unchanged. The precise nature of the solution method may be changed—the restrictions mandate a constrained optimization. However, the criterion is essentially unchanged. It follows then that

$$nq_R \xrightarrow{d} \chi^2[L - (K - J)].$$

This result suggests a method of testing the restrictions, although the distribution theory is not obvious. The weighted sum of squares with the restrictions imposed,  $nq_R$ , must be larger than the weighted sum of squares obtained without the restrictions,  $nq$ . The difference is

$$(nq_R - nq) \xrightarrow{d} \chi^2[J]. \quad (13-13)$$

The test is attributed to Newey and West (1987b). This provides one method of testing a set of restrictions. (The small-sample properties of this test will be the central focus of the application discussed in Section 13.6.5.) We now consider several alternatives.

### 13.5.2 GMM COUNTERPARTS TO THE WALD, LM, AND LR TESTS

Section 14.6 describes a trio of testing procedures that can be applied to a hypothesis in the context of maximum likelihood estimation. To reiterate, let the hypothesis to be tested be a set of  $J$  possibly nonlinear restrictions on  $K$  parameters  $\theta$  in the form  $H_0: \mathbf{r}(\theta) = \mathbf{0}$ . Let  $\mathbf{c}_1$  be the maximum likelihood estimates of  $\theta$  estimated without the restrictions, and let  $\mathbf{c}_0$  denote the restricted maximum likelihood estimates, that is, the estimates obtained while imposing the null hypothesis. The three statistics, which are asymptotically equivalent, are obtained as follows:

$$\text{LR} = \text{likelihood ratio} = -2(\ln L_0 - \ln L_1),$$

where

$$\ln L_j = \log \text{ likelihood function evaluated at } \mathbf{c}_j, \quad j = 0, 1.$$

The **likelihood ratio statistic** requires that both estimates be computed. The Wald statistic is

$$W = \text{Wald} = [\mathbf{r}(\mathbf{c}_1)]' \{\text{Est. Asy. Var}[\mathbf{r}(\mathbf{c}_1)]\}^{-1} [\mathbf{r}(\mathbf{c}_1)]. \quad (13-14)$$

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 481

The **Wald statistic** is the distance measure for the degree to which the unrestricted estimator fails to satisfy the restrictions. The usual estimator for the asymptotic covariance matrix would be

$$\text{Est. Asy. Var}[\mathbf{r}(\mathbf{c}_1)] = \mathbf{R}_1 \{ \text{Est. Asy. Var}[\mathbf{c}_1] \} \mathbf{R}'_1, \quad (13-15)$$

where

$$\mathbf{R}_1 = \partial \mathbf{r}(\mathbf{c}_1) / \partial \mathbf{c}'_1 \quad (\mathbf{R}_1 \text{ is a } J \times K \text{ matrix}).$$

The Wald statistic can be computed using only the unrestricted estimate. The LM statistic is

$$\text{LM} = \text{Lagrange multiplier} = \mathbf{g}'_1(\mathbf{c}_0) \{ \text{Est. Asy. Var}[\mathbf{g}_1(\mathbf{c}_0)] \}^{-1} \mathbf{g}_1(\mathbf{c}_0), \quad (13-16)$$

where

$$\mathbf{g}_1(\mathbf{c}_0) = \partial \ln L_1(\mathbf{c}_0) / \partial \mathbf{c}_0,$$

that is, the first derivatives of the *unconstrained* log-likelihood computed at the *restricted* estimates. The term  $\text{Est. Asy. Var}[\mathbf{g}_1(\mathbf{c}_0)]$  is the inverse of any of the usual estimators of the asymptotic covariance matrix of the maximum likelihood estimators of the parameters, computed using the restricted estimates. The most convenient choice is usually the BHHH estimator. The LM statistic is based on the restricted estimates.

Newey and West (1987b) have devised counterparts to these test statistics for the GMM estimator. The Wald statistic is computed identically, using the results of GMM estimation rather than maximum likelihood.<sup>10</sup> That is, in (13-14), we would use the unrestricted GMM estimator of  $\boldsymbol{\theta}$ . The appropriate asymptotic covariance matrix is (13-12). The computation is exactly the same. The counterpart to the LR statistic is the difference in the values of  $nq$  in (13-13). It is necessary to use the same weighting matrix,  $\mathbf{W}$  in both restricted and unrestricted estimators. Because the unrestricted estimator is consistent under both  $H_0$  and  $H_1$ , a consistent, unrestricted estimator of  $\boldsymbol{\theta}$  is used to compute  $\mathbf{W}$ . Label this  $\hat{\Phi}_1^{-1} = \{ \text{Asy. Var}[\sqrt{n} \bar{\mathbf{m}}_1(\mathbf{c}_1)] \}^{-1}$ . In each occurrence, the subscript 1 indicates reference to the unrestricted estimator. Then  $q$  is minimized without restrictions to obtain  $q_1$  and then subject to the restrictions to obtain  $q_0$ . The statistic is then  $(nq_0 - nq_1)$ .<sup>11</sup> Because we are using the same  $\mathbf{W}$  in both cases, this statistic is necessarily nonnegative. (This is the statistic discussed in Section 13.5.1.)

Finally, the counterpart to the LM statistic would be

$$\text{LM}_{GMM} = n [\bar{\mathbf{m}}_1(\mathbf{c}_0)' \hat{\Phi}_1^{-1} \bar{\mathbf{G}}_1(\mathbf{c}_0)] [\bar{\mathbf{G}}_1(\mathbf{c}_0)' \hat{\Phi}_1^{-1} \bar{\mathbf{G}}_1(\mathbf{c}_0)]^{-1} [\bar{\mathbf{G}}_1(\mathbf{c}_0)' \hat{\Phi}_1^{-1} \bar{\mathbf{m}}_1(\mathbf{c}_0)].$$

The logic for this LM statistic is the same as that for the MLE. The derivatives of the minimized criterion  $q$  in (13-3) evaluated at the restricted estimator are

$$\mathbf{g}_1(\mathbf{c}_0) = \frac{\partial q}{\partial \mathbf{c}_0} = 2 \bar{\mathbf{G}}_1(\mathbf{c}_0)' \hat{\Phi}_1^{-1} \bar{\mathbf{m}}_1(\mathbf{c}_0).$$

<sup>10</sup>See Burnside and Eichenbaum (1996) for some small-sample results on this procedure. Newey and McFadden (1994) have shown the asymptotic equivalence of the three procedures.

<sup>11</sup>Newey and West label this test the *D* test.

## 482 PART III ♦ Estimation Methodology

The **LM statistic**,  $\text{LM}_{GMM}$ , is a Wald statistic for testing the hypothesis that this vector equals zero under the restrictions of the null hypothesis. From our earlier results, we would have

$$\text{Est. Asy. Var}[\mathbf{g}_1(\mathbf{c}_0)] = \frac{4}{n} \bar{\mathbf{G}}_1(\mathbf{c}_0)' \hat{\Phi}_1^{-1} \{ \text{Est. Asy. Var}[\sqrt{n} \bar{\mathbf{m}}(\mathbf{c}_0)] \} \hat{\Phi}_1^{-1} \bar{\mathbf{G}}_1(\mathbf{c}_0).$$

The estimated asymptotic variance of  $\sqrt{n} \bar{\mathbf{m}}(\mathbf{c}_0)$  is  $\hat{\Phi}_1$ , so

$$\text{Est. Asy. Var}[\mathbf{g}_1(\mathbf{c}_0)] = \frac{4}{n} \bar{\mathbf{G}}_1(\mathbf{c}_0)' \hat{\Phi}_1^{-1} \bar{\mathbf{G}}_1(\mathbf{c}_0).$$

The Wald statistic would be

$$\begin{aligned} \text{Wald} &= \mathbf{g}_1(\mathbf{c}_0)' \{ \text{Est. Asy. Var}[\mathbf{g}_1(\mathbf{c}_0)] \}^{-1} \mathbf{g}_1(\mathbf{c}_0) \\ &= n \bar{\mathbf{m}}_1'(\mathbf{c}_0) \hat{\Phi}_1^{-1} \bar{\mathbf{G}}_1(\mathbf{c}_0) \{ \bar{\mathbf{G}}_1(\mathbf{c}_0)' \hat{\Phi}_1^{-1} \bar{\mathbf{G}}_1(\mathbf{c}_0) \}^{-1} \bar{\mathbf{G}}_1(\mathbf{c}_0)' \hat{\Phi}_1^{-1} \bar{\mathbf{m}}_1(\mathbf{c}_0). \end{aligned} \quad (13-17)$$

## 13.6 GMM ESTIMATION OF ECONOMETRIC MODELS

The preceding has suggested that the GMM approach to estimation broadly encompasses most of the estimators we will encounter in this book. We have implicitly examined least squares and the general method of instrumental variables in the process. In this section, we will formalize more specifically the GMM estimators for several of the estimators that appear in the earlier chapters. Section 13.6.1 examines the generalized regression model of Chapter 9. Section 13.6.2 describes a relatively minor extension of the GMM/IV estimator to nonlinear regressions. Sections 13.6.3 and 13.6.4 describe the GMM estimators for our models of systems of equations, the seemingly unrelated regressions (SUR) model and models of simultaneous equations. In the latter, as we did in Chapter 10, we consider both limited (single-equation) and full information (multiple-equation) estimators. Finally, in Section 13.6.5, we develop one of the major applications of GMM estimation, the Arellano–Bond–Bover estimator for dynamic panel data models.

### 13.6.1 SINGLE-EQUATION LINEAR MODELS

It is useful to confine attention to the instrumental variables case, as it is fairly general and we can easily specialize it to the simpler regression models if that is appropriate. Thus, we depart from the usual linear model (8-1), but we no longer require that  $E[\varepsilon_i | \mathbf{x}_i] = 0$ . Instead, we adopt the instrumental variables formulation in Chapter 8. That is, our model is

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i \\ E[\mathbf{z}_i \varepsilon_i] &= \mathbf{0} \end{aligned}$$

for  $K$  variables in  $\mathbf{x}_i$  and for some set of  $L$  instrumental variables,  $\mathbf{z}_i$ , where  $L \geq K$ . The earlier case of the generalized regression model arises if  $\mathbf{z}_i = \mathbf{x}_i$ , and the classical regression form results if we add  $\boldsymbol{\Omega} = \mathbf{I}$  as well, so this is a convenient encompassing model framework.

In Chapter 9 on generalized least squares estimation, we considered two cases, first one with a known  $\boldsymbol{\Omega}$ , then one with an unknown  $\boldsymbol{\Omega}$  that must be estimated. In estimation

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 483

by the generalized method of moments, neither of these approaches is relevant because we begin with much less (assumed) knowledge about the data generating process. We will consider three cases:

- Classical regression:  $\text{Var}[\varepsilon_i | \mathbf{X}, \mathbf{Z}] = \sigma^2$ ,
- Heteroscedasticity:  $\text{Var}[\varepsilon_i | \mathbf{X}, \mathbf{Z}] = \sigma_i^2$ ,
- Generalized model:  $\text{Cov}[\varepsilon_t, \varepsilon_s | \mathbf{X}, \mathbf{Z}] = \sigma^2 \omega_{ts}$ ,

where  $\mathbf{Z}$  and  $\mathbf{X}$  are the  $n \times L$  and  $n \times K$  observed data matrices. (We assume, as will often be true, that the fully general case will apply in a time-series setting. Hence the change in the subscripts.) *No specific distribution is assumed for the disturbances, conditional or unconditional.*

The assumption  $E[\mathbf{z}_i \varepsilon_i] = \mathbf{0}$  implies the following **orthogonality condition**:

$$\text{Cov}[\mathbf{z}_i, \varepsilon_i] = \mathbf{0}, \quad \text{or} \quad E[\mathbf{z}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta})] = \mathbf{0}.$$

By summing the terms, we find that this further implies the **population moment equation**,

$$E\left[\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta})\right] = E[\bar{\mathbf{m}}(\boldsymbol{\beta})] = \mathbf{0}. \quad (13-18)$$

This relationship suggests how we might now proceed to estimate  $\boldsymbol{\beta}$ . Note, in fact, that if  $\mathbf{z}_i = \mathbf{x}_i$ , then this is just the population counterpart to the least squares normal equations. So, as a guide to estimation, this would return us to least squares. Suppose, we now translate this population expectation into a sample analog and use that as our guide for estimation. That is, if the population relationship holds for the true parameter vector,  $\boldsymbol{\beta}$ , suppose we attempt to mimic this result with a sample counterpart, or empirical moment equation,

$$\left[\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i(y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}})\right] = \left[\frac{1}{n} \sum_{i=1}^n \mathbf{m}_i(\hat{\boldsymbol{\beta}})\right] = \bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \mathbf{0}. \quad (13-19)$$

In the absence of other information about the data generating process, we can use the empirical moment equation as the basis of our estimation strategy.

The empirical moment condition is  $L$  equations (the number of variables in  $\mathbf{Z}$ ) in  $K$  unknowns (the number of parameters we seek to estimate). There are three possibilities to consider:

1. **Underidentified.**  $L < K$ . If there are fewer moment equations than there are parameters, then it will not be possible to find a solution to the equation system in (13-19). With no other information, such as restrictions that would reduce the number of free parameters, there is no need to proceed any further with this case.

For the identified cases, it is convenient to write (13-19) as

$$\bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \left(\frac{1}{n} \mathbf{Z}' \mathbf{y}\right) - \left(\frac{1}{n} \mathbf{Z}' \mathbf{X}\right) \hat{\boldsymbol{\beta}}. \quad (13-20)$$

2. **Exactly identified.** If  $L = K$ , then you can easily show (we leave it as an exercise) that the single solution to the equation system is the familiar instrumental variables estimator from Section 12.5.2.

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \mathbf{y}. \quad (13-21)$$

## 484 PART III ♦ Estimation Methodology

**3. Overidentified.** If  $L > K$ , then there is no unique solution to the equation system  $\bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ . In this instance, we need to formulate some strategy to choose an estimator. One intuitively appealing possibility which has served well thus far is “least squares.” In this instance, that would mean choosing the estimator based on the criterion function

$$\text{Min}_{\boldsymbol{\beta}} q = \bar{\mathbf{m}}(\hat{\boldsymbol{\beta}})' \bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}).$$

We do keep in mind that we will only be able to minimize this at some positive value; there is no exact solution to (13-19) in the overidentified case. Also, you can verify that if we treat the exactly identified case as if it were overidentified, that is, use least squares anyway, we will still obtain the IV estimator shown in (13-21) for the solution to case (2). For the overidentified case, the first-order conditions are

$$\begin{aligned} \frac{\partial q}{\partial \boldsymbol{\beta}} &= 2 \left( \frac{\partial \bar{\mathbf{m}}'(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \right) \bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = 2 \bar{\mathbf{G}}(\hat{\boldsymbol{\beta}})' \bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}) \\ &= 2 \left( \frac{1}{n} \mathbf{X}' \mathbf{Z} \right) \left( \frac{1}{n} \mathbf{Z}' \mathbf{y} - \frac{1}{n} \mathbf{Z}' \mathbf{X} \hat{\boldsymbol{\beta}} \right) = \mathbf{0}. \end{aligned} \quad (13-22)$$

We leave as exercise to show that the solution in both cases (2) and (3) is now

$$\hat{\boldsymbol{\beta}} = [(\mathbf{X}' \mathbf{Z})(\mathbf{Z}' \mathbf{X})]^{-1} (\mathbf{X}' \mathbf{Z})(\mathbf{Z}' \mathbf{y}). \quad (13-23)$$

The estimator in (13-23) is a hybrid that we have not encountered before, though if  $L = K$ , then it does reduce to the earlier one in (13-21). (In the overidentified case, (13-21) is not an IV estimator, it is, as we have sought, a **method of moments estimator**.)

It remains to establish consistency and to obtain the asymptotic distribution and an asymptotic covariance matrix for the estimator. The intermediate results we need are Assumptions 13.1, 13.2 and 13.3 in Section 13.4.3:

- **Convergence of the moments.** The sample moment converges in probability to its population counterpart. That is,  $\bar{\mathbf{m}}(\boldsymbol{\beta}) \rightarrow \mathbf{0}$ . Different circumstances will produce different kinds of convergence, but we will require it in some form. For the simplest cases, such as a model of heteroscedasticity, this will be convergence in mean square. Certain time-series models that involve correlated observations will necessitate some other form of convergence. But, in any of the cases we consider, we will require the general result:  $\text{plim } \bar{\mathbf{m}}(\boldsymbol{\beta}) = \mathbf{0}$ .
- **Identification.** The parameters are identified in terms of the moment equations. Identification means, essentially, that a large enough sample will contain sufficient information for us actually to estimate  $\boldsymbol{\beta}$  consistently using the sample moments. There are two conditions which must be met—an **order condition**, which we have already assumed ( $L \geq K$ ), and a **rank condition**, which states that the moment equations are not redundant. The rank condition implies the order condition, so we need only formalize it:
- **Identification condition for GMM estimation.** The  $L \times K$  matrix

$$\boldsymbol{\Gamma}(\boldsymbol{\beta}) = E[\bar{\mathbf{G}}(\boldsymbol{\beta})] = \text{plim } \bar{\mathbf{G}}(\boldsymbol{\beta}) = \text{plim } \frac{\partial \bar{\mathbf{m}}}{\partial \boldsymbol{\beta}'} = \text{plim } \frac{1}{n} \sum_{i=1}^n \frac{\partial \mathbf{m}_i}{\partial \boldsymbol{\beta}'}$$

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 485

must have row rank equal to  $K$ .<sup>12</sup> Because this requires  $L \geq K$ , this implies the order condition. This assumption means that this derivative matrix converges in probability to its expectation. Note that we have assumed, in addition, that the derivatives, like the moments themselves, obey a law of large numbers—they converge in probability to their expectations.

- **Limiting Normal Distribution for the Sample Moments.** The population moment obeys a central limit theorem or some similar variant. Since we are studying a generalized regression model, Lindeberg–Levy (D.18.) will be too narrow—the observations will have different variances. Lindeberg–Feller (D.19.A) suffices in the heteroscedasticity case, but in the general case, we will ultimately require something more general. See Section 13.4.3.

It will follow from Assumptions 13.1–13.3 (again, at this point we do this without proof) that the GMM estimators that we obtain are, in fact, consistent. By virtue of the Slutsky theorem, we can transfer our limiting results to the empirical moment equations.

To obtain the asymptotic covariance matrix we will simply invoke the general result for GMM estimators in Section 13.4.3. That is,

$$\text{Asy. Var}[\hat{\beta}] = \frac{1}{n} [\Gamma' \Gamma]^{-1} \Gamma' \{\text{Asy. Var}[\sqrt{n} \bar{\mathbf{m}}(\boldsymbol{\beta})]\} \Gamma [\Gamma' \Gamma]^{-1}.$$

For the particular model we are studying here,

$$\begin{aligned}\bar{\mathbf{m}}(\boldsymbol{\beta}) &= (1/n)(\mathbf{Z}' \mathbf{y} - \mathbf{Z}' \mathbf{X} \boldsymbol{\beta}), \\ \bar{\mathbf{G}}(\boldsymbol{\beta}) &= (1/n)\mathbf{Z}' \mathbf{X}, \\ \Gamma(\boldsymbol{\beta}) &= \mathbf{Q}_{\mathbf{Z} \mathbf{X}} \text{ (see Section 8.3.2).}\end{aligned}$$

(You should check in the preceding expression that the dimensions of the particular matrices and the dimensions of the various products produce the correctly configured matrix that we seek.) The remaining detail, which is the crucial one for the model we are examining, is for us to determine

$$\mathbf{V} = \text{Asy. Var}[\sqrt{n} \bar{\mathbf{m}}(\boldsymbol{\beta})].$$

Given the form of  $\bar{\mathbf{m}}(\boldsymbol{\beta})$ ,

$$\mathbf{V} = \frac{1}{n} \text{Var} \left[ \sum_{i=1}^n \mathbf{z}_i \varepsilon_i \right] = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \sigma^2 \omega_{ij} \mathbf{z}_i \mathbf{z}'_j = \sigma^2 \frac{\mathbf{Z}' \Omega \mathbf{Z}}{n}$$

for the most general case. Note that this is precisely the expression that appears in (8.6), so the question that arose there arises here once again. That is, under what conditions will this converge to a constant matrix? We take the discussion there as given. The remaining detail is how to estimate this matrix. The answer appears in Section 8.2.3, where we pursued this same question in connection with robust estimation of the asymptotic covariance matrix of the least squares estimator. To review then, what we have achieved to this point is to provide a theoretical foundation for the instrumental

<sup>12</sup>We require that the row rank be at least as large as  $K$ . There could be redundant, that is, functionally dependent, moments, so long as there are at least  $K$  that are functionally independent.

### 486 PART III ♦ Estimation Methodology

variables estimator. As noted earlier, this specializes to the least squares estimator. The estimators of  $\mathbf{V}$  for our three cases will be

- Classical regression:

$$\hat{\mathbf{V}} = \frac{(\mathbf{e}'\mathbf{e}/n)}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}'_i = \frac{(\mathbf{e}'\mathbf{e}/n)}{n} \mathbf{Z}' \mathbf{Z}.$$

- Heteroscedastic regression:

$$\hat{\mathbf{V}} = \frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{z}_i \mathbf{z}'_i. \quad (13-24)$$

- Generalized regression:

$$\hat{\mathbf{V}} = \frac{1}{n} \left[ \sum_{t=1}^n e_t^2 \mathbf{z}_t \mathbf{z}'_t + \sum_{\ell=1}^p \left( 1 - \frac{\ell}{(p+1)} \right) \sum_{t=\ell+1}^n e_t e_{t-\ell} (\mathbf{z}_t \mathbf{z}'_{t-\ell} + \mathbf{z}_{t-\ell} \mathbf{z}'_t) \right].$$

We should observe that in each of these cases, we have actually used some information about the structure of  $\Omega$ . If it is known only that the terms in  $\bar{\mathbf{m}}(\beta)$  are uncorrelated, then there is a convenient estimator available,

$$\hat{\mathbf{V}} = \frac{1}{n} \sum_{i=1}^n \mathbf{m}_i(\hat{\beta}) \mathbf{m}_i(\hat{\beta})',$$

that is, the natural, empirical variance estimator. Note that this is what is being used in the heteroscedasticity case directly preceding.

Collecting all the terms so far, then, we have

$$\begin{aligned} \text{Est. Asy. Var}[\hat{\beta}] &= \frac{1}{n} [\bar{\mathbf{G}}(\hat{\beta})' \bar{\mathbf{G}}(\hat{\beta})]^{-1} \bar{\mathbf{G}}(\hat{\beta})' \hat{\mathbf{V}} \bar{\mathbf{G}}(\hat{\beta}) [\bar{\mathbf{G}}(\hat{\beta})' \bar{\mathbf{G}}(\hat{\beta})]^{-1} \\ &= n[(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{X})]^{-1} (\mathbf{X}'\mathbf{Z}) \hat{\mathbf{V}} (\mathbf{Z}'\mathbf{X}) [(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{X})]^{-1}. \end{aligned} \quad (13-25)$$

The preceding might seem to endow the least squares or method of moments estimators with some degree of optimality, but that is not the case. We have only provided them with a different statistical motivation (and established consistency). We now consider the question of whether, because this is the generalized regression model, there is some better (more efficient) means of using the data.

The class of minimum distance estimators for this model is defined by the solutions to the criterion function

$$\text{Min}_{\beta} q = \bar{\mathbf{m}}(\beta)' \mathbf{W} \bar{\mathbf{m}}(\beta),$$

where  $\mathbf{W}$  is *any* positive definite **weighting matrix**. Based on the assumptions just made, we can invoke Theorem 13.1 to obtain

$$\text{Asy. Var}[\hat{\beta}_{MD}] = \frac{1}{n} [\bar{\mathbf{G}}' \mathbf{W} \bar{\mathbf{G}}]^{-1} \bar{\mathbf{G}}' \mathbf{W} \mathbf{V} \mathbf{W} \bar{\mathbf{G}} [\bar{\mathbf{G}}' \mathbf{W} \bar{\mathbf{G}}]^{-1}.$$

Note that our entire preceding analysis was of the simplest minimum distance estimator, which has  $\mathbf{W} = \mathbf{I}$ . The obvious question now arises, if any  $\mathbf{W}$  produces a consistent estimator, is any  $\mathbf{W}$  better than any other one, or is it simply arbitrary? There is a firm answer, for which we have to consider two cases separately:

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 487

- **Exactly identified case.** If  $L = K$ ; that is, if the number of moment conditions is the same as the number of parameters being estimated, then  $\mathbf{W}$  is irrelevant to the solution, so on the basis of simplicity alone, the optimal  $\mathbf{W}$  is  $\mathbf{I}$ .
- **Overidentified case.** In this case, the “optimal” weighting matrix, that is, the  $\mathbf{W}$  that produces the most efficient estimator, is  $\mathbf{W} = \mathbf{V}^{-1}$ . The best weighting matrix is the inverse of the asymptotic covariance of the moment vector. In this case, the MDE will be the GMM estimator with

$$\hat{\boldsymbol{\beta}}_{GMM} = [(\mathbf{X}'\mathbf{Z})\hat{\mathbf{V}}^{-1}(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})\hat{\mathbf{V}}^{-1}(\mathbf{Z}'\mathbf{y}),$$

and

$$\begin{aligned}\text{Asy. Var}[\hat{\boldsymbol{\beta}}_{GMM}] &= \frac{1}{n}[\tilde{\mathbf{G}}'\mathbf{V}^{-1}\tilde{\mathbf{G}}]^{-1} \\ &= n[(\mathbf{X}'\mathbf{Z})\mathbf{V}^{-1}(\mathbf{Z}'\mathbf{X})]^{-1}.\end{aligned}$$

We conclude this discussion by tying together what should seem to be a loose end. The GMM estimator is computed as the solution to

$$\text{Min}_{\boldsymbol{\beta}} q = \bar{\mathbf{m}}(\boldsymbol{\beta})' \{\text{Asy. Var}[\sqrt{n} \bar{\mathbf{m}}(\boldsymbol{\beta})]\}^{-1} \bar{\mathbf{m}}(\boldsymbol{\beta}),$$

which might suggest that the weighting matrix is a function of the thing we are trying to estimate. The process of GMM estimation will have to proceed in two steps: Step 1 is to obtain an estimate of  $\mathbf{V}$ ; Step 2 will consist of using the inverse of this  $\mathbf{V}$  as the weighting matrix in computing the GMM estimator. The following is a common strategy:

**Step 1.** Use  $\mathbf{W} = \mathbf{I}$  to obtain a consistent estimator of  $\boldsymbol{\beta}$ . Then, estimate  $\mathbf{V}$  with

$$\hat{\mathbf{V}} = \frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{z}_i \mathbf{z}_i'$$

in the heteroscedasticity case (i.e., the White estimator) or, for the more general case, the Newey-West estimator.

**Step 2.** Use  $\mathbf{W} = \hat{\mathbf{V}}^{-1}$  to compute the GMM estimator.

By this point, the observant reader should have noticed that in all of the preceding, we have never actually encountered the two-stage least squares estimator that we introduced in Section 12.10. To obtain this estimator, we must revert back to the classical, that is, homoscedastic, and nonautocorrelated disturbances case. In that instance, the weighting matrix in Theorem 13.2 will be  $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$  and we will obtain the apparently missing result.

The **GMM estimator** in the heteroscedastic regression model is produced by the empirical moment equations

$$\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i (y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{GMM}) = \frac{1}{n} \mathbf{X}' \hat{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\beta}}_{GMM}) = \bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}_{GMM}) = \mathbf{0}. \quad (13-26)$$

## 488 PART III ♦ Estimation Methodology

The estimator is obtained by minimizing

$$q = \bar{\mathbf{m}}'(\hat{\beta}_{GMM}) \mathbf{W} \bar{\mathbf{m}}(\hat{\beta}_{GMM}),$$

where  $\mathbf{W}$  is a positive definite weighting matrix. The optimal weighting matrix would be

$$\mathbf{W} = \left\{ \text{Asy. Var}[\sqrt{n} \bar{\mathbf{m}}(\boldsymbol{\beta})] \right\}^{-1},$$

which is the inverse of

$$\text{Asy. Var}[\sqrt{n} \bar{\mathbf{m}}(\boldsymbol{\beta})] = \text{Asy. Var} \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i \boldsymbol{\epsilon}_i \right] = \plim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \sigma^2 \omega_i \mathbf{x}_i \mathbf{x}'_i = \sigma^2 \mathbf{Q}^*.$$

[See Section 9.4.1.] The optimal weighting matrix would be  $[\sigma^2 \mathbf{Q}^*]^{-1}$ . But recall that this minimization problem is an exactly identified case, so the weighting matrix is irrelevant to the solution. You can see the result in the moment equation—that equation is simply the normal equations for ordinary least squares. We can solve the moment equations exactly, so there is no need for the weighting matrix. *Regardless of the covariance matrix of the moments, the GMM estimator for the heteroscedastic regression model is ordinary least squares.* We can use the results we have already obtained to find its asymptotic covariance matrix. The implied estimator is the White estimator in (9-27). [Once again, see Theorem 13.2.] The conclusion to be drawn at this point is that until we make some specific assumptions about the variances, we do not have a more efficient estimator than least squares, but we do have to modify the estimated asymptotic covariance matrix.

### 13.6.2 SINGLE-EQUATION NONLINEAR MODELS

Suppose that the theory specifies a relationship

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \boldsymbol{\epsilon}_i,$$

where  $\boldsymbol{\beta}$  is a  $K \times 1$  parameter vector that we wish to estimate. This may not be a regression relationship, because it is possible that

$$\text{Cov}[\boldsymbol{\epsilon}_i, h(\mathbf{x}_i, \boldsymbol{\beta})] \neq 0,$$

or even

$$\text{Cov}[\boldsymbol{\epsilon}_i, \mathbf{x}_j] \neq 0 \text{ for all } i \text{ and } j.$$

Consider, for example, a model that contains lagged dependent variables and autocorrelated disturbances. (See Section 20.9.3.) For the present, we assume that

$$E[\boldsymbol{\epsilon} | \mathbf{X}] \neq \mathbf{0},$$

and

$$E[\boldsymbol{\epsilon} \boldsymbol{\epsilon}' | \mathbf{X}] = \sigma^2 \boldsymbol{\Omega} = \boldsymbol{\Sigma},$$

where  $\boldsymbol{\Sigma}$  is symmetric and positive definite but otherwise unrestricted. The disturbances may be heteroscedastic and/or autocorrelated. But for the possibility of correlation between regressors and disturbances, this model would be a generalized, possibly nonlinear, regression model. Suppose that at each observation  $i$  we observe a vector of  $L$  variables,  $\mathbf{z}_i$ , such that  $\mathbf{z}_i$  is uncorrelated with  $\boldsymbol{\epsilon}_i$ . You will recognize  $\mathbf{z}_i$  as a set of

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 489

**instrumental variables.** The assumptions thus far have implied a set of orthogonality conditions,

$$E[\mathbf{z}_i \varepsilon_i] = \mathbf{0},$$

which may be sufficient to identify (if  $L = K$ ) or even overidentify (if  $L > K$ ) the parameters of the model. (See Section 8.3.4.)

For convenience, define

$$\mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}}) = y_i - h(\mathbf{x}_i, \hat{\boldsymbol{\beta}}), \quad i = 1, \dots, n,$$

and

$$\mathbf{Z} = n \times L \text{ matrix whose } i\text{th row is } \mathbf{z}'_i.$$

By a straightforward extension of our earlier results, we can produce a GMM estimator of  $\boldsymbol{\beta}$ . The sample moments will be

$$\bar{\mathbf{m}}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{e}(\mathbf{x}_i, \boldsymbol{\beta}) = \frac{1}{n} \mathbf{Z}' \mathbf{e}(\mathbf{X}, \boldsymbol{\beta}).$$

The minimum distance estimator will be the  $\hat{\boldsymbol{\beta}}$  that minimizes

$$q = \bar{\mathbf{m}}_n(\hat{\boldsymbol{\beta}})' \mathbf{W} \bar{\mathbf{m}}_n(\hat{\boldsymbol{\beta}}) = \left( \frac{1}{n} [\mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}})' \mathbf{Z}] \right) \mathbf{W} \left( \frac{1}{n} [\mathbf{Z}' \mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}})] \right) \quad (13-27)$$

for some choice of  $\mathbf{W}$  that we have yet to determine. The criterion given earlier produces the **nonlinear instrumental variable estimator**. If we use  $\mathbf{W} = (\mathbf{Z}' \mathbf{Z})^{-1}$ , then we have exactly the estimation criterion we used in Section 8.7, where we defined the nonlinear instrumental variables estimator. Apparently (13-27) is more general, because we are not limited to this choice of  $\mathbf{W}$ . For any given choice of  $\mathbf{W}$ , as long as there are enough orthogonality conditions to identify the parameters, estimation by minimizing  $q$  is, at least in principle, a straightforward problem in nonlinear optimization. The optimal choice of  $\mathbf{W}$  for this estimator is

$$\begin{aligned} \mathbf{W}_{\text{GMM}} &= \left\{ \text{Asy. Var}[\sqrt{n} \bar{\mathbf{m}}_n(\boldsymbol{\beta})] \right\}^{-1} \\ &= \left\{ \text{Asy. Var} \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{z}_i \varepsilon_i \right] \right\}^{-1} = \left\{ \text{Asy. Var} \left[ \frac{1}{\sqrt{n}} \mathbf{Z}' \mathbf{e}(\mathbf{X}, \boldsymbol{\beta}) \right] \right\}^{-1}. \end{aligned} \quad (13-28)$$

For our model, this is

$$\mathbf{W} = \left[ \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \text{Cov}[\mathbf{z}_i \varepsilon_i, \mathbf{z}_j \varepsilon_j] \right]^{-1} = \left[ \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \sigma_{ij} \mathbf{z}_i \mathbf{z}'_j \right]^{-1} = \left[ \frac{\mathbf{Z}' \boldsymbol{\Sigma} \mathbf{Z}}{n} \right]^{-1}.$$

If we insert this result in (13-27), we obtain the criterion for the GMM estimator:

$$q = \left[ \left( \frac{1}{n} \right) \mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}})' \mathbf{Z} \right] \left( \frac{\mathbf{Z}' \boldsymbol{\Sigma} \mathbf{Z}}{n} \right)^{-1} \left[ \left( \frac{1}{n} \right) \mathbf{Z}' \mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}}) \right].$$

There is a possibly difficult detail to be considered. The GMM estimator involves

$$\frac{1}{n} \mathbf{Z}' \boldsymbol{\Sigma} \mathbf{Z} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \mathbf{z}_i \mathbf{z}'_j \text{Cov}[\varepsilon_i, \varepsilon_j] = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \mathbf{z}_i \mathbf{z}'_j \text{Cov}[(y_i - h(\mathbf{x}_i, \boldsymbol{\beta})), (y_j - h(\mathbf{x}_j, \boldsymbol{\beta}))].$$

## 490 PART III ♦ Estimation Methodology

The conditions under which such a double sum might converge to a positive definite matrix are sketched in Section 9.2.2. Assuming that they do hold, estimation appears to require that an estimate of  $\beta$  be in hand already, even though it is the object of estimation. It may be that a consistent but inefficient estimator of  $\beta$  is available. Suppose for the present that one is. If observations are uncorrelated, then the cross-observation terms may be omitted, and what is required is

$$\frac{1}{n} \mathbf{Z}' \Sigma \mathbf{Z} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i' \text{Var}[(y_i - h(\mathbf{x}_i, \beta))].$$

We can use a counterpart to the White (1980) estimator discussed in Section 9.4.4 for this case:

$$\mathbf{S}_0 = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i' (y_i - h(\mathbf{x}_i, \hat{\beta}))^2. \quad (13-29)$$

If the disturbances are autocorrelated but the process is stationary, then Newey and West's (1987a) estimator is available (assuming that the autocorrelations are sufficiently small at a reasonable lag,  $p$ ):

$$\mathbf{S} = \left[ \mathbf{S}_0 + \frac{1}{n} \sum_{\ell=1}^p w(\ell) \sum_{i=\ell+1}^n e_i e_{i-\ell} (\mathbf{z}_i \mathbf{z}_{i-\ell}' + \mathbf{z}_{i-\ell} \mathbf{z}_i') \right] = \sum_{\ell=0}^p w(\ell) \mathbf{S}_\ell, \quad (13-30)$$

where

$$w(\ell) = 1 - \frac{\ell}{p+1}.$$

The maximum lag length  $p$  must be determined in advance. We will require that observations that are far apart in time—that is, for which  $|i - \ell|$  is large—must have increasingly smaller covariances for us to establish the convergence results that justify OLS, GLS, and now GMM estimation. The choice of  $p$  is a reflection of how far back in time one must go to consider the autocorrelation negligible for purposes of estimating  $(1/n) \mathbf{Z}' \Sigma \mathbf{Z}$ . Current practice suggests using the smallest integer greater than or equal to  $n^{1/4}$ .

Still left open is the question of where the initial consistent estimator should be obtained. One possibility is to obtain an inefficient but consistent GMM estimator by using  $\mathbf{W} = \mathbf{I}$  in (13-27). That is, use a nonlinear (or linear, if the equation is linear) instrumental variables estimator. This first-step estimator can then be used to construct  $\mathbf{W}$ , which, in turn, can then be used in the GMM estimator. Another possibility is that  $\beta$  may be consistently estimable by some straightforward procedure other than GMM.

Once the GMM estimator has been computed, its asymptotic covariance matrix and asymptotic distribution can be estimated based on Theorem 13.2. Recall that

$$\bar{\mathbf{m}}_n(\beta) = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \varepsilon_i,$$

which is a sum of  $L \times 1$  vectors. The derivative,  $\partial \bar{\mathbf{m}}_n(\beta) / \partial \beta'$ , is a sum of  $L \times K$  matrices, so

$$\bar{\mathbf{G}}(\beta) = \partial \bar{\mathbf{m}}_n(\beta) / \partial \beta' = \frac{1}{n} \sum_{i=1}^n \mathbf{G}_i(\beta) = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \left[ \frac{\partial \varepsilon_i}{\partial \beta'} \right]. \quad (13-31)$$

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 491

In the model we are considering here,

$$\frac{\partial \varepsilon_i}{\partial \beta'} = \frac{-\partial h(\mathbf{x}_i, \beta)}{\partial \beta'}.$$

The derivatives are the pseudoregressors in the linearized regression model that we examined in Section 7.2.3. Using the notation defined there,

$$\frac{\partial \varepsilon_i}{\partial \beta} = -\mathbf{x}_i^0,$$

so

$$\bar{\mathbf{G}}(\beta) = \frac{1}{n} \sum_{i=1}^n \mathbf{G}_i(\beta) = \frac{1}{n} \sum_{i=1}^n -\mathbf{z}_i \mathbf{x}_i^0 = -\frac{1}{n} \mathbf{Z}' \mathbf{X}^0. \quad (13-32)$$

With this matrix in hand, the estimated asymptotic covariance matrix for the GMM estimator is

$$\text{Est. Asy. Var}[\hat{\beta}] = \left[ \bar{\mathbf{G}}(\hat{\beta})' \left( \frac{1}{n} \mathbf{Z}' \hat{\Sigma} \mathbf{Z} \right)^{-1} \bar{\mathbf{G}}(\hat{\beta}) \right]^{-1} = [(\mathbf{X}^0)' \mathbf{Z} (\mathbf{Z}' \hat{\Sigma} \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{X}^0)]^{-1}. \quad (13-33)$$

(The two minus signs, a  $1/n^2$ , and an  $n^2$ , all fall out of the result.)

If the  $\Sigma$  that appears in (13-33) were  $\sigma^2 \mathbf{I}$ , then (13-33) would be precisely the asymptotic covariance matrix that appears in Theorem 8.7 for linear models and Theorem 8.7 for nonlinear models. But there is an interesting distinction between this estimator and the IV estimators discussed earlier. In the earlier cases, when there were more instrumental variables than parameters, we resolved the overidentification by specifically choosing a set of  $K$  instruments, the  $K$  projections of the columns of  $\mathbf{X}$  or  $\mathbf{X}^0$  into the column space of  $\mathbf{Z}$ . Here, in contrast, we do not attempt to resolve the overidentification; we simply use all the instruments and minimize the GMM criterion. Now, you should be able to show that when  $\Sigma = \sigma^2 \mathbf{I}$  and we use this information, when all is said and done, the same parameter estimates will be obtained. But, if we use a weighting matrix that differs from  $\mathbf{W} = (\mathbf{Z}' \mathbf{Z}/n)^{-1}$ , then they are not.

### 13.6.3 SEEMINGLY UNRELATED REGRESSION MODELS

In Section 10.5, we considered FGLS estimation of the equation system

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{h}_1(\mathbf{X}, \beta) + \boldsymbol{\varepsilon}_1, \\ \mathbf{y}_2 &= \mathbf{h}_2(\mathbf{X}, \beta) + \boldsymbol{\varepsilon}_2, \\ &\vdots \\ \mathbf{y}_M &= \mathbf{h}_M(\mathbf{X}, \beta) + \boldsymbol{\varepsilon}_M. \end{aligned}$$

The development there extends backwards to the linear system as well. However, none of the estimators considered are consistent if the pseudoregressors,  $\mathbf{x}_{tm}^0$ , or the actual regressors,  $\mathbf{x}_{tm}$  for the linear model, are correlated with the disturbances,  $\boldsymbol{\varepsilon}_{tm}$ . Suppose we allow for this correlation both within and across equations. (If it is, in fact, absent, then the GMM estimator developed here will remain consistent.) For simplicity in this section, we will denote observations with subscript  $t$  and equations with subscripts  $i$  and  $j$ . Suppose, as well, that there are a set of instrumental variables,  $\mathbf{z}_t$ , such that

$$E[\mathbf{z}_t \boldsymbol{\varepsilon}_{tm}] = \mathbf{0}, t = 1, \dots, T \text{ and } m = 1, \dots, M. \quad (13-34)$$

## 492 PART III ♦ Estimation Methodology

(We could allow a separate set of instrumental variables for each equation, but it would needlessly complicate the presentation.)

Under these assumptions, the nonlinear FGLS and ML estimators given earlier will be inconsistent. But a relatively minor extension of the instrumental variables technique developed for the single-equation case in Section 8.4 can be used instead. The sample analog to (13-34) is

$$\frac{1}{T} \sum_{t=1}^T \mathbf{z}_t [y_{ti} - h_i(\mathbf{x}_t, \boldsymbol{\beta})] = \mathbf{0}, \quad i = 1, \dots, M.$$

If we use this result for each equation in the system, one at a time, then we obtain exactly the GMM estimator discussed in Section 13.6.2. But, in addition to the efficiency loss that results from not imposing the cross-equation constraints in  $\boldsymbol{\beta}$ , we would also neglect the correlation between the disturbances. Let

$$\frac{1}{T} \mathbf{Z}' \boldsymbol{\Omega}_{ij} \mathbf{Z} = E \left[ \frac{\mathbf{Z}' \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_j' \mathbf{Z}}{T} \right]. \quad (13-35)$$

The GMM criterion for estimation in this setting is

$$\begin{aligned} q &= \sum_{i=1}^M \sum_{j=1}^M [(\mathbf{y}_i - \mathbf{h}_i(\mathbf{X}, \boldsymbol{\beta}))' \mathbf{Z}/T] [\mathbf{Z}' \boldsymbol{\Omega}_{ij} \mathbf{Z}/T]^{ij} [\mathbf{Z}' (\mathbf{y}_j - \mathbf{h}_j(\mathbf{X}, \boldsymbol{\beta}))/T] \\ &= \sum_{i=1}^M \sum_{j=1}^M [\boldsymbol{\varepsilon}_i(\boldsymbol{\beta})' \mathbf{Z}/T] [\mathbf{Z}' \boldsymbol{\Omega}_{ij} \mathbf{Z}/T]^{ij} [\mathbf{Z}' \boldsymbol{\varepsilon}_j(\boldsymbol{\beta})/T], \end{aligned} \quad (13-36)$$

where  $[\mathbf{Z}' \boldsymbol{\Omega}_{ij} \mathbf{Z}/T]^{ij}$  denotes the  $ij$ th block of the inverse of the matrix with the  $ij$ th block equal to  $\mathbf{Z}' \boldsymbol{\Omega}_{ij} \mathbf{Z}/T$ . (This matrix is laid out in full in Section 13.6.4.)

GMM estimation would proceed in several passes. To compute any of the variance parameters, we will require an initial consistent estimator of  $\boldsymbol{\beta}$ . This step  can be done with equation-by-equation nonlinear instrumental variables—see Section 8.7—although if equations have parameters in common, then a choice must be made as to which to use. At the next step, the familiar White or Newey-West technique is used to compute, block by block, the matrix in (13-35). Because it is based on a consistent estimator of  $\boldsymbol{\beta}$  (we assume), this matrix need not be recomputed. Now, with this result in hand, an iterative solution to the maximization problem in (13-36) can be sought, for example, using the methods of Appendix E. The first-order conditions are

$$\frac{\partial q}{\partial \boldsymbol{\beta}} = -2 \sum_{i=1}^M \sum_{j=1}^M [\mathbf{X}_i^0(\boldsymbol{\beta})' \mathbf{Z}/T] [\mathbf{Z}' \mathbf{W}_{ij} \mathbf{Z}/T]^{ij} [\mathbf{Z}' \boldsymbol{\varepsilon}_j(\boldsymbol{\beta})/T] = \mathbf{0}. \quad (13-37)$$

Note again that the blocks of the inverse matrix in the center are extracted from the larger constructed matrix *after inversion*. [This brief discussion might understate the complexity of the optimization problem in (13-36), but that is inherent in the procedure.] At completion, the asymptotic covariance matrix for the GMM estimator is estimated with

$$\mathbf{V}_{\text{GMM}} = \frac{1}{T} \left[ \sum_{i=1}^M \sum_{j=1}^M [\mathbf{X}_i^0(\boldsymbol{\beta})' \mathbf{Z}/T] [\mathbf{Z}' \mathbf{W}_{ij} \mathbf{Z}/T]^{ij} [\mathbf{Z}' \mathbf{X}_j^0(\boldsymbol{\beta})/T] \right]^{-1}.$$

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 493

## 13.6.4 SIMULTANEOUS EQUATIONS MODELS WITH HETROSCEDECSTICITY

The GMM estimator in Section 13.6.1 is, with a minor change of notation, precisely the set of procedures we used in Section 10.7.4 and 10.7.5 to estimate the equations in a simultaneous equations model. Using a GMM estimator, however, will allow us to generalize the covariance structure for the disturbances. We assume that

$$y_{tj} = \mathbf{z}'_{tj} \boldsymbol{\delta}_j + \varepsilon_{tj}, \quad t = 1, \dots, T,$$

where  $\mathbf{z}_{tj} = [\mathbf{Y}_{tj}, \mathbf{x}_{tj}]$ . (We use the capital  $\mathbf{Y}_{tj}$  to denote the  $L_j$  included endogenous variables. Note, as well, that to maintain consistency with Chapter 10, the roles of the symbols  $\mathbf{x}$  and  $\mathbf{z}$  are reversed here;  $\mathbf{x}$  is now the vector of exogenous variables.) We have assumed that  $\varepsilon_{tj}$  in the  $j$ th equation is neither heteroscedastic nor autocorrelated. There is no need to impose those assumptions at this point. Autocorrelation in the context of a simultaneous equations model is a substantial complication, however. For the present, we will consider the heteroscedastic case only.

The assumptions of the model provide the orthogonality conditions,

$$E[\mathbf{x}_t \varepsilon_{tj}] = E[\mathbf{x}_t (y_{tj} - \mathbf{z}'_{tj} \boldsymbol{\delta}_j)] = \mathbf{0}.$$

If  $\mathbf{x}_t$  is taken to be the full set of exogenous variables in the model, then we obtain the criterion for the GMM estimator for the  $j$ th equation,

$$\begin{aligned} q &= \left[ \frac{\mathbf{e}(\mathbf{z}_t, \boldsymbol{\delta}_j)' \mathbf{X}}{T} \right] \mathbf{W}_{jj}^{-1} \left[ \frac{\mathbf{X}' \mathbf{e}(\mathbf{z}_t, \boldsymbol{\delta}_j)}{T} \right] \\ &= \bar{\mathbf{m}}(\boldsymbol{\delta}_j)' \mathbf{W}_{jj}^{-1} \bar{\mathbf{m}}(\boldsymbol{\delta}_j), \end{aligned}$$

where

$$\bar{\mathbf{m}}(\boldsymbol{\delta}_j) = \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t (y_{tj} - \mathbf{z}'_{tj} \boldsymbol{\delta}_j) \quad \text{and} \quad \mathbf{W}_{jj}^{-1} = \text{the GMM weighting matrix.}$$

Once again, this is precisely the estimator defined in Section 13.6.1. If the disturbances are assumed to be homoscedastic and nonautocorrelated, then the optimal weighting matrix will be an estimator of the inverse of

$$\begin{aligned} \mathbf{W}_{jj} &= \text{Asy. Var}[\sqrt{T} \bar{\mathbf{m}}(\boldsymbol{\delta}_j)] \\ &= \text{plim} \left[ \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t (y_{tj} - \mathbf{z}'_{tj} \boldsymbol{\delta}_j)^2 \right] \\ &= \text{plim} \frac{1}{T} \sum_{t=1}^T \sigma_{jj} \mathbf{x}_t \mathbf{x}'_t \\ &= \text{plim} \sigma_{jj} \left( \frac{\mathbf{X}' \mathbf{X}}{T} \right). \end{aligned}$$

The constant  $\sigma_{jj}$  is irrelevant to the solution. If we use  $(\mathbf{X}' \mathbf{X})^{-1}$  as the weighting matrix, then the GMM estimator that minimizes  $q$  is the 2SLS estimator.

The extension that we can obtain here is to allow for heteroscedasticity of unknown form. There is no need to rederive the earlier result. If the disturbances are

## 494 PART III ♦ Estimation Methodology

heteroscedastic, then

$$\mathbf{W}_{jj} = \text{plim} \frac{1}{T} \sum_{t=1}^T \omega_{jj,t} \mathbf{x}_t \mathbf{x}'_t = \text{plim} \frac{\mathbf{X}' \boldsymbol{\Omega}_{jj} \mathbf{X}}{T}.$$

The weighting matrix can be estimated with White's heteroscedasticity consistent estimator—see (13-24)—if a consistent estimator of  $\delta_j$  is in hand with which to compute the residuals. One is, because 2SLS ignoring the heteroscedasticity is consistent, albeit inefficient. The conclusion then is that under these assumptions, there is a way to improve on 2SLS by adding another step. The name 3SLS is reserved for the systems estimator of this sort. When choosing between 2.5-stage least squares and Davidson and MacKinnon's suggested “heteroscedastic 2SLS,” or **H2SLS**, we chose to opt for the latter. The estimator is based on the initial two-stage least squares procedure. Thus,

$$\hat{\delta}_{j,\text{H2SLS}} = [\mathbf{Z}'_j \mathbf{X} (\mathbf{S}_{0,jj})^{-1} \mathbf{X}' \mathbf{Z}_j]^{-1} [\mathbf{Z}'_j \mathbf{X} (\mathbf{S}_{0,jj})^{-1} \mathbf{X}' \mathbf{y}_j],$$

where

$$\mathbf{S}_{0,jj} = \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t (y_{tj} - \mathbf{z}'_{tj} \hat{\delta}_{j,\text{2SLS}})^2.$$

The asymptotic covariance matrix is estimated with

$$\text{Est. Asy. Var}[\hat{\delta}_{j,\text{H2SLS}}] = [\mathbf{Z}'_j \mathbf{X} (\mathbf{S}_{0,jj})^{-1} \mathbf{X}' \mathbf{Z}_j]^{-1}.$$

Extensions of this estimator were suggested by Cragg (1983) and Cumby, Huizinga, and Obstfeld (1983).

The GMM estimator for a system of equations is described in Section 13.6.3. As in the single-equation case, a minor change in notation produces the estimators for a simultaneous equations model. As before, we will consider the case of unknown heteroscedasticity only. The extension to autocorrelation is quite complicated. [See Cumby, Huizinga, and Obstfeld (1983).] The orthogonality conditions defined in (13-34) are

$$E[\mathbf{x}_t \varepsilon_{tj}] = E[\mathbf{x}_t (y_{tj} - \mathbf{z}'_{tj} \delta_j)] = \mathbf{0}.$$

If we consider all the equations jointly, then we obtain the criterion for estimation of all the model's parameters,

$$\begin{aligned} q &= \sum_{j=1}^M \sum_{l=1}^M \left[ \frac{\mathbf{e}(\mathbf{z}_t, \delta_j)' \mathbf{X}}{T} \right] [\mathbf{W}]^{jl} \left[ \frac{\mathbf{X}' \mathbf{e}(\mathbf{z}_t, \delta_l)}{T} \right] \\ &= \sum_{j=1}^M \sum_{l=1}^M \bar{\mathbf{m}}(\delta_j)' [\mathbf{W}]^{jl} \bar{\mathbf{m}}(\delta_l), \end{aligned}$$

where

$$\bar{\mathbf{m}}(\delta_j) = \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t (y_{tj} - \mathbf{z}'_{tj} \delta_j),$$

and

$$[\mathbf{W}]^{jl} = \text{block } jl \text{ of the weighting matrix, } \mathbf{W}^{-1}.$$

As before, we consider the optimal weighting matrix obtained as the asymptotic covariance matrix of the empirical moments,  $\bar{\mathbf{m}}(\delta_j)$ . These moments are stacked in a single

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 495

vector  $\bar{\mathbf{m}}(\delta)$ . Then, the  $jl$ th block of  $\text{Asy. Var}[\sqrt{T}\bar{\mathbf{m}}(\delta)]$  is

$$\Phi_{jl} = \text{plim} \left\{ \frac{1}{T} \sum_{t=1}^T [\mathbf{x}_t \mathbf{x}'_t (y_{tj} - \mathbf{z}'_{tj} \delta_j) (y_{tl} - \mathbf{z}'_{tl} \delta_l)] \right\} = \text{plim} \left( \frac{1}{T} \sum_{t=1}^T \omega_{jl,t} \mathbf{x}_t \mathbf{x}'_t \right).$$

If the disturbances are homoscedastic, then  $\Phi_{jl} = \sigma_{jl} [\text{plim}(\mathbf{X}'\mathbf{X}/T)]$  is produced. Otherwise, we obtain a matrix of the form  $\Phi_{jl} = \text{plim}[\mathbf{X}'\Omega_{jl}\mathbf{X}/T]$ . Collecting terms, then, the criterion function for GMM estimation is

$$q = \begin{bmatrix} [\mathbf{X}'(\mathbf{y}_1 - \mathbf{Z}_1 \delta_1)]/T \\ [\mathbf{X}'(\mathbf{y}_2 - \mathbf{Z}_2 \delta_2)]/T \\ \vdots \\ [\mathbf{X}'(\mathbf{y}_M - \mathbf{Z}_M \delta_M)]/T \end{bmatrix}' \begin{bmatrix} \Phi_{11} & \Phi_{12} & \cdots & \Phi_{1M} \\ \Phi_{21} & \Phi_{22} & \cdots & \Phi_{2M} \\ \vdots & \vdots & \cdots & \vdots \\ \Phi_{M1} & \Phi_{M2} & \cdots & \Phi_{MM} \end{bmatrix}^{-1} \begin{bmatrix} [\mathbf{X}'(\mathbf{y}_1 - \mathbf{Z}_1 \delta_1)]/T \\ [\mathbf{X}'(\mathbf{y}_2 - \mathbf{Z}_2 \delta_2)]/T \\ \vdots \\ [\mathbf{X}'(\mathbf{y}_M - \mathbf{Z}_M \delta_M)]/T \end{bmatrix}.$$

For implementation,  $\Phi_{jl}$  can be estimated with

$$\hat{\Phi}_{jl} = \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t (y_{tj} - \mathbf{z}'_{tj} \mathbf{d}_j) (y_{tl} - \mathbf{z}'_{tl} \mathbf{d}_l),$$

where  $\mathbf{d}_j$  is a consistent estimator of  $\delta_j$ . The two-stage least squares estimator is a natural choice. For the diagonal blocks, this choice is the White estimator as usual. For the off-diagonal blocks, it is a simple extension. With this result in hand, the first-order conditions for GMM estimation are

$$\frac{\partial \hat{q}}{\partial \delta_j} = -2 \sum_{l=1}^M \left( \frac{\mathbf{Z}'_j \mathbf{X}}{T} \right) \hat{\Phi}^{jl} \left[ \frac{\mathbf{X}'(\mathbf{y}_l - \mathbf{Z}_l \delta_l)}{T} \right],$$

where  $\hat{\Phi}^{jl}$  is the  $jl$ th block in the inverse of the estimate of the center matrix in  $q$ .

The solution is

$$\begin{bmatrix} \hat{\delta}_{1,\text{GMM}} \\ \hat{\delta}_{2,\text{GMM}} \\ \vdots \\ \hat{\delta}_{M,\text{GMM}} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}'_1 \mathbf{X} \hat{\Phi}^{11} \mathbf{X}' \mathbf{Z}_1 & \mathbf{Z}'_1 \mathbf{X} \hat{\Phi}^{12} \mathbf{X}' \mathbf{Z}_2 & \cdots & \mathbf{Z}'_1 \mathbf{X} \hat{\Phi}^{1M} \mathbf{X}' \mathbf{Z}_M \\ \mathbf{Z}'_2 \mathbf{X} \hat{\Phi}^{21} \mathbf{X}' \mathbf{Z}_1 & \mathbf{Z}'_2 \mathbf{X} \hat{\Phi}^{22} \mathbf{X}' \mathbf{Z}_2 & \cdots & \mathbf{Z}'_2 \mathbf{X} \hat{\Phi}^{2M} \mathbf{X}' \mathbf{Z}_M \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Z}'_M \mathbf{X} \hat{\Phi}^{M1} \mathbf{X}' \mathbf{Z}_1 & \mathbf{Z}'_M \mathbf{X} \hat{\Phi}^{M2} \mathbf{X}' \mathbf{Z}_2 & \cdots & \mathbf{Z}'_M \mathbf{X} \hat{\Phi}^{MM} \mathbf{X}' \mathbf{Z}_M \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^M \mathbf{Z}'_1 \mathbf{X} \hat{\Phi}^{1j} \mathbf{y}_j \\ \sum_{j=1}^M \mathbf{Z}'_2 \mathbf{X} \hat{\Phi}^{2j} \mathbf{y}_j \\ \vdots \\ \sum_{j=1}^M \mathbf{Z}'_M \mathbf{X} \hat{\Phi}^{Mj} \mathbf{y}_j \end{bmatrix}.$$

The asymptotic covariance matrix for the estimator would be estimated with  $T$  times the large inverse matrix in brackets.

Several of the estimators we have already considered are special cases:

- If  $\hat{\Phi}_{jj} = \hat{\sigma}_{jj}(\mathbf{X}'\mathbf{X}/T)$  and  $\hat{\Phi}_{jl} = \mathbf{0}$  for  $j \neq l$ , then  $\hat{\delta}_j$  is 2SLS.
- If  $\hat{\Phi}_{jl} = \mathbf{0}$  for  $j \neq l$ , then  $\hat{\delta}_j$  is H2SLS, the single-equation GMM estimator.
- If  $\hat{\Phi}_{jl} = \hat{\sigma}_{jl}(\mathbf{X}'\mathbf{X}/T)$ , then  $\hat{\delta}_j$  is 3SLS.

As before, the GMM estimator brings efficiency gains in the presence of heteroscedasticity. If the disturbances are homoscedastic, then it is asymptotically the same as 3SLS,

## 496 PART III ♦ Estimation Methodology

[although in a finite sample, it will differ numerically because  $\mathbf{S}_{jl}$  will not be identical to  $\hat{\sigma}_{jl}(\mathbf{X}'\mathbf{X})$ ].

### 13.6.5 GMM ESTIMATION OF DYNAMIC PANEL DATA MODELS

Panel data are well suited for examining dynamic effects, as in the first-order model,

$$\begin{aligned} y_{it} &= \mathbf{x}'_{it}\boldsymbol{\beta} + \delta y_{i,t-1} + c_i + \varepsilon_{it} \\ &= \mathbf{w}'_{it}\boldsymbol{\theta} + \alpha_i + \varepsilon_{it}, \end{aligned}$$

where the set of right-hand-side variables,  $\mathbf{w}_{it}$ , now includes the lagged dependent variable,  $y_{i,t-1}$ . Adding dynamics to a model in this fashion creates a major change in the interpretation of the equation. Without the lagged variable, the “independent variables” represent the full set of information that produce observed outcome  $y_{it}$ . With the lagged variable, we now have in the equation the entire history of the right-hand-side variables, so that any measured influence is conditioned on this history; in this case, any impact of  $\mathbf{x}_{it}$  represents the effect of *new* information. Substantial complications arise in estimation of such a model. In both the fixed and random effects settings, the difficulty is that the lagged dependent variable is correlated with the disturbance, even if it is assumed that  $\varepsilon_{it}$  is not itself autocorrelated. For the moment, consider the fixed effects model as an ordinary regression with a lagged dependent variable that is dependent across observations. In that dynamic regression model, the estimator based on  $T$  observations is biased in finite samples, but it is consistent in  $T$ . The finite sample bias is of order  $1/T$ . The same result applies here, but the difference is that whereas before we obtained our large sample results by allowing  $T$  to grow large, in this setting,  $T$  is assumed to be small and fixed, and large-sample results are obtained with respect to  $n$  growing large, not  $T$ . The fixed effects estimator of  $\boldsymbol{\theta} = [\boldsymbol{\beta}, \delta]$  can be viewed as an average of  $n$  such estimators. Assume for now that  $T \geq K + 1$  where  $K$  is the number of variables in  $\mathbf{x}_{it}$ . Then, from (11-13),

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \left[ \sum_{i=1}^n \mathbf{W}_i' \mathbf{M}^0 \mathbf{W}_i \right]^{-1} \left[ \sum_{i=1}^n \mathbf{W}_i' \mathbf{M}^0 \mathbf{y}_i \right] \\ &= \left[ \sum_{i=1}^n \mathbf{W}_i' \mathbf{M}^0 \mathbf{W}_i \right]^{-1} \left[ \sum_{i=1}^n \mathbf{W}_i' \mathbf{M}^0 \mathbf{W}_i \mathbf{d}_i \right] \\ &= \sum_{i=1}^n \mathbf{F}_i \mathbf{d}_i, \end{aligned}$$

where the rows of the  $T \times (K + 1)$  matrix  $\mathbf{W}_i$  are  $\mathbf{w}'_{it}$  and  $\mathbf{M}^0$  is the  $T \times T$  matrix that creates deviations from group means [see (11-14)]. Each group-specific estimator,  $\mathbf{d}_i$ , is inconsistent, as it is biased in finite samples and its variance does not go to zero as  $n$  increases. This matrix weighted average of  $n$  inconsistent estimators will also be inconsistent. (This analysis is only heuristic. If  $T < K + 1$ , then the individual coefficient vectors cannot be computed.<sup>13)</sup>

---

<sup>13</sup>Further discussion is given by Nickell (1981), Ridder and Wansbeek (1990), and Kiviet (1995).

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 497

The problem is more transparent in the random effects model. In the model

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \delta y_{i,t-1} + u_i + \varepsilon_{it},$$

the lagged dependent variable is correlated with the compound disturbance in the model, since the same  $u_i$  enters the equation for every observation in group  $i$ .

Neither of these results renders the model inestimable, but they do make necessary some technique other than our familiar LSDV or FGLS estimators. The general approach, which has been developed in several stages in the literature,<sup>14</sup> relies on instrumental variables estimators and, most recently [by **Arellano and Bond** (1991) and **Arellano and Bover** (1995)] on a GMM estimator. For example, in either the fixed or random effects cases, the heterogeneity can be swept from the model by taking first differences, which produces

$$y_{it} - y_{i,t-1} = (\mathbf{x}'_{it} - \mathbf{x}'_{i,t-1})'\boldsymbol{\beta} + \delta(y_{i,t-1} - y_{i,t-2}) + (\varepsilon_{it} - \varepsilon_{i,t-1}).$$

This model is still complicated by correlation between the lagged dependent variable and the disturbance (and by its first-order moving average disturbance). But without the group effects, there is a simple instrumental variables estimator available. Assuming that the time series is long enough, one could use the lagged differences,  $(y_{i,t-2} - y_{i,t-3})$ , or the lagged levels,  $y_{i,t-2}$  and  $y_{i,t-3}$ , as one or two instrumental variables for  $(y_{i,t-1} - y_{i,t-2})$ . (The other variables can serve as their own instruments.) This is the Anderson and Hsiao estimator developed for this model in Section 11.8.2. By this construction, then, the treatment of this model is a standard application of the instrumental variables technique that we developed in Section 11.8.<sup>15</sup> This illustrates the flavor of an instrumental variable approach to estimation. But, as Arellano et al. and Ahn and Schmidt (1995) have shown, there is still more information in the sample that can be brought to bear on estimation, in the context of a GMM estimator, which we now consider.

We can extend the Hausman and Taylor (HT) formulation of the random effects model in Section 11.7.5 to include the lagged dependent variable;

$$\begin{aligned} y_{it} &= \delta y_{i,t-1} + \mathbf{x}'_{1it}\boldsymbol{\beta}_1 + \mathbf{x}'_{2it}\boldsymbol{\beta}_2 + \mathbf{z}'_{1i}\boldsymbol{\alpha}_1 + \mathbf{z}'_{2i}\boldsymbol{\alpha}_2 + \varepsilon_{it} + \mathbf{u}_i \\ &= \boldsymbol{\theta}'\mathbf{w}_{it} + \varepsilon_{it} + \mathbf{u}_i \\ &= \boldsymbol{\theta}'\mathbf{w}_{it} + \eta_{it}, \end{aligned}$$

where

$$\mathbf{w}_{it} = [y_{i,t-1}, \mathbf{x}'_{1it}, \mathbf{x}'_{2it}, \mathbf{z}'_{1i}, \mathbf{z}'_{2i}]'$$

is now a  $(1+K_1+K_2+L_1+L_2)\times 1$  vector. The terms in the equation are the same as in the Hausman and Taylor model. Instrumental variables estimation of the model without the lagged dependent variable is discussed in Section 11.7.5 on the HT estimator. Moreover, by just including  $y_{i,t-1}$  in  $\mathbf{x}_{2it}$ , we see that the HT approach extends to this setting as well, essentially without modification. Arellano et al. suggest a GMM estimator and show that efficiency gains are available by using a larger set of moment conditions.

<sup>14</sup>The model was first proposed in this form by Balestra and Nerlove (1966). See, for example, Anderson and Hsiao (1981, 1982), Bhargava and Sargan (1983), Arellano (1989), Arellano and Bond (1991), Arellano and Bover (1995), Ahn and Schmidt (1995), and Nerlove (2003).

<sup>15</sup>There is a question as to whether one should use differences or levels as instruments. Arellano (1989) and Kiviet (1995) give evidence that the latter is preferable.

### 498 PART III ♦ Estimation Methodology

In the previous treatment, we used a GMM estimator constructed as follows: The set of moment conditions we used to formulate the instrumental variables were

$$E \left[ \begin{pmatrix} \mathbf{x}_{1it} \\ \mathbf{x}_{2it} \\ \mathbf{z}_{1i} \\ \bar{\mathbf{x}}_{1i.} \end{pmatrix} (\eta_{it} - \bar{\eta}_i) \right] = E \left[ \begin{pmatrix} \mathbf{x}_{1it} \\ \mathbf{x}_{2it} \\ \mathbf{z}_{1i} \\ \bar{\mathbf{x}}_{1i.} \end{pmatrix} (\varepsilon_{it} - \bar{\varepsilon}_i) \right] = \mathbf{0}.$$

This moment condition is used to produce the instrumental variable estimator. We could ignore the nonscalar variance of  $\eta_{it}$  and use simple instrumental variables at this point. However, by accounting for the random effects formulation and using the counterpart to feasible GLS, we obtain the more efficient estimator in Section 11.8. As usual, this can be done in two steps. The inefficient estimator is computed to obtain the residuals needed to estimate the variance components. This is Hausman and Taylor's steps 1 and 2. Steps 3 and 4 are the GMM estimator based on these estimated variance components.

Arellano et al. suggest that the preceding does not exploit all the information in the sample. In simple terms, within the  $T$  observations in group  $i$ , we have not used the fact that

$$E \left[ \begin{pmatrix} \mathbf{x}_{1it} \\ \mathbf{x}_{2it} \\ \mathbf{z}_{1i} \\ \bar{\mathbf{x}}_{1i.} \end{pmatrix} (\eta_{is} - \bar{\eta}_i) \right] = \mathbf{0} \quad \text{for some } s \neq t.$$

Thus, for example, not only are disturbances at time  $t$  uncorrelated with these variables at time  $t$ , arguably, they are uncorrelated with the same variables at time  $t-1, t-2$ , possibly  $t+1$ , and so on. In principle, the number of valid instruments is potentially enormous. Suppose, for example, that the set of instruments listed above is strictly exogenous with respect to  $\eta_{it}$  in every period including current, lagged, and future. Then, there are [T(K<sub>1</sub> + K<sub>2</sub>) + L<sub>1</sub> + K<sub>1</sub>] moment conditions for every observation. █

~~On this basis alone.~~ Consider, for example, a panel with two periods. We would have for the two periods,

$$E \left[ \begin{pmatrix} \mathbf{x}_{1i1} \\ \mathbf{x}_{2i1} \\ \mathbf{x}_{1i2} \\ \mathbf{x}_{2i2} \\ \mathbf{z}_{1i} \\ \bar{\mathbf{x}}_{1i.} \end{pmatrix} (\eta_{i1} - \bar{\eta}_i) \right] = \mathbf{0} \quad \text{and} \quad E \left[ \begin{pmatrix} \mathbf{x}_{1i1} \\ \mathbf{x}_{2i1} \\ \mathbf{x}_{1i2} \\ \mathbf{x}_{2i2} \\ \mathbf{z}_{1i} \\ \bar{\mathbf{x}}_{1i.} \end{pmatrix} (\eta_{i2} - \bar{\eta}_i) \right] = \mathbf{0}. \quad (13-38)$$

How much useful information is brought to bear on estimation of the parameters is uncertain, as it depends on the correlation of the instruments with the included exogenous variables in the equation. The farther apart in time these sets of variables become the less information is likely to be present. (The literature on this subject contains reference to “strong” versus “weak” instrumental variables.<sup>16</sup>) To proceed, as noted, we can include the lagged dependent variable in  $\mathbf{x}_{2i}$ . This set of instrumental variables can

<sup>16</sup>See West (2001).

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 499

be used to construct the estimator, actually whether the lagged variable is present or not. We note, at this point, that on this basis, Hausman and Taylor's estimator did not actually use all the information available in the sample. We now have the elements of the Arellano et al. estimator in hand; what remains is essentially the (unfortunately, fairly involved) algebra, which we now develop.

Let

$$\mathbf{W}_i = \begin{bmatrix} \mathbf{w}'_{i1} \\ \mathbf{w}'_{i2} \\ \vdots \\ \mathbf{w}'_{iT} \end{bmatrix} = \text{the full set of rhs data for group } i, \quad \text{and} \quad \mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix}.$$

Note that  $\mathbf{W}_i$  is assumed to be, a  $T \times (1 + K_1 + K_2 + L_1 + L_2)$  matrix. Because there is a lagged dependent variable in the model, it must be assumed that there are actually  $T + 1$  observations available on  $y_{it}$ . To avoid a cumbersome, cluttered notation, we will leave this distinction embedded in the notation for the moment. Later, when necessary, we will make it explicit. It will reappear in the formulation of the instrumental variables. A total of  $T$  observations will be available for constructing the IV estimators. We now form a matrix of instrumental variables. [Different approaches to this have been considered by Hausman and Taylor (1981), Arellano et al. (1991, 1995, 1999), Ahn and Schmidt (1995) and Amemiya and MacCurdy (1986), among others.] We will form a matrix  $\mathbf{V}_i$  consisting of  $T_i - 1$  rows constructed the same way for  $T_i - 1$  observations and a final row that will be different, as discussed later. [This is to exploit a useful algebraic result discussed by Arellano and Bover (1995).] The matrix will be of the form

$$\mathbf{V}_i = \begin{bmatrix} \mathbf{v}'_{i1} & \mathbf{0}' & \cdots & \mathbf{0}' \\ \mathbf{0}' & \mathbf{v}'_{i2} & \cdots & \mathbf{0}' \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \cdots & \mathbf{a}'_i \end{bmatrix}. \quad (13-39)$$

The instrumental variable sets contained in  $\mathbf{v}'_{it}$  which have been suggested might include the following from within the model:

- $\mathbf{x}_{it}$  and  $\mathbf{x}_{i,t-1}$  (i.e., current and one lag of all the time varying variables),
- $\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}$  (i.e., all current, past and future values of all the time varying variables),
- $\mathbf{x}_{i1}, \dots, \mathbf{x}_{it}$  (i.e., all current and past values of all the time varying variables).

The time-invariant variables that are uncorrelated with  $u_i$ , that is  $\mathbf{z}_{1i}$ , are appended at the end of the nonzero part of each of the first  $T - 1$  rows. It may seem that including  $\mathbf{x}_2$  in the instruments would be invalid. However, we will be converting the disturbances to deviations from group means which are free of the latent effects—that is, this set of moment conditions will ultimately be converted to what appears in (13-38). While the variables are correlated with  $u_i$  by construction, they are not correlated with  $\varepsilon_{it} - \bar{\varepsilon}_i$ . The final row of  $\mathbf{V}_i$  is important to the construction. Two possibilities have been suggested:

- $\mathbf{a}'_i = [\mathbf{z}'_{1i} \quad \bar{\mathbf{x}}_{i1}]$  (produces the Hausman and Taylor estimator),
- $\mathbf{a}'_i = [\mathbf{z}'_{1i} \quad \mathbf{x}'_{1i1}, \mathbf{x}'_{1i2}, \dots, \mathbf{x}'_{1iT}]$  (produces Amemiya and MacCurdy's estimator).

## 500 PART III ♦ Estimation Methodology

Note that the  $\mathbf{a}$  variables are exogenous time-invariant variables,  $\mathbf{z}_{1i}$  and the exogenous time-varying variables, either condensed into the single group mean or in the raw form, with the full set of  $T$  observations.

To construct the estimator, we will require a transformation matrix,  $\mathbf{H}$ , constructed as follows. Let  $\mathbf{M}^{01}$  denote the first  $T - 1$  rows of  $\mathbf{M}^0$ , the matrix that creates deviations from group means. Then,

$$\mathbf{H} = \begin{bmatrix} \mathbf{M}^{01} \\ \frac{1}{T} \mathbf{i}'_T \end{bmatrix}.$$

Thus,  $\mathbf{H}$  replaces the last row of  $\mathbf{M}^0$  with a row of  $1/T$ . The effect is as follows: if  $\mathbf{q}$  is  $T$  observations on a variable, then  $\mathbf{H}\mathbf{q}$  produces  $\mathbf{q}^*$  in which the first  $T - 1$  observations are converted to deviations from group means and the last observation is the group mean. In particular, let the  $T \times 1$  column vector of disturbances

$$\boldsymbol{\eta}_i = [\eta_{i1}, \eta_{i2}, \dots, \eta_{iT}] = [(\varepsilon_{i1} + u_i), (\varepsilon_{i2} + u_i), \dots, (\varepsilon_{iT} + u_i)]',$$

then

$$\mathbf{H}\boldsymbol{\eta} = \begin{bmatrix} \eta_{i1} - \bar{\eta}_i \\ \vdots \\ \eta_{iT-1} - \bar{\eta}_i \\ \bar{\eta}_i \end{bmatrix}.$$

We can now construct the moment conditions. With all this machinery in place, we have the result that appears in (13-40), that is

$$E[\mathbf{V}'_i \mathbf{H}\boldsymbol{\eta}_i] = E[\mathbf{g}_i] = \mathbf{0}.$$

It is useful to expand this for a particular case. Suppose  $T = 3$  and we use as instruments the current values in period 1, and the current and previous values in period 2 and the Hausman and Taylor form for the invariant variables. Then the preceding is

$$E \left[ \begin{pmatrix} \mathbf{x}_{1i1} & \mathbf{0} & \mathbf{0} \\ \mathbf{x}_{2i1} & \mathbf{0} & \mathbf{0} \\ \mathbf{z}_{1i} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{1i1} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2i1} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{1i2} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2i2} & \mathbf{0} \\ \mathbf{0} & \mathbf{z}_{1i} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{z}_{1i} \\ \mathbf{0} & \mathbf{0} & \bar{\mathbf{x}}_{1i} \end{pmatrix} \begin{pmatrix} \eta_{i1} - \bar{\eta}_i \\ \eta_{i2} - \bar{\eta}_i \\ \bar{\eta}_i \end{pmatrix} \right] = \mathbf{0}. \quad (13-40)$$

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 501

This is the same as (13-38).<sup>17</sup> The empirical moment condition that follows from this is

$$\begin{aligned} & \text{plim} \frac{1}{n} \sum_{i=1}^n \mathbf{V}_i' \mathbf{H} \boldsymbol{\eta}_i \\ &= \text{plim} \frac{1}{n} \sum_{i=1}^n \mathbf{V}_i' \mathbf{H} \begin{pmatrix} y_{i1} - \delta y_{i0} - \mathbf{x}'_{1i1} \boldsymbol{\beta}_1 - \mathbf{x}'_{2i1} \boldsymbol{\beta}_2 - \mathbf{z}'_{1i} \boldsymbol{\alpha}_1 - \mathbf{z}'_{2i} \boldsymbol{\alpha}_2 \\ y_{i2} - \delta y_{i1} - \mathbf{x}'_{1i2} \boldsymbol{\beta}_1 - \mathbf{x}'_{2i2} \boldsymbol{\beta}_2 - \mathbf{z}'_{1i} \boldsymbol{\alpha}_1 - \mathbf{z}'_{2i} \boldsymbol{\alpha}_2 \\ \vdots \\ y_{iT} - \delta y_{i,T-1} - \mathbf{x}'_{1iT} \boldsymbol{\beta}_1 - \mathbf{x}'_{2iT} \boldsymbol{\beta}_2 - \mathbf{z}'_{1i} \boldsymbol{\alpha}_1 - \mathbf{z}'_{2i} \boldsymbol{\alpha}_2 \end{pmatrix} = \mathbf{0}. \end{aligned}$$

Write this as

$$\text{plim} \frac{1}{n} \sum_{i=1}^n \mathbf{m}_i = \text{plim} \bar{\mathbf{m}} = \mathbf{0}.$$

The GMM estimator  $\hat{\boldsymbol{\theta}}$  is then obtained by minimizing

$$q = \bar{\mathbf{m}}' \mathbf{A} \bar{\mathbf{m}}$$

with an appropriate choice of the weighting matrix,  $\mathbf{A}$ . The optimal weighting matrix will be the inverse of the asymptotic covariance matrix of  $\sqrt{n} \bar{\mathbf{m}}$ . With a consistent estimator of  $\boldsymbol{\theta}$  in hand, this can be estimated empirically using

$$\text{Est. Asy. Var}[\sqrt{n} \bar{\mathbf{m}}] = \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{m}}_i \hat{\mathbf{m}}_i' = \frac{1}{n} \sum_{i=1}^n \mathbf{V}_i' \mathbf{H} \hat{\boldsymbol{\eta}}_i \hat{\boldsymbol{\eta}}_i' \mathbf{H}' \mathbf{V}_i.$$

This is a robust estimator that allows an unrestricted  $T \times T$  covariance matrix for the  $T$  distances,  $\varepsilon_{it} + u_i$ . But, we have assumed that this covariance matrix is the  $\Sigma$  defined in (9Z8) for the random effects model. To use this information we would, instead, use the residuals in

$$\hat{\boldsymbol{\eta}}_i = \mathbf{y}_i - \mathbf{W}_i \hat{\boldsymbol{\theta}}$$

to estimate  $\sigma_u^2$  and  $\sigma_\varepsilon^2$  and then  $\Sigma$ , which produces

$$\text{Est. Asy. Var}[\sqrt{n} \bar{\mathbf{m}}] = \frac{1}{n} \sum_{i=1}^n \mathbf{V}_i' \mathbf{H} \hat{\Sigma} \mathbf{H}' \mathbf{V}_i.$$

We now have the full set of results needed to compute the GMM estimator. The solution to the optimization problem of minimizing  $q$  with respect to the parameter vector  $\boldsymbol{\theta}$  is

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{GMM} &= \left[ \left( \sum_{i=1}^n \mathbf{W}_i' \mathbf{H} \mathbf{V}_i \right) \left( \sum_{i=1}^n \mathbf{V}_i' \mathbf{H}' \hat{\Sigma} \mathbf{H} \mathbf{V}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{V}_i' \mathbf{H}' \mathbf{W}_i \right) \right]^{-1} \\ &\quad \times \left( \sum_{i=1}^n \mathbf{W}_i' \mathbf{H} \mathbf{V}_i \right) \left( \sum_{i=1}^n \mathbf{V}_i' \mathbf{H}' \hat{\Sigma} \mathbf{H} \mathbf{V}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{V}_i' \mathbf{H}' \mathbf{y}_i \right). \end{aligned} \quad (13-41)$$

The estimator of the asymptotic covariance matrix for  $\hat{\boldsymbol{\theta}}_{GMM}$  is the inverse matrix in brackets.

<sup>17</sup>In some treatments [e.g., Blundell and Bond (1998)], an additional condition is assumed for the initial value,  $y_{i0}$ , namely  $E[y_{i0} | \text{exogenous data}] = \mu_0$ . This would add a row at the top of the matrix in (13-40) containing  $[y_{i0} - \mu_0, 0, 0]$ .

## 502 PART III ♦ Estimation Methodology

The remaining loose end is how to obtain the consistent estimator of  $\hat{\theta}$  to compute  $\Sigma$ . Recall that the GMM estimator is consistent with any positive definite weighting matrix,  $\mathbf{A}$ , in our preceding expression. Therefore, for an initial estimator, we could set  $\mathbf{A} = \mathbf{I}$  and use the simple instrumental variables estimator,

$$\hat{\theta}_{IV} = \left[ \left( \sum_{i=1}^n \mathbf{W}'_i \mathbf{H} \mathbf{V}_i \right) \left( \sum_{i=1}^n \mathbf{V}'_i \mathbf{H} \mathbf{W}_i \right) \right]^{-1} \left( \sum_{i=1}^n \mathbf{W}'_i \mathbf{H} \mathbf{V}_i \right) \left( \sum_{i=1}^n \mathbf{V}'_i \mathbf{H} \mathbf{y}_i \right).$$

It is more common to  refer directly to the “two-stage least squares” estimator (see Sections 8.3.4 and 11.1), which uses

$$\mathbf{A} = \left( \frac{1}{n} \sum_{i=1}^n \mathbf{V}'_i \mathbf{H}' \mathbf{H} \mathbf{V}_i \right)^{-1}.$$

The estimator is, then, the one given earlier in (13-41) with  $\hat{\Sigma}$  replaced by  $\mathbf{I}_T$ . Either estimator is a function of the sample data only and provides the initial estimator we need.

Ahn and Schmidt (among others) observed that the IV estimator proposed here, as extensive as it is, still neglects quite a lot of information and is therefore (relatively) inefficient. For example, in the first differenced model,

$$E[y_{is}(\varepsilon_{it} - \varepsilon_{i,t-1})] = 0, \quad s = 0, \dots, t-2, \quad t = 2, \dots, T.$$

That is, the *level* of  $y_{is}$  is uncorrelated with the differences of disturbances that are at least two periods subsequent.<sup>18</sup> (The differencing transformation, as the transformation to deviations from group means, removes the individual effect.) The corresponding moment equations that can enter the construction of a GMM estimator are

$$\frac{1}{n} \sum_{i=1}^n y_{is} [(y_{it} - y_{i,t-1}) - \delta(y_{i,t-1} - y_{i,t-2}) - (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})' \boldsymbol{\beta}] = 0 \\ s = 0, \dots, t-2, \quad t = 2, \dots, T.$$

Altogether, Ahn and Schmidt identify  $T(T-1)/2 + T-2$  such equations that involve mixtures of the levels and differences of the variables. The main conclusion that they demonstrate is that in the dynamic model, there is a large amount of information to be gleaned not only from the familiar relationships among the levels of the variables, but also from the implied relationships between the levels and the first differences. The issue of correlation between the transformed  $y_{it}$  and the deviations of  $\varepsilon_{it}$  is discussed in the papers cited. [As Ahn and Schmidt show, there are potentially huge numbers of additional orthogonality conditions in this model owing to the relationship between first differences and second moments. We do not consider those. The matrix  $\mathbf{V}_i$  could be huge. Consider a model with 10 time-varying right-hand-side variables and suppose  $T_i$  is 15. Then, there are 15 rows and roughly  $15 \times (10 \times 15)$  or 2,250 columns. The Ahn and Schmidt estimator, which involves potentially thousands of instruments in a model containing only a handful of parameters, may become a bit impractical at this point. The common approach is to use only a small subset of the available instrumental

<sup>18</sup>This is the approach suggested by Holtz-Eakin (1988) and Holtz-Eakin, Newey, and Rosen (1988).

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 503

variables. The order of the computation grows as the number of parameters times the square of  $T$ .]

The number of orthogonality conditions (instrumental variables) used to estimate the parameters of the model is determined by the number of variables in  $\mathbf{v}_{it}$  and  $\mathbf{a}_i$  in (13-39). In most cases, the model is vastly overidentified—there are far more orthogonality conditions than parameters. As usual in GMM estimation, a test of the overidentifying restrictions can be based on  $q$ , the estimation criterion. At its minimum, the limiting distribution of  $nq$  is chi-squared with degrees of freedom equal to the number of instrumental variables in total minus  $(1 + K_1 + K_2 + L_1 + L_2)$ .<sup>19</sup>

### **Example 13.10 GMM Estimation of a Dynamic Panel Data Model of Local Government Expenditures**

Dahlberg and Johansson (2000) estimated a model for the local government expenditure of several hundred municipalities in Sweden observed over the nine-year period  $t = 1979$  to 1987. The equation of interest is

$$S_{i,t} = \alpha_t + \sum_{j=1}^m \beta_j S_{i,t-j} + \sum_{j=1}^m \gamma_j R_{i,t-j} + \sum_{j=1}^m \delta_j G_{i,t-j} + f_i + \varepsilon_{it},$$

for  $i = 1, \dots, n = 265$ , and  $t = m+1, \dots, 9$ . (We have changed their notation slightly to make it more convenient.)  $S_{i,t}$ ,  $R_{i,t}$ , and  $G_{i,t}$  are municipal spending, receipts (taxes and fees), and central government grants, respectively. Analogous equations are specified for the current values of  $R_{i,t}$  and  $G_{i,t}$ . The appropriate lag length,  $m$ , is one of the features of interest to be determined by the empirical study. The model contains a municipality specific effect,  $f_i$ , which is not specified as being either “fixed” or “random.” To eliminate the individual effect, the model is converted to first differences. The resulting equation is

$$\Delta S_{i,t} = \lambda_t + \sum_{j=1}^m \beta_j \Delta S_{i,t-j} + \sum_{j=1}^m \gamma_j \Delta R_{i,t-j} + \sum_{j=1}^m \delta_j \Delta G_{i,t-j} + u_{it},$$

or

$$y_{i,t} = \mathbf{x}'_{i,t} \boldsymbol{\theta} + u_{it},$$

where  $\Delta S_{i,t} = S_{i,t} - S_{i,t-1}$  and so on and  $u_{i,t} = \varepsilon_{i,t} - \varepsilon_{i,t-1}$ . This removes the group effect and leaves the time effect. Because the time effect was unrestricted to begin with,  $\Delta \alpha_t = \lambda_t$  remains an unrestricted time effect, which is treated as “fixed” and modeled with a time-specific dummy variable. The maximum lag length is set at  $m = 3$ . With nine years of data, this leaves usable observations from 1983 to 1987 for estimation, that is,  $t = m+2, \dots, 9$ . Similar equations were fit for  $R_{i,t}$  and  $G_{i,t}$ .

The orthogonality conditions claimed by the authors are

$$E[S_{i,s}u_{i,t}] = E[R_{i,s}u_{i,t}] = E[G_{i,s}u_{i,t}] = 0, \quad s = 1, \dots, t-2.$$

The orthogonality conditions are stated in terms of the levels of the financial variables and the differences of the disturbances. The issue of this formulation as opposed to, for example,  $E[\Delta S_{i,s}\Delta \varepsilon_{i,t}] = 0$  (which is implied) is discussed by Ahn and Schmidt (1995). As we shall see, this set of orthogonality conditions implies a total of 80 instrumental variables. The authors use only the first of the three sets listed, which produces a total of 30. For the five observations, using the formulation developed in Section 13.6.5, we have the following matrix

---

<sup>19</sup>This is true generally in GMM estimation. It was proposed for the dynamic panel data model by Bhargava and Sargan (1983).

## 504 PART III ♦ Estimation Methodology

of instrumental variables for the orthogonality conditions

$$\mathbf{Z}_i = \begin{bmatrix} S_{81-79} & d_{83} & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & 0' & 1983 \\ \mathbf{0}' & 0 & S_{82-79} & d_{84} & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & 0' & 1984 \\ \mathbf{0}' & 0 & \mathbf{0}' & 0 & S_{83-79} & d_{85} & \mathbf{0}' & 0 & \mathbf{0}' & 0 & 0' & 1985 \\ \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & S_{84-79} & d_{86} & \mathbf{0}' & 0 & 0' & 1986 \\ \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & S_{85-79} & d_{87} & 0' & 1987 \end{bmatrix}$$

where the notation  $S_{t_1-t_0}$  indicates the range of years for that variable. For example,  $S_{83-79}$  denotes  $[S_{i,1983}, S_{i,1982}, S_{i,1981}, S_{i,1980}, S_{i,1979}]$  and  $d_{\text{year}}$  denotes the year-specific dummy variable. Counting columns in  $\mathbf{Z}_i$  we see that using only the lagged values of the dependent variable and the time dummy variables, we have  $(3+1)+(4+1)+(5+1)+(6+1)+(7+1)=30$  instrumental variables. Using the lagged values of the other two variables in each equation would add 50 more, for a total of 80 if all the orthogonality conditions suggested earlier were employed. Given the preceding construction, the orthogonality conditions are now

$$E[\mathbf{Z}_i' \mathbf{u}_i] = \mathbf{0},$$

where  $\mathbf{u}_i = [u_{i,1983}, u_{i,1984}, u_{i,1985}, u_{i,1986}, u_{i,1987}]'$ . The empirical moment equation is

$$\text{plim} \left[ \frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i' \mathbf{u}_i \right] = \text{plim} \bar{\mathbf{m}}(\theta) = \mathbf{0}.$$

The parameters are vastly overidentified. Using only the lagged values of the dependent variable in each of the three equations estimated, there are 30 moment conditions and 14 parameters being estimated when  $m = 3$ , 11 when  $m = 2$ , 8 when  $m = 1$ , and 5 when  $m = 0$ . (As we do our estimation of each of these, we will retain the same matrix of instrumental variables in each case.) GMM estimation proceeds in two steps. In the first step, basic, unweighted instrumental variables is computed using

$$\hat{\theta}'_{IV} = \left[ \left( \sum_{i=1}^n \mathbf{X}_i' \mathbf{Z}_i \right) \left( \sum_{i=1}^n \mathbf{Z}_i' \mathbf{Z}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{Z}_i' \mathbf{X}_i \right) \right]^{-1} \left( \sum_{i=1}^n \mathbf{X}_i' \mathbf{z}_i \right) \left( \sum_{i=1}^n \mathbf{Z}_i' \mathbf{z}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{Z}_i' \mathbf{y}_i \right),$$

where

$$\mathbf{y}_i' = (\Delta S_{83} \quad \Delta S_{84} \quad \Delta S_{85} \quad \Delta S_{86} \quad \Delta S_{87}),$$

and

$$\mathbf{X}_i = \begin{bmatrix} \Delta S_{82} & \Delta S_{81} & \Delta S_{80} & \Delta R_{82} & \Delta R_{81} & \Delta R_{80} & \Delta G_{82} & \Delta G_{81} & \Delta G_{80} & 1 & 0 & 0 & 0 & 0 \\ \Delta S_{83} & \Delta S_{82} & \Delta S_{81} & \Delta R_{83} & \Delta R_{82} & \Delta R_{81} & \Delta G_{83} & \Delta G_{82} & \Delta G_{81} & 0 & 1 & 0 & 0 & 0 \\ \Delta S_{84} & \Delta S_{83} & \Delta S_{82} & \Delta R_{84} & \Delta R_{83} & \Delta R_{82} & \Delta G_{84} & \Delta G_{83} & \Delta G_{82} & 0 & 0 & 1 & 0 & 0 \\ \Delta S_{85} & \Delta S_{84} & \Delta S_{83} & \Delta R_{85} & \Delta R_{84} & \Delta R_{83} & \Delta G_{85} & \Delta G_{84} & \Delta G_{83} & 0 & 0 & 0 & 1 & 0 \\ \Delta S_{86} & \Delta S_{85} & \Delta S_{84} & \Delta R_{86} & \Delta R_{85} & \Delta R_{84} & \Delta G_{86} & \Delta G_{85} & \Delta G_{84} & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The second step begins with the computation of the new weighting matrix,

$$\hat{\Phi} = \text{Est. Asy. Var}[\sqrt{n} \bar{\mathbf{m}}] = \frac{1}{N} \sum_{i=1}^n \mathbf{Z}_i' \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' \mathbf{Z}_i.$$

## CHAPTER 13 ♦ Minimum Distance Estimation and GMM 505

**TABLE 13.3** Descriptive Statistics for Local Expenditure Data

<i>Variable</i>	<i>Mean</i>	<i>Std. Deviation</i>	<i>Minimum</i>	<i>Maximum</i>
Spending	18478.51	3174.36	12225.68	33883.25
Revenues	13422.56	3004.16	6228.54	29141.62
Grants	5236.03	1260.97	1570.64	12589.14

After multiplying and dividing by the implicit  $(1/n)$  in the outside matrices, we obtain the estimator,

$$\begin{aligned}\theta'_{GMM} &= \left[ \left( \sum_{i=1}^n \mathbf{X}'_i \mathbf{Z}_i \right) \left( \sum_{i=1}^n \mathbf{Z}'_i \hat{\mathbf{u}}_i \hat{\mathbf{u}}'_i \mathbf{Z}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{Z}'_i \mathbf{X}_i \right) \right]^{-1} \\ &\quad \times \left( \sum_{i=1}^n \mathbf{X}'_i \mathbf{Z}_i \right) \left( \sum_{i=1}^n \mathbf{Z}'_i \hat{\mathbf{u}}_i \hat{\mathbf{u}}'_i \mathbf{Z}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{Z}'_i \mathbf{y}_i \right) \\ &= \left[ \left( \sum_{i=1}^n \mathbf{X}'_i \mathbf{Z}_i \right) \mathbf{W} \left( \sum_{i=1}^n \mathbf{Z}'_i \mathbf{X}_i \right) \right]^{-1} \left( \sum_{i=1}^n \mathbf{X}'_i \mathbf{Z}_i \right) \mathbf{W} \left( \sum_{i=1}^n \mathbf{Z}'_i \mathbf{y}_i \right).\end{aligned}$$

The estimator of the asymptotic covariance matrix for the estimator is the inverse matrix in square brackets in the first line of the result.

The primary focus of interest in the study was not the estimator itself, but the lag length and whether certain lagged values of the independent variables appeared in each equation. These restrictions would be tested by using the GMM criterion function, which in this formulation would be (based on recomputing the residuals after GMM estimation)

$$nq = \left( \sum_{i=1}^n \hat{\mathbf{u}}'_i \mathbf{Z}_i \right) \mathbf{W} \left( \sum_{i=1}^n \mathbf{Z}'_i \hat{\mathbf{u}}_i \right).$$

Note that the weighting matrix is not (necessarily) recomputed. For purposes of testing hypotheses, the same weighting matrix should be used.

At this point, we will consider the appropriate lag length,  $m$ . The specification can be reduced simply by redefining  $\mathbf{X}$  to change the lag length. To test the specification, the weighting matrix must be kept constant for all restricted versions ( $m = 2$  and  $m = 1$ ) of the model.

The Dahlberg and Johansson data may be downloaded from the *Journal of Applied Econometrics* Web site—see Appendix Table F13.1. The authors provide the summary statistics for the raw data that are given in Table 13.3. The data used in the study and provided in the internet source are nominal values in Swedish kroner, deflated by a municipality-specific price index then converted to per capita values. Descriptive statistics for the raw data appear in Table 13.3.<sup>20</sup> Equations were estimated for all three variables, with maximum lag lengths of  $m = 1$ , 2, and 3. (The authors did not provide the actual estimates.) Estimation is done using the methods developed by Ahn and Schmidt (1995), Arellano and Bover (1995), and Holtz-Eakin, Newey, and Rosen (1988), as described. The estimates of the first specification provided are given in Table 13.4.

Table 13.5 contains estimates of the model parameters for each of the three equations, and for the three lag lengths, as well as the value of the GMM criterion function for each model estimated. The base case for each model has  $m = 3$ . There are three restrictions implied by each reduction in the lag length. The critical chi-squared value for three degrees of freedom is 7.81 for 95 percent significance, so at this level, we find that the two-level model is just

<sup>20</sup>The data provided on the web site and used in our computations were further transformed by dividing by 100,000.

**506 PART III ♦ Estimation Methodology**
**TABLE 13.4** Estimated Spending Equation

<i>Variable</i>	<i>Estimate</i>	<i>Standard Error</i>	<i>t Ratio</i>
Year 1983	-0.0036578	0.0002969	-12.32
Year 1984	-0.00049670	0.0004128	-1.20
Year 1985	0.00038085	0.0003094	1.23
Year 1986	0.00031469	0.0003282	0.96
Year 1987	0.00086878	0.0001480	5.87
Spending ( $t - 1$ )	1.15493	0.34409	3.36
Revenues ( $t - 1$ )	-1.23801	0.36171	-3.42
Grants ( $t - 1$ )	0.016310	0.82419	0.02
Spending ( $t - 2$ )	-0.0376625	0.22676	-0.17
Revenues ( $t - 2$ )	0.0770075	0.27179	0.28
Grants ( $t - 2$ )	1.55379	0.75841	2.05
Spending ( $t - 3$ )	-0.56441	0.21796	-2.59
Revenues ( $t - 3$ )	0.64978	0.26930	2.41
Grants ( $t - 3$ )	1.78918	0.69297	2.58

**TABLE 13.5** Estimated Lag Equations for Spending, Revenue, and Grants

	<i>Expenditure Model</i>			<i>Revenue Model</i>			<i>Grant Model</i>		
	<i>m = 3</i>	<i>m = 2</i>	<i>m = 1</i>	<i>m = 3</i>	<i>m = 2</i>	<i>m = 1</i>	<i>m = 3</i>	<i>m = 2</i>	<i>m = 1</i>
$S_{t-1}$	1.155	0.8742	0.5562	-0.1715	-0.3117	-0.1242	-0.1675	-0.1461	-0.1958
$S_{t-2}$	-0.0377	0.2493	—	0.1621	-0.0773	—	-0.0303	-0.0304	—
$S_{t-3}$	-0.5644	—	—	-0.1772	—	—	-0.0955	—	—
$R_{t-1}$	-1.2380	-0.8745	-0.5328	-0.0176	0.1863	-0.0245	0.1578	0.1453	0.2343
$R_{t-2}$	0.0770	-0.2776	—	-0.0309	0.1368	—	0.0485	0.0175	—
$R_{t-3}$	0.6497	—	—	0.0034	—	—	0.0319	—	—
$G_{t-1}$	0.0163	-0.4203	0.1275	-0.3683	0.5425	-0.0808	-0.2381	-0.2066	-0.0559
$G_{t-2}$	1.5538	0.1866	—	2.7152	2.4621	—	-0.0492	-0.0804	—
$G_{t-3}$	1.7892	—	—	0.0948	—	—	0.0598	—	—
$n_q$	22.8287	30.4526	34.4986	30.5398	34.2590	53.2506	17.5810	20.5416	27.5927

barely accepted for the spending equation, but clearly appropriate for the other two—the difference between the two criteria is 7.62. Conditioned on  $m = 2$ , only the revenue model rejects the restriction of  $m = 1$ . As a final test, we might ask whether the data suggest that perhaps no lag structure at all is necessary. The GMM criterion value for the three equations with only the time dummy variables are 45.840, 57.908, and 62.042, respectively. Therefore, all three zero lag models are rejected.

Among the interests in this study were the appropriate critical values to use for the specification test of the moment restriction. With 16 degrees of freedom, the critical chi-squared value for 95 percent significance is 26.3, which would suggest that the revenues equation is misspecified. Using a bootstrap technique, the authors find that a more appropriate critical value leaves the specification intact. Finally, note that the three-equation model in the  $m = 3$  columns of Table 13.5 imply a **vector autoregression** of the form

$$\mathbf{y}_t = \boldsymbol{\Gamma}_1 \mathbf{y}_{t-1} + \boldsymbol{\Gamma}_2 \mathbf{y}_{t-2} + \boldsymbol{\Gamma}_3 \mathbf{y}_{t-3} + \mathbf{v}_t,$$

where  $\mathbf{y}_t = (\Delta S_t, \Delta R_t, \Delta G_t)'$ . We will explore the properties and characteristics of equation systems such as this in our discussion of time-series models in Chapter 22.

**CHAPTER 13 ♦ Minimum Distance Estimation and GMM 507****13.7 SUMMARY AND CONCLUSIONS**

The generalized method of moments provides an estimation framework that includes least squares, nonlinear least squares, instrumental variables, and maximum likelihood, and a general class of estimators that extends beyond these. But it is more than just a theoretical umbrella. The GMM provides a method of formulating models and implied estimators without making strong distributional assumptions. Hall's model of household consumption is a useful example that shows how the optimization conditions of an underlying economic theory produce a set of distribution-free estimating equations. In this chapter, we first examined the classical method of moments. GMM as an estimator is an extension of this strategy that allows the analyst to use additional information beyond that necessary to identify the model, in an optimal fashion. After defining and establishing the properties of the estimator, we then turned to inference procedures. It is convenient that the GMM procedure provides counterparts to the familiar trio of test statistics: Wald, LM, and LR. In the final section, we specialized the GMM estimator for linear and nonlinear equations and multiple-equation models. We then developed an example that appears at many points in the recent applied literature, the dynamic panel data model with individual specific effects, and lagged values of the dependent variable.

**Key Terms and Concepts**

- Analog estimation
- Arellano and Bond
- Arellano and Bover estimator
- Central limit theorem
- Central moments
- Criterion function
- Dynamic panel data model
- Empirical moment equation
- Ergodic theorem
- Euler equation
- Exactly identified cases
- Exactly defined cases
- Exponential family
- Generalized method of moments
- GMM estimator
- H2SLS
- Identification
- Instrumental variables
- Likelihood ratio statistic
- LM statistic
- Martingale difference series
- Maximum likelihood estimator
- Mean value theorem
- Method of moment generating functions
- Method of moments
- Method of moments estimators
- Minimum distance estimator (MDE)
- Moment equation
- Newey-West estimator
- Nonlinear instrumental variable estimator
- Optimal weighting matrix
- Order condition
- Orthogonality conditions
- Overidentifying restrictions
- Overidentified cases
- Population moment equation
- Probability limit
- Random sample
- Rank condition
- Slutsky theorem
- Specification test
- Sufficient statistic
- Taylor series
- Uncentered moment
- Vector autoregression
- Wald statistic
- Weighted least squares
- Weighting matrix

**Exercises**

1. For the normal distribution  $\mu_{2k} = \sigma^{2k}(2k)!/(k!2^k)$  and  $\mu_{2k+1} = 0$ ,  $k = 0, 1, \dots$ . Use this result to analyze the two estimators

$$\sqrt{b_1} = \frac{m_3}{m_2^{3/2}} \quad \text{and} \quad b_2 = \frac{m_4}{m_2^2},$$

## 508 PART III ♦ Estimation Methodology

where  $m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$ . The following result will be useful:

$$\text{Asy. Cov}[\sqrt{n}m_j, \sqrt{n}m_k] = \mu_{j+k} - \mu_j \mu_k + jk\mu_2 \mu_{j-1} \mu_{k-1} - j\mu_{j-1} \mu_{k+1} - k\mu_{k-1} \mu_{j+1}.$$

Use the delta method to obtain the asymptotic variances and covariance of these two functions, assuming the data are drawn from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . (*Hint:* Under the assumptions, the sample mean is a consistent estimator of  $\mu$ , so for purposes of deriving asymptotic results, the difference between  $\bar{x}$  and  $\mu$  may be ignored. As such, no generality is lost by assuming the mean is zero, and proceeding from there.) Obtain  $\mathbf{V}$ , the  $3 \times 3$  covariance matrix for the three moments and then use the delta method to show that the covariance matrix for the two estimators is

$$\mathbf{J} \mathbf{V} \mathbf{J}' = \begin{bmatrix} 6/n & 0 \\ 0 & 24/n \end{bmatrix},$$

where  $\mathbf{J}$  is the  $2 \times 3$  matrix of derivatives.

2. Using the results in Example 13.7, estimate the asymptotic covariance matrix of the method of moments estimators of  $P$  and  $\lambda$  based on  $m'_1$  and  $m'_2$ . [Note: You will need to use the data in Example C.1 to estimate  $\mathbf{V}$ .]
3. **Exponential Families of Distributions.** For each of the following distributions, determine whether it is an exponential family by examining the log-likelihood function. Then, identify the sufficient statistics.
  - a. Normal distribution with mean  $\mu$  and variance  $\sigma^2$ .
  - b. The Weibull distribution in Exercise 4 in Chapter 14.
  - c. The mixture distribution in Exercise 3 in Chapter 14.
4. In the classical regression model with heteroscedasticity, which is more efficient, ordinary least squares or GMM? Obtain the two estimators and their respective asymptotic covariance matrices, then prove your assertion.
5. Consider the probit model analyzed in Chapter 17. The model states that for given vector of independent variables,

$$\text{Prob}[y_i = 1 | \mathbf{x}_i] = \Phi[\mathbf{x}'_i \boldsymbol{\beta}], \quad \text{Prob}[y_i = 0 | \mathbf{x}_i] = 1 - \text{Prob}[y_i = 1 | \mathbf{x}_i].$$

Consider a GMM estimator based on the result that

$$E[y_i | \mathbf{x}_i] = \Phi(\mathbf{x}'_i \boldsymbol{\beta}).$$

This suggests that we might base estimation on the orthogonality conditions

$$E[(y_i - \Phi(\mathbf{x}'_i \boldsymbol{\beta})) \mathbf{x}_i] = \mathbf{0}.$$

Construct a GMM estimator based on these results. Note that this is not the nonlinear least squares estimator. Explain—what would the orthogonality conditions be for nonlinear least squares estimation of this model?

6. Consider GMM estimation of a regression model as shown at the beginning of Example 13.8. Let  $\mathbf{W}_1$  be the optimal weighting matrix based on the moment equations. Let  $\mathbf{W}_2$  be some other positive definite matrix. Compare the asymptotic covariance matrices of the two proposed estimators. Show conclusively that the asymptotic covariance matrix of the estimator based on  $\mathbf{W}_1$  is not larger than that based on  $\mathbf{W}_2$ .

## 14

# MAXIMUM LIKELIHOOD ESTIMATION

---

## 14.1 INTRODUCTION

The generalized method of moments discussed in Chapter 13 and the semiparametric, nonparametric, and Bayesian estimators discussed in Chapters 12 and 16 are becoming widely used by model builders. Nonetheless, the maximum likelihood estimator discussed in this chapter remains the preferred estimator in many more settings than the others listed. As such, we focus our discussion of generally applied estimation methods on this technique. Sections 14.2 through 14.6 present basic statistical results for estimation and hypothesis testing based on the maximum likelihood principle. Sections 14.7 and 14.8 present two extensions of the method, two-step estimation and pseudo maximum likelihood estimation. After establishing the general results for this method of estimation, we will then apply them in the more familiar setting of econometric models. The applications presented in Section 14.9 apply the maximum likelihood method to most of the models in the preceding chapters and several others that illustrate different uses of the technique.

## 14.2 THE LIKELIHOOD FUNCTION AND IDENTIFICATION OF THE PARAMETERS

The probability density function, or pdf, for a random variable,  $y$ , conditioned on a set of parameters,  $\theta$ , is denoted  $f(y | \theta)$ .<sup>1</sup> This function identifies the data-generating process that underlies an observed sample of data and, at the same time, provides a mathematical description of the data that the process will produce. The joint density of  $n$  independent and identically distributed (i.i.d.) observations from this process is the product of the individual densities;

$$f(y_1, \dots, y_n | \theta) = \prod_{i=1}^n f(y_i | \theta) = L(\theta | \mathbf{y}). \quad (14-1)$$

This joint density is the **likelihood function**, defined as a function of the unknown parameter vector,  $\theta$ , where  $\mathbf{y}$  is used to indicate the collection of sample data. Note that we write the joint density as a function of the data conditioned on the parameters whereas when we form the likelihood function, we will write this function in reverse, as a function of the parameters, conditioned on the data. Though the two functions are the same, it is to be emphasized that the likelihood function is written in this fashion

---

<sup>1</sup>Later we will extend this to the case of a random vector,  $\mathbf{y}$ , with a multivariate density, but at this point, that would complicate the notation without adding anything of substance to the discussion.

## 510 PART III ♦ Estimation Methodology

to highlight our interest in the parameters and the information about them that is contained in the observed data. However, it is understood that the likelihood function is not meant to represent a probability density for the parameters as it is in Chapter IV. In this classical estimation framework, the parameters are assumed to be fixed constants that we hope to learn about from the data.

It is usually simpler to work with the log of the likelihood function:

$$\ln L(\boldsymbol{\theta} | \mathbf{y}) = \sum_{i=1}^n \ln f(y_i | \boldsymbol{\theta}). \quad (14-2)$$

Again, to emphasize our interest in the parameters, given the observed data, we denote this function  $L(\boldsymbol{\theta} | \text{data}) = L(\boldsymbol{\theta} | \mathbf{y})$ . The likelihood function and its logarithm, evaluated at  $\boldsymbol{\theta}$ , are sometimes denoted simply  $L(\boldsymbol{\theta})$  and  $\ln L(\boldsymbol{\theta})$ , respectively, or, where no ambiguity can arise, just  $L$  or  $\ln L$ .

It will usually be necessary to generalize the concept of the likelihood function to allow the density to depend on other conditioning variables. To jump immediately to one of our central applications, suppose the disturbance in the classical linear regression model is normally distributed. Then, conditioned on its specific  $\mathbf{x}_i$ ,  $y_i$  is normally distributed with mean  $\mu_i = \mathbf{x}'_i \boldsymbol{\beta}$  and variance  $\sigma^2$ . That means that the observed random variables are not i.i.d.; they have different means. Nonetheless, the observations are independent, and as we will examine in closer detail,

$$\ln L(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X}) = \sum_{i=1}^n \ln f(y_i | \mathbf{x}_i, \boldsymbol{\theta}) = -\frac{1}{2} \sum_{i=1}^n [\ln \sigma^2 + \ln(2\pi) + (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 / \sigma^2], \quad (14-3)$$

where  $\mathbf{X}$  is the  $n \times K$  matrix of data with  $i$ th row equal to  $\mathbf{x}'_i$ .

The rest of this chapter will be concerned with obtaining estimates of the parameters,  $\boldsymbol{\theta}$ , and in testing hypotheses about them and about the data-generating process. Before we begin that study, we consider the question of whether estimation of the parameters is possible at all—the question of **identification**. Identification is an issue related to the formulation of the model. The issue of identification must be resolved before estimation can even be considered. The question posed is essentially this: Suppose we had an infinitely large sample—that is, for current purposes, all the information there is to be had about the parameters. Could we uniquely determine the values of  $\boldsymbol{\theta}$  from such a sample? As will be clear shortly, the answer is sometimes no.

### DEFINITION 14.1 Identification

*The parameter vector  $\boldsymbol{\theta}$  is identified (**estimable**) if for any other parameter vector,  $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}$ , for some data  $\mathbf{y}$ ,  $L(\boldsymbol{\theta}^* | \mathbf{y}) \neq L(\boldsymbol{\theta} | \mathbf{y})$ .*

This result will be crucial at several points in what follows. We consider two examples, the first of which will be very familiar to you by now.

#### Example 14.1 Identification of Parameters

For the regression model specified in (14-3), suppose that there is a nonzero vector  $\mathbf{a}$  such that  $\mathbf{x}'_i \mathbf{a} = 0$  for every  $\mathbf{x}_i$ . Then there is another “parameter” vector,  $\boldsymbol{\gamma} = \boldsymbol{\beta} + \mathbf{a} \neq \boldsymbol{\beta}$  such that  $\mathbf{x}'_i \boldsymbol{\beta} = \mathbf{x}'_i \boldsymbol{\gamma}$  for every  $\mathbf{x}_i$ . You can see in (14-3) that if this is the case, then the log-likelihood

## CHAPTER 14 ♦ Maximum Likelihood Estimation 511

is the same whether it is evaluated at  $\beta$  or at  $\gamma$ . As such, it is not possible to consider estimation of  $\beta$  in this model because  $\beta$  cannot be distinguished from  $\gamma$ . This is the case of perfect collinearity in the regression model, which we ruled out when we first proposed the linear regression model with "Assumption 2. Identifiability of the Model Parameters."

The preceding dealt with a necessary characteristic of the sample data. We now consider a model in which identification is secured by the specification of the parameters in the model. (We will study this model in detail in Chapter 17.) Consider a simple form of the regression model considered earlier,  $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ , where  $\varepsilon_i | x_i$  has a normal distribution with zero mean and variance  $\sigma^2$ . To put the model in a context, consider a consumer's purchases of a large commodity such as a car where  $x_i$  is the consumer's income and  $y_i$  is the difference between what the consumer is willing to pay for the car,  $p_i^*$ , and the price tag on the car,  $p_i$ . Suppose rather than observing  $p_i^*$  or  $p_i$ , we observe only whether the consumer actually purchases the car, which, we assume, occurs when  $y_i = p_i^* - p_i$  is positive. Collecting this information, our model states that they will purchase the car if  $y_i > 0$  and not purchase it if  $y_i \leq 0$ . Let us form the likelihood function for the observed data, which are purchase (or not) and income. The random variable in this model is "purchase" or "not purchase"—there are only two outcomes. The probability of a purchase is

$$\begin{aligned} \text{Prob}(\text{purchase} | \beta_1, \beta_2, \sigma, x_i) &= \text{Prob}(y_i > 0 | \beta_1, \beta_2, \sigma, x_i) \\ &= \text{Prob}(\beta_1 + \beta_2 x_i + \varepsilon_i > 0 | \beta_1, \beta_2, \sigma, x_i) \\ &= \text{Prob}[\varepsilon_i > -(\beta_1 + \beta_2 x_i) | \beta_1, \beta_2, \sigma, x_i] \\ &= \text{Prob}[\varepsilon_i / \sigma > -(\beta_1 + \beta_2 x_i) / \sigma | \beta_1, \beta_2, \sigma, x_i] \\ &= \text{Prob}[z_i > -(\beta_1 + \beta_2 x_i) / \sigma | \beta_1, \beta_2, \sigma, x_i] \end{aligned}$$

where  $z_i$  has a standard normal distribution. The probability of not purchase is just one minus this probability. The likelihood function is

$$\prod_{i=\text{purchased}} [\text{Prob}(\text{purchase} | \beta_1, \beta_2, \sigma, x_i)] \prod_{i=\text{not purchased}} [1 - \text{Prob}(\text{purchase} | \beta_1, \beta_2, \sigma, x_i)].$$

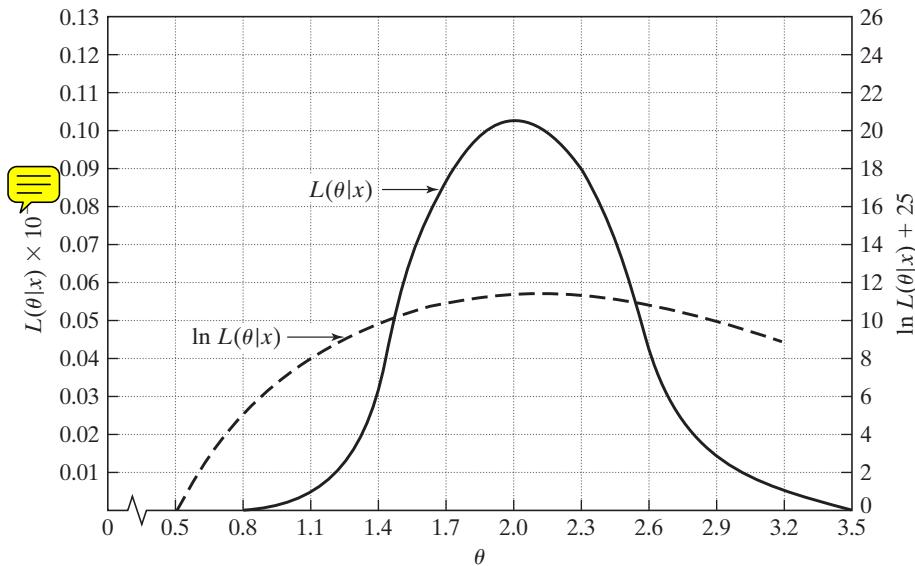
We need go no further to see that the parameters of this model are not identified. If  $\beta_1$ ,  $\beta_2$ , and  $\sigma$  are all multiplied by the same nonzero constant, regardless of what it is, then  $\text{Prob}(\text{purchase})$  is unchanged,  $1 - \text{Prob}(\text{purchase})$  is also, and the likelihood function does not change. This model requires a **normalization**. The one usually used is  $\sigma = 1$ , but some authors [e.g., Horowitz (1993)] have used  $\beta_1 = 1$  instead.

### 14.3 EFFICIENT ESTIMATION: THE PRINCIPLE OF MAXIMUM LIKELIHOOD

The principle of **maximum likelihood** provides a means of choosing an asymptotically efficient estimator for a parameter or a set of parameters. The logic of the technique is easily illustrated in the setting of a discrete distribution. Consider a random sample of the following 10 observations from a Poisson distribution: 5, 0, 1, 1, 0, 3, 2, 3, 4, and 1. The density for each observation is

$$f(y_i | \theta) = \frac{e^{-\theta} \theta^{y_i}}{y_i!}.$$

## 512 PART III ♦ Estimation Methodology



**FIGURE 14.1** Likelihood and Log-Likelihood Functions for a Poisson Distribution.

Because the observations are independent, their joint density, which is the likelihood for this sample, is

$$f(y_1, y_2, \dots, y_{10} | \theta) = \prod_{i=1}^{10} f(y_i | \theta) = \frac{e^{-10\theta} \theta^{\sum_{i=1}^{10} y_i}}{\prod_{i=1}^{10} y_i!} = \frac{e^{-10\theta} \theta^{20}}{207,360}.$$

The last result gives the probability of observing *this particular sample*, assuming that a Poisson distribution with as yet unknown parameter  $\theta$  generated the data. What value of  $\theta$  would make this sample most probable? Figure 14.1 plots this function for various values of  $\theta$ . It has a single mode at  $\theta = 2$ , which would be the **maximum likelihood estimate**, or MLE, of  $\theta$ .

Consider maximizing  $L(\theta | \mathbf{y})$  with respect to  $\theta$ . Because the log function is monotonically increasing and easier to work with, we usually maximize  $\ln L(\theta | \mathbf{y})$  instead; in sampling from a Poisson population,

$$\ln L(\theta | \mathbf{y}) = -n\theta + \ln \theta \sum_{i=1}^n y_i - \sum_{i=1}^n \ln(y_i!),$$

$$\frac{\partial \ln L(\theta | \mathbf{y})}{\partial \theta} = -n + \frac{1}{\theta} \sum_{i=1}^n y_i = 0 \Rightarrow \hat{\theta}_{\text{ML}} = \bar{y}_n.$$

For the assumed sample of observations,

$$\ln L(\theta | \mathbf{y}) = -10\theta + 20 \ln \theta - 12.242,$$

$$\frac{d \ln L(\theta | \mathbf{y})}{d \theta} = -10 + \frac{20}{\theta} = 0 \Rightarrow \hat{\theta} = 2,$$

## CHAPTER 14 ♦ Maximum Likelihood Estimation 513

and

$$\frac{d^2 \ln L(\theta | \mathbf{y})}{d\theta^2} = \frac{-20}{\theta^2} < 0 \Rightarrow \text{this is a maximum.}$$

The solution is the same as before. Figure 14.1 also plots the log of  $L(\theta | \mathbf{y})$  to illustrate the result.

The reference to the probability of observing the given sample is not exact in a continuous distribution, because a particular sample has probability zero. Nonetheless, the principle is the same. The values of the parameters that maximize  $L(\theta | \mathbf{data})$  or its log are the maximum likelihood estimates, denoted  $\hat{\theta}$ . The logarithm is a monotonic function, so the values that maximize  $L(\theta | \mathbf{data})$  are the same as those that maximize  $\ln L(\theta | \mathbf{data})$ . The necessary condition for maximizing  $\ln L(\theta | \mathbf{data})$  is

$$\frac{\partial \ln L(\theta | \mathbf{data})}{\partial \theta} = 0. \quad (14-4)$$

This is called the **likelihood equation**. The general result then is that the MLE is a root of the likelihood equation. The application to the parameters of the dgp for a discrete random variable are suggestive that maximum likelihood is a “good” use of the data. It remains to establish this as a general principle. We turn to that issue in the next section.

**Example 14.2 Log-Likelihood Function and Likelihood Equations for the Normal Distribution**

In sampling from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , the log-likelihood function and the likelihood equations for  $\mu$  and  $\sigma^2$  are

$$\ln L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2} \sum_{i=1}^n \left[ \frac{(y_i - \mu)^2}{\sigma^2} \right], \quad (14-5)$$

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu) = 0, \quad (14-6)$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \mu)^2 = 0. \quad (14-7)$$

To solve the likelihood equations, multiply (14-6) by  $\sigma^2$  and solve for  $\hat{\mu}$ , then insert this solution in (14-7) and solve for  $\sigma^2$ . The solutions are

$$\hat{\mu}_{\text{ML}} = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}_n \quad \text{and} \quad \hat{\sigma}_{\text{ML}}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_n)^2. \quad (14-8)$$

#### 14.4 PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATORS

Maximum likelihood estimators (MLEs) are most attractive because of their large-sample or asymptotic properties.

## 514 PART III ♦ Estimation Methodology

### DEFINITION 14.2 Asymptotic Efficiency

An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed (CAN), and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.<sup>2</sup>

If certain regularity conditions are met, the MLE will have these properties. The finite sample properties are sometimes less than optimal. For example, the MLE may be biased; the MLE of  $\sigma^2$  in Example 14.2 is biased downward. The occasional statement that the properties of the MLE are *only* optimal in large samples is not true, however. It can be shown that when sampling is from an exponential family of distributions (see Definition 13.1), there will exist sufficient statistics. If so, MLEs will be functions of them, which means that when minimum variance unbiased estimators exist, they will be MLEs. [See Stuart and Ord (1989).] Most applications in econometrics do not involve exponential families, so the appeal of the MLE remains primarily its asymptotic properties.

We use the following notation:  $\hat{\theta}$  is the maximum likelihood estimator;  $\theta_0$  denotes the true value of the parameter vector;  $\theta$  denotes another possible value of the parameter vector, not the MLE and not necessarily the true values. Expectation based on the true values of the parameters is denoted  $E_0[\cdot]$ . If we assume that the regularity conditions discussed momentarily are met by  $f(\mathbf{x}, \theta_0)$ , then we have the following theorem.

### THEOREM 14.1 Properties of an MLE

*Under regularity, the maximum likelihood estimator (MLE) has the following asymptotic properties:*

- M1. **Consistency:**  $\text{plim } \hat{\theta} = \theta_0$ .
- M2. **Asymptotic normality:**  $\hat{\theta} \stackrel{d}{\sim} N[\theta_0, \{\mathbf{I}(\theta_0)\}^{-1}]$ , where  

$$\mathbf{I}(\theta_0) = -E_0[\partial^2 \ln L / \partial \theta_0 \partial \theta'_0].$$
- M3. **Asymptotic efficiency:**  $\hat{\theta}$  is asymptotically efficient and achieves the **Cramér-Rao lower bound** for consistent estimators, given in M2 and Theorem C.2.
- M4. **Invariance:** The maximum likelihood estimator of  $\gamma_0 = \mathbf{c}(\theta_0)$  is  $\mathbf{c}(\hat{\theta})$  if  $\mathbf{c}(\theta_0)$  is a continuous and continuously differentiable function.

#### 14.4.1 REGULARITY CONDITIONS

To sketch proofs of these results, we first obtain some useful properties of probability density functions. We assume that  $(y_1, \dots, y_n)$  is a random sample from the population with density function  $f(y_i | \theta_0)$  and that the following **regularity conditions** hold. [Our

<sup>2</sup>Not larger is defined in the sense of (A-118): The covariance matrix of the less efficient estimator equals that of the efficient estimator plus a nonnegative definite matrix.

## CHAPTER 14 ♦ Maximum Likelihood Estimation 515

statement of these is informal. A more rigorous treatment may be found in Stuart and Ord (1989) or Davidson and MacKinnon (2004).]

**DEFINITION 14.3 Regularity Conditions**

- R1.** *The first three derivatives of  $\ln f(y_i | \theta)$  with respect to  $\theta$  are continuous and finite for almost all  $y_i$  and for all  $\theta$ . This condition ensures the existence of a certain Taylor series approximation [ ... ] the finite variance of the derivatives of  $\ln L$ .*
- R2.** *The conditions necessary to obtain the expectations of the first and second derivatives of  $\ln f(y_i | \theta)$  are met.*
- R3.** *For all values of  $\theta$ ,  $|\partial^3 \ln f(y_i | \theta)/\partial \theta_j \partial \theta_k \partial \theta_l|$  is less than a function that has a finite expectation. This condition will allow us to truncate the Taylor series.*

With these regularity conditions, we will obtain the following fundamental characteristics of  $f(y_i | \theta)$ : D1 is simply a consequence of the definition of the likelihood function. D2 leads to the moment condition which defines the maximum likelihood estimator. On the one hand, the MLE is found as the maximizer of a function, which mandates finding the vector that equates the gradient to zero. On the other, D2 is a more fundamental relationship that places the MLE in the class of generalized method of moments estimators. D3 produces what is known as the **information matrix equality**. This relationship shows how to obtain the asymptotic covariance matrix of the MLE.

**14.4.2 PROPERTIES OF REGULAR DENSITIES**

Densities that are “regular” by Definition 14.3 have three properties that are used in establishing the properties of maximum likelihood estimators:

**THEOREM 14.2 Moments of the Derivatives of the Log-Likelihood**

- D1.**  *$\ln f(y_i | \theta)$ ,  $\mathbf{g}_i = \partial \ln f(y_i | \theta)/\partial \theta$ , and  $\mathbf{H}_i = \partial^2 \ln f(y_i | \theta)/\partial \theta \partial \theta'$ ,  $i = 1, \dots, n$ , are all random samples of random variables. This statement follows from our assumption of random sampling. The notation  $\mathbf{g}_i(\theta_0)$  and  $\mathbf{H}_i(\theta_0)$  indicates the derivative evaluated at  $\theta_0$ .*
- D2.**  *$E_0[\mathbf{g}_i(\theta_0)] = \mathbf{0}$ .*
- D3.**  *$\text{Var}[\mathbf{g}_i(\theta_0)] = -E[\mathbf{H}_i(\theta_0)]$ .*

*Condition D1 is simply a consequence of the definition of the density.*

For the moment, we allow the range of  $y_i$  to depend on the parameters;  $A(\theta_0) \leq y_i \leq B(\theta_0)$ . (Consider, for example, finding the maximum likelihood estimator of  $\theta_0$  for a continuous uniform distribution with range  $[0, \theta_0]$ .) (In the following, the single integral

### 516 PART III ♦ Estimation Methodology

$\int \dots dy_i$ , where  be used to indicate the multiple integration over all the elements of a multivariate of  $y_i$  if that were necessary.) By definition,

$$\int_{A(\theta_0)}^{B(\theta_0)} f(y_i | \theta_0) dy_i = 1.$$

Now, differentiate this expression with respect to  $\theta_0$ . Leibnitz's theorem gives

$$\begin{aligned} \frac{\partial \int_{A(\theta_0)}^{B(\theta_0)} f(y_i | \theta_0) dy_i}{\partial \theta_0} &= \int_{A(\theta_0)}^{B(\theta_0)} \frac{\partial f(y_i | \theta_0)}{\partial \theta_0} dy_i + f(B(\theta_0) | \theta_0) \frac{\partial B(\theta_0)}{\partial \theta_0} \\ &\quad - f(A(\theta_0) | \theta_0) \frac{\partial A(\theta_0)}{\partial \theta_0} \\ &= \mathbf{0}. \end{aligned}$$

If the second and third terms go to zero, then we may interchange the operations of differentiation and integration. The necessary condition is that  $\lim_{y_i \rightarrow A(\theta_0)} f(y_i | \theta_0) = \lim_{y_i \rightarrow B(\theta_0)} f(y_i | \theta_0) = 0$ . (Note that the uniform distribution suggested earlier violates this condition.) Sufficient conditions are that the range of the observed random variable,  $y_i$ , does not depend on the parameters, which means that  $\partial A(\theta_0)/\partial \theta_0 = \partial B(\theta_0)/\partial \theta_0 = \mathbf{0}$  or that the density is zero at the terminal points. This condition, then, is regularity condition R2. The latter is usually assumed, and we will assume it in what follows. So,

$$\begin{aligned} \frac{\partial \int f(y_i | \theta_0) dy_i}{\partial \theta_0} &= \int \frac{\partial f(y_i | \theta_0)}{\partial \theta_0} dy_i = \int \frac{\partial \ln f(y_i | \theta_0)}{\partial \theta_0} f(y_i | \theta_0) dy_i \\ &= E_0 \left[ \frac{\partial \ln f(y_i | \theta_0)}{\partial \theta_0} \right] = \mathbf{0}. \end{aligned}$$

This proves D2.

Because we may interchange the operations of integration and differentiation, we differentiate under the integral once again to obtain

$$\int \left[ \frac{\partial^2 \ln f(y_i | \theta_0)}{\partial \theta_0 \partial \theta'_0} f(y_i | \theta_0) + \frac{\partial \ln f(y_i | \theta_0)}{\partial \theta_0} \frac{\partial f(y_i | \theta_0)}{\partial \theta'_0} \right] dy_i = \mathbf{0}.$$

But

$$\frac{\partial f(y_i | \theta_0)}{\partial \theta'_0} = f(y_i | \theta_0) \frac{\partial \ln f(y_i | \theta_0)}{\partial \theta'_0},$$

and the integral of a sum is the sum of integrals. Therefore,

$$-\int \left[ \frac{\partial^2 \ln f(y_i | \theta_0)}{\partial \theta_0 \partial \theta'_0} \right] f(y_i | \theta_0) dy_i = \int \left[ \frac{\partial \ln f(y_i | \theta_0)}{\partial \theta_0} \frac{\partial \ln f(y_i | \theta_0)}{\partial \theta'_0} \right] f(y_i | \theta_0) dy_i.$$

The left-hand side of the equation is the negative of the expected second derivatives matrix. The right-hand side is the expected square (outer product) of the first derivative vector. But, because this vector has expected value  $\mathbf{0}$  (we just showed this), the right-hand side is the variance of the first derivative vector, which proves D3:

$$\text{Var}_0 \left[ \frac{\partial \ln f(y_i | \theta_0)}{\partial \theta_0} \right] = E_0 \left[ \left( \frac{\partial \ln f(y_i | \theta_0)}{\partial \theta_0} \right) \left( \frac{\partial \ln f(y_i | \theta_0)}{\partial \theta'_0} \right) \right] = -E \left[ \frac{\partial^2 \ln f(y_i | \theta_0)}{\partial \theta_0 \partial \theta'_0} \right].$$

## CHAPTER 14 ♦ Maximum Likelihood Estimation 517

### 14.4.3 THE LIKELIHOOD EQUATION

The log-likelihood function is

$$\ln L(\boldsymbol{\theta} | \mathbf{y}) = \sum_{i=1}^n \ln f(y_i | \boldsymbol{\theta}).$$

The first derivative vector, or **score vector**, is

$$\mathbf{g} = \frac{\partial \ln L(\boldsymbol{\theta} | \mathbf{y})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^n \frac{\partial \ln f(y_i | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^n \mathbf{g}_i. \quad (14-9)$$

Because we are just adding terms, it follows from D1 and D2 that at  $\boldsymbol{\theta}_0$ ,

$$E_0 \left[ \frac{\partial \ln L(\boldsymbol{\theta}_0 | \mathbf{y})}{\partial \boldsymbol{\theta}_0} \right] = E_0[\mathbf{g}_0] = \mathbf{0}. \quad (14-10)$$

which is the **likelihood equation** mentioned earlier.

### 14.4.4 THE INFORMATION MATRIX EQUALITY

The Hessian of the log-likelihood is

$$\mathbf{H} = \frac{\partial^2 \ln L(\boldsymbol{\theta} | \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{i=1}^n \frac{\partial^2 \ln f(y_i | \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{i=1}^n \mathbf{H}_i.$$

Evaluating once again at  $\boldsymbol{\theta}_0$ , by taking

$$E_0[\mathbf{g}_0 \mathbf{g}'_0] = E_0 \left[ \sum_{i=1}^n \sum_{j=1}^n \mathbf{g}_{0i} \mathbf{g}'_{0j} \right],$$

and, because of D1, dropping terms with unequal subscripts we obtain

$$E_0[\mathbf{g}_0 \mathbf{g}'_0] = E_0 \left[ \sum_{i=1}^n \mathbf{g}_{0i} \mathbf{g}'_{0i} \right] = E_0 \left[ \sum_{i=1}^n (-\mathbf{H}_{0i}) \right] = -E_0[\mathbf{H}_0],$$

so that

$$\begin{aligned} \text{Var}_0 \left[ \frac{\partial \ln L(\boldsymbol{\theta}_0 | \mathbf{y})}{\partial \boldsymbol{\theta}_0} \right] &= E_0 \left[ \left( \frac{\partial \ln L(\boldsymbol{\theta}_0 | \mathbf{y})}{\partial \boldsymbol{\theta}_0} \right) \left( \frac{\partial \ln L(\boldsymbol{\theta}_0 | \mathbf{y})}{\partial \boldsymbol{\theta}'_0} \right) \right] \\ &= -E_0 \left[ \frac{\partial^2 \ln L(\boldsymbol{\theta}_0 | \mathbf{y})}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}'_0} \right]. \end{aligned} \quad (14-11)$$

This very useful result is known as the **information matrix equality**.

### 14.4.5 ASYMPTOTIC PROPERTIES OF THE MAXIMUM LIKELIHOOD ESTIMATOR

We can now sketch a derivation of the asymptotic properties of the MLE. Formal proofs of these results require some fairly intricate mathematics. Two widely cited derivations are those of Cramér (1948) and Amemiya (1985). To suggest the flavor of the exercise, we will sketch an analysis provided by Stuart and Ord (1989) for a simple case, and indicate where it will be necessary to extend the derivation if it were to be fully general.

## 518 PART III ♦ Estimation Methodology

### 14.4.5.a Consistency

We assume that  $f(y_i | \theta_0)$  is a possibly multivariate density that at this point does not depend on covariates,  $x_i$ . Thus, this is the i.i.d., random sampling case. Because  $\hat{\theta}$  is the MLE, in any finite sample, for any  $\theta \neq \hat{\theta}$  (including the true  $\theta_0$ ) it must be true that

$$\ln L(\hat{\theta}) \geq \ln L(\theta). \quad (14-12)$$

Consider, then, the random variable  $L(\theta)/L(\theta_0)$ . Because the log function is strictly concave, from Jensen's Inequality (Theorem D.13.), we have

$$E_0 \left[ \ln \frac{L(\theta)}{L(\theta_0)} \right] < \ln E_0 \left[ \frac{L(\theta)}{L(\theta_0)} \right]. \quad (14-13)$$

The expectation on the right-hand side is exactly equal to one, as

$$E_0 \left[ \frac{L(\theta)}{L(\theta_0)} \right] = \int \left( \frac{L(\theta)}{L(\theta_0)} \right) L(\theta_0) d\theta = 1 \quad (14-14)$$

~~is simply the integral of a joint density. Now, take logs on both sides of (14-13), insert the result of (14-14), then divide by  $n$  to produce~~

$$E_0[1/n \ln L(\theta)] - E_0[1/n \ln L(\theta_0)] < 0.$$

This produces a central result:

### THEOREM 14.3 Likelihood Inequality

$$E_0[(1/n) \ln L(\theta_0)] > E_0[(1/n) \ln L(\theta)] \quad \text{for any } \theta \neq \theta_0 \text{ (including } \hat{\theta}).$$

~~This result is (14-15).~~

In words, *the expected value of the log-likelihood is maximized at the true value of the parameters.*

For any  $\theta$ , including  $\hat{\theta}$ ,

$$[(1/n) \ln L(\theta)] = (1/n) \sum_{i=1}^n \ln f(y_i | \theta)$$

is the sample mean of  $n$  i.i.d. random variables, with expectation  $E_0[(1/n) \ln L(\theta)]$ . Because the sampling is i.i.d. by the regularity conditions, we can invoke the Khinchine theorem, D.5; the sample mean converges in probability to the population mean. Using  $\theta = \hat{\theta}$ , it follows from Theorem 14.3 that as  $n \rightarrow \infty$ ,  $\lim \text{Prob}\{[(1/n) \ln L(\hat{\theta})] < [(1/n) \ln L(\theta_0)]\} = 1$  if  $\hat{\theta} \neq \theta_0$ . But,  $\hat{\theta}$  is the MLE, so for every  $n$ ,  $(1/n) \ln L(\hat{\theta}) \geq (1/n) \ln L(\theta_0)$ . The only way these can both be true is if  $(1/n)$  times the sample log-likelihood evaluated at the MLE converges to the population expectation of  $(1/n)$  times the log-likelihood evaluated at the true parameters. There remains one final step. Does  $(1/n) \ln L(\hat{\theta}) \rightarrow (1/n) \ln L(\theta_0)$  imply that  $\hat{\theta} \rightarrow \theta_0$ ? If there is a single parameter and the likelihood function is one to one, then clearly so. For more general cases, this requires a further characterization of the likelihood function. If the likelihood is strictly continuous and twice differentiable, which we assumed in the regularity conditions, and if the parameters of the model are identified which we assumed at the beginning of this discussion, then yes, it does, so we have the result.

## CHAPTER 14 ♦ Maximum Likelihood Estimation 519

This is a heuristic proof. As noted, formal presentations appear in more advanced treatises than this one. We should also note, we have assumed at several points that sample means converge to the population expectations. This is likely to be true for the sorts of applications usually encountered in econometrics, but a fully general set of results would look more closely at this condition. Second, we have assumed i.i.d. sampling in the preceding—that is, the density for  $y_i$  does not depend on any other variables,  $x_i$ . This will almost never be true in practice. Assumptions about the behavior of these variables will enter the proofs as well. For example, in assessing the large sample behavior of the least squares estimator, we have invoked an assumption that the data are “well behaved.” The same sort of consideration will apply here as well. We will return to this issue shortly. With all this in place, we have property M1,  $\text{plim } \hat{\theta} = \theta_0$ .

**14.4.5.b Asymptotic Normality**

At the maximum likelihood estimator, the gradient of the log-likelihood equals zero (by definition), so

$$\mathbf{g}(\hat{\theta}) = \mathbf{0}.$$

(This is the sample statistic, not the expectation.) Expand this set of equations in a Taylor series around the true parameters  $\theta_0$ . We will use the mean value theorem to truncate the Taylor series at the second term,

$$\mathbf{g}(\hat{\theta}) = \mathbf{g}(\theta_0) + \mathbf{H}(\bar{\theta})(\hat{\theta} - \theta_0) = \mathbf{0}.$$

The Hessian is evaluated at a point  $\bar{\theta}$  that is between  $\hat{\theta}$  and  $\theta_0$  [ $\bar{\theta} = w\hat{\theta} + (1-w)\theta_0$  for some  $0 < w < 1$ ]. We then rearrange this function and multiply the result by  $\sqrt{n}$  to obtain

$$\sqrt{n}(\hat{\theta} - \theta_0) = [-\mathbf{H}(\bar{\theta})]^{-1}[\sqrt{n}\mathbf{g}(\theta_0)].$$

Because  $\text{plim}(\hat{\theta} - \theta_0) = \mathbf{0}$ ,  $\text{plim}(\hat{\theta} - \bar{\theta}) = \mathbf{0}$  as well. The second derivatives are continuous functions. Therefore, if the limiting distribution exists, then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} [-\mathbf{H}(\theta_0)]^{-1}[\sqrt{n}\mathbf{g}(\theta_0)].$$

By dividing  $\mathbf{H}(\theta_0)$  and  $\mathbf{g}(\theta_0)$  by  $n$ , we obtain

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \left[-\frac{1}{n}\mathbf{H}(\theta_0)\right]^{-1}[\sqrt{n}\bar{\mathbf{g}}(\theta_0)]. \quad (14-15)$$

We may apply the Lindeberg–Levy central limit theorem (D.18) to  $[\sqrt{n}\bar{\mathbf{g}}(\theta_0)]$ , because it is  $\sqrt{n}$  times the mean of a random sample; we have invoked D1 again. The limiting variance of  $[\sqrt{n}\bar{\mathbf{g}}(\theta_0)]$  is  $-E_0[(1/n)\mathbf{H}(\theta_0)]$ , so

$$\sqrt{n}\bar{\mathbf{g}}(\theta_0) \xrightarrow{d} N\{\mathbf{0}, -E_0\left[\frac{1}{n}\mathbf{H}(\theta_0)\right]\}.$$

By virtue of Theorem D.2,  $\text{plim}[-(1/n)\mathbf{H}(\theta_0)] = -E_0[(1/n)\mathbf{H}(\theta_0)]$ . This result is a constant matrix, so we can combine results to obtain

$$\left[-\frac{1}{n}\mathbf{H}(\theta_0)\right]^{-1}\sqrt{n}\bar{\mathbf{g}}(\theta_0) \xrightarrow{d} N\left[\mathbf{0}, \left\{-E_0\left[\frac{1}{n}\mathbf{H}(\theta_0)\right]\right\}^{-1}\left\{-E_0\left[\frac{1}{n}\mathbf{H}(\theta_0)\right]\right\}\left\{-E_0\left[\frac{1}{n}\mathbf{H}(\theta_0)\right]\right\}^{-1}\right],$$

or

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left[\mathbf{0}, \left\{-E_0\left[\frac{1}{n}\mathbf{H}(\theta_0)\right]\right\}^{-1}\right],$$

## 520 PART III ♦ Estimation Methodology

which gives the asymptotic distribution of the MLE:

$$\hat{\theta} \xrightarrow{a} N[\theta_0, \{I(\theta_0)\}^{-1}].$$

This last step completes M2.

### **Example 14.3 Information Matrix for the Normal Distribution**

For the likelihood function in Example 14.2, the second derivatives are

$$\begin{aligned}\frac{\partial^2 \ln L}{\partial \mu^2} &= \frac{-n}{\sigma^2}, \\ \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n (y_i - \mu)^2, \\ \frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2} &= \frac{-1}{\sigma^4} \sum_{i=1}^n (y_i - \mu).\end{aligned}$$

For the **asymptotic variance** of the maximum likelihood estimator, we need the expectations of these derivatives. The first is nonstochastic, and the third has expectation 0, as  $E[y_i] = \mu$ . That leaves the second, which you can verify has expectation  $-n/(2\sigma^4)$  because each of the  $n$  terms  $(y_i - \mu)^2$  has expected value  $\sigma^2$ . Collecting these in the information matrix, reversing the sign, and inverting the matrix gives the asymptotic covariance matrix for the maximum likelihood estimators:

$$\left\{ -E_0 \left[ \frac{\partial^2 \ln L}{\partial \theta_0 \partial \theta_0'} \right] \right\}^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{bmatrix}.$$

#### **14.4.5.c Asymptotic Efficiency**

Theorem C.2 provides the lower bound for the variance of an unbiased estimator. Because the asymptotic variance of the MLE achieves this bound, it seems natural to extend the result directly. There is, however, a loose end in that the MLE is almost never unbiased. As such, we need an asymptotic version of the bound, which was provided by Cramér (1948) and Rao (1945) (hence the name):

### **THEOREM 14.4 Cramér–Rao Lower Bound**

*Assuming that the density of  $y_i$  satisfies the regularity conditions R1–R3, the asymptotic variance of a consistent and asymptotically normally distributed estimator of the parameter vector  $\theta_0$  will always be at least as large as*

$$[I(\theta_0)]^{-1} = \left( -E_0 \left[ \frac{\partial^2 \ln L(\theta_0)}{\partial \theta_0 \partial \theta_0'} \right] \right)^{-1} = \left( E_0 \left[ \left( \frac{\partial \ln L(\theta_0)}{\partial \theta_0} \right) \left( \frac{\partial \ln L(\theta_0)}{\partial \theta_0} \right)' \right] \right)^{-1}.$$

The asymptotic variance of the MLE is, in fact, equal to the Cramér–Rao Lower Bound for the variance of a consistent, asymptotically normally distributed estimator, so this completes the argument.<sup>3</sup>

<sup>3</sup>A result reported by LeCam (1953) and recounted in Amemiya (1985, p. 124) suggests that, in principle, there do exist CAN functions of the data with smaller variances than the MLE. But, the finding is a narrow result with no practical implications. For practical purposes, the statement may be taken as given.

## CHAPTER 14 ♦ Maximum Likelihood Estimation 521

**14.4.5.d Invariance**

Last, the invariance property, M4, is a mathematical result of the method of computing MLEs; it is not a statistical result as such. More formally, the MLE is invariant to *one-to-one* transformations of  $\theta$ . Any transformation that is not one to one either renders the model inestimable if it is one to many or imposes restrictions if it is many to one. Some theoretical aspects of this feature are discussed in Davidson and MacKinnon (2004, pp. 446, 539–540). For the practitioner, the result can be extremely useful. For example, when a parameter appears in a likelihood function in the form  $1/\theta_j$ , it is usually worthwhile to reparameterize the model in terms of  $\gamma_j = 1/\theta_j$ . In an important application, Olsen (1978) used this result to great advantage. (See Section 18.5.) Suppose that the normal log-likelihood in Example 14.2 is parameterized in terms of the **precision parameter**,  $\theta^2 = 1/\sigma^2$ . The log-likelihood becomes

$$\ln L(\mu, \theta^2) = -(n/2) \ln(2\pi) + (n/2) \ln \theta^2 - \frac{\theta^2}{2} \sum_{i=1}^n (y_i - \mu)^2.$$

The MLE for  $\mu$  is clearly still  $\bar{x}$ . But the likelihood equation for  $\theta^2$  is now

$$\partial \ln L(\mu, \theta^2) / \partial \theta^2 = \frac{1}{2} \left[ n/\theta^2 - \sum_{i=1}^n (y_i - \mu)^2 \right] = 0,$$

which has solution  $\hat{\theta}^2 = n / \sum_{i=1}^n (y_i - \hat{\mu})^2 = 1/\hat{\sigma}^2$ , as expected. There is a second implication. If it is desired to analyze a function of an MLE, then the function of  $\hat{\theta}$  will, itself, be the MLE.

**14.4.5.e Conclusion**

These four properties explain the prevalence of the maximum likelihood technique in econometrics. The second greatly facilitates hypothesis testing and the construction of interval estimates. The third is a particularly powerful result. The MLE has the minimum variance achievable by a consistent and asymptotically normally distributed estimator.

**14.4.6 ESTIMATING THE ASYMPTOTIC VARIANCE OF THE MAXIMUM LIKELIHOOD ESTIMATOR**

The asymptotic covariance matrix of the maximum likelihood estimator is a matrix of parameters that must be estimated (i.e., it is a function of the  $\theta_0$  that is being estimated). If the form of the expected values of the second derivatives of the log-likelihood is known, then

$$[\mathbf{I}(\theta_0)]^{-1} = \left\{ -E_0 \left[ \frac{\partial^2 \ln L(\theta_0)}{\partial \theta_0 \partial \theta'_0} \right] \right\}^{-1} \quad (14-16)$$

can be evaluated at  $\hat{\theta}$  to estimate the covariance matrix for the MLE. This estimator will rarely be available. The second derivatives of the log-likelihood will almost always be complicated nonlinear functions of the data whose exact expected values will be unknown. There are, however, two alternatives. A second estimator is

$$[\hat{\mathbf{I}}(\hat{\theta})]^{-1} = \left( -\frac{\partial^2 \ln L(\hat{\theta})}{\partial \hat{\theta} \partial \hat{\theta}' } \right)^{-1}. \quad (14-17)$$

This estimator is computed simply by evaluating the actual (not expected) second derivatives matrix of the log-likelihood function at the maximum likelihood estimates. It is

## 522 PART III ♦ Estimation Methodology

straightforward to show that this amounts to estimating the expected second derivatives of the density with the sample mean of this quantity. Theorem D.4 and Result (D-5) can be used to justify the computation. The only shortcoming of this estimator is that the second derivatives can be complicated to derive and program for a computer. A third estimator based on result D3 in Theorem 14.2, that the expected second derivatives matrix is the covariance matrix of the first derivatives vector, is

$$[\hat{\mathbf{I}}(\hat{\theta})]^{-1} = \left[ \sum_{i=1}^n \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i' \right]^{-1} = [\hat{\mathbf{G}}' \hat{\mathbf{G}}]^{-1}, \quad (14-18)$$

where

$$\hat{\mathbf{g}}_i = \frac{\partial \ln f(\mathbf{x}_i, \hat{\theta})}{\partial \hat{\theta}}, \quad \text{[Note: } f(y_i | x_i, \hat{\theta}) \text{ is crossed out]}$$

and

$$\hat{\mathbf{G}} = [\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2, \dots, \hat{\mathbf{g}}_n]'.$$

$\hat{\mathbf{G}}$  is an  $n \times K$  matrix with  $i$ th row equal to the transpose of the  $i$ th vector of derivatives in the terms of the log-likelihood function. For a single parameter, this estimator is just the reciprocal of the sum of squares of the first derivatives. This estimator is extremely convenient, in most cases, because it does not require any computations beyond those required to solve the likelihood equation. It has the added virtue that it is always non-negative definite. For some extremely complicated log-likelihood functions, sometimes because of rounding error, the *observed* Hessian can be indefinite, even at the maximum of the function. The estimator in (14-18) is known as the **BHHH estimator**<sup>4</sup> and the **outer product of gradients**, or **OPG**, estimator.

None of the three estimators given here is preferable to the others on statistical grounds; all are asymptotically equivalent. In most cases, the BHHH estimator will be the easiest to compute. One caution is in order. As the following example illustrates, these estimators can give different results in a finite sample. This is an unavoidable finite sample problem that can, in some cases, lead to different statistical conclusions. The example is a case in point. Using the usual procedures, we would reject the hypothesis that  $\beta = 0$  if either of the first two variance estimators were used, but not if the third were used. The estimator in (14-16) is usually unavailable, as the exact expectation of the Hessian is rarely known. Available evidence suggests that in small or moderate-sized samples, (14-17) (the Hessian) is preferable.

### Example 14.4 Variance Estimators for an MLE

The sample data in Example C.1 are generated by a model of the form

$$f(y_i, x_i, \beta) = \frac{1}{\beta + x_i} e^{-y_i / (\beta + x_i)},$$

where  $y$  = income and  $x$  = education. To find the maximum likelihood estimate of  $\beta$ , we maximize

$$\ln L(\beta) = - \sum_{i=1}^n \ln(\beta + x_i) - \sum_{i=1}^n \frac{y_i}{\beta + x_i}.$$

<sup>4</sup>It appears to have been advocated first in the econometrics literature in Berndt et al. (1974).

## CHAPTER 14 ♦ Maximum Likelihood Estimation 523

The likelihood equation is

$$\frac{\partial \ln L(\beta)}{\partial \beta} = - \sum_{i=1}^n \frac{1}{\beta + x_i} + \sum_{i=1}^n \frac{y_i}{(\beta + x_i)^2} = 0, \quad (14-19)$$

which has the solution  $\hat{\beta} = 15.602727$ . To compute the asymptotic variance of the MLE, we require

$$\frac{\partial^2 \ln L(\beta)}{\partial \beta^2} = \sum_{i=1}^n \frac{1}{(\beta + x_i)^2} - 2 \sum_{i=1}^n \frac{y_i}{(\beta + x_i)^3}. \quad (14-20)$$

Because the function  $E(y_i) = \beta + x_i$  is known, the exact form of the expected value in (14-20) is known. Inserting  $\hat{\beta} + x_i$  for  $y_i$  in (14-20) and taking the negative of the reciprocal yields the first variance estimate, 44.2546. Simply inserting  $\hat{\beta} = 15.602727$  in (14-20) and taking the negative of the reciprocal gives the second estimate, 46.16337. Finally, by computing the reciprocal of the sum of squares of first derivatives of the densities evaluated at  $\hat{\beta}$ ,

$$[\hat{\mathbf{I}}(\hat{\beta})]^{-1} = \frac{1}{\sum_{i=1}^n [-1/(\hat{\beta} + x_i) + y_i/(\hat{\beta} + x_i)^2]^2},$$

we obtain the BHHH estimate, 100.5116.

#### 14.5 CONDITIONAL LIKELIHOODS, ECONOMETRIC MODELS, AND THE GMM ESTIMATOR

All of the preceding results form the statistical underpinnings of the technique of maximum likelihood estimation. But, for our purposes, a crucial element is missing. We have done the analysis in terms of the density of an observed random variable and a vector of parameters,  $f(y_i | \alpha)$ . But econometric models will involve exogenous or predetermined variables,  $\mathbf{x}_i$ , so the results must be extended. A workable approach is to treat this modeling framework the same as the one in Chapter 4, where we considered the large sample properties of the linear regression model. Thus, we will allow  $\mathbf{x}_i$  to denote a mix of random variables and constants that enter the conditional density of  $y_i$ . By partitioning the joint density of  $y_i$  and  $\mathbf{x}_i$  into the product of the conditional and the marginal, the log-likelihood function may be written

$$\ln L(\alpha | \text{data}) = \sum_{i=1}^n \ln f(y_i, \mathbf{x}_i | \alpha) = \sum_{i=1}^n \ln f(y_i | \mathbf{x}_i, \alpha) + \sum_{i=1}^n \ln g(\mathbf{x}_i | \alpha),$$

where any nonstochastic elements in  $\mathbf{x}_i$  such as a time trend or dummy variable are being carried as constants. To proceed, we will assume as we did before that the process generating  $\mathbf{x}_i$  takes place outside the model of interest. For present purposes, that means that the parameters that appear in  $g(\mathbf{x}_i | \alpha)$  do not overlap with those that appear in  $f(y_i | \mathbf{x}_i, \alpha)$ . Thus, we partition  $\alpha$  into  $[\theta, \delta]$  so that the log-likelihood function may be written

$$\ln L(\theta, \delta | \text{data}) = \sum_{i=1}^n \ln f(y_i, \mathbf{x}_i | \alpha) = \sum_{i=1}^n \ln f(y_i | \mathbf{x}_i, \theta) + \sum_{i=1}^n \ln g(\mathbf{x}_i | \delta).$$

As long as  $\theta$  and  $\delta$  have no elements in common and no restrictions connect them (such as  $\theta + \delta = 1$ ), then the two parts of the log likelihood may be analyzed separately. In most cases, the marginal distribution of  $\mathbf{x}_i$  will be of secondary (or no) interest.

## 524 PART III ♦ Estimation Methodology

Asymptotic results for the maximum conditional likelihood estimator must now account for the presence of  $\mathbf{x}_i$  in the functions and derivatives of  $\ln f(y_i | \mathbf{x}_i, \theta)$ . We will proceed under the assumption of well-behaved data so that sample averages such as

$$(1/n) \ln L(\theta | \mathbf{y}, \mathbf{X}) = \frac{1}{n} \sum_{i=1}^n \ln f(y_i | \mathbf{x}_i, \theta)$$

and its gradient with respect to  $\theta$  will converge in probability to their population expectations. We will also need to invoke central limit theorems to establish the asymptotic normality of the gradient of the log likelihood, so as to be able to characterize the MLE itself. We will leave it to more advanced treatises such as Amemiya (1985) and Newey and McFadden (1994) to establish specific conditions and fine points that must be assumed to claim the “usual” properties for maximum likelihood estimators. For present purposes (and the vast bulk of empirical applications), the following minimal assumptions should suffice:

- **Parameter space.** Parameter spaces that have gaps and nonconvexities in them will generally disable these procedures. An estimation problem that produces this failure is that of “estimating” a parameter that can take only one among a discrete set of values. For example, this set of procedures does not include “estimating” the timing of a structural change in a model. The likelihood function must be a continuous function of a convex parameter space. We allow unbounded parameter spaces, such as  $\sigma > 0$  in the regression model, for example.
- **Identifiability.** Estimation must be feasible. This is the subject of Definition 16.1 concerning identification and the surrounding discussion.
- **Well-behaved data.** Laws of large numbers apply to sample means involving the data and some form of central limit theorem (generally Lyapounov) can be applied to the gradient. Ergodic stationarity is broad enough to encompass any situation that is likely to arise in practice, though it is probably more general than we need for most applications, because we will not encounter dependent observations specifically until later in the book. The definitions in Chapter 4 are assumed to hold generally.

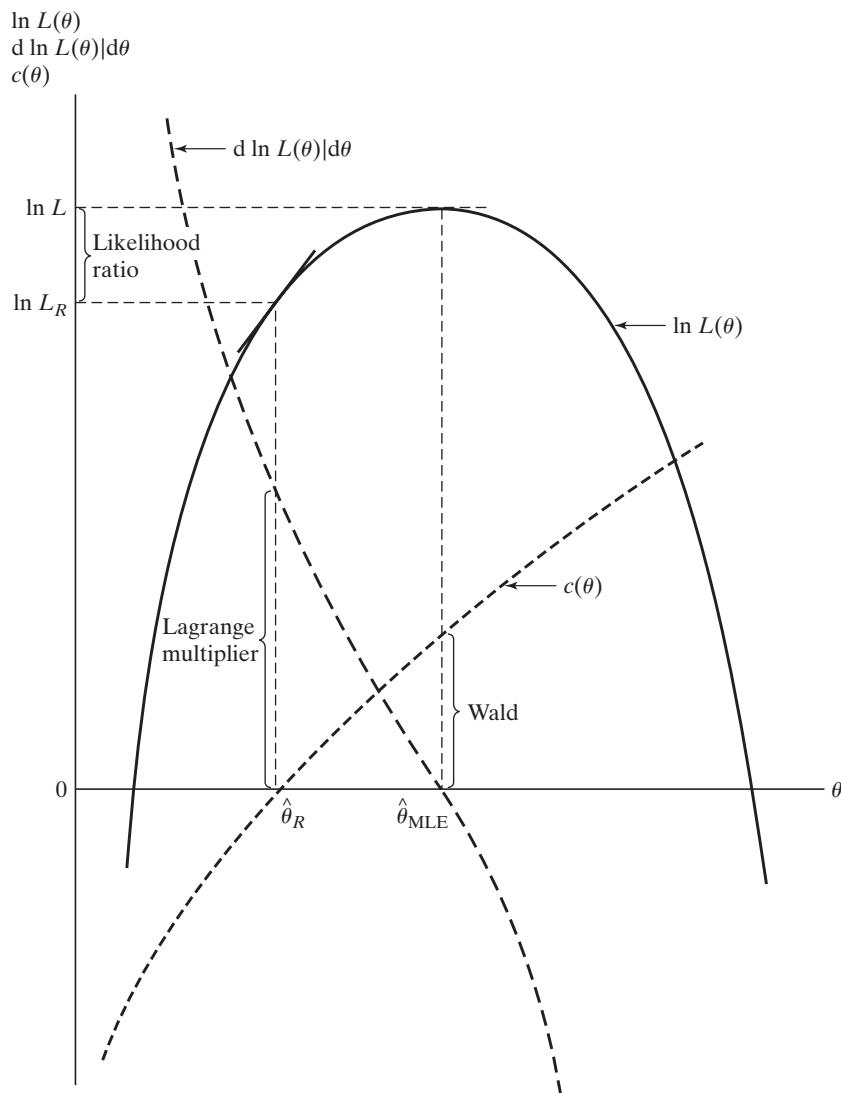
With these in place, analysis is essentially the same in character as that we used in the linear regression model in Chapter 4 and follows precisely along the lines of Section 12.5.

### 14.6 HYPOTHESIS AND SPECIFICATION TESTS AND FIT MEASURES

The next several sections will discuss the most commonly used test procedures: the likelihood ratio, Wald, and Lagrange multiplier tests. [Extensive discussion of these procedures is given in Godfrey (1988).] We consider maximum likelihood estimation of a parameter  $\theta$  and a test of the hypothesis  $H_0: c(\theta) = 0$ . The logic of the tests can be seen in Figure 14.2.<sup>5</sup> The figure plots the log-likelihood function  $\ln L(\theta)$ , its derivative with respect to  $\theta$ ,  $d \ln L(\theta) / d\theta$ , and the constraint  $c(\theta)$ . There are three approaches to

<sup>5</sup>See Buse (1982). Note that the scale of the vertical axis would be different for each curve. As such, the points of intersection have no significance.

## CHAPTER 14 ♦ Maximum Likelihood Estimation 525

**FIGURE 14.2** Three Bases for Hypothesis Tests.

testing the hypothesis suggested in the figure:

- **Likelihood ratio test.** If the restriction  $c(\theta) = 0$  is valid, then imposing it should not lead to a large reduction in the log-likelihood function. Therefore, we base the test on the difference,  $\ln L_U - \ln L_R$ , where  $L_U$  is the value of the likelihood function at the unconstrained value of  $\theta$  and  $L_R$  is the value of the likelihood function at the restricted estimate.
- **Wald test.** If the restriction is valid, then  $c(\hat{\theta}_{MLE})$  should be close to zero because the MLE is consistent. Therefore, the test is based on  $c(\hat{\theta}_{MLE})$ . We reject the hypothesis if this value is significantly different from zero.

## 526 PART III ♦ Estimation Methodology

- **Lagrange multiplier test.** If the restriction is valid, then the restricted estimator should be near the point that maximizes the log-likelihood. Therefore, the slope of the log-likelihood function should be near zero at the restricted estimator. The test is based on the slope of the log-likelihood at the point where the function is maximized subject to the restriction.

These three tests are asymptotically equivalent under the null hypothesis, but they can behave rather differently in a small sample. Unfortunately, their small-sample properties are unknown, except in a few special cases. As a consequence, the choice among them is typically made on the basis of ease of computation. The likelihood ratio test requires calculation of both restricted and unrestricted estimators. If both are simple to compute, then this way to proceed is convenient. The Wald test requires only the unrestricted estimator, and the Lagrange multiplier test requires only the restricted estimator. In some problems, one of these estimators may be much easier to compute than the other. For example, a linear model is simple to estimate but becomes nonlinear and cumbersome if a nonlinear constraint is imposed. In this case, the Wald statistic might be preferable. Alternatively, restrictions sometimes amount to the removal of nonlinearities, which would make the Lagrange multiplier test the simpler procedure.

### 14.6.1 THE LIKELIHOOD RATIO TEST

Let  $\theta$  be a vector of parameters to be estimated, and let  $H_0$  specify some sort of restriction on these parameters. Let  $\hat{\theta}_U$  be the maximum likelihood estimator of  $\theta$  obtained without regard to the constraints, and let  $\hat{\theta}_R$  be the constrained maximum likelihood estimator. If  $\hat{L}_U$  and  $\hat{L}_R$  are the likelihood functions evaluated at these two estimates, then the **likelihood ratio** is

$$\lambda = \frac{\hat{L}_R}{\hat{L}_U}. \quad (14-21)$$

This function must be between zero and one. Both likelihoods are positive, and  $\hat{L}_R$  cannot be larger than  $\hat{L}_U$ . (A restricted optimum is never superior to an unrestricted one.) If  $\lambda$  is too small, then doubt is cast on the restrictions.

An example from a discrete distribution helps to fix these ideas. In estimating from a sample of 10 from a Poisson distribution at the beginning of Section 14.3, we found the MLE of the parameter  $\theta$  to be 2. At this value, the likelihood, which is the probability of observing the sample we did, is  $0.104 \times 10^{-7}$ . Are these data consistent with  $H_0: \theta = 1.8$ ?  $L_R = 0.936 \times 10^{-8}$ , which is, as expected, smaller. This particular sample is somewhat less probable under the hypothesis.

The formal test procedure is based on the following result.

#### THEOREM 14.5 Limiting Distribution of the Likelihood Ratio Test Statistic

*Under regularity and under  $H_0$ , the large-sample distribution of  $-2 \ln \lambda$  is chi-squared, with degrees of freedom equal to the number of restrictions imposed.*

## CHAPTER 14 ♦ Maximum Likelihood Estimation 527

The null hypothesis is rejected if this value exceeds the appropriate critical value from the chi-squared tables. Thus, for the Poisson example,

$$-2 \ln \lambda = -2 \ln \left( \frac{0.0936}{0.104} \right) = 0.21072.$$

This chi-squared statistic with one degree of freedom is not significant at any conventional level, so we would not reject the hypothesis that  $\theta = 1.8$  on the basis of this test.<sup>6</sup>

It is tempting to use the likelihood ratio test to test a simple null hypothesis against a simple alternative. For example, we might be interested in the Poisson setting in testing  $H_0: \theta = 1.8$  against  $H_1: \theta = 2.2$ . But the test cannot be used in this fashion. The degrees of freedom of the chi-squared statistic for the likelihood ratio test equals the reduction in the number of dimensions in the parameter space that results from imposing the restrictions. In testing a simple null hypothesis against a simple alternative, this value is zero.<sup>7</sup> Second, one sometimes encounters an attempt to test one distributional assumption against another with a likelihood ratio test; for example, a certain model will be estimated assuming a normal distribution and then assuming a  $t$  distribution. The ratio of the two likelihoods is then compared to determine which distribution is preferred. This comparison is also inappropriate. The parameter spaces, and hence the likelihood functions of the two cases, are unrelated.

#### 14.6.2 THE WALD TEST

A practical shortcoming of the likelihood ratio test is that it usually requires estimation of both the restricted and unrestricted parameter vectors. In complex models, one or the other of these estimates may be very difficult to compute. Fortunately, there are two alternative testing procedures, the Wald test and the Lagrange multiplier test, that circumvent this problem. Both tests are based on an estimator that is asymptotically normally distributed.

These two tests are based on the distribution of the full rank quadratic form considered in Section B.11.6. Specifically,

$$\text{If } \mathbf{x} \sim N_J[\boldsymbol{\mu}, \boldsymbol{\Sigma}], \text{ then } (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \text{chi-squared}[J]. \quad (14-22)$$

In the setting of a hypothesis test, under the hypothesis that  $E(\mathbf{x}) = \boldsymbol{\mu}$ , the quadratic form has the chi-squared distribution. If the hypothesis that  $E(\mathbf{x}) = \boldsymbol{\mu}$  is false, however, then the quadratic form just given will, on average, be larger than it would be if the hypothesis were true.<sup>8</sup> This condition forms the basis for the test statistics discussed in this and the next section.

Let  $\hat{\boldsymbol{\theta}}$  be the vector of parameter estimates obtained without restrictions. We hypothesize a set of restrictions

$$H_0: \mathbf{c}(\boldsymbol{\theta}) = \mathbf{q}.$$

<sup>6</sup>Of course, our use of the large-sample result in a sample of 10 might be questionable.

<sup>7</sup>Note that because both likelihoods are restricted in this instance, there is nothing to prevent  $-2 \ln \lambda$  from being negative.

<sup>8</sup>If the mean is not  $\boldsymbol{\mu}$ , then the statistic in (14-22) will have a **noncentral chi-squared distribution**. This distribution has the same basic shape as the central chi-squared distribution, with the same degrees of freedom, but lies to the right of it. Thus, a random draw from the noncentral distribution will tend, on average, to be larger than a random observation from the central distribution.

## 528 PART III ♦ Estimation Methodology

If the restrictions are valid, then at least approximately  $\hat{\theta}$  should satisfy them. If the hypothesis is erroneous, however, then  $\mathbf{c}(\hat{\theta}) - \mathbf{q}$  should be farther from  $\mathbf{0}$  than would be explained by sampling variability alone. The device we use to formalize this idea is the Wald test.

### THEOREM 14.6 Limiting Distribution of the Wald Test Statistic

The Wald statistic is

$$W = [\mathbf{c}(\hat{\theta}) - \mathbf{q}]' (\text{Asy.Var}[\mathbf{c}(\hat{\theta}) - \mathbf{q}])^{-1} [\mathbf{c}(\hat{\theta}) - \mathbf{q}].$$

Under  $H_0$ , ~~in large samples~~,  $W$  has a ~~a~~ squared distribution with degrees of freedom equal to the number of restrictions [i.e., the number of equations in  $\mathbf{c}(\hat{\theta}) - \mathbf{q} = \mathbf{0}$ ]. A derivation of the limiting distribution of the Wald statistic appears in Theorem 5.1.

This test is analogous to the chi-squared statistic in (14-22) if  $\mathbf{c}(\hat{\theta}) - \mathbf{q}$  is normally distributed with the hypothesized mean of  $\mathbf{0}$ . A large value of  $W$  leads to rejection of the hypothesis. Note, finally, that  $W$  only requires computation of the unrestricted model. One must still compute the covariance matrix appearing in the preceding quadratic form. This result is the variance of a possibly nonlinear function, which we treated earlier.

$$\begin{aligned} \text{Est. Asy. Var}[\mathbf{c}(\hat{\theta}) - \mathbf{q}] &= \hat{\mathbf{C}} \text{Est. Asy. Var}[\hat{\theta}] \hat{\mathbf{C}}', \\ \hat{\mathbf{C}} &= \left[ \frac{\partial \mathbf{c}(\hat{\theta})}{\partial \hat{\theta}'} \right]. \end{aligned} \quad (14-23)$$

That is,  $\mathbf{C}$  is the  $J \times K$  matrix whose  $j$ th row is the derivatives of the  $j$ th constraint with respect to the  $K$  elements of  $\theta$ . A common application occurs in testing a set of linear restrictions.

For testing a set of linear restrictions  $\mathbf{R}\theta = \mathbf{q}$ , the Wald test would be based on

$$\begin{aligned} H_0: \mathbf{c}(\theta) - \mathbf{q} &= \mathbf{R}\theta - \mathbf{q} = \mathbf{0}, \\ \hat{\mathbf{C}} &= \left[ \frac{\partial \mathbf{c}(\hat{\theta})}{\partial \hat{\theta}'} \right] = \mathbf{R}' \end{aligned} \quad (14-24)$$

$$\text{Est. Asy. Var}[\mathbf{c}(\hat{\theta}) - \mathbf{q}] = \mathbf{R} \text{Est. Asy. Var}[\hat{\theta}] \mathbf{R}',$$

and

$$W = [\mathbf{R}\hat{\theta} - \mathbf{q}]' [\mathbf{R} \text{Est. Asy. Var}(\hat{\theta}) \mathbf{R}']^{-1} [\mathbf{R}\hat{\theta} - \mathbf{q}].$$

The degrees of freedom is the number of rows in  $\mathbf{R}$ .

If  $\mathbf{c}(\theta) = \mathbf{q}$  is a single restriction, then the Wald test will be the same as the test based on the confidence interval developed previously. If the test is

$$H_0: \theta = \theta_0 \quad \text{versus} \quad H_1: \theta \neq \theta_0,$$

then the earlier test is based on

$$z = \frac{|\hat{\theta} - \theta_0|}{s(\hat{\theta})}, \quad (14-25)$$

## CHAPTER 14 ♦ Maximum Likelihood Estimation 529

where  $s(\hat{\theta})$  is the estimated asymptotic standard error. The test statistic is compared to the appropriate value from the standard normal table. The Wald test will be based on

$$W = [(\hat{\theta} - \theta_0) - 0] (\text{Asy. Var}[(\hat{\theta} - \theta_0) - 0])^{-1} [(\hat{\theta} - \theta_0) - 0] = \frac{(\hat{\theta} - \theta_0)^2}{\text{Asy. Var}[\hat{\theta}]} = z^2. \quad (14-26)$$

Here  $W$  has a  squared distribution with one degree of freedom, which is the distribution of the square of the standard normal test statistic in (14-25).

To summarize, the Wald test is based on measuring the extent to which the unrestricted estimates fail to satisfy the hypothesized restrictions. There are two shortcomings of the Wald test. First, it is a pure significance test against the null hypothesis, not necessarily for a specific alternative hypothesis. As such, its power may be limited in some settings. In fact, the test statistic tends to be rather large in applications. The second shortcoming is not shared by either of the other test statistics discussed here. The Wald statistic is not invariant to the formulation of the restrictions. For example, for a test of the hypothesis that a function  $\theta = \beta/(1 - \gamma)$  equals a specific value  $q$  there are two approaches one might choose. A Wald test based directly on  $\theta - q = 0$  would use a statistic based on the variance of this nonlinear function. An alternative approach would be to analyze the linear restriction  $\beta - q(1 - \gamma) = 0$ , which is an equivalent, but linear, restriction. The Wald statistics for these two tests could be different and might lead to different inferences. These two shortcomings have been widely viewed as compelling arguments against use of the Wald test. But, in its favor, the Wald test does not rely on a strong distributional assumption, as do the likelihood ratio and Lagrange multiplier tests. The recent econometrics literature is replete with applications that are based on distribution free estimation procedures, such as the GMM method. As such, in recent years, the Wald test has enjoyed a redemption of sorts.

#### 14.6.3 THE LAGRANGE MULTIPLIER TEST

The third test procedure is the **Lagrange multiplier (LM) or efficient score (or just score) test**. It is based on the restricted model instead of the unrestricted model. Suppose that we maximize the log-likelihood subject to the set of constraints  $\mathbf{c}(\boldsymbol{\theta}) - \mathbf{q} = \mathbf{0}$ . Let  $\lambda$  be a vector of Lagrange multipliers and define the Lagrangean function

$$\ln L^*(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) + \lambda'(\mathbf{c}(\boldsymbol{\theta}) - \mathbf{q}).$$

The solution to the constrained maximization problem is the root of

$$\begin{aligned} \frac{\partial \ln L^*}{\partial \boldsymbol{\theta}} &= \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \mathbf{C}'\lambda = \mathbf{0}, \\ \frac{\partial \ln L^*}{\partial \lambda} &= \mathbf{c}(\boldsymbol{\theta}) - \mathbf{q} = \mathbf{0}, \end{aligned} \quad (14-27)$$

where  $\mathbf{C}'$  is the transpose of the derivatives matrix in the second line of (14-23). If the restrictions are valid, then imposing them will not lead to a significant difference in the maximized value of the likelihood function. In the first-order conditions, the meaning is that the second term in the derivative vector will be small. In particular,  $\lambda$  will be small. We could test this directly, that is, test  $H_0: \lambda = \mathbf{0}$ , which leads to the Lagrange multiplier test. There is an equivalent simpler formulation, however. At the restricted maximum,

## 530 PART III ♦ Estimation Methodology

the derivatives of the log-likelihood function are

$$\frac{\partial \ln L(\hat{\theta}_R)}{\partial \hat{\theta}_R} = -\hat{\mathbf{C}}'\hat{\boldsymbol{\lambda}} = \hat{\mathbf{g}}_R. \quad (14-28)$$

If the restrictions are valid, at least within the range of sampling variability, then  $\hat{\mathbf{g}}_R = \mathbf{0}$ . That is, the derivatives of the log-likelihood evaluated at the restricted parameter vector will be approximately zero. The vector of first derivatives of the log-likelihood is the vector of **efficient scores**. Because the test is based on this vector, it is called the **score test** as well as the Lagrange multiplier test. The variance of the first derivative vector is the information matrix, which we have used to compute the asymptotic covariance matrix of the MLE. The test statistic is based on reasoning analogous to that underlying the Wald test statistic.

### THEOREM 14.7 Limiting Distribution of the Lagrange Multiplier Statistic

The Lagrange multiplier test statistic is

$$LM = \left( \frac{\partial \ln L(\hat{\theta}_R)}{\partial \hat{\theta}_R} \right)' [\mathbf{I}(\hat{\theta}_R)]^{-1} \left( \frac{\partial \ln L(\hat{\theta}_R)}{\partial \hat{\theta}_R} \right).$$

Under the null hypothesis, LM has a limiting chi-squared distribution with degrees of freedom equal to the number of restrictions. All terms are computed at the restricted estimator.

The LM statistic has a useful form. Let  $\hat{\mathbf{g}}_{iR}$  denote the  $i$ th term in the gradient of the log-likelihood function. Then,

$$\hat{\mathbf{g}}_R = \sum_{i=1}^n \hat{\mathbf{g}}_{iR} = \hat{\mathbf{G}}'_R \mathbf{i},$$

where  $\hat{\mathbf{G}}_R$  is the  $n \times K$  matrix with  $i$ th row equal to  $\hat{\mathbf{g}}'_{iR}$  and  $\mathbf{i}$  is a column of 1s. If we use the BHHH (outer product of gradients) estimator in (14-18) to estimate the Hessian, then

$$[\hat{\mathbf{I}}(\hat{\theta})]^{-1} = [\hat{\mathbf{G}}'_R \hat{\mathbf{G}}_R]^{-1},$$

and

$$LM = \mathbf{i}' \hat{\mathbf{G}}_R [\hat{\mathbf{G}}'_R \hat{\mathbf{G}}_R]^{-1} \hat{\mathbf{G}}'_R \mathbf{i}.$$

Now, because  $\mathbf{i}'\mathbf{i}$  equals  $n$ ,  $LM = n(\mathbf{i}' \hat{\mathbf{G}}_R [\hat{\mathbf{G}}'_R \hat{\mathbf{G}}_R]^{-1} \hat{\mathbf{G}}'_R \mathbf{i}/n) = nR^2$ , which is  $n$  times the uncentered squared multiple correlation coefficient in a linear regression of a column of 1s on the derivatives of the log-likelihood function computed at the restricted estimator. We will encounter this result in various forms at several points in the book.

## CHAPTER 14 ♦ Maximum Likelihood Estimation 531

## 14.6.4 AN APPLICATION OF THE LIKELIHOOD-BASED TEST PROCEDURES

Consider, again, the data in Example C.1. In Example 14.4, the parameter  $\beta$  in the model

$$f(y_i | x_i, \beta) = \frac{1}{\beta + x_i} e^{-y_i/(\beta + x_i)} \quad (14-29)$$

was estimated by maximum likelihood. For convenience, let  $\beta_i = 1/(\beta + x_i)$ . This exponential density is a restricted form of a more general gamma distribution,

$$f(y_i | x_i, \beta, \rho) = \frac{\beta_i^\rho}{\Gamma(\rho)} y_i^{\rho-1} e^{-y_i \beta_i}. \quad (14-30)$$

The restriction is  $\rho = 1$ .<sup>9</sup> We consider testing the hypothesis

$$H_0: \rho = 1 \quad \text{versus} \quad H_1: \rho \neq 1$$

using the various procedures described previously. The log-likelihood and its derivatives are

$$\begin{aligned} \ln L(\beta, \rho) &= \rho \sum_{i=1}^n \ln \beta_i - n \ln \Gamma(\rho) + (\rho - 1) \sum_{i=1}^n \ln y_i - \sum_{i=1}^n y_i \beta_i, \\ \frac{\partial \ln L}{\partial \beta} &= -\rho \sum_{i=1}^n \beta_i + \sum_{i=1}^n y_i \beta_i^2, \quad \frac{\partial \ln L}{\partial \rho} = \sum_{i=1}^n \ln \beta_i - n \Psi(\rho) + \sum_{i=1}^n \ln y_i, \quad (14-31) \\ \frac{\partial^2 \ln L}{\partial \beta^2} &= \rho \sum_{i=1}^n \beta_i^2 - 2 \sum_{i=1}^n y_i \beta_i^3, \quad \frac{\partial^2 \ln L}{\partial \rho^2} = -n \Psi'(\rho), \quad \frac{\partial^2 \ln L}{\partial \beta \partial \rho} = -\sum_{i=1}^n \beta_i. \end{aligned}$$

[Recall that  $\Psi(\rho) = d \ln \Gamma(\rho) / d\rho$  and  $\Psi'(\rho) = d^2 \ln \Gamma(\rho) / d\rho^2$ .] Unrestricted maximum likelihood estimates of  $\beta$  and  $\rho$  are obtained by equating the two first derivatives to zero. The restricted maximum likelihood estimate of  $\beta$  is obtained by equating  $\partial \ln L / \partial \beta$  to zero while fixing  $\rho$  at one. The results are shown in Table 14.1. Three estimators are available for the asymptotic covariance matrix of the estimators of  $\theta = (\beta, \rho)'$ . Using the actual Hessian as in (14-15), we compute  $\mathbf{V} = [-\Sigma_i \partial^2 \ln f(y_i | x_i, \beta, \rho) / \partial \theta \partial \theta']^{-1}$  at the maximum likelihood estimates. For this model, it is easy to show that  $E[y_i | x_i] = \rho(\beta + x_i)$  (either by direct integration or, more simply, by using the result that  $E[\partial \ln L / \partial \beta] = 0$  to deduce it). Therefore, we can also use the expected Hessian as in (14-16) to compute  $\mathbf{V}_E = [-\Sigma_i E[\partial^2 \ln f(y_i | x_i, \beta, \rho) / \partial \theta \partial \theta']]^{-1}$ . Finally, by using the sums of squares and cross products of the first derivatives, we obtain the BHHH estimator in (14-18),  $\mathbf{V}_B = [\Sigma_i (\partial \ln f(y_i | x_i, \beta, \rho) / \partial \theta)(\partial \ln f(y_i | x_i, \beta, \rho) / \partial \theta')]^{-1}$ . Results in Table 14.1 are based on  $\mathbf{V}$ .

The three estimators of the asymptotic covariance matrix produce notably different results:

$$\mathbf{V} = \begin{bmatrix} 5.499 & -1.653 \\ -1.653 & 0.6309 \end{bmatrix}, \quad \mathbf{V}_E = \begin{bmatrix} 4.900 & -1.473 \\ -1.473 & 0.5768 \end{bmatrix}, \quad \mathbf{V}_B = \begin{bmatrix} 13.37 & -4.322 \\ -4.322 & 1.537 \end{bmatrix}.$$

<sup>9</sup>The gamma function  $\Gamma(\rho)$  and the gamma distribution are described in Sections B.4.5 and E2.3.

**532 PART III ♦ Estimation Methodology**
**TABLE 14.1** Maximum Likelihood Estimates

<i>Quantity</i>	<i>Unrestricted Estimate<sup>a</sup></i>	<i>Restricted Estimate</i>
$\beta$	-4.7185 (2.345)	15.6027 (6.794)
$\rho$	3.1509 (0.794)	1.0000 (0.000)
$\ln L$	-82.91605	-88.43626
$\partial \ln L / \partial \beta$	0.0000	0.0000
$\partial \ln L / \partial \rho$	0.0000	7.9145
$\partial^2 \ln L / \partial \beta^2$	-0.85570	-0.02166
$\partial^2 \ln L / \partial \rho^2$	-7.4592	-32.8987
$\partial^2 \ln L / \partial \beta \partial \rho$	-2.2420	-0.66891

<sup>a</sup>Estimated asymptotic standard errors based on  $\mathbf{V}$  are given in parentheses.

Given the small sample size, the differences are to be expected. Nonetheless, the striking difference of the BHHH estimator is typical of its erratic performance in small samples.

- **Confidence interval test:** A 95 percent confidence interval for  $\rho$  based on the unrestricted estimates is  $3.1509 \pm 1.96\sqrt{0.6309} = [1.5941, 4.7076]$ . This interval does not contain  $\rho = 1$ , so the hypothesis is rejected.
- **Likelihood ratio test:** The LR statistic is  $\lambda = -2[-88.437 - (-82.914)] = 11.0404$ . The table value for the test, with one degree of freedom, is 3.842. The computed value is larger than this critical value, so the hypothesis is again rejected.
- **Wald test:** The Wald test is based on the unrestricted estimates. For this restriction,  $c(\theta) - q = \rho - 1$ ,  $dc(\hat{\rho})/d\hat{\rho} = 1$ ,  $\text{Est. Asy. Var}[c(\hat{\rho}) - q] = \text{Est. Asy. Var}[\hat{\rho}] = 0.6309$ , so  $W = (3.1517 - 1)^2/[0.6309] = 7.3384$ . The critical value is the same as the previous one. Hence,  $H_0$  is once again rejected. Note that the Wald statistic is the square of the corresponding test statistic that would be used in the confidence interval test,  $|3.1509 - 1|/\sqrt{0.6309} = 2.7335$ .
- **Lagrange multiplier test:** The Lagrange multiplier test is based on the restricted estimators. The estimated asymptotic covariance matrix of the derivatives used to compute the statistic can be any of the three estimators discussed earlier. The BHHH estimator,  $\mathbf{V}_B$ , is the empirical estimator of the variance of the gradient and is the one usually used in practice. This computation produces

$$\text{LM} = [0.0000 \quad 7.9145] \begin{bmatrix} 0.00995 & 0.26776 \\ 0.26776 & 11.199 \end{bmatrix}^{-1} \begin{bmatrix} 0.0000 \\ 7.9145 \end{bmatrix} = 15.687.$$

The conclusion is the same as before. Note that the same computation done using  $\mathbf{V}$  rather than  $\mathbf{V}_B$  produces a value of 5.1162. As before, we observe substantial small sample variation produced by the different estimators.

The latter three test statistics have substantially different values. It is possible to reach different conclusions, depending on which one is used. For example, if the test had been carried out at the 1 percent level of significance instead of 5 percent and LM had been computed using  $\mathbf{V}$ , then the critical value from the chi-squared statistic would have been 6.635 and the hypothesis would not have been rejected by the LM test. Asymptotically, all three tests are equivalent. But, in a finite sample such as this one,

## CHAPTER 14 ♦ Maximum Likelihood Estimation 533

differences are to be expected.<sup>10</sup> Unfortunately, there is no clear rule for how to proceed in such a case, which highlights the problem of relying on a particular significance level and drawing a firm reject or accept conclusion based on sample evidence.

#### 14.6.5 COMPARING MODELS AND COMPUTING MODEL FIT

The test statistics described in Sections 14.6.1–14.6.3 are available for assessing the validity of restrictions on the parameters in a model. When the models are nested, any of the three mentioned testing procedures can be used. For nonnested models, the computation is a comparison of one model to another based on an estimation criterion to discern which is to be preferred. Two common measures that are based on the same logic as the adjusted  $R^2$  for the linear model are

$$\text{Akaike information criterion (AIC)} = -2 \ln L + 2K,$$

$$\text{Bayes (Schwarz) information criterion (BIC)} = -2 \ln L + K \ln n,$$

where  $K$  is the number of parameters in the model. Choosing a model based on the lowest AIC is logically the same as using  $R^2$  in the linear model; nonstatistical, albeit widely accepted.

The AIC and BIC are information criteria, not fit measures as such. This does leave open the question of how to assess the “fit” of the model. Only the case of a linear least squares regression in a model with a constant term produces an  $R^2$ , which measures the proportion of variation explained by the regression. The ambiguity in  $R^2$  as a fit measure arose immediately when we moved from the linear regression model to the generalized regression model in Chapter 9. The problem is yet more acute in the context of the models we consider in this chapter. For example, the estimators of the models for count data in Example 14.10 make no use of the “variation” in the dependent variable and there is no obvious measure of “explained variation.”

A measure of “fit” that was originally proposed for discrete choice models in McFadden (1974), but surprisingly has gained wide currency throughout the empirical literature is the **likelihood ratio index**, which has come to be known as the **Pseudo  $R^2$** . It is computed as

$$\text{Pseudo } R^2 = 1 - (\ln L) / (\ln L_0)$$


where  $\ln L$  is the log-likelihood for the model estimated and  $\ln L_0$  is the log-likelihood for the same model with only a constant term. The statistic does resemble the  $R^2$  in a linear regression. The choice of name is for this statistic is unfortunate, however, because even in the discrete choice context for which it was proposed, it has no connection to the fit of the model to the data. In discrete choice settings in which log-likelihoods must be negative, the pseudo  $R^2$  must be between zero and one and rises as variables are added to the model. It can obviously be zero, but is usually bounded below one. In the linear model with normally distributed disturbances, the maximized log-likelihood is

$$\ln L = (-n/2)[1 + \ln 2\pi + \ln(\mathbf{e}'\mathbf{e}/n)].$$

<sup>10</sup>For further discussion of this problem, see Berndt and Savin (1977).

## 534 PART III ♦ Estimation Methodology

With a small amount of manipulation, we find that the pseudo  $R^2$  for the linear regression model is

$$\text{Pseudo } R^2 = \frac{-\ln(1 - R^2)}{1 + \ln 2\pi + \ln s_y^2},$$

while the “true”  $R^2$  is  $1 - \mathbf{e}'\mathbf{e}/\mathbf{e}'\mathbf{e}_0$ . Because  $s_y^2$  can vary independently of  $R^2$ —multiplying  $\mathbf{y}$  by any scalar,  $A$ , leaves  $R^2$  unchanged but multiplies  $s_y^2$  by  $A^2$ —although the upper limit is one, there is no lower limit on this measure. This same problem arises in any model that uses information on the scale of a dependent variable, such as the tobit model (Chapter 16). The computation makes even less sense as a fit measure in multinomial models such as the ordered probit model (Chapter 17) or the multinomial logit model. For discrete choice models, there are a variety of such measures discussed in Chapter 17. For limited dependent variable and many loglinear models, some other measure that is related to a correlation between a prediction and the actual value would be more useable. Nonetheless, the measure seems to have gained currency in the contemporary literature. [The popular software package, *Stata*, reports the pseudo  $R^2$  with every model fit by MLE, but at the same time, admonishes its users not to interpret it as anything meaningful. See, for example, <http://www.stata.com/support/faqs/stat/pseudor2.html>. Cameron and Trivedi (2005) document the pseudo  $R^2$  at length and then give similar cautions about it and urge their readers to seek a more meaningful measure of the correlation between model predictions and the outcome variable of interest. Wooldridge (2002a) dismisses it summarily, and argues that coefficients are more interesting.]

### 14.6.6 VUONG'S TEST AND THE KULLBACK-LEIBLER INFORMATION CRITERION

Vuong's (1989) approach to testing **nonnested models** is also based on the likelihood ratio statistic. The logic of the test is similar to that which motivates the likelihood ratio test in general. Suppose that  $f(y_i | \mathbf{Z}_i, \boldsymbol{\theta})$  and  $g(y_i | \mathbf{Z}_i, \boldsymbol{\gamma})$  are two competing models for the density of the random variable  $y_i$ , with  $f$  being the null model,  $H_0$ , and  $g$  being the alternative,  $H_1$ . For instance, in Example 5.7, both densities are (by assumption now) normal,  $y_i$  is consumption,  $C_t$ ,  $\mathbf{Z}_i$  is  $[1, Y_t, Y_{t-1}, C_{t-1}]$ ,  $\boldsymbol{\theta}$  is  $(\beta_1, \beta_2, \beta_3, 0, \sigma^2)$ ,  $\boldsymbol{\gamma}$  is  $(\gamma_1, \gamma_2, 0, \gamma_3, \omega^2)$ , and  $\sigma^2$  and  $\omega^2$  are the respective conditional variances of the disturbances,  $\varepsilon_{0t}$  and  $\varepsilon_{1t}$ . The crucial element of Vuong's analysis is that it need not be the case that either competing model is “true”; they may both be incorrect. What we want to do is attempt to use the data to determine which competitor is closer to the truth, that is, closer to the correct (unknown) model.

We assume that observations in the sample (disturbances) are conditionally independent. Let  $L_{i,0}$  denote the  $i$ th contribution to the likelihood function under the null hypothesis. Thus, the log likelihood function under the null hypothesis is  $\Sigma_i \ln L_{i,0}$ . Define  $L_{i,1}$  likewise for the alternative model. Now, let  $m_i$  equal  $\ln L_{i,1} - \ln L_{i,0}$ . If we were using the familiar likelihood ratio test, then, the likelihood ratio statistic would be simply  $LR = 2\Sigma_i m_i = 2n \bar{m}$  when  $L_{i,0}$  and  $L_{i,1}$  are computed at the respective maximum likelihood estimators. When the competing models are nested— $H_0$  is a restriction on  $H_1$ —we know that  $\Sigma_i m_i \geq 0$ . The restrictions of the null hypothesis will never increase the likelihood function. (In the linear regression model with normally distributed disturbances

## CHAPTER 14 ♦ Maximum Likelihood Estimation 535

that we have examined so far, the log likelihood and these results are all based on the sum of squared residuals, and as we have seen, imposing restrictions never reduces the sum of squares.) The limiting distribution of the  $LR$  statistic under the assumption of the null hypothesis is chi squared with degrees of freedom equal to the reduction in the number of dimensions of the parameter space of the alternative hypothesis that results from imposing the restrictions.

Vuong's analysis is concerned with nonnested models for which  $\Sigma_i m_i$  need not be positive. Formalizing the test requires us to look more closely at what is meant by the "right" model (and provides a convenient departure point for the discussion in the next two sections). In the context of nonnested models, Vuong allows for the possibility that neither model is "true" in the absolute sense. We maintain the classical assumption that there does exist a "true" model,  $h(y_i | \mathbf{Z}_i, \boldsymbol{\alpha})$  where  $\boldsymbol{\alpha}$  is the "true" parameter vector, but possibly neither hypothesized model is that true model. The **Kullback–Leibler Information Criterion** (KLIC) measures the distance between the true model (distribution) and a hypothesized model in terms of the likelihood function. Loosely, the KLIC is the log likelihood function under the hypothesis of the true model minus the log-likelihood function for the (misspecified) hypothesized model under the assumption of the true model. Formally, for the model of the null hypothesis,

$$\text{KLIC} = E[\ln h(y_i | \mathbf{Z}_i, \boldsymbol{\alpha}) | h \text{ is true}] - E[\ln f(y_i | \mathbf{Z}_i, \boldsymbol{\theta}) | h \text{ is true}].$$

The first term on the right hand side is what we would estimate with  $(1/n)\ln L$  if we maximized the log likelihood for the true model,  $h(y_i | \mathbf{Z}_i, \boldsymbol{\alpha})$ . The second term is what is estimated by  $(1/n)\ln L$  assuming (incorrectly) that  $f(y_i | \mathbf{Z}_i, \boldsymbol{\theta})$  is the correct model. Notice that  $f(y_i | \mathbf{Z}_i, \boldsymbol{\theta})$  is written in terms of a parameter vector,  $\boldsymbol{\theta}$ . Because  $\boldsymbol{\alpha}$  is the "true" parameter vector, it is perhaps ambiguous what is meant by the parameterization,  $\boldsymbol{\theta}$ . Vuong (p. 310) calls this the "pseudottrue" parameter vector. It is the vector of constants that the estimator converges to when one uses the estimator implied by  $f(y_i | \mathbf{Z}_i, \boldsymbol{\theta})$ . In Example 5.2, if  $H_0$  gives the correct model, this formulation assumes that the least squares estimator in  $H_1$  would converge to some vector of pseudo-true parameters. But, these are not the parameters of the correct model—they would be the slopes in the population linear projection of  $C_t$  on  $[1, Y_t, C_{t-1}]$ .

Suppose the "true" model is  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ , with normally distributed disturbances and  $\mathbf{y} = \mathbf{Z}\boldsymbol{\delta} + \mathbf{w}$  is the proposed competing model. The KLIC would be the expected log likelihood function for the true model minus the expected log likelihood function for the second model, still assuming that the first one is the truth. By construction, the KLIC is positive. We will now say that one model is "better" than another if it is closer to the "truth" based on the KLIC. If we take the difference of the two KLICs for two models, the true log likelihood function falls out, and we are left with

$$\text{KLIC}_1 - \text{KLIC}_0 = E[\ln f(y_i | \mathbf{Z}_i, \boldsymbol{\theta}) | h \text{ is true}] - E[\ln g(y_i | \mathbf{Z}_i, \boldsymbol{\gamma}) | h \text{ is true}].$$

To compute this using a sample, we would simply compute the likelihood ratio statistic,  $n\bar{m}$  (without multiplying by 2) again. Thus, this provides an interpretation of the LR statistic. But, in this context, the statistic can be negative—we don't know which competing model is closer to the truth.

## 536 PART III ♦ Estimation Methodology

Vuong's general result for nonnested models (his Theorem 5.1) describes the behavior of the statistic

$$V = \frac{\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n m_i \right)}{\sqrt{\frac{1}{n} \sum_{i=1}^n (m_i - \bar{m})^2}} = \sqrt{n}(\bar{m}/s_m), \quad m_i = \ln L_{i,0} - \ln L_{i,1}.$$

He finds:

1. Under the hypothesis that the models are "equivalent",  $V \xrightarrow{D} N[0, 1]$
2. Under the hypothesis that  $f(y_i | \mathbf{Z}_i, \theta)$  is "better",  $V \xrightarrow{A.S.} +\infty$
3. Under the hypothesis that  $g(y_i | \mathbf{Z}_i, \gamma)$  is "better",  $V \xrightarrow{A.S.} -\infty$ .

This test is directional. Large positive values favor the null model while large negative values favor the alternative. The intermediate values (e.g., between  $-1.96$  and  $+1.96$  for 95% cent significance) are an inconclusive region. An application appears in Example 19.10.

## 14.7 TWO-STEP MAXIMUM LIKELIHOOD ESTIMATION

The applied literature contains a large and increasing number of applications in which elements of one model are embedded in another, which produces what are known as "two-step" estimation problems. [Among the best known of these is Heckman's (1979) model of sample selection discussed in Example 1.1 and in Chapter 14.] There are two parameter vectors,  $\theta_1$  and  $\theta_2$ . The first appears in the second model, but not the reverse. In such a situation, there are two ways to proceed. **Full information maximum likelihood (FIML)** estimation would involve forming the joint distribution  $f(y_1, y_2 | \mathbf{x}_1, \mathbf{x}_2, \theta_1, \theta_2)$  of the two random variables and then maximizing the full log-likelihood function,

$$\ln L(\theta_1, \theta_2) = \sum_{i=1}^n \ln f(y_{i1}, y_{i2} | \mathbf{x}_{i1}, \mathbf{x}_{i2}, \theta_1, \theta_2).$$

A two-step procedure for this kind of model could be used by estimating the parameters of model 1 first by maximizing

$$\ln L_1(\theta_1) = \sum_{i=1}^n \ln f_1(y_{i1} | \mathbf{x}_{i1}, \theta_1)$$

and then maximizing the marginal likelihood function for  $y_2$  while embedding the consistent estimator of  $\theta_1$ , treating it as given. The second step involves maximizing

$$\ln L_2(\hat{\theta}_1, \theta_2) = \sum_{i=1}^n \ln f_2(y_{i2} | \mathbf{x}_{i2}, \hat{\theta}_1, \theta_2).$$

There are at least two reasons one might proceed in this fashion. First, it may be straightforward to formulate the two separate log-likelihoods, but very complicated to derive the joint distribution. This situation frequently arises when the two variables being modeled are from different kinds of populations, such as one discrete and one continuous (which is a very common case in this framework). The second reason is that maximizing the separate log-likelihoods may be fairly straightforward, but maximizing the joint

## CHAPTER 14 ♦ Maximum Likelihood Estimation 537

log-likelihood may be numerically complicated or difficult.<sup>11</sup> The results given here can be found in an important reference on the subject, Murphy and Topel (2002, first published in 1985).

Suppose, then, that our model consists of the two marginal distributions,  $f_1(y_1 | \mathbf{x}_1, \boldsymbol{\theta}_1)$  and  $f_2(y_2 | \mathbf{x}_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ . Estimation proceeds in two steps.

1. Estimate  $\boldsymbol{\theta}_1$  by maximum likelihood in model 1. Let  $\hat{\mathbf{V}}_1$  be  $n$  times any of the estimators of the asymptotic covariance matrix of this estimator that were discussed in Section 14.4.6.
2. Estimate  $\boldsymbol{\theta}_2$  by maximum likelihood in model 2, with  $\hat{\boldsymbol{\theta}}_1$  inserted in place of  $\boldsymbol{\theta}_1$  as if it were known. Let  $\hat{\mathbf{V}}_2$  be  $n$  times any appropriate estimator of the asymptotic covariance matrix of  $\hat{\boldsymbol{\theta}}_2$ .

The argument for consistency of  $\hat{\boldsymbol{\theta}}_2$  is essentially that if  $\boldsymbol{\theta}_1$  were known, then all our results for MLEs would apply for estimation of  $\boldsymbol{\theta}_2$ , and because  $\text{plim } \hat{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}_1$ , asymptotically, this line of reasoning is correct. (See point 3 Theorem D.16.) But the same line of reasoning is not sufficient to justify using  $(1/n)\hat{\mathbf{V}}_2$  as the estimator of the asymptotic covariance matrix of  $\hat{\boldsymbol{\theta}}_2$ . Some correction is necessary to account for an estimate of  $\boldsymbol{\theta}_1$  being used in estimation of  $\boldsymbol{\theta}_2$ . The essential result is the following.

**THEOREM 14.8 Asymptotic Distribution of the Two-Step MLE  
[Murphy and Topel (2002)]**

If the standard regularity conditions are met for both log-likelihood functions, then the second-step maximum likelihood estimator of  $\boldsymbol{\theta}_2$  is consistent and asymptotically normally distributed with asymptotic covariance matrix

$$\mathbf{V}_2^* = \frac{1}{n} [\mathbf{V}_2 + \mathbf{V}_2[\mathbf{C}\mathbf{V}_1\mathbf{C}' - \mathbf{R}\mathbf{V}_1\mathbf{C}' - \mathbf{C}\mathbf{V}_1\mathbf{R}']\mathbf{V}_2],$$

where

$$\mathbf{V}_1 = \text{Asy. Var}[\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1)] \text{ based on } \ln L_1,$$

$$\mathbf{V}_2 = \text{Asy. Var}[\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2)] \text{ based on } \ln L_2 | \boldsymbol{\theta}_1,$$

$$\mathbf{C} = E\left[\frac{1}{n}\left(\frac{\partial \ln L_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial \ln L_2}{\partial \boldsymbol{\theta}'_1}\right)\right], \quad \mathbf{R} = E\left[\frac{1}{n}\left(\frac{\partial \ln L_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial \ln L_1}{\partial \boldsymbol{\theta}'_1}\right)\right].$$

The correction of the asymptotic covariance matrix at the second step requires some additional computation. Matrices  $\mathbf{V}_1$  and  $\mathbf{V}_2$  are estimated by the respective uncorrected covariance matrices. Typically, the BHHH estimators,

$$\hat{\mathbf{V}}_1 = \left[ \frac{1}{n} \sum_{i=1}^n \left( \frac{\partial \ln f_{i1}}{\partial \hat{\boldsymbol{\theta}}_1} \right) \left( \frac{\partial \ln f_{i1}}{\partial \hat{\boldsymbol{\theta}}_1} \right)' \right]^{-1}$$

<sup>11</sup>There is a third possible motivation. If either model is misspecified, then the FIML estimates of both models will be inconsistent. But if only the second is misspecified, at least the first will be estimated consistently. Of course, this result is only "half a loaf," but it may be better than none.

**538 PART III ♦ Estimation Methodology**
**THEOREM 14.8 (Continued)**

and

$$\hat{\mathbf{V}}_2 = \left[ \frac{1}{n} \sum_{i=1}^n \left( \frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2} \right) \left( \frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}'_2} \right) \right]^{-1}$$

are used. The matrices  $\mathbf{R}$  and  $\mathbf{C}$  are obtained by summing the individual observations on the cross products of the derivatives. These are estimated with

$$\hat{\mathbf{C}} = \frac{1}{n} \sum_{i=1}^n \left( \frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2} \right) \left( \frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}'_1} \right)$$

and

$$\hat{\mathbf{R}} = \frac{1}{n} \sum_{i=1}^n \left( \frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2} \right) \left( \frac{\partial \ln f_{i1}}{\partial \hat{\boldsymbol{\theta}}'_1} \right).$$

A derivation of this useful result is instructive. We will rely on (14-11) and the results of Section 14.4.5.b where the asymptotic normality of the maximum likelihood estimator is developed. The first step MLE of  $\boldsymbol{\theta}_1$  is defined by

$$\begin{aligned} \frac{1}{n} \frac{\partial \ln L_1(\hat{\boldsymbol{\theta}}_1)}{\partial \hat{\boldsymbol{\theta}}_1} &= \frac{1}{n} \sum_{i=1}^n \frac{\partial \ln f_1(y_{i1} | \mathbf{x}_{i1}, \hat{\boldsymbol{\theta}}_1)}{\partial \hat{\boldsymbol{\theta}}_1} \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{g}_{i1}(\hat{\boldsymbol{\theta}}_1) = \bar{\mathbf{g}}_1(\hat{\boldsymbol{\theta}}_1) = \mathbf{0}. \end{aligned}$$

Using the results in that section, we obtained the asymptotic distribution from (14-15),

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \xrightarrow{d} \left[ -\mathbf{H}_{11}^{(1)}(\boldsymbol{\theta}_1) \right]^{-1} \sqrt{n}\bar{\mathbf{g}}_1(\boldsymbol{\theta}_1),$$

where the expression means that the limiting distribution of the two random vectors is the same, and

$$\mathbf{H}_{11}^{(1)} = E \left[ \frac{1}{n} \frac{\partial^2 \ln L_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}'_1} \right].$$

The second step MLE of  $\boldsymbol{\theta}_2$  is defined by

$$\begin{aligned} \frac{1}{n} \frac{\partial \ln L_2(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)}{\partial \hat{\boldsymbol{\theta}}_2} &= \frac{1}{n} \sum_{i=1}^n \frac{\partial \ln f_2(y_{i2} | \mathbf{x}_{i1}, \mathbf{x}_{i2}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)}{\partial \hat{\boldsymbol{\theta}}_2} \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{g}_{i2}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \bar{\mathbf{g}}_2(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \mathbf{0}. \end{aligned}$$

Expand the derivative vector,  $\bar{\mathbf{g}}_2(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ , in a linear Taylor series as usual, and use the results in Section 16.4.5.b once again;

$$\begin{aligned} \bar{\mathbf{g}}_2(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) &= \bar{\mathbf{g}}_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + \left[ \mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \right] (\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2) \\ &\quad + \left[ \mathbf{H}_{21}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \right] (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) + o(1/n) = \mathbf{0}. \end{aligned}$$


## CHAPTER 14 ♦ Maximum Likelihood Estimation 539

where

$$\mathbf{H}_{21}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = E\left[\frac{1}{n} \frac{\partial^2 \ln L_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}'_1}\right] \text{ and } \mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = E\left[\frac{1}{n} \frac{\partial^2 \ln L_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}'_2}\right].$$

To obtain the asymptotic distribution, we use the same device as before,

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2) &\xrightarrow{d} \left[ -\mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \right]^{-1} \sqrt{n}\bar{\mathbf{g}}_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &+ \left[ -\mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \right]^{-1} \left[ \mathbf{H}_{21}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \right] \sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1). \end{aligned}$$

For convenience, denote  $\mathbf{H}_{22}^{(2)} = \mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ ,  $\mathbf{H}_{21}^{(2)} = \mathbf{H}_{21}^{(2)}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$  and  $\mathbf{H}_{11}^{(1)} = \mathbf{H}_{11}^{(1)}(\boldsymbol{\theta}_1)$ . Now substitute the first step estimator of  $\boldsymbol{\theta}_1$  in this expression to obtain

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2) &\xrightarrow{d} \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1} \sqrt{n}\bar{\mathbf{g}}_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &+ \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1} \left[ \mathbf{H}_{21}^{(2)} \right] \left[ -\mathbf{H}_{11}^{(1)} \right]^{-1} \sqrt{n}\bar{\mathbf{g}}_1(\boldsymbol{\theta}_1). \end{aligned}$$

Consistency and asymptotic normality of the two estimators follow from our earlier results. To obtain the asymptotic covariance matrix for  $\hat{\boldsymbol{\theta}}_2$  we will obtain the limiting variance of the random vector in the preceding expression. The joint normal distribution of the two first derivative vectors has zero means and

$$Var \begin{bmatrix} \sqrt{n}\bar{\mathbf{g}}_1(\boldsymbol{\theta}_1) \\ \sqrt{n}\bar{\mathbf{g}}_2(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Then, the asymptotic covariance matrix we seek is

$$\begin{aligned} Var [\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2)] &= \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1} \boldsymbol{\Sigma}_{22} \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1} \\ &+ \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1} \left[ \mathbf{H}_{21}^{(2)} \right] \left[ -\mathbf{H}_{11}^{(1)} \right]^{-1} \boldsymbol{\Sigma}_{11} \left[ -\mathbf{H}_{11}^{(1)} \right]^{-1} \left[ \mathbf{H}_{21}^{(2)} \right]' \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1} \\ &+ \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1} \boldsymbol{\Sigma}_{21} \left[ -\mathbf{H}_{11}^{(1)} \right]^{-1} \left[ \mathbf{H}_{21}^{(2)} \right]' \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1} \\ &+ \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1} \left[ \mathbf{H}_{21}^{(2)} \right] \left[ -\mathbf{H}_{11}^{(1)} \right]^{-1} \boldsymbol{\Sigma}_{12} \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1}. \end{aligned}$$

As we found earlier, the variance of the first derivative vector of the log likelihood is the negative of the expected second derivative matrix [see (14-11)]. Therefore  $\boldsymbol{\Sigma}_{22} = [-\mathbf{H}_{22}^{(2)}]$  and  $\boldsymbol{\Sigma}_{11} = [-\mathbf{H}_{11}^{(1)}]$ . Making the substitution we obtain

$$\begin{aligned} Var [\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2)] &= \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1} + \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1} \left[ \mathbf{H}_{21}^{(2)} \right] \left[ -\mathbf{H}_{11}^{(1)} \right]^{-1} \left[ \mathbf{H}_{21}^{(2)} \right]' \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1} \\ &+ \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1} \boldsymbol{\Sigma}_{21} \left[ -\mathbf{H}_{11}^{(1)} \right]^{-1} \left[ \mathbf{H}_{21}^{(2)} \right]' \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1} \\ &+ \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1} \left[ \mathbf{H}_{21}^{(2)} \right] \left[ -\mathbf{H}_{11}^{(1)} \right]^{-1} \boldsymbol{\Sigma}_{12} \left[ -\mathbf{H}_{22}^{(2)} \right]^{-1}. \end{aligned}$$

## 540 PART III ♦ Estimation Methodology

From (14-15),  $[-\mathbf{H}_{11}^{(1)}]^{-1}$  and  $[-\mathbf{H}_{22}^{(2)}]^{-1}$  are the  $\mathbf{V}_1$  and  $\mathbf{V}_2$  that appear in Theorem 14.8, which further reduces the expression to

$$\begin{aligned} & \text{Var} [\sqrt{n}(\hat{\theta}_2 - \theta_2)] \\ &= V_2 \cancel{+} V_2 \left[ \mathbf{H}_{21}^{(2)} \right] \mathbf{V}_1 \left[ \mathbf{H}_{21}^{(2)} \right]' \mathbf{V}_2 - \mathbf{V}_2 \Sigma_{21} \mathbf{V}_1 \left[ \mathbf{H}_{21}^{(2)} \right]' \mathbf{V}_2 - \mathbf{V}_2 \left[ \mathbf{H}_{21}^{(2)} \right] \mathbf{V}_1 \Sigma_{12} \mathbf{V}_2. \end{aligned}$$

Two remaining terms are  $\mathbf{H}_{21}^{(2)}$  which is the  $E[\partial^2 \ln L_2(\theta_1, \theta_2)/\partial \theta_2 \partial \theta_1]$ , which is being estimated by  $-\mathbf{C}$  in the statement of the theorem [note (14-11) again for the change of sign] and  $\Sigma_{21}$  which is the covariance of the two first derivative vectors. This is being estimated by  $\mathbf{R}$  in Theorem 14.8. Making these last two substitutions produces

$$\text{Var} [\sqrt{n}(\hat{\theta}_2 - \theta_2)] = \mathbf{V}_2 + \mathbf{V}_2 \mathbf{C} \mathbf{V}_1 \mathbf{C}' \mathbf{V}_2 - \mathbf{V}_2 \mathbf{R} \mathbf{V}_1 \mathbf{C}' \mathbf{V}_2 - \mathbf{V}_2 \mathbf{C} \mathbf{V}_1 \mathbf{R}' \mathbf{V}_2,$$

which completes the derivation.

### Example 14.5 Two-Step ML Estimation

A common application of the two-step method is accounting for the variation in a constructed regressor in a second step model. In this instance, the constructed variable is often an estimate of an expected value of a variable that is likely to be endogenous in the second step model. In this example, we will construct a rudimentary model that illustrates the computations.

In Riphahn, Wambach and Million (RWM, 2003), the authors studied whether individuals' use of the German health care system was at least partly explained by whether or not they had purchased a particular type of supplementary health insurance. We have used their data set, German Socioeconomic Panel (GSOEP) at several points. (See, e.g., Example 7.6.) One of the variables of interest in the study is *DocVis*, the number of times  $i$  an individual visits the doctor during the survey year. RWM considered the possibility that the presence of supplementary (*Addon*) insurance had an influence on the number of visits. Our simple model is as follows: The model for the number of visits is a Poisson regression (see Section 19.2). This is a loglinear model that we will specify as

$$E[\text{DocVis}|\mathbf{x}_2, P_{\text{Addon}}] = \mu(\mathbf{x}_2'\beta + \mathbf{x}_1'\alpha) = \exp[\mathbf{x}_2'\beta + \mathbf{x}_1'\alpha].$$

The model contains not the dummy variable 1 if the individual has *Addon* insurance and 0 otherwise, which is likely to be endogenous in the equation, but an estimate of  $E[\text{Addon}|\mathbf{x}_1]$  from a logistic probability model (see Section 17.5) for whether the individual has insurance,

$$\Lambda(\mathbf{x}_1'\alpha) = \frac{\exp(\mathbf{x}_1'\alpha)}{1 + \exp(\mathbf{x}_1'\alpha)} = \text{Prob}[\text{Individual has purchased Addon insurance} | \mathbf{x}_1].$$

For purposes of the exercise we will specify

$$(y_1 = \text{Addon}) \mathbf{x}_1 = (\text{constant}, \text{Age}, \text{Education}, \text{Married}, \text{Kids}),$$

$$(y_2 = \text{DocVis}) \mathbf{x}_2 = (\text{constant}, \text{Age}, \text{Education}, \text{Income}, \text{Female}).$$

As before, to sidestep issues related to the panel data nature of the data set, we will use the 4483 observations in the 1988 wave of the data set, and drop the two observations for which *Income* is zero.

The log likelihood for the logistic probability model is

$$\ln L_1(\alpha) = \sum_i \{(1 - y_{i1}) \ln[1 - \Lambda(\mathbf{x}_{i1}'\alpha)] + y_{i1} \ln \Lambda(\mathbf{x}_{i1}'\alpha)\}.$$

The derivatives of this log-likelihood are

$$\mathbf{g}_{i1}(\alpha) = \partial \ln f_1(y_{i1}|\mathbf{x}_{i1}, \alpha) / \partial \alpha = [y_{i1} - \Lambda(\mathbf{x}_{i1}'\alpha)] \mathbf{x}_{i1}.$$

## CHAPTER 14 ♦ Maximum Likelihood Estimation 541

We will maximize this log likelihood with respect to  $\beta$  and then compute  $\mathbf{V}_1$  using the BHHH estimator, as in Theorem 14.8. We will also use  $\mathbf{g}_{i1}(\alpha)$  for computing  $\mathbf{R}$ .

The log-likelihood for the Poisson regression model is

$$\ln L_2 = \sum_i [-\mu(\mathbf{x}'_{i2}\beta, \gamma, \mathbf{x}'_{i1}\alpha) + y_{i2} \ln \mu(\mathbf{x}'_{i2}\beta, \gamma, \mathbf{x}'_{i1}\alpha) - \ln y'_{i2}].$$

The derivatives of this log likelihood are

$$\begin{aligned}\mathbf{g}_{i2}^{(2)}(\beta, \gamma, \alpha) &= \partial \ln f_2(y_{i2}, \mathbf{x}'_{i1}, \mathbf{x}'_{i2}, \beta, \gamma, \alpha) / \partial (\beta', \gamma)' = [y_{i2} - \mu(\mathbf{x}'_{i2}\beta, \gamma, \mathbf{x}'_{i1}\alpha)] [\mathbf{x}'_{i2}, \Lambda(\mathbf{x}'_{i1}\alpha)]' \\ \mathbf{g}_{i1}^{(2)}(\beta, \gamma, \alpha) &= \partial \ln f_2(y_{i2}, \mathbf{x}'_{i1}, \mathbf{x}'_{i2}, \beta, \gamma, \alpha) / \partial \alpha = [y_{i2} - \mu(\mathbf{x}'_{i2}\beta, \gamma, \mathbf{x}'_{i1}\alpha)] \gamma \Lambda(\mathbf{x}'_{i1}\alpha) [1 - \Lambda(\mathbf{x}'_{i1}\alpha)] \mathbf{x}'_{i1}.\end{aligned}$$

We will use  $\mathbf{g}_{i2}^{(2)}$  for computing  $\mathbf{V}_2$  and in computing  $\mathbf{R}$  and  $\mathbf{C}$  and  $\mathbf{g}_{i1}^{(2)}$  in computing  $\mathbf{C}$ . In particular,

$$\begin{aligned}\mathbf{V}_1 &= [(1/n) \sum_i \mathbf{g}_{i1}(\alpha) \mathbf{g}_{i1}(\alpha)']^{-1}, \\ \mathbf{V}_2 &= [(1/n) \sum_i \mathbf{g}_{i2}^{(2)}(\beta, \gamma, \alpha) \mathbf{g}_{i2}^{(2)}(\beta, \gamma, \alpha)']^{-1}, \\ \mathbf{C} &= [(1/n) \sum_i \mathbf{g}_{i2}^{(2)}(\beta, \gamma, \alpha) \mathbf{g}_{i1}^{(2)}(\beta, \gamma, \alpha)'], \\ \mathbf{R} &= [(1/n) \sum_i \mathbf{g}_{i2}^{(2)}(\beta, \gamma, \alpha) \mathbf{g}_{i1}(\alpha)'].\end{aligned}$$

Table 14.2 presents the two-step maximum likelihood estimates of the model parameters and estimated standard errors. For the first-step logistic model, the standard errors marked  $\mathbf{H}_1$  vs.  $\mathbf{V}_1$  compares the values computed using the negative inverse of the second derivatives matrix ( $\mathbf{H}_1$ ) vs. the outer products of the first derivatives ( $\mathbf{V}_1$ ). As expected with a sample this large, the difference is minor. The latter were used in computing the corrected covariance matrix at the second step. In the Poisson model, the comparison of  $\mathbf{V}_2$  to  $\mathbf{V}_2^*$  shows distinctly that accounting for the presence of  $\hat{\alpha}$  in the constructed regressor has a substantial impact on the standard errors, even in this relatively large sample. Note that the effect of the correction is to double the standard errors on the coefficients for the variables that the equations have in common, but it is quite minor for *Income* and *Female*, which are unique to the second step model.

The covariance of the two gradients,  $\mathbf{R}$ , may converge to zero in a particular application. When the first- and second-step estimates are based on different samples,  $\mathbf{R}$  is exactly zero. For example, in our earlier application,  $\mathbf{R}$  is based on two residuals,

$$\mathbf{g}_{i1} = \{Addon_i - E[Addon_i | \mathbf{x}'_{i1}]\} \text{ and } \mathbf{g}_{i2}^{(2)} = \{DocVis_i - E[DocVis_i | \mathbf{x}'_{i2}, \Lambda_{i1}]\}.$$

The two residuals may well be uncorrelated. This assumption would be checked on a model-by-model basis, but in such an instance, the third and fourth terms in  $\mathbf{V}_2$  vanish

**TABLE 14.2** Estimated Logistic and Poisson Models

	Logistic Model for Addon		Poisson Model for DocVis			
	Coefficient	Standard Error ( $\mathbf{H}_1$ )	Standard Error ( $\mathbf{V}_1$ )	Coefficient	Standard Error ( $\mathbf{V}_2$ )	Standard Error ( $\mathbf{V}_2^*$ )
Constant	-6.19246	0.60228	0.58287	0.77808	0.04884	0.09319
Age	0.01486	0.00912	0.00924	0.01752	0.00044	0.00111
Education	0.16091	0.03003	0.03326	-0.03858	0.00462	0.00980
Married	0.22206	0.23584	0.23523			
Kids	-0.10822	0.21591	0.21993			
Income				-0.80298	0.02339	0.02719
Female				0.16409	0.00601	0.00770
$\Lambda(\mathbf{x}'_{i1}\alpha)$				3.91140	0.77283	1.87014

## 542 PART III ♦ Estimation Methodology

asymptotically and what remains is the simpler alternative,

$$\mathbf{V}_2^{**} = (1/n)[\mathbf{V}_2 + \mathbf{V}_2 \mathbf{C} \mathbf{V}_1 \mathbf{C}' \mathbf{V}_2].$$

(In our application, the sample correlation between  $\mathbf{g}_{i1}$  and  $\mathbf{g}_{i2}^{(2)}$  is only 0.015658 and the elements of the estimate of  $\mathbf{R}$  are only about 0.01 times the corresponding elements of  $\mathbf{C}$ —essentially about 99 percent of the correction in  $\mathbf{V}_2^*$  is accounted for by  $\mathbf{C}$ .)

It has been suggested that this set of procedures might be more complicated than necessary. [E.g., Cameron and Trivedi (2005, p. 202).] There are two alternative approaches one might take. First, under general circumstances, the asymptotic covariance matrix of the second-step estimator could be approximated using the bootstrapping procedure discussed in Section 15.1. We would note, however, if this approach is taken, then it is essential that both steps be “bootstrapped.” Otherwise, taking  $\hat{\theta}_1$  as given and fixed, we will end up estimating  $(1/n)\mathbf{V}_2$ , not the appropriate covariance matrix. The point of the exercise is to account for the variation in  $\hat{\theta}_1$ . The second possibility is to fit the full model at once. That is, use a one-step, full information maximum likelihood estimator and estimate  $\theta_1$  and  $\theta_2$  simultaneously. Of course, this is usually the procedure we sought to avoid in the first place. And with modern software, this two-step method is often quite straightforward. Nonetheless, this is occasionally a possibility. Once again, Heckman’s (1979) famous sample selection model provides an illuminating case. The two-step and full information estimators for Heckman’s model are developed in Section 18.5.3.

### 14.8 PSEUDO-MAXIMUM LIKELIHOOD ESTIMATION AND ROBUST ASYMPTOTIC COVARIANCE MATRICES

Maximum likelihood estimation requires complete specification of the distribution of the observed random variable. If the correct distribution is something other than what we assume, then the likelihood function is misspecified and the desirable properties of the MLE might not hold. This section considers a set of results on an estimation approach that is robust to some kinds of model misspecification. For example, we have found that in a model, if the conditional mean function is  $E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$ , then certain estimators, such as least squares, are “robust” to specifying the wrong distribution of the disturbances. That is, LS is MLE if the disturbances are normally distributed, but we can still claim some desirable properties for LS, including consistency, even if the disturbances are not normally distributed. This section will discuss some results that relate to what happens if we maximize the “wrong” log-likelihood function, and for those cases in which the estimator is consistent despite this, how to compute an appropriate asymptotic covariance matrix for it.<sup>12</sup>

<sup>12</sup>The following will sketch a set of results related to this estimation problem. The important references on this subject are White (1982a); Gourieroux, Monfort, and Trognon (1984); Huber (1967); and Amemiya (1985). A recent work with a large amount of discussion on the subject is Mittelhammer et al. (2000). The derivations in these works are complex, and we will only attempt to provide an intuitive introduction to the topic.

## CHAPTER 14 ♦ Maximum Likelihood Estimation 543

## 14.8.1 MAXIMUM LIKELIHOOD AND GMM ESTIMATION

Let  $f(y_i | \mathbf{x}_i, \boldsymbol{\beta})$  be the true probability density for a random variable  $y_i$  given a set of covariates  $\mathbf{x}_i$  and parameter vector  $\boldsymbol{\beta}$ . The log-likelihood function is  $(1/n) \ln L(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) = (1/n) \sum_{i=1}^n \ln f(y_i | \mathbf{x}_i, \boldsymbol{\beta})$ . The MLE,  $\hat{\boldsymbol{\beta}}_{ML}$ , is the sample statistic that maximizes this function. (The division of  $\ln L$  by  $n$  does not affect the solution.) We maximize the log-likelihood function by equating its derivatives to zero, so the MLE is obtained by solving the set of empirical moment equations

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial \ln f(y_i | \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{ML})}{\partial \hat{\boldsymbol{\beta}}_{ML}} = \frac{1}{n} \sum_{i=1}^n \mathbf{d}_i(\hat{\boldsymbol{\beta}}_{ML}) = \bar{\mathbf{d}}(\hat{\boldsymbol{\beta}}_{ML}) = \mathbf{0}.$$

The population counterpart to the sample moment equation is

$$E \left[ \frac{1}{n} \frac{\partial \ln L}{\partial \boldsymbol{\beta}} \right] = E \left[ \frac{1}{n} \sum_{i=1}^n \mathbf{d}_i(\boldsymbol{\beta}) \right] = E[\bar{\mathbf{d}}(\boldsymbol{\beta})] = \mathbf{0}.$$

Using what we know about GMM estimators, if  $E[\bar{\mathbf{d}}(\boldsymbol{\beta})] = \mathbf{0}$ , then  $\hat{\boldsymbol{\beta}}_{ML}$  is consistent and asymptotically normally distributed, with asymptotic covariance matrix equal to

$$\mathbf{V}_{ML} = [\mathbf{G}(\boldsymbol{\beta})' \mathbf{G}(\boldsymbol{\beta})]^{-1} \mathbf{G}(\boldsymbol{\beta})' \{ \text{Var}[\bar{\mathbf{d}}(\boldsymbol{\beta})] \} \mathbf{G}(\boldsymbol{\beta}) [\mathbf{G}(\boldsymbol{\beta})' \mathbf{G}(\boldsymbol{\beta})]^{-1},$$

where  $\mathbf{G}(\boldsymbol{\beta}) = \text{plim } \partial \bar{\mathbf{d}}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}'$ . Because  $\bar{\mathbf{d}}(\boldsymbol{\beta})$  is the derivative vector,  $\mathbf{G}(\boldsymbol{\beta})$  is  $1/n$  times the expected Hessian of  $\ln L$ ; that is,  $(1/n) E[\mathbf{H}(\boldsymbol{\beta})] = \bar{\mathbf{H}}(\boldsymbol{\beta})$ . As we saw earlier,  $\text{Var}[\partial \ln L / \partial \boldsymbol{\beta}] = -E[\mathbf{H}(\boldsymbol{\beta})]$ . Collecting all seven appearances of  $(1/n) E[\mathbf{H}(\boldsymbol{\beta})]$ , we obtain the familiar result  $\mathbf{V}_{ML} = \{-E[\mathbf{H}(\boldsymbol{\beta})]\}^{-1}$ . [All the  $n$ 's cancel and  $\text{Var}[\bar{\mathbf{d}}] = (1/n) \bar{\mathbf{H}}(\boldsymbol{\beta})$ .] Note that this result depends crucially on the result  $\text{Var}[\partial \ln L / \partial \boldsymbol{\beta}] = -E[\mathbf{H}(\boldsymbol{\beta})]$ .

## 14.8.2 MAXIMUM LIKELIHOOD AND M ESTIMATION

The maximum likelihood estimator is obtained by maximizing the function  $\bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = (1/n) \sum_{i=1}^n \ln f(y_i, \mathbf{x}_i, \boldsymbol{\beta})$ . This function converges to its expectation as  $n \rightarrow \infty$ . Because this function is the log-likelihood for the sample, it is also the case (not proven here) that as  $n \rightarrow \infty$ , it attains its unique maximum at the true parameter vector,  $\boldsymbol{\beta}$ . (We used this result in proving the consistency of the maximum likelihood estimator.) Since  $\text{plim } \bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = E[\bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})]$ , it follows (by interchanging differentiation and the expectation operation) that  $\text{plim } \partial \bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) / \partial \boldsymbol{\beta} = E[\partial \bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}]$ . But, if this function achieves its *maximum* at  $\boldsymbol{\beta}$ , then it must be the case that  $\text{plim } \partial \bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) / \partial \boldsymbol{\beta} = \mathbf{0}$ .

An estimator that is obtained by maximizing a criterion function is called an **M estimator** [Huber (1967)] or an extremum estimator [Amemiya (1985)]. Suppose that we obtain an estimator by maximizing some other function,  $M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})$  that, although not the log-likelihood function, also attains its unique maximum at the true  $\boldsymbol{\beta}$  as  $n \rightarrow \infty$ . Then the preceding argument might produce a consistent estimator with a known asymptotic distribution. For example, the log-likelihood for a linear regression model with normally distributed disturbances with *different* variances,  $\sigma^2 \omega_i$ , is

$$\bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{-1}{2} \left[ \ln(2\pi\sigma^2 \omega_i) + \frac{(y_i - \mathbf{x}_i' \boldsymbol{\beta})^2}{\sigma^2 \omega_i} \right] \right\}.$$

### 544 PART III ♦ Estimation Methodology

By maximizing this function, we obtain the maximum likelihood estimator. But we also examined another estimator, simple least squares, which maximizes  $M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = -(1/n) \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2$ . As we showed earlier, least squares is consistent and asymptotically normally distributed even with this extension, so it qualifies as an  $M$  estimator of the sort we are considering here.

Now consider the general case. Suppose that we estimate  $\boldsymbol{\beta}$  by maximizing a criterion function

$$M_n(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \ln g(y_i | \mathbf{x}_i, \boldsymbol{\beta}).$$

Suppose as well that  $\text{plim } M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = E[M_n(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta})]$  and that as  $n \rightarrow \infty$ ,  $E[M_n(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta})]$  attains its unique maximum at  $\boldsymbol{\beta}$ . Then, by the argument we used earlier for the MLE,  $\text{plim } \partial M_n(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}) / \partial \boldsymbol{\beta} = E[\partial M_n(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}] = \mathbf{0}$ . Once again, we have a set of moment equations for estimation. Let  $\hat{\boldsymbol{\beta}}_E$  be the estimator that maximizes  $M_n(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta})$ . Then the estimator is defined by

$$\frac{\partial M_n(\mathbf{y} | \mathbf{X}, \hat{\boldsymbol{\beta}}_E)}{\partial \hat{\boldsymbol{\beta}}_E} = \frac{1}{n} \sum_{i=1}^n \frac{\partial \ln g(y_i | \mathbf{x}_i, \hat{\boldsymbol{\beta}}_E)}{\partial \hat{\boldsymbol{\beta}}_E} = \bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}_E) = \mathbf{0}.$$

Thus,  $\hat{\boldsymbol{\beta}}_E$  is a GMM estimator. Using the notation of our earlier discussion,  $\mathbf{G}(\hat{\boldsymbol{\beta}}_E)$  is the symmetric Hessian of  $E[M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})]$ , which we will denote  $(1/n)E[\mathbf{H}_M(\hat{\boldsymbol{\beta}}_E)] = \bar{\mathbf{H}}_M(\hat{\boldsymbol{\beta}}_E)$ . Proceeding as we did above to obtain  $\mathbf{V}_{ML}$ , we find that the appropriate asymptotic covariance matrix for the extremum estimator would be

$$\mathbf{V}_E = [\bar{\mathbf{H}}_M(\boldsymbol{\beta})]^{-1} \left( \frac{1}{n} \boldsymbol{\Phi} \right) [\bar{\mathbf{H}}_M(\boldsymbol{\beta})]^{-1},$$

where  $\boldsymbol{\Phi} = \text{Var}[\partial \log g(y_i | \mathbf{x}_i, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}]$ , and, as before, the asymptotic distribution is normal.

The Hessian in  $\mathbf{V}_E$  can easily be estimated by using its empirical counterpart,

$$\text{Est.}[\bar{\mathbf{H}}_M(\hat{\boldsymbol{\beta}}_E)] = \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \ln g(y_i | \mathbf{x}_i, \hat{\boldsymbol{\beta}}_E)}{\partial \hat{\boldsymbol{\beta}}_E \partial \hat{\boldsymbol{\beta}}'_E}.$$

But,  $\boldsymbol{\Phi}$  remains to be specified, and it is unlikely that we would know what function to use. The important difference is that in this case, the variance of the first derivatives vector need not equal the Hessian, so  $\mathbf{V}_E$  does not simplify. We can, however, consistently estimate  $\boldsymbol{\Phi}$  by using the sample variance of the first derivatives,

$$\hat{\boldsymbol{\Phi}} = \frac{1}{n} \sum_{i=1}^n \left[ \frac{\partial \ln g(y_i | \mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} \right] \left[ \frac{\partial \ln g(y_i | \mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}'} \right].$$

If this were the maximum likelihood estimator, then  $\hat{\boldsymbol{\Phi}}$  would be the OPG estimator that we have used at several points. For example, for the least squares estimator in the heteroscedastic linear regression model, the criterion is  $M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = -(1/n) \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2$ , the solution is  $\mathbf{b}$ ,  $\mathbf{G}(\mathbf{b}) = (-2/n)\mathbf{X}'\mathbf{X}$ , and

$$\hat{\boldsymbol{\Phi}} = \frac{1}{n} \sum_{i=1}^n [2\mathbf{x}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta})][2\mathbf{x}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta})]' = \frac{4}{n} \sum_{i=1}^n e_i^2 \mathbf{x}_i \mathbf{x}'_i.$$

Collecting terms, the 4s cancel and we are left precisely with the White estimator of (9-27)!

## CHAPTER 14 ♦ Maximum Likelihood Estimation 545

## 14.8.3 SANDWICH ESTIMATORS

At this point, we consider the motivation for all this weighty theory. One disadvantage of maximum likelihood estimation is its requirement that the density of the observed random variable(s) be fully specified. The preceding discussion suggests that in some situations, we can make somewhat fewer assumptions about the distribution than a full specification would require. The extremum estimator is robust to some kinds of specification errors. One useful result to emerge from this derivation is an estimator for the asymptotic covariance matrix of the extremum estimator that is robust at least to some misspecification. In particular, if we obtain  $\hat{\beta}_E$  by maximizing a criterion function that satisfies the other assumptions, then the appropriate estimator of the asymptotic covariance matrix is

$$\text{Est. } \mathbf{V}_E = \frac{1}{n} [\bar{\mathbf{H}}(\hat{\beta}_E)]^{-1} \hat{\Phi}(\hat{\beta}_E) [\bar{\mathbf{H}}(\hat{\beta}_E)]^{-1}.$$

If  $\hat{\beta}_E$  is the true MLE, then  $\mathbf{V}_E$  simplifies to  $\{-[\mathbf{H}(\hat{\beta}_E)]\}^{-1}$ . In the current literature, this estimator has been called the **sandwich estimator**. There is a trend in the current literature to compute this estimator routinely, regardless of the likelihood function. It is worth noting that if the log-likelihood is not specified correctly, then the parameter estimators are likely to be inconsistent, save for the cases such as those noted later, so robust estimation of the asymptotic covariance matrix may be misdirected effort. But if the likelihood function is correct, then the sandwich estimator is unnecessary. This method is not a general patch for misspecified models. Not every likelihood function qualifies as a consistent extremum estimator *for the parameters of interest in the model*.

One might wonder at this point how likely it is that the conditions needed for all this to work will be met. There are applications in the literature in which this machinery has been used that probably do not meet these conditions, such as the tobit model of Chapter 16. We have seen one important case. Least squares in the generalized regression model passes the test. Another important application is models of “individual heterogeneity” in cross-section data. Evidence suggests that simple models often overlook unobserved sources of variation across individuals in cross sections, such as unmeasurable “family effects” in studies of earnings or employment. Suppose that the correct model for a variable is  $h(y_i | \mathbf{x}_i, \mathbf{v}_i, \boldsymbol{\beta}, \theta)$ , where  $\mathbf{v}_i$  is a random term that is not observed and  $\theta$  is a parameter of the distribution of  $\mathbf{v}$ . The correct log-likelihood function is  $\sum_i \ln f(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \theta) = \sum_i \ln f_v h(y_i | \mathbf{x}_i, \mathbf{v}_i, \boldsymbol{\beta}, \theta) f(\mathbf{v}_i) dv_i$ . Suppose that we maximize some other **pseudo-log-likelihood function**,  $\sum_i \ln g(y_i | \mathbf{x}_i, \boldsymbol{\beta})$  and then use the sandwich estimator to estimate the asymptotic covariance matrix of  $\hat{\boldsymbol{\beta}}$ . Does this produce a consistent estimator of the true parameter vector? Surprisingly, sometimes it does, even though it has ignored the nuisance parameter,  $\theta$ . We saw one case, using OLS in the GR model with heteroscedastic disturbances. Inappropriately fitting a Poisson model when the negative binomial model is correct—see Chapter 19—is another case. For some specifications, using the wrong likelihood function in the probit model with proportions data is a third. [These examples are suggested, with several others, by Gourieroux, Monfort, and Trognon (1984).] We do emphasize once again that the sandwich estimator, in and of itself, is not necessarily of any virtue if the likelihood function is misspecified and the other conditions for the  $M$  estimator are not met.

## 546 PART III ♦ Estimation Methodology

### 14.8.4 CLUSTER ESTIMATORS

Micro-level, or individual, data are often grouped or “clustered.” A model of production or economic success at the firm level might be based on a group of industries, with multiple firms in each industry. Analyses of student educational attainment might be based on samples of entire classes, or schools, or statewide averages of schools within school districts. And, of course, such “clustering” is the defining feature of a panel data set. We considered several of these types of applications in our analysis of panel data in Chapter 11. The recent literature contains many studies of clustered data in which the analyst has estimated a pooled model but sought to accommodate the expected correlation across observations with a correction to the asymptotic covariance matrix. We used this approach in computing a robust covariance matrix for the pooled least squares estimator in a panel data model [see (11-3) and Example 11.1 in Section 11.1].

For the normal linear regression model, the log-likelihood that we maximize with the pooled least squares estimator is

$$\ln L = \sum_{i=1}^n \sum_{t=1}^{T_i} \left[ -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{1}{2} \frac{(y_{it} - \mathbf{x}'_{it}\beta)^2}{\sigma^2} \right].$$

[See (14-34).] The “cluster-robust” estimator in (11-3) can be written

$$\begin{aligned} \mathbf{W} &= \left( \sum_{i=1}^n \mathbf{X}'_i \mathbf{X}_i \right)^{-1} \left[ \sum_{i=1}^n (\mathbf{X}'_i \mathbf{e}_i)(\mathbf{e}'_i \mathbf{X}_i) \right] \left( \sum_{i=1}^n \mathbf{X}'_i \mathbf{X}_i \right)^{-1} \\ &= \left( -\frac{1}{\hat{\sigma}^2} \sum_{i=1}^n \sum_{t=1}^{T_i} \mathbf{x}'_{it} \mathbf{x}'_{it} \right)^{-1} \left[ \sum_{i=1}^n \left( \sum_{t=1}^{T_i} \frac{1}{\hat{\sigma}^2} \mathbf{x}'_{it} e_{it} \right) \left( \sum_{t=1}^{T_i} \frac{1}{\hat{\sigma}^2} e_{it} \mathbf{x}'_{it} \right) \right] \left( -\frac{1}{\hat{\sigma}^2} \sum_{i=1}^n \sum_{t=1}^{T_i} \mathbf{x}'_{it} \mathbf{x}'_{it} \right)^{-1} \\ &= \left( \sum_{i=1}^n \sum_{t=1}^{T_i} \frac{\partial^2 \ln f_{it}}{\partial \beta \partial \beta'} \right)^{-1} \left[ \sum_{i=1}^n \left( \sum_{t=1}^{T_i} \frac{\partial \ln f_{it}}{\partial \beta} \right) \left( \sum_{t=1}^{T_i} \frac{\partial \ln f_{it}}{\partial \beta'} \right) \right] \left( \sum_{i=1}^n \sum_{t=1}^{T_i} \frac{\partial^2 \ln f_{it}}{\partial \beta \partial \beta'} \right)^{-1}, \end{aligned}$$

where  $f_{it}$  is the normal density with mean  $\mathbf{x}'_{it}\beta$  and variance  $\sigma^2$ . This is precisely the “cluster-corrected” robust covariance matrix that appears elsewhere in the literature [minus an ad hoc “finite population correction” as in (11-4)].

In the generalized linear regression model (as in others), the OLS estimator is consistent, and will have asymptotic covariance matrix equal to

$$\text{Asy. Var}[\mathbf{b}] = (\mathbf{X}' \mathbf{X})^{-1} [\mathbf{X}' (\sigma^2 \boldsymbol{\Omega}) \mathbf{X}] (\mathbf{X}' \mathbf{X})^{-1}.$$

(See Theorem 9.1.) The center matrix in the sandwich for the panel data case can be written

$$\mathbf{X}' (\sigma^2 \boldsymbol{\Omega}) \mathbf{X} = \sum_{i=1}^n \mathbf{X}'_i \boldsymbol{\Sigma} \mathbf{X}_i,$$

which motivates the preceding robust estimator. Whereas when we first encountered it, we motivated the cluster estimator with an appeal to the same logic that leads to the White estimator for heteroscedasticity, we now have an additional result that appears to justify the estimator in terms of the likelihood function.

Consider the specification error that the estimator is intended to accommodate. Suppose that the observations in group  $i$  were multivariate normally distributed with

## CHAPTER 14 ♦ Maximum Likelihood Estimation 547

disturbance mean vector  $\mathbf{0}$  and unrestricted  $T_i \times T_i$  covariance matrix,  $\Sigma_i$ . Then, the appropriate log-likelihood function would be

$$\ln L = \sum_{i=1}^n \left( -T_i/2 \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| - \frac{1}{2} \boldsymbol{\varepsilon}'_i \Sigma_i^{-1} \boldsymbol{\varepsilon}_i \right),$$

where  $\boldsymbol{\varepsilon}_i$  is the  $T_i \times 1$  vector of disturbances for individual  $i$ . Therefore, we have maximized the wrong likelihood function. Indeed, the  $\beta$  that maximizes this log likelihood function is the GLS estimator, not the OLS estimator. OLS, and the cluster corrected estimator given earlier, “work” in the sense that (1) the least squares estimator is consistent in spite of the misspecification and (2) the robust estimator does, indeed, estimate the appropriate asymptotic covariance matrix.

Now, consider the more general case. Suppose the data set consists of  $n$  multivariate observations,  $[y_{i,1}, \dots, y_{i,T_i}]$ ,  $i = 1, \dots, n$ . Each cluster is a draw from joint density  $f_i(\mathbf{y}_i | \mathbf{X}_i, \theta)$ . Once again, to preserve the generality of the result, we will allow the cluster sizes to differ. The appropriate log likelihood for the sample is

$$\ln L = \sum_{i=1}^n \ln f_i(\mathbf{y}_i | \mathbf{X}_i, \theta).$$

Instead of maximizing  $\ln L$ , we maximize a pseudo-log-likelihood

$$\ln L_P = \sum_{i=1}^n \sum_{t=1}^{T_i} \ln g(y_{it} | \mathbf{x}_{it}, \theta),$$

where we make the possibly unreasonable assumption that the same parameter vector,  $\theta$  enters the pseudo-log-likelihood as enters the correct one. Assume that it does. Using our familiar first-order asymptotics, the **pseudo-maximum likelihood estimator** (MLE) will satisfy

$$\begin{aligned} (\hat{\theta}_{P,ML} - \theta) &\approx \left( \frac{1}{\sum_{i=1}^n T_i} \sum_{i=1}^n \sum_{t=1}^{T_i} \frac{\partial^2 \ln f_{it}}{\partial \theta \partial \theta'} \right)^{-1} \left( \frac{1}{\sum_{i=1}^n T_i} \sum_{i=1}^n \sum_{t=1}^{T_i} \frac{\partial \ln f_{it}}{\partial \theta} \right) + (\theta - \beta) \\ &= \left( \frac{1}{\sum_{i=1}^n T_i} \sum_{i=1}^n \sum_{t=1}^{T_i} H_{it} \right)^{-1} \left( \sum_{i=1}^n w_i \bar{\mathbf{g}}_i \right) + (\theta - \beta), \end{aligned}$$

where  $w_i = T_i / \sum_{i=1}^n T_i$  and  $\bar{\mathbf{g}}_i = (1/T_i) \sum_{t=1}^{T_i} \partial \ln f_{it} / \partial \theta$ . The trailing term in the expression is included to allow for the possibility that  $\text{plim } \hat{\theta}_{P,ML} = \beta$ , which may not equal  $\theta$ . [Note, for example, Cameron and Trivedi (2005, p. 842) specifically assume consistency in the generic model they describe.] Taking the expected outer product of this expression to estimate the asymptotic mean squared deviation will produce two terms—the cross term vanishes. The first will be the cluster-corrected matrix that is ubiquitous in the current literature. The second will be the squared error that may persist as  $n$  increases because the pseudo-MLE need not estimate the parameters of the model of interest.

We draw two conclusions. We can justify the cluster estimator based on this approximation. In general, it will estimate the expected squared variation of the pseudo-MLE around its probability limit. Whether it measures the variation around the appropriate

## 548 PART III ♦ Estimation Methodology

parameters of the model hangs on whether the second term equals zero. In words, perhaps not surprisingly, this apparatus only works if the estimator is consistent. Is that likely? Certainly not if the pooled model is ignoring unobservable fixed effects. Moreover, it will be inconsistent in most cases in which the misspecification is to ignore latent random effects as well. The pseudo-MLE is only consistent for random effects in a few special cases, such as the linear model and Poisson and negative binomial models discussed in Chapter 17. It is not consistent in the probit and logit models in which this approach often used. In the end, the cases in which the estimator are consistent are rarely, if ever, enumerated. The upshot is stated succinctly by Freedman (2006, p. 302): “The sandwich algorithm, under stringent regularity conditions, yields variances for the MLE that are asymptotically correct even when the specification—and hence the likelihood function—are incorrect. However, it is quite another thing to ignore bias. It remains unclear why applied workers should care about the variance of an estimator for the wrong parameter.”

### 14.9 APPLICATIONS OF MAXIMUM LIKELIHOOD ESTIMATION

We will now examine several applications of the maximum likelihood estimator (MLE). We begin by developing the ML counterparts to most of the estimators for the classical and generalized regression models in Chapters 4 through 11. (Generally, the development for dynamic models becomes more involved than we are able to pursue here. The one exception we will consider is the standard model of autocorrelation.) We emphasize, in each of these cases, that we have already developed an efficient, generalized method of moments estimator that has the same asymptotic properties as the MLE under the assumption of normality. In more general cases, we will sometimes find that the GMM estimator is actually preferred to the MLE because of its robustness to failures of the distributional assumptions or its freedom from the necessity to make those assumptions in the first place. However, for the extensions of the classical model based on generalized least squares that are treated here, that is not the case. It might be argued that in these cases, the MLE is superfluous. There are occasions when the MLE will be preferred for other reasons, such as its invariance to transformation in nonlinear models and, possibly, its small sample behavior (although that is usually not the case). And, we will examine some nonlinear models in which there is no linear, method of moments counterpart, so the MLE is the natural estimator. Finally, in each case, we will find some useful aspect of the estimator, itself, including the development of algorithms such as Newton’s method and the EM method for latent class models.

#### 14.9.1 THE NORMAL LINEAR REGRESSION MODEL

The linear regression model is

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i.$$

The likelihood function for a sample of  $n$  independent, identically and normally distributed disturbances is

$$L = (2\pi\sigma^2)^{-n/2} e^{-\mathbf{e}' \mathbf{e}/(2\sigma^2)}. \quad (14-32)$$

## CHAPTER 14 ♦ Maximum Likelihood Estimation 549

The transformation from  $\varepsilon_i$  to  $y_i$  is  $\varepsilon_i = y_i - \mathbf{x}_i' \boldsymbol{\beta}$ , so the **Jacobian** for each observation,  $|\partial \varepsilon_i / \partial y_i|$ , is one.<sup>13</sup> Making the transformation, we find that the likelihood function for the  $n$  observations on the observed random variables is

$$L = (2\pi\sigma^2)^{-n/2} e^{(-1/(2\sigma^2))(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}. \quad (14-33)$$

To maximize this function with respect to  $\boldsymbol{\beta}$ , it will be necessary to maximize the exponent or minimize the familiar sum of squares. Taking logs, we obtain the log-likelihood function for the classical regression model:

$$\ln L = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}. \quad (14-34)$$

The necessary conditions for maximizing this log-likelihood are

$$\begin{bmatrix} \frac{\partial \ln L}{\partial \boldsymbol{\beta}} \\ \frac{\partial \ln L}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} \\ \frac{-n}{2\sigma^2} + \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}. \quad (14-35)$$

The values that satisfy these equations are

$$\hat{\boldsymbol{\beta}}_{ML} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{b} \quad \text{and} \quad \hat{\sigma}_{ML}^2 = \frac{\mathbf{e}'\mathbf{e}}{n}. \quad (14-36)$$

The slope estimator is the familiar one, whereas the variance estimator differs from the least squares value by the divisor of  $n$  instead of  $n - K$ .<sup>14</sup>

The Cramér–Rao bound for the variance of an unbiased estimator is the negative inverse of the expectation of

$$\begin{bmatrix} \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} & \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \sigma^2} \\ \frac{\partial^2 \ln L}{\partial \sigma^2 \partial \boldsymbol{\beta}'} & \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} -\frac{\mathbf{X}'\mathbf{X}}{\sigma^2} & -\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{\sigma^4} \\ -\frac{\boldsymbol{\varepsilon}'\mathbf{X}}{\sigma^4} & \frac{n}{2\sigma^4} - \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\sigma^6} \end{bmatrix}. \quad (14-37)$$

In taking expected values, the off-diagonal term vanishes, leaving

$$[\mathbf{I}(\boldsymbol{\beta}, \sigma^2)]^{-1} = \begin{bmatrix} \sigma^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0}' & 2\sigma^4/n \end{bmatrix}. \quad (14-38)$$

The least squares slope estimator is the maximum likelihood estimator for this model. Therefore, it inherits all the desirable *asymptotic* properties of maximum likelihood estimators.

We showed earlier that  $s^2 = \mathbf{e}'\mathbf{e}/(n - K)$  is an unbiased estimator of  $\sigma^2$ . Therefore, the maximum likelihood estimator is biased toward zero:

$$E[\hat{\sigma}_{ML}^2] = \frac{n - K}{n} \sigma^2 = \left(1 - \frac{K}{n}\right) \sigma^2 < \sigma^2. \quad (14-39)$$

<sup>13</sup>See (B-41) in Section B.5. The analysis to follow is conditioned on  $\mathbf{X}$ . To avoid cluttering the notation, we will leave this aspect of the model implicit in the results. As noted earlier, we assume that the data generating process for  $\mathbf{X}$  does not involve  $\boldsymbol{\beta}$  or  $\sigma^2$  and that the data are well behaved as discussed in Chapter 4.

<sup>14</sup>As a general rule, maximum likelihood estimators do not make corrections for degrees of freedom.

## 550 PART III ♦ Estimation Methodology

Despite its small-sample bias, the maximum likelihood estimator of  $\sigma^2$  has the same desirable asymptotic properties. We see in (14-39) that  $s^2$  and  $\hat{\sigma}^2$  differ only by a factor  $-K/n$ , which vanishes in large samples. It is instructive to formalize the asymptotic equivalence of the two. From (14-38), we know that

$$\sqrt{n}(\hat{\sigma}_{\text{ML}}^2 - \sigma^2) \xrightarrow{d} N[0, 2\sigma^4].$$

It follows that

$$z_n = \left(1 - \frac{K}{n}\right)\sqrt{n}(\hat{\sigma}_{\text{ML}}^2 - \sigma^2) + \frac{K}{\sqrt{n}}\sigma^2 \xrightarrow{d} \left(1 - \frac{K}{n}\right)N[0, 2\sigma^4] + \frac{K}{\sqrt{n}}\sigma^2.$$

But  $K/\sqrt{n}$  and  $K/n$  vanish as  $n \rightarrow \infty$ , so the limiting distribution of  $z_n$  is also  $N[0, 2\sigma^4]$ . Because  $z_n = \sqrt{n}(s^2 - \sigma^2)$ , we have shown that the asymptotic distribution of  $s^2$  is the same as that of the maximum likelihood estimator.

The standard test statistic for assessing the validity of a set of linear restrictions in the linear model,  $\mathbf{R}\beta - \mathbf{q} = \mathbf{0}$ , is the  $F$  ratio,

$$F[J, n - K] = \frac{(\mathbf{e}'_* \mathbf{e}_* - \mathbf{e}' \mathbf{e})/J}{\mathbf{e}' \mathbf{e}/(n - K)} = \frac{(\mathbf{R}\mathbf{b} - \mathbf{q})' [\mathbf{R}s^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q})}{J}.$$

With normally distributed disturbances, the  $F$  test is valid in any sample size. There remains a problem with nonlinear restrictions of the form  $\mathbf{c}(\beta) = \mathbf{0}$ , since the counterpart to  $F$ , which we will examine here, has validity only asymptotically even with normally distributed disturbances. In this section, we will reconsider the Wald statistic and examine two related statistics, the likelihood ratio statistic and the Lagrange multiplier statistic. These statistics are both based on the likelihood function and, like the Wald statistic, are generally valid only asymptotically.

No simplicity is gained by restricting ourselves to linear restrictions at this point, so we will consider general hypotheses of the form

$$H_0: \mathbf{c}(\beta) = \mathbf{0},$$

$$H_1: \mathbf{c}(\beta) \neq \mathbf{0}.$$

The **Wald statistic** for testing this hypothesis and its limiting distribution under  $H_0$  would be

$$W = \mathbf{c}(\mathbf{b})' \{ \mathbf{C}(\mathbf{b})[\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}] \mathbf{C}(\mathbf{b})' \}^{-1} \mathbf{c}(\mathbf{b}) \xrightarrow{d} \chi^2[J], \quad (14-40)$$

where

$$\mathbf{C}(\mathbf{b}) = [\partial \mathbf{c}(\mathbf{b}) / \partial \mathbf{b}']. \quad (14-41)$$

The **likelihood ratio (LR) test** is carried out by comparing the values of the log-likelihood function with and without the restrictions imposed. We leave aside for the present how the restricted estimator  $\mathbf{b}_*$  is computed (except for the linear model, which we saw earlier). The test statistic and its limiting distribution under  $H_0$  are

$$\text{LR} = -2[\ln L_* - \ln L] \xrightarrow{d} \chi^2[J]. \quad (14-42)$$

The log-likelihood for the regression model is given in (14-34). The first-order conditions imply that regardless of how the slopes are computed, the estimator of  $\sigma^2$  without restrictions on  $\beta$  will be  $\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})/n$  and likewise for a restricted estimator

## CHAPTER 14 ♦ Maximum Likelihood Estimation 551

$\hat{\sigma}_*^2 = (\mathbf{y} - \mathbf{X}\mathbf{b}_*)'(\mathbf{y} - \mathbf{X}\mathbf{b}_*)/n = \mathbf{e}'_*\mathbf{e}_*/n$ . The **concentrated log-likelihood**<sup>15</sup> will be

$$\ln L_c = -\frac{n}{2}[1 + \ln 2\pi + \ln(\mathbf{e}'\mathbf{e}/n)]$$

and likewise for the restricted case. If we insert these in the definition of LR, then we obtain

$$LR = n \ln[\mathbf{e}'_*\mathbf{e}_*/\mathbf{e}'\mathbf{e}] = n(\ln \hat{\sigma}_*^2 - \ln \hat{\sigma}^2) = n \ln(\hat{\sigma}_*^2/\hat{\sigma}^2). \quad (14-43)$$

The **Lagrange multiplier (LM)** test is based on the gradient of the log-likelihood function. The principle of the test is that if the hypothesis is valid, then at the restricted estimator, the derivatives of the log-likelihood function should be close to zero. There are two ways to carry out the LM test. The log-likelihood function can be maximized subject to a set of restrictions by using

$$\ln L_{LM} = -\frac{n}{2} \left[ \ln 2\pi + \ln \sigma^2 + \frac{[(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)]/n}{\sigma^2} \right] + \lambda'\mathbf{c}(\beta).$$

The first-order conditions for a solution are

$$\begin{bmatrix} \frac{\partial \ln L_{LM}}{\partial \beta} \\ \frac{\partial \ln L_{LM}}{\partial \sigma^2} \\ \frac{\partial \ln L_{LM}}{\partial \lambda} \end{bmatrix} = \begin{bmatrix} \frac{\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta)}{\sigma^2} + \mathbf{C}(\beta)'\lambda \\ \frac{-n}{2\sigma^2} + \frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{2\sigma^4} \\ \mathbf{c}(\beta) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \quad (14-44)$$

The solutions to these equations give the restricted least squares estimator,  $\mathbf{b}_*$ ; the usual variance estimator, now  $\mathbf{e}'_*\mathbf{e}_*/n$ ; and the Lagrange multipliers. There are now two ways to compute the test statistic. In the setting of the classical linear regression model, when we actually compute the Lagrange multipliers, a convenient way to proceed is to test the hypothesis that the multipliers equal zero. For this model, the solution for  $\lambda_*$  is  $\lambda_* = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q})$ . This equation is a linear function of the least squares estimator. If we carry out a *Wald* test of the hypothesis that  $\lambda_*$  equals  $\mathbf{0}$ , then the statistic will be

$$LM = \lambda'_* \{ \text{Est. Var}[\lambda_*] \}^{-1} \lambda_* = (\mathbf{R}\mathbf{b} - \mathbf{q})' [\mathbf{R} s_*^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q}). \quad (14-45)$$

The disturbance variance estimator,  $s_*^2$ , based on the restricted slopes is  $\mathbf{e}'_*\mathbf{e}_*/n$ .

An alternative way to compute the LM statistic often produces interesting results. In most situations, we maximize the log-likelihood function without actually computing the vector of Lagrange multipliers. (The restrictions are usually imposed some other way.) An alternative way to compute the statistic is based on the (general) result that under the hypothesis being tested,

$$E[\partial \ln L / \partial \beta] = E[(1/\sigma^2)\mathbf{X}'\mathbf{e}] = \mathbf{0}$$

and<sup>16</sup>

$$\text{Asy. Var}[\partial \ln L / \partial \beta] = -E[\partial^2 \ln L / \partial \beta \partial \beta']^{-1} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}. \quad (14-46)$$

<sup>15</sup>See Section E4.3.

<sup>16</sup>This makes use of the fact that the Hessian is block diagonal.

## 552 PART III ♦ Estimation Methodology

We can test the hypothesis that at the restricted estimator, the derivatives are equal to zero. The statistic would be

$$\text{LM} = \frac{\mathbf{e}_*'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}_*}{\mathbf{e}_*'\mathbf{e}_*/n} = nR_*^2. \quad (14-47)$$

In this form, the LM statistic is  $n$  times the coefficient of determination in a regression of the residuals  $e_{i*} = (y_i - \mathbf{x}_i'\mathbf{b}_*)$  on the full set of regressors.

With some manipulation we can show that  $W = [n/(n - K)]JF$  and LR and LM are approximately equal to this function of  $F$ .<sup>17</sup> All three statistics converge to  $JF$  as  $n$  increases. The linear model is a special case in that the LR statistic is based only on the unrestricted estimator and does not actually require computation of the restricted least squares estimator, although computation of  $F$  does involve most of the computation of  $\mathbf{b}_*$ . Because the log function is concave, and  $W/n \geq \ln(1 + W/n)$ , Godfrey (1988) also shows that  $W \geq \text{LR} \geq \text{LM}$ , so for the linear model, we have a firm ranking of the three statistics.

There is ample evidence that the asymptotic results for these statistics are problematic in small or moderately sized samples. [See, e.g., Davidson and MacKinnon (2004, pp. 424–428).] The true distributions of all three statistics involve the data and the unknown parameters and, as suggested by the algebra, converge to the  $F$  distribution *from above*. The implication is that critical values from the chi-squared distribution are likely to be too small; that is, using the limiting chi-squared distribution in small or moderately sized samples is likely to exaggerate the significance of empirical results. Thus, in applications, the more conservative  $F$  statistic (or  $t$  for one restriction) is likely to be preferable unless one's data are plentiful.

### 14.9.2 THE GENERALIZED REGRESSION MODEL

For the generalized regression model of Section 14.1,

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, i = 1, \dots, n, \\ E[\boldsymbol{\varepsilon} | \mathbf{X}] &= \mathbf{0}, \\ E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}] &= \sigma^2\boldsymbol{\Omega}, \end{aligned}$$

as before, we first assume that  $\boldsymbol{\Omega}$  is a matrix of known constants. If the disturbances are multivariate normally distributed, then the log-likelihood function for the sample is

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\ln|\boldsymbol{\Omega}|. \quad (14-48)$$

Because  $\boldsymbol{\Omega}$  is a matrix of known constants, the maximum likelihood estimator of  $\boldsymbol{\beta}$  is the vector that minimizes the **generalized sum of squares**,

$$S_*(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

<sup>17</sup>See Godfrey (1988, pp. 49–51).

## CHAPTER 14 ♦ Maximum Likelihood Estimation 553

(hence the name *generalized least squares*). The necessary conditions for maximizing  $L$  are

$$\begin{aligned}\frac{\partial \ln L}{\partial \beta} &= \frac{1}{\sigma^2} \mathbf{X}' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\beta) = \frac{1}{\sigma^2} \mathbf{X}'_* (\mathbf{y}_* - \mathbf{X}_*\beta) = \mathbf{0}, \\ \frac{\partial \ln L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{y} - \mathbf{X}\beta)' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\beta) \\ &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{y}_* - \mathbf{X}_*\beta)' (\mathbf{y}_* - \mathbf{X}_*\beta) = 0.\end{aligned}\tag{14-49}$$

The solutions are the OLS estimators using the transformed data:

$$\hat{\beta}_{ML} = (\mathbf{X}'_* \mathbf{X}_*)^{-1} \mathbf{X}'_* \mathbf{y}_* = (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{y},\tag{14-50}$$

$$\begin{aligned}\hat{\sigma}_{ML}^2 &= \frac{1}{n} (\mathbf{y}_* - \mathbf{X}_*\hat{\beta})' (\mathbf{y}_* - \mathbf{X}_*\hat{\beta}) \\ &= \frac{1}{n} (\mathbf{y} - \mathbf{X}\hat{\beta})' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta}),\end{aligned}\tag{14-51}$$

which implies that with normally distributed disturbances, generalized least squares is also maximum likelihood. As in the classical regression model, the maximum likelihood estimator of  $\sigma^2$  is biased. An unbiased estimator is the one in (9-14). The conclusion, which would be expected, is that when  $\boldsymbol{\Omega}$  is known, the maximum likelihood estimator is generalized least squares.

When  $\boldsymbol{\Omega}$  is unknown and must be estimated, then it is necessary to maximize the log-likelihood in (14-48) with respect to the full set of parameters  $[\beta, \sigma^2, \boldsymbol{\Omega}]$  simultaneously. Because an unrestricted  $\boldsymbol{\Omega}$  alone contains  $n(n+1)/2 - 1$  parameters, it is clear that some restriction will have to be placed on the structure of  $\boldsymbol{\Omega}$  for estimation to proceed. We will examine several applications in which  $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\theta)$  for some smaller vector of parameters in the next several sections. We note only a few general results at this point.

1. For a given value of  $\theta$  the estimator of  $\beta$  would be feasible GLS and the estimator of  $\sigma^2$  would be the estimator in (14-51).
2. The likelihood equations for  $\theta$  will generally be complicated functions of  $\beta$  and  $\sigma^2$ , so joint estimation will be necessary. However, in many cases, for given values of  $\beta$  and  $\sigma^2$ , the estimator of  $\theta$  is straightforward. For example, in the model of (9-15), the iterated estimator of  $\theta$  when  $\beta$  and  $\sigma^2$  and a prior value of  $\theta$  are given is the prior value plus the slope in the regression of  $(e_i^2/\hat{\sigma}_i^2 - 1)$  on  $z_i$ .

The second step suggests a sort of back and forth iteration for this model that will work in many situations—starting with, say, OLS, iterating back and forth between 1 and 2 until convergence will produce the joint maximum likelihood estimator. This situation was examined by Oberhofer and Kmenta (1974), who showed that under some fairly weak requirements, most importantly that  $\theta$  not involve  $\sigma^2$  or any of the parameters in  $\beta$ , this procedure would produce the maximum likelihood estimator. Another implication of this formulation which is simple to show (we leave it as an exercise) is that under the Oberhofer and Kmenta assumption, the asymptotic covariance matrix of the estimator is the same as the GLS estimator. This is the same whether  $\boldsymbol{\Omega}$  is known or estimated, which means that if  $\theta$  and  $\beta$  have no parameters in common, then *exact knowledge of*

## 554 PART III ♦ Estimation Methodology

$\Omega$  brings no gain in asymptotic efficiency in the estimation of  $\beta$  over estimation of  $\beta$  with a consistent estimator of  $\Omega$ .

We will now examine the two primary, single-equation applications: heteroscedasticity and autocorrelation.

### 14.9.2.a Multiplicative Heteroscedasticity

Harvey's (1976) model of multiplicative heteroscedasticity is a very flexible, general model that includes most of the useful formulations as special cases. The general formulation is

$$\sigma_i^2 = \sigma^2 \exp(\mathbf{z}'_i \boldsymbol{\alpha}). \quad (14-52)$$

A model with heteroscedasticity of the form

$$\sigma_i^2 = \sigma^2 \prod_{m=1}^M z_{im}^{\alpha_m} \quad (14-53)$$

results if the logs of the variables are placed in  $z_i$ . The groupwise heteroscedasticity model described in Section 9.8.2 is produced by making  $\mathbf{z}_i$  a set of group dummy variables (one must be omitted). In this case,  $\sigma^2$  is the disturbance variance for the base group whereas for the other groups,  $\sigma_g^2 = \sigma^2 \exp(\alpha_g)$ .

We begin with a useful simplification. Let  $\mathbf{z}_i$  include a constant term so that  $\mathbf{z}'_i = [1, \mathbf{q}'_i]$ , where  $\mathbf{q}_i$  is the original set of variables, and let  $\boldsymbol{\gamma}' = [\ln \sigma^2, \boldsymbol{\alpha}']$ . Then, the model is simply  $\sigma_i^2 = \exp(\mathbf{z}'_i \boldsymbol{\gamma})$ . Once the full parameter vector is estimated,  $\exp(\gamma_1)$  provides the estimator of  $\sigma^2$ . (This estimator uses the invariance result for maximum likelihood estimation. See Section 14.4.5.d.)

The log-likelihood is

$$\begin{aligned} \ln L &= -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^n \ln \sigma_i^2 - \frac{1}{2} \sum_{i=1}^n \frac{\varepsilon_i^2}{\sigma_i^2} \\ &= -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^n \mathbf{z}'_i \boldsymbol{\gamma} - \frac{1}{2} \sum_{i=1}^n \frac{\varepsilon_i^2}{\exp(\mathbf{z}'_i \boldsymbol{\gamma})}. \end{aligned} \quad (14-54)$$

The likelihood equations are

$$\begin{aligned} \frac{\partial \ln L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^n \mathbf{x}_i \frac{\varepsilon_i}{\exp(\mathbf{z}'_i \boldsymbol{\gamma})} = \mathbf{X}' \boldsymbol{\Omega}^{-1} \boldsymbol{\epsilon} = \mathbf{0}, \\ \frac{\partial \ln L}{\partial \boldsymbol{\gamma}} &= \frac{1}{2} \sum_{i=1}^n \mathbf{z}_i \left( \frac{\varepsilon_i^2}{\exp(\mathbf{z}'_i \boldsymbol{\gamma})} - 1 \right) = \mathbf{0}. \end{aligned} \quad (14-55)$$

For this model, the method of scoring turns out to be a particularly convenient way to maximize the log-likelihood function. The terms in the Hessian are

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = - \sum_{i=1}^n \frac{1}{\exp(\mathbf{z}'_i \boldsymbol{\gamma})} \mathbf{x}_i \mathbf{x}'_i = -\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X}, \quad (14-56)$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}'} = - \sum_{i=1}^n \frac{\varepsilon_i}{\exp(\mathbf{z}'_i \boldsymbol{\gamma})} \mathbf{x}_i \mathbf{z}'_i, \quad (14-57)$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} = -\frac{1}{2} \sum_{i=1}^n \frac{\varepsilon_i^2}{\exp(\mathbf{z}'_i \boldsymbol{\gamma})} \mathbf{z}_i \mathbf{z}'_i. \quad (14-58)$$

## CHAPTER 14 ♦ Maximum Likelihood Estimation 555

The expected value of  $\partial^2 \ln L / \partial \beta \partial \gamma'$  is  $\mathbf{0}$  because  $E[\varepsilon_i | \mathbf{x}_i, \mathbf{z}_i] = 0$ . The expected value of the fraction in  $\partial^2 \ln L / \partial \gamma \partial \gamma'$  is  $E[\varepsilon_i^2 / \sigma_i^2 | \mathbf{x}_i, \mathbf{z}_i] = 1$ . Let  $\delta = [\beta, \gamma]$ . Then

$$-E\left(\frac{\partial^2 \ln L}{\partial \delta \partial \delta'}\right) = \begin{bmatrix} \mathbf{X}' \Omega^{-1} \mathbf{X} & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2} \mathbf{Z}' \mathbf{Z} \end{bmatrix} = -\bar{\mathbf{H}}. \quad (14-59)$$

The **method of scoring** is an algorithm for finding an iterative solution to the likelihood equations. The iteration is

$$\delta_{t+1} = \delta_t - \bar{\mathbf{H}}^{-1} \mathbf{g}_t,$$

where  $\delta_t$  (i.e.,  $\beta_t$ ,  $\gamma_t$ , and  $\Omega_t$ ) is the estimate at iteration  $t$ ,  $\mathbf{g}_t$  is the two-part vector of first derivatives  $[\partial \ln L / \partial \beta_t', \partial \ln L / \partial \gamma_t']'$ , and  $\bar{\mathbf{H}}$  is partitioned likewise. [Newton's method uses the actual second derivatives in (14-56)–(14-58) rather than their expectations in (14-59). The scoring method exploits the convenience of the zero expectation of the off-diagonal block (cross derivative) in (14-57).] Because  $\bar{\mathbf{H}}$  is block diagonal, the iteration can be written as separate equations:

$$\begin{aligned} \beta_{t+1} &= \beta_t + (\mathbf{X}' \Omega_t^{-1} \mathbf{X})^{-1} (\mathbf{X}' \Omega_t^{-1} \boldsymbol{\epsilon}_t) \\ &= \beta_t + (\mathbf{X}' \Omega_t^{-1} \mathbf{X})^{-1} \mathbf{X}' \Omega_t^{-1} (\mathbf{y} - \mathbf{X}\beta_t) \\ &= (\mathbf{X}' \Omega_t^{-1} \mathbf{X})^{-1} \mathbf{X}' \Omega_t^{-1} \mathbf{y} \text{ (of course).} \end{aligned} \quad (14-60)$$

Therefore, the updated coefficient vector  $\beta_{t+1}$  is computed by FGLS using the previously computed estimate of  $\gamma$  to compute  $\Omega$ . We use the same approach for  $\gamma$ :

$$\gamma_{t+1} = \gamma_t + [2(\mathbf{Z}' \mathbf{Z})^{-1}] \left[ \frac{1}{2} \sum_{i=1}^n \mathbf{z}_i \left( \frac{\varepsilon_i^2}{\exp(\mathbf{z}_i' \gamma)} - 1 \right) \right]. \quad (14-61)$$

The 2 and  $\frac{1}{2}$  cancel. The updated value of  $\gamma$  is computed by adding the vector of coefficients in the least squares regression of  $[\varepsilon_i^2 / \exp(\mathbf{z}_i' \gamma) - 1]$  on  $\mathbf{z}_i$  to the old one. Note that the correction is  $2(\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' (\partial \ln L / \partial \gamma)$ , so convergence occurs when the derivative is zero.

The remaining detail is to determine the starting value for the iteration. Because any consistent estimator will do, the simplest procedure is to use OLS for  $\beta$  and the slopes in a regression of the logs of the squares of the least squares residuals on  $\mathbf{z}_i$  for  $\gamma$ . Harvey (1976) shows that this method will produce an inconsistent estimator of  $\gamma_1 = \ln \sigma^2$ , but the inconsistency can be corrected just by adding 1.2704 to the value obtained.<sup>18</sup> Thereafter, the iteration is simply:

1. Estimate the disturbance variance  $\sigma_i^2$  with  $\exp(\mathbf{z}_i' \gamma)$ .
2. Compute  $\beta_{t+1}$  by FGLS.<sup>19</sup>
3. Update  $\gamma_t$  using the regression described in the preceding paragraph.
4. Compute  $\mathbf{d}_{t+1} = [\beta_{t+1}, \gamma_{t+1}] - [\beta_t, \gamma_t]$ . If  $\mathbf{d}_{t+1}$  is large, then return to step 1.

<sup>18</sup>He also presents a correction for the asymptotic covariance matrix for this first step estimator of  $\gamma$ .

<sup>19</sup>The two-step estimator obtained by stopping here would be fully efficient if the starting value for  $\gamma$  were consistent, but it would not be the maximum likelihood estimator.

## 556 PART III ♦ Estimation Methodology

If  $\mathbf{d}_{t+1}$  at step 4 is sufficiently small, then exit the iteration. The asymptotic covariance matrix is simply  $-\mathbf{H}^{-1}$ , which is block diagonal with blocks

$$\text{Asy. Var}[\hat{\beta}_{\text{ML}}] = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1},$$

$$\text{Asy. Var}[\hat{\gamma}_{\text{ML}}] = 2(\mathbf{Z}'\mathbf{Z})^{-1}.$$

If desired, then  $\hat{\sigma}^2 = \exp(\hat{\gamma}_1)$  can be computed. The asymptotic variance would be  $[\exp(\hat{\gamma}_1)]^2(\text{Asy. Var}[\hat{\gamma}_{1,\text{ML}}])$ .

Testing the null hypothesis of homoscedasticity in this model,

$$H_0: \boldsymbol{\alpha} = \mathbf{0}$$

in (14-52), is particularly simple. The Wald test will be carried out by testing the hypothesis that the last  $M$  elements of  $\boldsymbol{\gamma}$  are zero. Thus, the statistic will be

$$\lambda_{WALD} = \hat{\alpha}' \left\{ [\mathbf{0} \quad \mathbf{I}] [2(\mathbf{Z}'\mathbf{Z})]^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \right\} \hat{\alpha}.$$

Because the first column in  $\mathbf{Z}$  is a constant term, this reduces to

$$\lambda_{WALD} = \frac{1}{2} \hat{\alpha}' (\mathbf{Z}'_1 \mathbf{M}^0 \mathbf{Z}_1) \hat{\alpha} \quad \text{💡}$$

where  $\mathbf{Z}_1$  is the last  $M$  columns of  $\mathbf{Z}$ , not including the column of ones, and  $\mathbf{M}^0$  creates deviations from means. The likelihood ratio statistic is computed based on (14-54). Under both the null hypothesis (homoscedastic—using OLS) and the alternative (heteroscedastic—using MLE), the third term in  $\ln L$  reduces to  $-n/2$ . Therefore, the statistic is simply

$$\lambda_{LR} = 2(\ln L_1 - \ln L_0) = n \ln s^2 - \sum_{i=1}^n \ln \hat{\sigma}_i^2,$$

where  $s^2 = \mathbf{e}'\mathbf{e}/n$  using the OLS residuals. To compute the LM statistic, we will use the expected Hessian in (14-59). Under the null hypothesis, the part of the derivative vector in (14-55) that corresponds to  $\boldsymbol{\beta}$  is  $(1/s^2)\mathbf{X}'\mathbf{e} = \mathbf{0}$ . Therefore, using (14-55), the LM statistic is

$$\lambda_{LM} = \left[ \frac{1}{2} \sum_{i=1}^n \left( \frac{e_i^2}{s^2} - 1 \right) \begin{pmatrix} 1 \\ \mathbf{z}_{i1} \end{pmatrix} \right]' \left[ \frac{1}{2} (\mathbf{Z}'\mathbf{Z}) \right]^{-1} \left[ \frac{1}{2} \sum_{i=1}^n \left( \frac{e_i^2}{s^2} - 1 \right) \begin{pmatrix} 1 \\ \mathbf{z}_{i1} \end{pmatrix} \right].$$

The first element in the derivative vector is zero, because  $\sum_i e_i^2 = ns^2$ . Therefore, the expression reduces to

$$\lambda_{LM} = \frac{1}{2} \left[ \sum_{i=1}^n \left( \frac{e_i^2}{s^2} - 1 \right) \mathbf{z}_{i1} \right]' (\mathbf{Z}'_1 \mathbf{M}^0 \mathbf{Z}_1)^{-1} \left[ \sum_{i=1}^n \left( \frac{e_i^2}{s^2} - 1 \right) \mathbf{z}_{i1} \right].$$

This is one-half times the explained sum of squares in the linear regression of the variable  $h_i = (e_i^2/s^2 - 1)$  on  $\mathbf{Z}$ , which is the Breusch–Pagan/Godfrey LM statistic from Section 9.5.2.

## CHAPTER 14 ♦ Maximum Likelihood Estimation 557

**Example 14.6 Multiplicative Heteroscedasticity**

In Example 6.1, we fit a cost function for the U.S. airline industry of the form

$$\ln C_{it} = \beta_1 + \beta_2 \ln Q_{it} + \beta_3 [\ln Q_{it}]^2 + \beta_4 \ln P_{fuel,i,t} + \beta_5 \text{Loadfactor}_{i,t} + \varepsilon_{i,t},$$

where  $C_{i,t}$  is total cost,  $Q_{i,t}$  is output, and  $P_{fuel,i,t}$  is the price of fuel and the 90 observations in the data set are for six firms observed for 15 years. (The model also included dummy variables for firm and year, which we will omit for simplicity.) In Example 14.6, we fit a revised model in which the load factor appears in the variance of  $\varepsilon_{i,t}$  rather than in the regression function. The model is

$$\begin{aligned}\sigma_{i,t}^2 &= \sigma^2 \exp(\alpha \text{Loadfactor}_{i,t}) \\ &= \exp(\gamma_1 + \gamma_2 \text{Loadfactor}_{i,t}).\end{aligned}$$

Estimates were obtained by iterating the weighted least squares procedure using weights  $W_{i,t} = \exp(-c_1 - c_2 \text{Loadfactor}_{i,t})$ . The estimates of  $\gamma_1$  and  $\gamma_2$  were obtained at each iteration by regressing the logs of the squared residuals on a constant and  $\text{Loadfactor}_{i,t}$ . It was noted at the end of the example [and is evident in (14-61)] that these would be the wrong weights to use for the iterated weighted least if we wish to compute the MLE. Table 14.3 reproduces the results from Example 9.4 and adds the MLEs produced using Harvey's method. The MLE of  $\gamma_2$  is substantially different from the earlier result. The Wald statistic for testing the homoscedasticity restriction ( $\alpha = 0$ ) is  $(9.78076/2.839)^2 = 11.869$ , which is greater than 3.84, so the null hypothesis would be rejected. The likelihood ratio statistic is  $-2(54.2747 - 57.3122) = 6.075$ , which produces the same conclusion. However, the LM statistic is 2.96, which conflicts. This is a finite sample result that is not uncommon.

**14.9.2.b Autocorrelation**

At various points in the preceding sections, we have considered models in which there is correlation across observations, including the spatial autocorrelation case in Section 11.7.2, autocorrelated disturbances in panel data models [Section 11.6.3 and in (11-28)], and in the seemingly unrelated regressions model in Section 9.3. The first order autoregression model examined there will be formalized in detail in Chapter 20.

**TABLE 14.3** Multiplicative Heteroscedasticity Model

	<i>Constant</i>	<i>Ln Q</i>	<i>Ln<sup>2</sup> Q</i>	<i>Ln P<sub>f</sub></i>	<i>R</i> <sup>2</sup>	<i>Sum of Squares</i>
OLS	9.1382	0.92615	0.029145	0.41006		
ln L = 54.2747	0.24507 <sup>a</sup> 0.22595 <sup>b</sup>	0.032306 0.030128	0.012304 0.011346	0.018807 0.017524	0.9861674 <sup>c</sup>	1.577479 <sup>d</sup>
Two-step	9.2463 0.21896	0.92136 0.033028	0.024450 0.011412	0.40352 0.016974	0.986119	1.612938
Iterated <sup>e</sup>	9.2774 0.20977	0.91609 0.032993	0.021643 0.011017	0.40174 0.016332	0.986071	1.645693
MLE <sup>f</sup>	9.2611	0.91931	0.023281	0.40266		
ln L = 57.3122	0.2099	0.032295	0.010987	0.016304	0.986100	1.626301

<sup>a</sup>Conventional OLS standard errors<sup>b</sup>White robust standard errors<sup>c</sup>Squared correlation between actual and fitted values<sup>d</sup>Sum of squared residuals<sup>e</sup>Values of  $c_2$  by iteration: 8.254344, 11.622473, 11.705029, 11.710618, 11.711012,

11.711040, 11.711042

<sup>f</sup>Estimate of  $\gamma_2$  is 9.78076 (2.839).

### 558 PART III ♦ Estimation Methodology

We will briefly examine it here to highlight some useful results about the maximum likelihood estimator.

The linear regression model with first order autoregressive [AR(1)] disturbances is

$$\begin{aligned}y_t &= \mathbf{x}'_t \boldsymbol{\beta} + \varepsilon_t, t = 1, \dots, T, \\ \varepsilon_t &= \rho \varepsilon_{t-1} + u_t, |\rho| < 1, \\ E[u_t | \mathbf{X}] &= 0 \\ E[u_t u_s | \mathbf{X}] &= \sigma_u^2 \quad \text{if } t = s \quad \text{and 0 otherwise.}\end{aligned}$$

Feasible GLS estimation of the parameters of this model is examined in detail in Chapter 20. We now add the assumption of normality;  $u_t \sim N[0, \sigma_u^2]$ , and construct the maximum likelihood estimator.

Because every observation on  $y_t$  is correlated with every other observation, in principle, to form the likelihood function, we have the joint density of one  $T$ -variate observation. The Prais and Winsten (1954) transformation in (20-28) suggests a useful way to reformulate this density. We can write

$$f(y_1, y_2, \dots, y_T) = f(y_1) f(y_2 | y_1), f(y_3 | y_2) \dots, f(y_T | y_{T-1}).$$

Because

$$\begin{aligned}\sqrt{1 - \rho^2} y_1 &= \sqrt{1 - \rho^2} \mathbf{x}'_1 \boldsymbol{\beta} + u_1 \\ y_t | y_{t-1} &= \rho y_{t-1} + (\mathbf{x}_t - \rho \mathbf{x}_{t-1})' \boldsymbol{\beta} + u_t,\end{aligned}\tag{14-62}$$

and the observations on  $u_t$  are independently normally distributed, we can use these results to form the log-likelihood function,

$$\begin{aligned}\ln L &= \left[ -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma_u^2 + \frac{1}{2} \ln(1 - \rho^2) - \frac{(1 - \rho^2)(y_1 - \mathbf{x}'_1 \boldsymbol{\beta})^2}{2\sigma_u^2} \right] \\ &\quad + \sum_{t=2}^T \left[ -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma_u^2 - \frac{[(y_t - \rho y_{t-1}) - (\mathbf{x}_t - \rho \mathbf{x}_{t-1})' \boldsymbol{\beta}]^2}{2\sigma_u^2} \right].\end{aligned}\tag{14-63}$$

As usual, the MLE of  $\boldsymbol{\beta}$  is GLS based on the MLEs of  $\sigma_u^2$  and  $\rho$ , and the MLE for  $\sigma_u^2$  will be  $\mathbf{u}'\mathbf{u}/T$  given  $\boldsymbol{\beta}$  and  $\rho$ . The complication is how to compute  $\rho$ . As we will note in Chapter 20, there is a strikingly large number of choices for consistently estimating  $\rho$  in the AR(1) model. It is tempting to choose the most convenient, and then begin the back and forth iterations between  $\boldsymbol{\beta}$  and  $(\sigma_u^2, \rho)$  to obtain the MLE. However, this strategy will not (in general) locate the MLE unless the intermediate estimates of the variance parameters also satisfy the likelihood equation, which for  $\rho$  is

$$\frac{\partial \ln L}{\partial \rho} = \frac{\rho \varepsilon_1^2}{\sigma_u^2} - \frac{\rho}{1 - \rho^2} + \sum_{t=2}^T \frac{u_t \varepsilon_{t-1}}{\sigma_u^2}.$$

One could sidestep the problem simply by scanning the range of  $\rho$  of  $(-1, +1)$  and computing the other estimators at every point, to locate the maximum of the likelihood function by brute force. With modern computers, even with long time series, the amount of computation involved would be minor (if a bit inelegant and inefficient). Beach and MacKinnon (1978a) developed a more systematic algorithm for searching for  $\rho$  in this model. The iteration is then defined between  $\rho$  and  $(\boldsymbol{\beta}, \sigma_u^2)$  as usual.

## CHAPTER 14 ♦ Maximum Likelihood Estimation 559

The information matrix for this log-likelihood is

$$-E \left[ \frac{\partial^2 \ln L}{\partial \begin{pmatrix} \beta \\ \sigma_u^2 \\ \rho \end{pmatrix} \partial (\beta' \sigma_u^2 \rho)} \right] = \begin{bmatrix} \frac{1}{\sigma_u^2} \mathbf{X}' \Omega^{-1} \mathbf{X} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & \frac{T}{2\sigma_u^4} & \frac{\rho}{\sigma_u^2(1-\rho^2)} \\ \mathbf{0}' & \frac{\rho}{\sigma_u^2(1-\rho^2)} & \frac{T-2}{1-\rho^2} + \frac{1+\rho^2}{(1-\rho^2)^2} \end{bmatrix}. \quad (14-64)$$

Note that the diagonal elements in the matrix are  $O(T)$ . But the (2, 3) and (3, 2) elements are constants of  $O(1)$  that will, like the second part of the (3, 3) element, become minimal as  $T$  increases. Dropping these “end effects” (and treating  $T-2$  as the same as  $T$  when  $T$  increases) produces a diagonal matrix from which we extract the standard approximations for the MLEs in this model:

$$\begin{aligned} \text{Asy. Var}[\hat{\beta}] &= \sigma_u^2 (\mathbf{X}' \Omega^{-1} \mathbf{X})^{-1}, \\ \text{Asy. Var}[\hat{\sigma}_u^2] &= \frac{2\sigma_u^4}{T}, \\ \text{Asy. Var}[\hat{\rho}] &= \frac{1-\rho^2}{T}. \end{aligned} \quad (14-65)$$

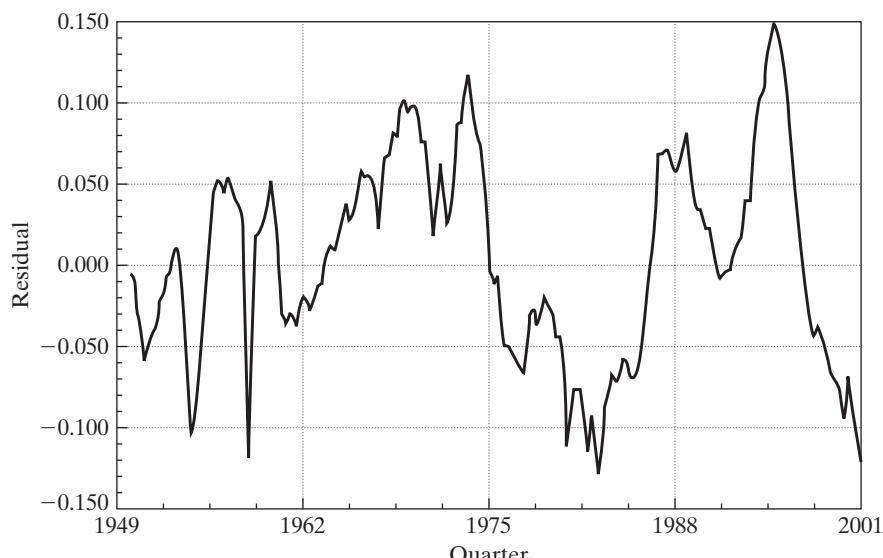
**Example 14.7 Autocorrelation in a Money Demand Equation**

Using the macroeconomic data in Table F5.2, we fit a money demand equation,

$$\ln(M1/CPI\_u)_t = \beta_1 + \beta_2 \ln \text{Real GDP}_t + \beta_3 \ln \text{T-bill rate}_t + \varepsilon_t.$$

The least squares residuals shown in Figure 14.3 display the typical pattern for a highly autocorrelated series.

**FIGURE 14.3** Residuals from Estimated Money Demand Equation.



## 560 PART III ♦ Estimation Methodology

**TABLE 14.4** Estimates of Money Demand Equation:  $T = 204$

<b>Variable</b>	<b>OLS</b>		<b>Prais and Winsten</b>		<b>Maximum Likelihood</b>	
	<b>Estimate</b>	<b>Std. Error</b>	<b>Estimate</b>	<b>Std. Error</b>	<b>Estimate</b>	<b>Std. Error</b>
Constant	-2.1316	0.09100	-1.4755	0.2550	-1.6319	0.4296
Ln real GDP	0.3519	0.01205	0.2549	0.03097	0.2731	0.0518
Ln T-bill rate	-0.1249	0.009841	-0.02666	0.007007	-0.02522	0.006941
$\sigma_\epsilon$	0.06185		0.07767		0.07571	
$\sigma_u$	0.06185		0.01298		0.01273	
$\rho$	0.	0.	0.9557	0.02061	0.9858	0.01180

The simple first-order autocorrelation of the ordinary least squares residuals is  $r = 1 - d/2 = 0.9557$ , where  $d$  is the Durbin-Watson Statistic in (20-23). We then refit the model using the Prais and Winsten FGLS estimator and the maximum likelihood estimator using the Beach and MacKinnon algorithm. The results are shown in Table 14.4. Although the OLS estimator is consistent in this model, nonetheless, the FGLS and ML estimates are quite different.

### 14.9.3 SEEMINGLY UNRELATED REGRESSION MODELS

The general form of the seemingly unrelated regression (SUR) model is given in (10-1)–(10-3);

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, M, \\ E[\boldsymbol{\epsilon}_i | \mathbf{X}_1, \dots, \mathbf{X}_M] &= 0, \\ E[\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}'_j | \mathbf{X}_1, \dots, \mathbf{X}_M] &= \sigma_{ij} \mathbf{I}. \end{aligned} \tag{14-66}$$

FGLS estimation of this model is examined in detail in Section 10.2.3. We will now add the assumption of normally distributed disturbances to the model and develop the maximum likelihood estimators. Given the covariance structure defined in (14-66), the joint normality assumption applies to the vector of  $M$  disturbances observed at time  $t$ , which we write as

$$\boldsymbol{\epsilon}_t | \mathbf{X}_1, \dots, \mathbf{X}_M \sim N[\mathbf{0}, \Sigma], \quad t = 1, \dots, T. \tag{14-67}$$

#### 14.9.3.a The Pooled Model

The pooled model, in which all coefficient vectors are equal, provides a convenient starting point. With the assumption of equal coefficient vectors, the regression model becomes

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}' \boldsymbol{\beta} + \epsilon_{it}, \\ E[\epsilon_{it} | \mathbf{X}_1, \dots, \mathbf{X}_M] &= 0, \\ E[\epsilon_{it} \epsilon_{js} | \mathbf{X}_1, \dots, \mathbf{X}_M] &= \sigma_{ij} \quad \text{if } t = s, \quad \text{and} \quad 0 \quad \text{if } t \neq s. \end{aligned} \tag{14-68}$$

This is a model of heteroscedasticity and cross-sectional correlation. With multivariate normality, the log likelihood is

$$\ln L = \sum_{t=1}^T \left[ -\frac{M}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} \boldsymbol{\epsilon}'_t \Sigma^{-1} \boldsymbol{\epsilon}_t \right]. \tag{14-69}$$

## CHAPTER 14 ♦ Maximum Likelihood Estimation 561

As we saw earlier, the efficient estimator for this model is GLS as shown in (10-21). Because the elements of  $\Sigma$  must be estimated, the FGLS estimator based on (10-9) is used.

As we have seen in several applications now, the maximum likelihood estimator of  $\beta$ , given  $\Sigma$ , is GLS, based on (10-21). The maximum likelihood estimator of  $\Sigma$  is

$$\hat{\sigma}_{ij} = \frac{(\mathbf{y}'_i - \mathbf{X}_i \hat{\beta}_{ML})' (\mathbf{y}_j - \mathbf{X}_j \hat{\beta}_{ML})}{T} = \frac{\hat{\epsilon}'_i \hat{\epsilon}_j}{T} \quad (14-70)$$

based on the MLE of  $\beta$ . If each MLE requires the other, how can we proceed to obtain both? The answer is provided by **Oberhofer and Kmenta** (1974), who show that for certain models, including this one, one can iterate back and forth between the two estimators. Thus, the MLEs are obtained by iterating to convergence between (14-70) and

$$\hat{\beta} = [\mathbf{X}' \hat{\Omega}^{-1} \mathbf{X}]^{-1} [\mathbf{X}' \hat{\Omega}^{-1} \mathbf{y}]. \quad (14-71)$$

The process may begin with the (consistent) ordinary least squares estimator, then (14-70), and so on. The computations are simple, using basic matrix algebra. Hypothesis tests about  $\beta$  may be done using the familiar Wald statistic. The appropriate estimator of the asymptotic covariance matrix is the inverse matrix in brackets in (10-21).

For testing the hypothesis that the off-diagonal elements of  $\Sigma$  are zero—that is, that there is no correlation across firms—there are three approaches. The likelihood ratio test is based on the statistic

$$\lambda_{LR} = T(\ln |\hat{\Sigma}_{heteroscedastic}| - \ln |\hat{\Sigma}_{general}|) = T \left( \sum_{i=1}^M \ln \hat{\sigma}_i^2 - \ln |\hat{\Sigma}| \right), \quad (14-72)$$

where  $\hat{\sigma}_i^2$  are the estimates of  $\sigma_i^2$  obtained from the maximum likelihood estimates of the groupwise heteroscedastic model and  $\hat{\Sigma}$  is the maximum likelihood estimator in the unrestricted model. (Note how the excess variation produced by the restrictive model is used to construct the test.) The large-sample distribution of the statistic is chi-squared with  $M(M - 1)/2$  degrees of freedom. The Lagrange multiplier test developed by Breusch and Pagan (1980) provides an alternative. The general form of the statistic is

$$\lambda_{LM} = T \sum_{i=2}^n \sum_{j=1}^{i-1} r_{ij}^2, \quad (14-73)$$

where  $r_{ij}^2$  is the  $ij$ th residual correlation coefficient. If every equation had a different parameter vector, then equation specific ordinary least squares would be efficient (and ML) and we would compute  $r_{ij}$  from the OLS residuals (assuming that there are sufficient observations for the computation). Here, however, we are assuming only a single-parameter vector. Therefore, the appropriate basis for computing the correlations is the residuals from the iterated estimator in the groupwise heteroscedastic model, that is, the same residuals used to compute  $\hat{\sigma}_i^2$ . (An asymptotically valid approximation to the test can be based on the FGLS residuals instead.) Note that this is not a procedure for testing all the way down to the classical, homoscedastic regression model. That case involves different LM and LR statistics based on the groupwise heteroscedasticity model. If either the LR statistic in (14-72) or the LM statistic in (14-73) are smaller than the critical value from the table, the conclusion, based on this test, is that the appropriate model is the groupwise heteroscedastic model.

## 562 PART III ♦ Estimation Methodology

### 14.9.3.b The SUR Model

The Oberhofer-Kmenta (1974) conditions are met for the seemingly unrelated regressions model, so maximum likelihood estimates can be obtained by iterating the FGLS procedure. We note, once again, that this procedure presumes the use of (10-9) for estimation of  $\sigma_{ij}$  at each iteration. Maximum likelihood enjoys no advantages over FGLS in its asymptotic properties.<sup>20</sup> Whether it would be preferable in a small sample is an open question whose answer will depend on the particular data set.

### 14.9.3.c Exclusion Restrictions

By simply inserting the special form of  $\Omega$  in the log-likelihood function for the generalized regression model in (14-48), we can consider direct maximization instead of iterated FGLS. It is useful, however, to reexamine the model in a somewhat different formulation. This alternative construction of the likelihood function appears in many other related models in a number of literatures.

Consider one observation on each of the  $M$  dependent variables and their associated regressors. We wish to arrange this observation horizontally instead of vertically. The model for this observation can be written

$$\begin{aligned}[y_1 & \quad y_2 \quad \cdots \quad y_M]_t &= [\mathbf{x}_t^*]'[\boldsymbol{\pi}_1 \quad \boldsymbol{\pi}_2 \quad \cdots \quad \boldsymbol{\pi}_M] + [\varepsilon_1 \quad \varepsilon_2 \quad \cdots \quad \varepsilon_M]_t \\ &= [\mathbf{x}_t^*]' \boldsymbol{\Pi} + \mathbf{E},\end{aligned}\tag{14-74}$$

where  $\mathbf{x}_t^*$  is the full set of all  $K^*$  *different* independent variables that appear in the model. The parameter matrix then has one column for each equation, but the columns are not the same as  $\boldsymbol{\beta}_i$  in (14-66) unless every variable happens to appear in every equation. Otherwise, in the  $i$ th equation,  $\boldsymbol{\pi}_i$  will have a number of zeros in it, each one imposing an **exclusion restriction**. For example, consider a two-equation model for production costs for two airlines,

$$\begin{aligned}C_{1t} &= \alpha_1 + \beta_{1P} P_{1t} + \beta_{1L} LF_{1t} + \varepsilon_{1t}, \\ C_{2t} &= \alpha_2 + \beta_{2P} P_{2t} + \beta_{2L} LF_{2t} + \varepsilon_{2t},\end{aligned}$$

where  $C$  is cost,  $P$  is fuel price, and  $LF$  is load factor. The  $t$ th observation would be

$$[C_1 \quad C_2]_t = [1 \quad P_1 \quad LF_1 \quad P_2 \quad LF_2]_t \begin{bmatrix} \alpha_1 & \alpha_2 \\ \beta_{1P} & 0 \\ \beta_{1L} & 0 \\ 0 & \beta_{2P} \\ 0 & \beta_{2L} \end{bmatrix} + [\varepsilon_1 \quad \varepsilon_2]_t.$$

This vector is one observation. Let  $\boldsymbol{\varepsilon}_t$  be the vector of  $M$  disturbances for this observation arranged, for now, in a column. Then  $E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t'] = \boldsymbol{\Sigma}$ . The log of the joint normal density of these  $M$  disturbances is

$$\ln L_t = -\frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln|\boldsymbol{\Sigma}| - \frac{1}{2} \boldsymbol{\varepsilon}_t' \boldsymbol{\Sigma}^{-1} \boldsymbol{\varepsilon}_t.\tag{14-75}$$

<sup>20</sup>Jensen (1995) considers some variation on the computation of the asymptotic covariance matrix for the estimator that allows for the possibility that the normality assumption might be violated.

## CHAPTER 14 ♦ Maximum Likelihood Estimation 563

The log-likelihood for a sample of  $T$  joint observations is the sum of these over  $t$ :

$$\ln L = \sum_{t=1}^T \ln L_t = -\frac{MT}{2} \ln(2\pi) - \frac{T}{2} \ln|\Sigma| - \frac{1}{2} \sum_{t=1}^T \boldsymbol{\varepsilon}'_t \Sigma^{-1} \boldsymbol{\varepsilon}_t. \quad (14-76)$$

The term in the summation in (14-76) is a scalar that equals its trace. We can always permute the matrices in a trace, so

$$\sum_{t=1}^T \boldsymbol{\varepsilon}'_t \Sigma^{-1} \boldsymbol{\varepsilon}_t = \sum_{t=1}^T \text{tr}(\boldsymbol{\varepsilon}'_t \Sigma^{-1} \boldsymbol{\varepsilon}_t) = \sum_{t=1}^T \text{tr}(\Sigma^{-1} \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_t). \quad (14-77)$$

This can be further simplified. The sum of the traces of  $T$  matrices equals the trace of the sum of the matrices [see (A-91)]. We will now also be able to move the constant matrix,  $\Sigma^{-1}$ , outside the summation. Finally, it will prove useful to multiply and divide by  $T$ . Combining all three steps, we obtain

$$\sum_{t=1}^T \text{tr}(\Sigma^{-1} \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_t) = T \text{tr}\left[\Sigma^{-1} \left(\frac{1}{T} \sum_{t=1}^T \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_t\right)\right] = T \text{tr}(\Sigma^{-1} \mathbf{W}), \quad (14-78)$$

where

$$\mathbf{W}_{ij} = \frac{1}{T} \sum_{t=1}^T \varepsilon_{ti} \varepsilon_{tj}.$$

Because this step uses actual disturbances,  $E[\mathbf{W}_{ij}] = \sigma_{ij}$ ;  $\mathbf{W}$  is the  $M \times M$  matrix we would use to estimate  $\Sigma$  if the  $\varepsilon$ 's were actually observed. Inserting this result in the log-likelihood, we have

$$\ln L = -\frac{T}{2} [M \ln(2\pi) + \ln|\Sigma| + \text{tr}(\Sigma^{-1} \mathbf{W})]. \quad (14-79)$$

We now consider maximizing this function.

It has been shown<sup>21</sup> that

$$\begin{aligned} \frac{\partial \ln L}{\partial \Pi'} &= \frac{T}{2} \mathbf{X}^{*\prime} \mathbf{E} \Sigma^{-1}, \\ \frac{\partial \ln L}{\partial \Sigma} &= -\frac{T}{2} \Sigma^{-1} (\Sigma - \mathbf{W}) \Sigma^{-1}. \end{aligned} \quad (14-80)$$

where the  $\mathbf{x}_t^{*\prime}$  in (14-74) is row  $t$  of  $\mathbf{X}^*$ . Equating the second of these derivatives to a zero matrix, we see that given the maximum likelihood estimates of the slope parameters, the maximum likelihood estimator of  $\Sigma$  is  $\mathbf{W}$ , the matrix of mean residual sums of squares and cross products—that is, the matrix we have used for FGLS. [Notice that there is no correction for degrees of freedom;  $\partial \ln L / \partial \Sigma = \mathbf{0}$  implies (10-9).]

We also know that because this model is a generalized regression model, the maximum likelihood estimator of the parameter matrix  $[\beta]$  must be equivalent to the FGLS estimator we discussed earlier.<sup>22</sup> It is useful to go a step further. If we insert our solution

<sup>21</sup>See, for example, Joreskog (1973).

<sup>22</sup>This equivalence establishes the Oberhofer–Kmenta conditions.

## 564 PART III ♦ Estimation Methodology

for  $\Sigma$  in the likelihood function, then we obtain the **concentrated log-likelihood**,

$$\ln L_c = -\frac{T}{2}[M(1 + \ln(2\pi)) + \ln|\mathbf{W}|]. \quad (14-81)$$

We have shown, therefore, that the criterion for choosing the maximum likelihood estimator of  $\beta$  is

$$\hat{\beta}_{ML} = \text{Min}_{\beta} \frac{1}{2} \ln|\mathbf{W}|, \quad (14-82)$$

*subject to the exclusion restrictions.* This important result reappears in many other models and settings. This minimization must be done subject to the constraints in the parameter matrix. In our two-equation example, there are two blocks of zeros in the parameter matrix, which must be present in the MLE as well. The estimator of  $\beta$  is the set of nonzero elements in the parameter matrix in (14-74).

The **likelihood ratio statistic** is an alternative to the  $F$  statistic discussed earlier for testing hypotheses about  $\beta$ . The likelihood ratio statistic is<sup>23</sup>

$$\lambda = -2(\log L_r - \log L_u) = T(\log|\hat{\mathbf{W}}_r| - \log|\hat{\mathbf{W}}_u|), \quad (14-83)$$

where  $\hat{\mathbf{W}}_r$  and  $\hat{\mathbf{W}}_u$  are the residual sums of squares and cross-product matrices using the constrained and unconstrained estimators, respectively. Under the null hypothesis of the restrictions, the limiting distribution of the likelihood ratio statistic is chi-squared with degrees of freedom equal to the number of restrictions. This procedure can also be used to test the homogeneity restriction in the multivariate regression model. The restricted model is the pooled model discussed in the preceding section.

It may also be of interest to test whether  $\Sigma$  is a diagonal matrix. Two possible approaches were suggested in Section 14.9.3a [see (14-72) and (14-73)]. The unrestricted model is the one we are using here, whereas the restricted model is the groupwise heteroscedastic model of Section 9.8.2 (Example 9.5), without the restriction of equal-parameter vectors. As such, the restricted model reduces to separate regression models, estimable by ordinary least squares. The likelihood ratio statistic would be

$$\lambda_{LR} = T \left[ \sum_{i=1}^M \log \hat{\sigma}_i^2 - \log |\hat{\Sigma}| \right], \quad (14-84)$$

where  $\hat{\sigma}_i^2$  is  $\mathbf{e}'_i \mathbf{e}_i / T$  from the individual least squares regressions and  $\hat{\Sigma}$  is the maximum likelihood estimate of  $\Sigma$ . This statistic has a limiting chi-squared distribution with  $M(M - 1)/2$  degrees of freedom under the hypothesis. The alternative suggested by Breusch and Pagan (1980) is the **Lagrange multiplier statistic**,

$$\lambda_{LM} = T \sum_{i=2}^M \sum_{j=1}^{i-1} r_{ij}^2, \quad (14-85)$$

where  $r_{ij}$  is the estimated correlation  $\hat{\sigma}_{ij}/[\hat{\sigma}_{ii}\hat{\sigma}_{jj}]^{1/2}$ . This statistic also has a limiting chi-squared distribution with  $M(M - 1)/2$  degrees of freedom. This test has the advantage that it does not require computation of the maximum likelihood estimator of  $\Sigma$ , because it is based on the OLS residuals.

<sup>23</sup>See Attfield (1998) for refinements of this calculation to improve the small sample performance.

## CHAPTER 14 ♦ Maximum Likelihood Estimation 565

**Example 14.8 ML Estimates of a Seemingly Unrelated Regressions Model**

Although a bit dated, the Grunfeld data used in Application 14.1 have withstood the test of time and are still the standard data set used to demonstrate the SUR model. The data in Appendix Table F10.4 are for 10 firms and 20 years (1935–1954). For the purpose of this illustration, we will use the first four firms. [The data are downloaded from the web site for Baltagi (2005), at <http://www.wiley.com/legacy/wileychi/baltagi/supp/Grunfeld.xls>.]

The model is an investment equation:

$$I_{it} = \beta_{1i} + \beta_{2i}F_{it} + \beta_{3i}C_{it} + \varepsilon_{it}, \quad t = 1, \dots, 20, i = 1, \dots, 10,$$

where

$I_{it}$  = real gross investment for firm  $i$  in year  $t$ ,

$F_{it}$  = real value of the firm-shares outstanding,

$C_{it}$  = real value of the capital stock.

The OLS estimates for the four equations are shown in the left panel of Table 14.5. The correlation matrix for the four OLS residual vectors is

$$\mathbf{R}_e = \begin{bmatrix} 1 & -0.261 & 0.279 & -0.273 \\ -0.261 & 1 & 0.428 & 0.338 \\ 0.279 & 0.428 & 1 & -0.0679 \\ -0.273 & 0.338 & -0.0679 & 1 \end{bmatrix}.$$

Before turning to the FGLS and MLE estimates, we carry out the LM test against the null hypothesis that the regressions are actually unrelated. We leave as an exercise to show that the LM statistic in (14-85) can be computed as

$$\lambda_{LM} = (T/2)[\text{trace}(\mathbf{R}'_e \mathbf{R}_e) - M] = 10.451.$$

The 95 percent critical value from the chi squared distribution with 6 degrees of freedom is 12.59, so at this point, it appears that the null hypothesis is not rejected. We will proceed in spite of this finding.

**TABLE 14.5** Estimated Investment Equations

Firm	Variable	OLS		FGLS		MLE	
		Estimate	St. Er.	Estimate	St. Er.	Estimate	St. Er.
1	Constant	-149.78	97.58	-160.68	90.41	-179.41	86.66
	$F$	0.1192	0.02382	0.1205	0.02187	0.1248	0.02086
	$C$	0.3714	0.03418	0.3800	0.03311	0.3802	0.03266
2	Constant	-49.19	136.52	21.16	116.18	36.46	106.18
	$F$	0.1749	0.06841	0.1304	0.05737	0.1244	0.05191
	$C$	0.3896	0.1312	0.4485	0.1225	0.4367	0.1171
3	Constant	-9.956	28.92	-19.72	26.58	-24.10	25.80
	$F$	0.02655	0.01435	0.03464	0.01279	0.03808	0.01217
	$C$	0.1517	0.02370	0.1368	0.02249	0.1311	0.02223
4	Constant	-6.190	12.45	0.9366	11.59	2.581	11.54
	$F$	0.07795	0.01841	0.06785	0.01705	0.06564	0.01698
	$C$	0.3157	0.02656	0.3146	0.02606	0.3137	0.02617

## 566 PART III ♦ Estimation Methodology

The next step is to compute the covariance matrix for the OLS residuals using

$$\mathbf{W} = (1/T)\mathbf{E}'\mathbf{E} = \begin{bmatrix} \mathbf{7160.29} & -1967.05 & 607.533 & -282.756 \\ -1967.05 & \mathbf{7904.66} & 978.45 & 367.84 \\ 607.533 & 978.45 & \mathbf{660.829} & -21.3757 \\ -282.756 & 367.84 & -21.3757 & \mathbf{149.872} \end{bmatrix},$$

where  $\mathbf{E}$  is the  $20 \times 4$  matrix of OLS residuals. Stacking the data in the partitioned matrices

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{X}_4 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \mathbf{y}_4 \end{bmatrix},$$

we now compute  $\hat{\Omega} = \mathbf{W} \otimes \mathbf{I}_{20}$  and the FGLS estimates,

$$\hat{\beta} = [\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\hat{\Omega}^{-1}\mathbf{y}.$$

The estimated asymptotic covariance matrix for the FGLS estimates is the bracketed inverse matrix. These results are shown in the center panel in Table 14.5.

To compute the MLE, we will take advantage of the Oberhofer and Kmenta (1974) result and iterate the FGLS estimator. Using the FGLS coefficient vector, we recompute the residuals, then recompute  $\mathbf{W}$ , then reestimate  $\beta$ . The iteration is repeated until the estimated parameter vector converges. We use as our convergence measure the following criterion based on the change in the estimated parameter from iteration  $(s-1)$  to iteration  $(s)$ :

$$\delta = [\hat{\beta}(s) - \hat{\beta}(s-1)]'\hat{\Omega}(s)^{-1}\mathbf{X}[\hat{\beta}(s) - \hat{\beta}(s-1)].$$

The sequence of values of this criterion function are: 0.21922, 0.16318, 0.00662, 0.00037, 0.00002367825, 0.000001563348,  $0.1041980 \times 10^{-6}$ . We exit the iterations after iteration 7. The ML estimates are shown in the right panel of Table 14.5.

We then carry out the likelihood ratio test of the null hypothesis of a diagonal covariance matrix. The maximum likelihood estimate of  $\Sigma$  is

$$\hat{\Sigma} = \begin{bmatrix} \mathbf{7235.46} & -2455.13 & 615.167 & -325.413 \\ -2455.13 & \mathbf{8146.41} & 1288.66 & 427.011 \\ 615.167 & 1288.66 & \mathbf{702.268} & 2.51786 \\ -325.413 & 427.011 & 2.51786 & \mathbf{153.889} \end{bmatrix}$$

The estimate for the constrained model is the diagonal matrix formed from the diagonals of  $\mathbf{W}$  shown earlier for the OLS results. (The estimates are shown in boldface in the preceding matrix.) The test statistic is then

$$LR = T(\ln |\text{diag}(\mathbf{W})| - \ln |\hat{\Sigma}|) = 18.55.$$

Recall that the critical value is 12.59. The results contradict the LM statistic. The hypothesis of diagonal covariance matrix is now rejected.

Note that aside from the constants, the four sets of coefficient estimates are fairly similar. Because of the constants, there seems little doubt that the pooling restriction will be rejected. To find out, we compute the Wald statistic based on the MLE results. For testing

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4,$$

we can formulate the hypothesis as

$$H_0: \beta_1 - \beta_4 = \mathbf{0}, \beta_2 - \beta_4 = \mathbf{0}, \beta_3 - \beta_4 = \mathbf{0}.$$

The Wald statistic is

$$\lambda_W = (\mathbf{R}\hat{\beta} - \mathbf{q})'[\mathbf{R}\mathbf{V}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta} - \mathbf{q}) = 2190.96$$

## CHAPTER 14 ♦ Maximum Likelihood Estimation 567

where  $\mathbf{R} = \begin{bmatrix} \mathbf{I}_3 & \mathbf{0} & \mathbf{0} & -\mathbf{I}_3 \\ \mathbf{0} & \mathbf{I}_3 & \mathbf{0} & -\mathbf{I}_3 \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_3 & -\mathbf{I}_3 \end{bmatrix}$ ,  $\mathbf{q} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$ , and  $\mathbf{V} = [\mathbf{X}' \hat{\boldsymbol{\Omega}}^{-1} \mathbf{X}]^{-1}$ . Under the null hypothesis, the

Wald statistic has a limiting chi-squared distribution with 9 degrees of freedom. The critical value is 16.92, so, as expected, the hypothesis is rejected. It may be that the difference is due to the different constant terms. To test the hypothesis that the four pairs of slope coefficients are equal, we replaced the  $\mathbf{I}_3$  in  $\mathbf{R}$  with  $[\mathbf{0}, \mathbf{I}_2]$ , the  $\mathbf{0}$ s with  $2 \times 3$  zero matrices and  $\mathbf{q}$  with a  $6 \times 1$  zero vector. The resulting chi-squared statistic equals 229.005. The critical value is 12.59, so this hypothesis is rejected also.

## 14.9.4 SIMULTANEOUS EQUATIONS MODELS

In Chapter 10, we noted two approaches to maximum likelihood estimation in the equation system

$$\begin{aligned} \mathbf{y}_t' \boldsymbol{\Gamma} + \mathbf{x}_t' \mathbf{B} &= \boldsymbol{\varepsilon}_t', \\ \boldsymbol{\varepsilon}_t | \mathbf{X} &\sim N[\mathbf{0}, \boldsymbol{\Sigma}]. \end{aligned} \tag{14-86}$$

The limited information maximum likelihood (LIML) estimator is a single-equation approach that estimates the parameters one equation at a time. The full information maximum likelihood (FIML) estimator analyzes the full set of equations at one step.

Derivation of the LIML estimator is quite complicated. Lengthy treatments appear in Anderson and Rubin (1948), Theil (1971), and Davidson and MacKinnon (1993, Chapter 18). The mechanics of the computation are surprisingly simple, as shown earlier (Section 10.5.4). The LIML estimates for Klein's Model I appear in Example 10.5.4 with the other single-equation and system estimators. For the practitioner, a useful result is that the asymptotic variance of the two-stage least squares (2SLS) estimator, which is yet simpler to compute, is the same as that of the LIML estimator. For practical purposes, this would generally render the LIML estimator, with its additional normality assumption, moot. The virtue of the LIML is largely theoretical—it provides a useful benchmark for the analysis of the properties of single-equation estimators. The single exception would be the invariance of the estimator to normalization of the equation (i.e., which variable appears on the left of the equals sign). This turns out to be useful in the context of analysis in the presence of weak instruments. (See Sections 8.7 and 10.5.4.)

The FIML estimator is much simpler to derive than the LIML and considerably more difficult to implement. To obtain the needed results, we first operated on the reduced form

$$\begin{aligned} \mathbf{y}_t' &= \mathbf{x}_t' \boldsymbol{\Pi} + \mathbf{v}_t', \\ \mathbf{v}_t | \mathbf{X} &\sim N[\mathbf{0}, \boldsymbol{\Omega}], \end{aligned} \tag{14-87}$$

which is the seemingly unrelated regressions model analyzed at length in Chapter 10 and in Section 14.9.3. The complication is the restrictions imposed on the parameters,

$$\boldsymbol{\Pi} = -\mathbf{B} \boldsymbol{\Gamma}^{-1} \quad \text{and} \quad \boldsymbol{\Omega} = (\boldsymbol{\Gamma}^{-1})' \boldsymbol{\Sigma} (\boldsymbol{\Gamma}^{-1}). \tag{14-88}$$

As is now familiar from several applications, given estimates of  $\boldsymbol{\Gamma}$  and  $\mathbf{B}$  in (14-86), the estimator of  $\boldsymbol{\Sigma}$  is  $(1/T)\mathbf{E}'\mathbf{E}$  based on the residuals. We can even show fairly easily that given  $\boldsymbol{\Gamma}$  and  $\boldsymbol{\Sigma}$ , the estimator of  $(-\mathbf{B})$  in (14-86) would be provided by the results for the SUR model in Section 14.9.3.c (where we estimate the model subject to the zero restrictions in the coefficient matrix). The complication in estimation is brought by

## 568 PART III ♦ Estimation Methodology

$\Gamma$ ; this is a Jacobian. The term  $\ln |\Gamma|$  appears in the log-likelihood function. Nonlinear optimization over the nonzero elements in a function that includes this term is exceedingly complicated. However, three-stage least squares (3SLS) has the same asymptotic efficiency as the FIML estimator, again without the normality assumption and without the practical complications.

The end result is that for the practitioner, the LIML and FIML estimators have been supplanted in the literature by much simpler GMM estimators, 2SLS, H2SLS, 3SLS, and H3SLS. Interest remains in these estimators, but largely as a component of the ongoing theoretical development.

### 14.9.5 MAXIMUM LIKELIHOOD ESTIMATION OF NONLINEAR REGRESSION MODELS

In Chapter 7, we considered nonlinear regression models in which the nonlinearity in the parameters appeared entirely on the right-hand side of the equation. Maximum likelihood is used when the disturbances in a regression, or the dependent variable, more generally, is not normally distributed. The geometric regression model provides an application.

#### **Example 14.9 Identification in a Loglinear Regression Model**

In Example 7.6, we estimated an exponential regression model, of the form

$$E[Income|Age, Education, Female] = \exp(\gamma_1^* + \gamma_2 Age + \gamma_3 Education + \gamma_4 Female).$$

This loglinear conditional mean is consistent with several different distributions, including the lognormal, Weibull, gamma, and exponential models. In each of these cases, the conditional mean function is of the form

$$\begin{aligned} E[Income|\mathbf{x}] &= g(\theta) \exp(\gamma_1 + \mathbf{x}'\gamma_2) \\ &= \exp(\gamma_1^* + \mathbf{x}'\gamma_2), \end{aligned}$$

where  $\theta$  is an additional parameter of the distribution and  $\gamma_1^* = \ln g(\theta) + \gamma_1$ . Two implications are:

1. Nonlinear least squares (NLS) is robust at least to some failures of the distributional assumption. The nonlinear least squares estimator of  $\gamma_2$  will be consistent and asymptotically normally distributed in all cases for which  $E[Income|\mathbf{x}] = \exp(\gamma_1^* + \mathbf{x}'\gamma_2)$ .
2. The NLS estimator cannot produce a consistent estimator of  $\gamma_1$ ;  $\text{plim}_{\mathbf{x}} \gamma_1 = \gamma_1^*$ , which varies depending on the specific distribution. In the conditional mean function, any pair of values for which  $\gamma_1' = \ln g(\theta) + \gamma_1$  the same will lead to the same sum of squares. This is a form of multicollinearity; the pseudoregressor for  $\theta$  is  $\partial E[Income|\mathbf{x}]/\partial\theta = \exp(\gamma_1^* + \mathbf{x}'\gamma_2)[g'(\theta)/g(\theta)]$  while that for  $\gamma_1$  is  $\partial E[Income|\mathbf{x}]/\partial\gamma_1 = \exp(\gamma_1^* + \mathbf{x}'\gamma_2)$ . The first is a constant multiple of the second.

NLS cannot provide separate estimates of  $\theta$  and  $\gamma_1$  while MLE can—see the example to follow. Second, NLS might be less efficient than MLE since it does not use the information about the distribution of the dependent variable. This second consideration is uncertain. For estimation of  $\gamma_2$ , the NLS estimator is less efficient for not using the distributional information. However, that shortcoming might be offset because the NLS estimator does not attempt to compute an independent estimator of the additional parameter,  $\theta$ .

To illustrate, we reconsider the estimator in Example 7.6. The gamma regression model specifies

$$f(y|\mathbf{x}) = \frac{\mu(\mathbf{x})^\theta}{\Gamma(\theta)} \exp[-\mu(\mathbf{x})y]y^{\theta-1}, y > 0, \theta > 0, \mu(\mathbf{x}) = \exp(-\gamma_1 - \mathbf{x}'\gamma_2).$$

## CHAPTER 14 ♦ Maximum Likelihood Estimation 569

**TABLE 14.6** Estimated Gamma Regression Model

	(1) NLS	(2) Constrained NLS	(3) MLE	(4) NLS/MLE
Constant	1.22468 (47722.5)	1.69331 (0.04408)	3.36826 (0.05048)	3.36380 (0.04408)
Age	-0.00207 (0.00061)	-0.00207 (0.00061)	-0.00153 (0.00061)	-0.00207 (0.00061)
Education	-0.04792 (0.00247)	-0.04792 (0.00247)	-0.04975 (0.00286)	-0.04792 (0.00247)
Female	0.00658 (0.01373)	0.00658 (0.01373)	0.00696 (0.01322)	0.00658 (0.08677)
P	0.62699 (29921.3)	—	5.31474 (0.10894)	5.31474 (0.00000)

The conditional mean function for this model is

$$E[y|\mathbf{x}] = \theta/\mu(\mathbf{x}) = \theta \exp(\gamma_1 + \mathbf{x}'\gamma_2) = \exp(\gamma_1^* + \mathbf{x}'\gamma_2).$$

Table 14.6 presents estimates of  $\theta$  and  $(\gamma_1, \gamma_2)$ . Estimated standard errors appear in parentheses. The estimates in columns (1), (2) and (4) are all computed using nonlinear least squares. In (1), an attempt is made to estimate  $\theta$  and  $\gamma_1$  separately. The estimator “converged” on two values. However, the estimated standard errors are essentially infinite. The convergence to anything at all is due to rounding error in the computer. The results in column (2) are for  $\gamma_1^*$  and  $\gamma_2$ . The sums of squares for these two estimates as well as for those in (4) are all 112.19688, indicating that the three results merely show three different sets of results for which  $\gamma_1^*$  is the same. The full maximum likelihood estimates are presented in (3). Note that an estimate of  $\theta$  is obtained here because the assumed gamma distribution provides another independent moment equation for this parameter,  $\partial \ln L / \partial \theta = -n \ln \Psi(\theta) + \sum_i (\ln y_i - \ln \mu(\mathbf{x})) = 0$ , while the normal equations for the sum of squares provides the same normal equation for  $\theta$  and  $\gamma_1$ .

The standard approach to modeling counts of events begins with the Poisson regression model,

$$\text{Prob}[Y = y_i | \mathbf{x}_i] = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}), y_i = 0, 1, \dots$$

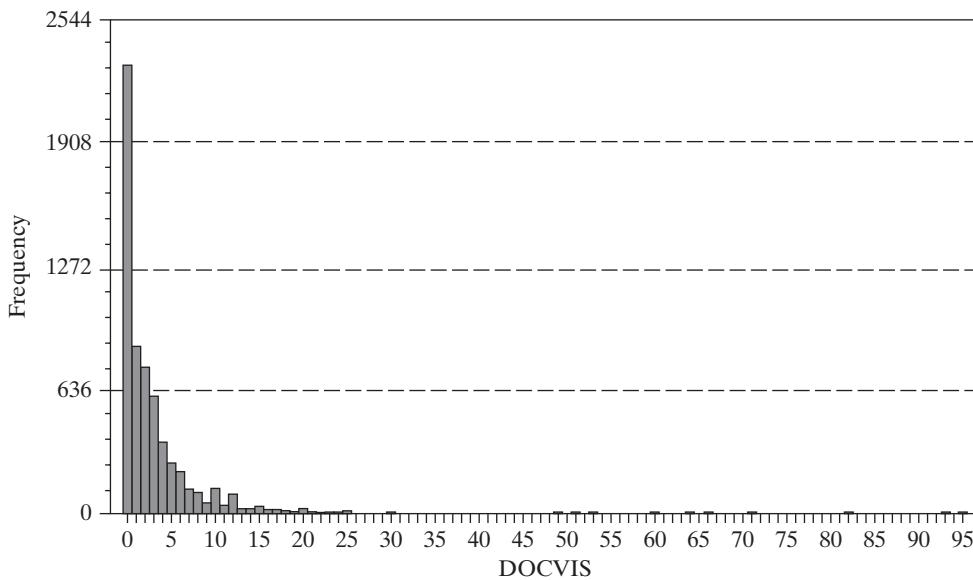
which has **loglinear conditional mean** function  $E[y_i | \mathbf{x}_i] = \lambda_i$ . (The Poisson regression model and other specifications for data on counts are discussed at length in Chapter 19.) We introduce the topic here to begin development of the MLE in a fairly straightforward, typical nonlinear setting.) Appendix Table F7.1 presents the Riphahn et al. (2003) data, which we will use to analyze a count variable, *DocVis*, the number of visits to physicians in the survey year. The histogram in Figure 14.4 shows a distinct spike at zero followed by rapidly declining frequencies. While the Poisson distribution, which is typically hump-shaped, can accommodate this configuration if  $\lambda_i$  is less than one, the shape is nonetheless somewhat “non-Poisson.” [So-called Zero Inflation models (discussed in Chapter 19) are often used for this situation.]

The geometric distribution,

$$f(y_i | \mathbf{x}_i) = \theta_i(1 - \theta_i)^{y_i}, \theta_i = 1/(1 + \lambda_i), \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}), y_i = 0, 1, \dots,$$

is a convenient specification that produces the effect shown in Figure 14.4. (Note that, formally, the specification is used to model the number of failures before the first success

## 570 PART III ♦ Estimation Methodology



**FIGURE 14.4** Histogram for Doctor Visits.

in successive independent trials each with success probability  $\theta_i$ , so in fact, it is misspecified as a model for counts. The model does provide a convenient and useful illustration, however.) The conditional mean function is also  $E[y_i | \mathbf{x}_i] = \lambda_i$ . The partial effects in the model are

$$\frac{\partial E[y_i | \mathbf{x}_i]}{\partial \mathbf{x}_i} = \lambda_i \boldsymbol{\beta},$$

so this is a distinctly nonlinear regression model. We will construct a maximum likelihood estimator, then compare the MLE to the **nonlinear least squares** and (misspecified) linear least squares estimates.

The log-likelihood function is

$$\ln L = \sum_{i=1}^n \ln f(y_i | \mathbf{x}_i, \boldsymbol{\beta}) = \sum_{i=1}^n \ln \theta_i + y_i \ln(1 - \theta_i).$$

The likelihood equations are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \left( \frac{1}{\theta_i} - \frac{y_i}{1 - \theta_i} \right) \frac{d\theta_i}{d\lambda_i} \frac{\partial \lambda_i}{\partial \boldsymbol{\beta}}.$$

Because

$$\frac{d\theta_i}{d\lambda_i} \frac{\partial \lambda_i}{\partial \boldsymbol{\beta}} = \left( \frac{-1}{(1 + \lambda_i)^2} \right) \lambda_i \mathbf{x}_i = -\theta_i(1 - \theta_i) \mathbf{x}_i,$$

the likelihood equations simplify to

$$\begin{aligned} \frac{\partial \ln L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^n (\theta_i y_i - (1 - \theta_i)) \mathbf{x}_i \\ &= \sum_{i=1}^n (\theta_i(1 + y_i) - 1) \mathbf{x}_i. \end{aligned}$$

## CHAPTER 14 ♦ Maximum Likelihood Estimation 571

To estimate the asymptotic covariance matrix, we can use any of the three estimators of  $\text{Asy. Var} [\hat{\beta}_{\text{MLE}}]$ . The BHHH estimator would be

$$\begin{aligned}\text{Est. Asy. Var}_{\text{BHHH}}[\hat{\beta}_{\text{MLE}}] &= \left[ \sum_{i=1}^n \left( \frac{\partial \ln f(y_i | \mathbf{x}_i, \hat{\beta})}{\partial \hat{\beta}} \right) \left( \frac{\partial \ln f(y_i | \mathbf{x}_i, \hat{\beta})}{\partial \hat{\beta}} \right)' \right]^{-1} \\ &= \left[ \sum_{i=1}^n (\hat{\theta}_i(1+y_i) - 1)^2 \mathbf{x}_i \mathbf{x}'_i \right].\end{aligned}$$

The negative inverse of the second derivatives matrix evaluated at the MLE is

$$\left[ -\frac{\partial^2 \ln L}{\partial \hat{\beta} \partial \hat{\beta}'} \right]^{-1} = \left[ \sum_{i=1}^n (1+y_i)\hat{\theta}_i(1-\hat{\theta}_i) \mathbf{x}_i \mathbf{x}'_i \right]^{-1}.$$

Finally, as noted earlier,  $E[y_i | \mathbf{x}_i] = \lambda_i = (1-\theta_i)/\theta_i$ , is known, so we can also use the negative inverse of the expected second derivatives matrix,

$$\left[ -E \left( \frac{\partial^2 \ln L}{\partial \hat{\beta} \partial \hat{\beta}'} \right) \right]^{-1} = \left[ \sum_{i=1}^n (1-\hat{\theta}_i) \mathbf{x}_i \mathbf{x}'_i \right]^{-1}.$$

To compute the estimates of the parameters, either **Newton's method**,

$$\hat{\beta}^{t+1} = \hat{\beta}^t - [\hat{\mathbf{H}}^t]^{-1} \hat{\mathbf{g}}^t,$$

or the method of scoring,

$$\hat{\beta}^{t+1} = \hat{\beta}^t - \{E[\hat{\mathbf{H}}^t]\}^{-1} \hat{\mathbf{g}}^t,$$

can be used, where  $\mathbf{H}$  and  $\mathbf{g}$  are the second and first derivatives that will be evaluated at the current estimates of the parameters. Like many models of this sort, there is a convenient set of starting values, assuming the model contains a constant term. Because  $E[y_i | \mathbf{x}_i] = \lambda_i$ , if we start the slope parameters at zero, then a natural starting value for the constant term is the log of  $\bar{y}$ .

**Example 14.1**  Geometric Regression Model for Doctor Visits

In Example 11.14, we considered nonlinear least squares estimation of a loglinear model for the number of doctor visits variable shown in Figure 14.4. The data are drawn from the Riphahn et al. (2003) data set in Appendix Table F7.1. We will continue that analysis here by fitting a more detailed model for the count variable  $DocVis$ . The conditional mean analyzed here is

$$\ln E[DocVis_{it} | \mathbf{x}_{it}] = \beta_1 + \beta_2 Age_{it} + \beta_3 Educ_{it} + \beta_4 Income_{it} + \beta_5 Kids_{it}$$

(This differs slightly from the model in Example 11.14. For this exercise, with an eye toward the fixed effects model in Example 14.13), we have specified a model that does not contain any time-invariant variables, such as  $\text{Female}_{it}$ .  Example means for the variables in the model are given in Table 14.7. Note, these data are a panel. In this exercise, we are ignoring that fact, and fitting a pooled model. We will turn to panel data treatments in the next section, and revisit this application.

## 572 PART III ♦ Estimation Methodology

We used Newton's method for the optimization, with starting values as suggested earlier. The five iterations are as follows:

<b>Variable</b>	<b>Constant</b>	<b>Age</b>	<b>Educ</b>	<b>Income</b>	<b>Kids</b>
Start values:	.11580e+01	.00000e+00	.00000e+00	.00000e+00	.00000e+00
1st derivs.	-.25191e-08	-.61777e+05	.73202e+04	.42575e+04	.16464e+04
Parameters:	.11580e+01	.00000e+00	.00000e+00	.00000e+00	.00000e+00
Iteration 1 F =	.6287e+05	<b>g'inv(H)g =</b>	.4367e+02		
1st derivs.	.48616e+03	-.22449e+05	-.57162e+04	-.17112e+04	-.16521e+03
Parameters:	.11186e+01	.17563e-01	-.50263e-01	-.46274e-01	-.15609e+00
Iteration 2 F =	.6192e+05	<b>g'inv(H)g =</b>	.3547e+01		
1st derivs.	-.31284e+01	-.15595e+03	-.37197e+02	-.10630e+02	-.77186e+00
Parameters:	.10922e+01	.17981e-01	-.47303e-01	-.46739e-01	-.15683e+00
Iteration 3 F =	.6192e+05	<b>g'inv(H)g =</b>	.2598e-01		
1st derivs.	-.18417e-03	-.99368e-02	-.21992e-02	-.59354e-03	-.25994e-04
Parameters:	.10918e+01	.17988e-01	-.47274e-01	-.46751e-01	-.15686e+00
Iteration 4 F =	.6192e+05	<b>g'inv(H)g =</b>	.1831e-05		
1st derivs.	-.35727e-11	.86745e-10	-.26302e-10	-.61006e-11	-.15620e-11
Parameters:	.10918e+01	.17988e-01	-.47274e-01	-.46751e-01	-.15686e+00
Iteration 5 F =	.6192e+05	<b>g'inv(H)g =</b>	.1772e-12		

Convergence based on the LM criterion,  $\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$  is achieved after the fourth iteration. Note that the derivatives at this point are extremely small, albeit not absolutely zero. Table 14.7 presents the maximum likelihood estimates of the parameters. Several sets of standard errors are presented. The three sets based on different estimators of the information matrix are presented first. The fourth set are based on the cluster corrected covariance matrix discussed in Section 14.8.4. Because this is actually an (unbalanced) panel data set, we anticipate correlation across observations. Not surprisingly, the standard errors rise substantially. The partial effects listed next are computed in two ways. The "Average Partial Effect" is computed by averaging  $\lambda_i \beta$  across the individuals in the sample. The "Partial Effect" is computed for the average individual by computing  $\lambda$  at the means of the data. The next-to-last column contains the ordinary least squares coefficients. In this model, there is no reason to expect ordinary least squares to provide a consistent estimator of  $\beta$ . The question might arise, What does ordinary least squares estimate? The answer is the slopes of the linear projection of DocVis on  $\mathbf{x}_{it}$ . The resemblance of the OLS coefficients to the estimated partial effects is more than coincidental, and suggests an answer to the question.

The analysis in the table suggests three competing approaches to modeling DocVis. The results for the geometric regression model are given in Table 14.7. At the beginning of this section, we noted that the more conventional approach to modeling a count variable such as DocVis is with the Poisson regression model. The log-likelihood function and its derivatives

**TABLE 14.7** Estimated Geometric Regression Model Dependent Variable: DocVis:  
Mean = 3.18352, Standard Deviation = 5.68969

<b>Variable</b>	<b>Estimate</b>	<b>St. Er. <i>H</i></b>	<b>St. Er. <i>E[H]</i></b>	<b>St. Er. <i>BHHH</i></b>	<b>St. Er. <i>Cluster</i></b>	<b>APE</b>	<b>PE Mean</b>	<b>OLS</b>	<b>Mean</b>
Constant	1.0918	0.0524	0.0524	0.0354	0.1112	—	—	2.656	
Age	0.0180	0.0007	0.0007	0.0005	0.0013	0.0572	0.0547	0.061	43.52
Education	-0.0473	0.0033	0.0033	0.0023	0.0069	-0.150	-0.144	-0.121	11.32
Income	-0.0468	0.0041	0.0042	0.0023	0.0075	-0.149	-0.142	-0.162	3.52
Kids	-0.1569	0.0156	0.0155	0.0103	0.0319	-0.499	-0.477	-0.517	0.40

## CHAPTER 14 ♦ Maximum Likelihood Estimation 573

**TABLE 14.8** Estimates of Three Models for DOCVIS

Variable	Geometric Model		Poisson Model		Nonlinear Reg.	
	Estimate	St. Er.	Estimate	St. Er.	Estimate	St. Er.
Constant	1.0918	0.0524	1.0480	0.0272	0.9801	0.0893
Age	0.0180	0.0007	0.0184	0.0003	0.0187	0.0011
Education	-0.0473	0.0033	-0.0433	0.0017	-0.0361	0.0057
Income	-0.0468	0.0041	-0.0520	0.0022	-0.0591	0.0072
Kids	-0.1569	0.0156	-0.1609	0.0080	-0.1692	0.0264

are even simpler than the geometric model,

$$\begin{aligned}\ln L &= \sum_{i=1}^n y_i \ln \lambda_i - \lambda_i - \ln y_i!, \\ \partial \ln L / \partial \beta &= \sum_{i=1}^n (y_i - \lambda_i) \mathbf{x}_i, \\ \partial^2 \ln L / \partial \beta \partial \beta' &= \sum_{i=1}^n -\lambda_i \mathbf{x}_i \mathbf{x}_i'.\end{aligned}$$

A third approach might be a semiparametric, nonlinear regression model,

$$y_{it} = \exp(\mathbf{x}'_{it} \beta) + \varepsilon_{it}.$$

This is, in fact, the model that applies to both the geometric and Poisson cases. Under either distributional assumption, nonlinear least squares is inefficient compared to MLE. But, the distributional assumption can be dropped altogether, and the model fit as a simple exponential regression. Table 14.8 presents the three sets of estimates.

It is not obvious how to choose among the alternatives. Of the three, the Poisson model is used most often by far. The Poisson and geometric models are not nested, so we cannot use a simple parametric test to choose between them. However, these two models will surely fit the conditions for the Vuong test described in Section 14.6.6. To implement the test, we first computed

$$V_{it} = \ln f_{it} | \text{geometric} - \ln f_{it} | \text{Poisson}$$

using the respective MLEs of the parameters. The test statistic given in Section 14.6.6 is then

$$V = \frac{\left( \sqrt{\sum_{i=1}^n T_i} \right) \bar{V}}{s_V}.$$

This statistic converges to standard normal under the underlying assumptions. A large positive value favors the geometric model. The computed sample value is 37.885, which strongly favors the geometric model over the Poisson.

#### 14.9.6 PANEL DATA APPLICATIONS

Application of panel data methods to the linear panel data models we have considered so far is a fairly marginal extension. For the random effects linear model, considered in the following Section 14.9.6.a, the MLE of  $\beta$  is, as always, FGLS given the MLEs of the variance parameters. The latter produce a fairly substantial complication, as we shall

## 574 PART III ♦ Estimation Methodology

see. This extension does provide a convenient, interesting application to see the payoff to the invariance property of the MLE—we will reparameterize a fairly complicated log-likelihood function to turn it into a simple one. Where the method of maximum likelihood becomes essential is in analysis of fixed and random effects in nonlinear models. We will develop two general methods for handling these situations in generic terms in Sections 14.9.6.b and 14.9.6.c, then apply them in several models later in the book.

### 14.9.6.a ML Estimation of the Linear Random Effects Model

The contribution of the  $i$ th individual to the log-likelihood for the random effects model [(14-26) to (14-27)] with normally distributed disturbances is

$$\begin{aligned}\ln L_i(\boldsymbol{\beta}, \sigma_e^2, \sigma_u^2) &= \frac{-1}{2} [T_i \ln 2\pi + \ln |\boldsymbol{\Omega}_i| + (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})' \boldsymbol{\Omega}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})] \\ &= \frac{-1}{2} [T_i \ln 2\pi + \ln |\boldsymbol{\Omega}_i| + \boldsymbol{\varepsilon}_i' \boldsymbol{\Omega}_i^{-1} \boldsymbol{\varepsilon}_i],\end{aligned}\tag{14-89}$$

where

$$\boldsymbol{\Omega}_i = \sigma_e^2 \mathbf{I}_{T_i} + \sigma_u^2 \mathbf{i} \mathbf{i}'$$

and  $\mathbf{i}$  denotes a  $T_i \times 1$  column of ones. Note that the  $\boldsymbol{\Omega}_i$  varies over  $i$  because it is  $T_i \times T_i$ . Baltagi (2005, pp. 19–20) presents a convenient and compact estimator for this model that involves iteration between an estimator of  $\phi^2 = [\sigma_e^2 / (\sigma_e^2 + T\sigma_u^2)]$ , based on sums of squared residuals, and  $(\alpha, \boldsymbol{\beta}, \sigma_e^2)$  ( $\alpha$  is the constant term) using FGLS. Unfortunately, the convenience and compactness come unraveled in the unbalanced case. We consider, instead, what Baltagi labels a “brute force” approach, that is, direct maximization of the log-likelihood function in (14-89). (See, op. cit, pp. 169–170.)

Using (A-66), we find (in (11-28) that

$$\boldsymbol{\Omega}_i^{-1} = \frac{1}{\sigma_e^2} \left[ \mathbf{I}_{T_i} - \frac{\sigma_u^2}{\sigma_e^2 + T_i \sigma_u^2} \mathbf{i} \mathbf{i}' \right].$$

We will also need the determinant of  $\boldsymbol{\Omega}_i$ . To obtain this, we will use the product of its characteristic roots. First, write

$$|\boldsymbol{\Omega}_i| = (\sigma_e^2)^{T_i} |\mathbf{I} + \gamma \mathbf{i} \mathbf{i}'|,$$

where  $\gamma = \sigma_u^2 / \sigma_e^2$ . To find the characteristic roots of the matrix, use the definition

$$[\mathbf{I} + \gamma \mathbf{i} \mathbf{i}'] \mathbf{c} = \lambda \mathbf{c},$$

where  $\mathbf{c}$  is a characteristic vector and  $\lambda$  is the associated characteristic root. The equation implies that  $\gamma \mathbf{i} \mathbf{i}' \mathbf{c} = (\lambda - 1) \mathbf{c}$ . Premultiply by  $\mathbf{i}'$  to obtain  $\gamma (\mathbf{i}' \mathbf{i})(\mathbf{i}' \mathbf{c}) = (\lambda - 1)(\mathbf{i}' \mathbf{c})$ . Any vector  $\mathbf{c}$  with elements that sum to zero will satisfy this equality. There will be  $T_i - 1$  such vectors and the associated characteristic roots will be  $(\lambda - 1) = 0$  or  $\lambda = 1$ . For the remaining root, divide by the nonzero  $(\mathbf{i}' \mathbf{c})$  and note that  $\mathbf{i}' \mathbf{i} = T_i$ , so the last root is  $T_i \gamma = \lambda - 1$  or  $\lambda = (1 + T_i \gamma)$ .<sup>24</sup> It follows that the determinant is

$$\ln |\boldsymbol{\Omega}_i| = T_i \ln \sigma_e^2 + \ln(1 + T_i \gamma).$$

<sup>24</sup>By this derivation, we have established a useful general result. The characteristic roots of a  $T \times T$  matrix of the form  $\mathbf{A} = (\mathbf{I} + a \mathbf{b} \mathbf{b}')$  are 1 with multiplicity  $(T - 1)$  and  $a \mathbf{b}' \mathbf{b}$  with multiplicity 1. The proof follows precisely along the lines of our earlier derivation.

## CHAPTER 14 ♦ Maximum Likelihood Estimation 575

Expanding the parts and multiplying out the third term gives the log-likelihood function

$$\begin{aligned}\ln L &= \sum_{i=1}^n \ln L_i \\ &= -\frac{1}{2} \left[ (\ln 2\pi + \ln \sigma_\varepsilon^2) \sum_{i=1}^n T_i + \sum_{i=1}^n \ln(1 + T_i \gamma) \right] - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^n \left[ \boldsymbol{\varepsilon}'_i \boldsymbol{\varepsilon}_i - \frac{\sigma_u^2 (T_i \bar{\varepsilon}_i)^2}{\sigma_\varepsilon^2 + T_i \sigma_u^2} \right].\end{aligned}$$

Note that in the third term, we can write  $\sigma_\varepsilon^2 + T_i \sigma_u^2 = \sigma_\varepsilon^2 (1 + T_i \gamma)$  and  $\sigma_u^2 = \sigma_\varepsilon^2 \gamma$ . After inserting these, two appearances of  $\sigma_\varepsilon^2$  in the square brackets will cancel, leaving

$$\ln L = -\frac{1}{2} \sum_{i=1}^n \left( T_i (\ln 2\pi + \ln \sigma_\varepsilon^2) + \ln(1 + T_i \gamma) + \frac{1}{\sigma_\varepsilon^2} \left[ \boldsymbol{\varepsilon}'_i \boldsymbol{\varepsilon}_i - \frac{\gamma (T_i \bar{\varepsilon}_i)^2}{1 + T_i \gamma} \right] \right).$$

Now, let  $\theta = 1/\sigma_\varepsilon^2$ ,  $R_i = 1 + T_i \gamma$ , and  $Q_i = \gamma/R_i$ . The individual contribution to the log likelihood becomes

$$\ln L_i = -\frac{1}{2} [\theta(\boldsymbol{\varepsilon}'_i \boldsymbol{\varepsilon}_i - Q_i (T_i \bar{\varepsilon}_i)^2) + \ln R_i - T_i \ln \theta + T_i \ln 2\pi].$$

The likelihood equations are

$$\begin{aligned}\frac{\partial \ln L_i}{\partial \beta} &= \theta \left[ \sum_{t=1}^{T_i} \mathbf{x}_{it} \varepsilon_{it} \right] - \theta \left[ Q_i \left( \sum_{t=1}^{T_i} \mathbf{x}_{it} \right) \left( \sum_{t=1}^{T_i} \varepsilon_{it} \right) \right], \\ \frac{\partial \ln L_i}{\partial \theta} &= -\frac{1}{2} \left[ \left( \sum_{t=1}^{T_i} \varepsilon_{it}^2 \right) - Q_i \left( \sum_{t=1}^{T_i} \varepsilon_{it} \right)^2 - \frac{T_i}{\theta} \right], \\ \frac{\partial \ln L_i}{\partial \gamma} &= \frac{1}{2} \left[ \theta \left( \frac{1}{R_i^2} \left( \sum_{t=1}^{T_i} \varepsilon_{it} \right)^2 \right) - \frac{T_i}{R_i} \right].\end{aligned}$$

These will be sufficient for programming an optimization algorithm such as DFP or BFGS. (See Section E3.3.) We could continue to derive the second derivatives for computing the asymptotic covariance matrix, but this is unnecessary. For  $\hat{\beta}_{MLE}$ , we know that because this is a generalized regression model, the appropriate asymptotic covariance matrix is

$$\text{Asy. Var}[\hat{\beta}_{MLE}] = \left[ \sum_{i=1}^n \mathbf{X}'_i \hat{\Omega}_i^{-1} \mathbf{X}_i \right]^{-1}.$$

(See Section 11.5.1.) We also know that the MLEs of the variance components estimators will be asymptotically uncorrelated with that of  $\beta$ . In principle, we could continue to estimate the asymptotic variances of the MLEs of  $\sigma_\varepsilon^2$  and  $\sigma_u^2$ . It would be necessary to derive these from the estimators of  $\theta$  and  $\gamma$ , which one would typically do in any event. However, statistical inference about the disturbance variance,  $\sigma_\varepsilon^2$  in a regression model, is typically of no interest. On the other hand, one might want to test the hypothesis that  $\sigma_u^2$  equals zero, or  $\gamma = 0$ . Breusch and Pagan's (1979) LM statistic in (11-3), extended

## 576 PART III ♦ Estimation Methodology

to the unbalanced panel case considered here would be

$$\begin{aligned} LM &= \frac{\left(\sum_{i=1}^N T_i\right)^2}{\left[2 \sum_{i=1}^N T_i(T_i - 1)\right]} \left[ \frac{\sum_{i=1}^N (T_i \bar{e}_i)^2}{\sum_{i=1}^N \sum_{t=1}^{T_i} e_{it}^2} - 1 \right]^2 \\ &= \frac{\left(\sum_{i=1}^N T_i\right)^2}{\left[2 \sum_{i=1}^N T_i(T_i - 1)\right]} \left[ \frac{\sum_{i=1}^N [(T_i \bar{e}_i)^2 - \mathbf{e}'_i \mathbf{e}_i]}{\sum_{i=1}^N \mathbf{e}'_i \mathbf{e}_i} \right]^2. \end{aligned}$$

### **Example 14.11 Maximum Likelihood and FGLS Estimates of a Wage Equation**

Example 14.10 presented FGLS estimates of a wage equation using Cornwell and Rupert's panel data. We have reestimated the wage equation using maximum likelihood instead of FGLS. The parameter estimates appear in Table 14.9, with the FGLS and pooled OLS estimates. The estimates of the variance components are shown in the table as well. The similarity of the MLEs and FGLS estimates is to be expected given the large sample size. The LM statistic for testing for the presence of the common effects is 3,881.34, which is far larger than the critical value of 3.84. With the MLE, we can also use an LR test to test for random effects against the null hypothesis of no effects. The chi-squared statistic based on the two log-likelihoods is 4297.57, which leads to the same conclusion.

#### 14.9.6.b Nested Random Effects

Consider a data set on test scores for multiple school districts in a state. To establish a notation for this complex model, we define a four-level unbalanced structure,

$Z_{ijkt}$  = test score for student  $t$ , teacher  $k$ , school  $j$ , district  $i$ ,

$L$  = school districts,  $i = 1, \dots, L$ ,

$M_i$  = schools in each district,  $j = 1, \dots, M_i$ ,

$N_{ij}$  = teachers in each school,  $k = 1, \dots, N_{ij}$

$T_{ijk}$  = students in each class,  $t = 1, \dots, T_{ijk}$ .

**TABLE 14.9** Estimates of the Wage Equation

<b>Variable</b>	<b>Pooled Least Squares</b>		<b>Random Effects MLE</b>		<b>Random Effects FGLS</b>	
	<b>Estimate</b>	<b>Std. Error<sup>a</sup></b>	<b>Estimate</b>	<b>Std. Error</b>	<b>Estimate</b>	<b>Std. Error</b>
Exp	0.0361	0.004533	0.1078	0.002480	0.08906	0.002280
Exp <sup>2</sup>	-0.0006550	0.0001016	-0.0005054	0.00005452	-0.0007577	0.00005036
Wks	0.004461	0.001728	0.0008663	0.0006031	0.001066	0.0005939
Occ	-0.3176	0.02726	-0.03954	0.01374	-0.1067	0.01269
Ind	0.03213	0.02526	0.008807	0.01531	-0.01637	0.01391
South	-0.1137	0.02868	-0.01615	0.03201	-0.06899	0.02354
SMSA	0.1586	0.02602	-0.04019	0.01901	-0.01530	0.01649
MS	0.3203	0.03494	-0.03540	0.01880	-0.02398	0.01711
Union	0.06975	0.02667	0.03306	0.01482	0.03597	0.01367
Constant	5.8802	0.09673	4.8197	0.06035	5.3455	0.04361
$\sigma_e^2$	0.146119		0.023436 ( $\theta = 42.66926$ )		0.023102	
$\sigma_u^2$	0		0.876517 ( $\gamma = 37.40035$ )		0.838361	
ln $L$	-1899.537		249.25		—	

<sup>a</sup> Robust standard errors

## CHAPTER 14 ♦ Maximum Likelihood Estimation 577

Thus, from the outset, we allow the model to be unbalanced at all levels. In general terms, then, the random effects regression model would be

$$y_{ijkt} = \mathbf{x}'_{ijkt} \boldsymbol{\beta} + u_{ijk} + v_{ij} + w_i + \varepsilon_{ijkt}.$$

Strict exogeneity of the regressors is assumed at all levels. All parts of the disturbance are also assumed to be uncorrelated. (A normality assumption will be added later as well.) From the structure of the disturbances, we can see that the overall covariance matrix,  $\Omega$ , is block-diagonal over  $i$ , with each diagonal block itself block-diagonal in turn over  $j$ , each of these is block-diagonal over  $k$ , and, at the lowest level, the blocks, for example, for the class in our example, have the form for the random effects model that we saw earlier.

Generalized least squares has been well worked out for the balanced case. [See, e.g., Baltagi, Song, and Jung (2001), who also provide results for the three-level unbalanced case.] Define the following to be constructed from the variance components,  $\sigma_\varepsilon^2$ ,  $\sigma_u^2$ ,  $\sigma_v^2$ , and  $\sigma_w^2$ :

$$\begin{aligned}\sigma_1^2 &= T\sigma_u^2 + \sigma_\varepsilon^2, \\ \sigma_2^2 &= NT\sigma_v^2 + T\sigma_u^2 + \sigma_\varepsilon^2 = \sigma_1^2 + NT\sigma_v^2, \\ \sigma_3^2 &= MNT\sigma_w^2 + NT\sigma_v^2 + T\sigma_u^2 + \sigma_\varepsilon^2 = \sigma_2^2 + MNT\sigma_w^2.\end{aligned}$$

Then, full generalized least squares is equivalent to OLS regression of

$$\tilde{y}_{ijkt} = y_{ijkt} - \left(1 - \frac{\sigma_\varepsilon}{\sigma_1}\right) \bar{y}_{ijk} - \left(\frac{\sigma_\varepsilon}{\sigma_1} - \frac{\sigma_\varepsilon}{\sigma_2}\right) \bar{y}_{ij} - \left(\frac{\sigma_\varepsilon}{\sigma_2} - \frac{\sigma_\varepsilon}{\sigma_3}\right) \bar{y}_i \dots$$

on the same transformation of  $\mathbf{x}_{ijkt}$ . FGLS estimates are obtained by three groupwise between estimators and the within estimator for the innermost grouping.

The counterparts for the unbalanced case can be derived [see Baltagi et al. (2001)], but the degree of complexity rises dramatically. As Antwiler (2001) shows, however, if one is willing to assume normality of the distributions, then the log likelihood is very tractable. (We note an intersection of practicality with nonrobustness.) Define the variance ratios

$$\rho_u = \frac{\sigma_u^2}{\sigma_\varepsilon^2}, \rho_v = \frac{\sigma_v^2}{\sigma_\varepsilon^2}, \rho_w = \frac{\sigma_w^2}{\sigma_\varepsilon^2}.$$

Construct the following intermediate results:

$$\theta_{ijk} = 1 + T_{ijk}\rho_u, \phi_{ij} = \sum_{k=1}^{N_{ij}} \frac{T_{ijk}}{\theta_{ijk}}, \theta_{ij} = 1 + \phi_{ij}\rho_v, \phi_i = \sum_{j=1}^{M_i} \frac{\phi_{ij}}{\theta_{ij}}, \theta_i = 1 + \rho_w\phi_i$$

and sums of squares of the disturbances  $e_{ijkt} = y_{ijkt} - \mathbf{x}'_{ijkt} \boldsymbol{\beta}$ ,

$$\begin{aligned}A_{ijk} &= \sum_{t=1}^{T_{ijk}} e_{ijkt}^2, \\ B_{ijk} &= \sum_{t=1}^{T_{ijk}} e_{ijkt}, B_{ij} = \sum_{k=1}^{N_{ij}} \frac{B_{ijk}}{\theta_{ijk}}, B_i = \sum_{j=1}^{M_i} \frac{B_{ij}}{\theta_{ij}}.\end{aligned}$$

## 578 PART III ♦ Estimation Methodology

The log likelihood is

$$\ln L = -\frac{1}{2}H \ln(2\pi\sigma_\varepsilon^2) - \frac{1}{2} \left[ \sum_{i=1}^L \left\{ \ln \theta_i + \sum_{j=1}^{M_i} \left\{ \ln \theta_{ij} + \sum_{k=1}^{N_{ij}} \right. \right. \right. \\ \left. \left. \left. \left\{ \ln \theta_{ijk} + \frac{A_{ijk}}{\sigma_\varepsilon^2} - \frac{\rho_u}{\theta_{ijk}} \frac{B_{ijk}^2}{\sigma_\varepsilon^2} \right\} - \frac{\rho_v}{\theta_{ij}} \frac{B_{ij}^2}{\sigma_\varepsilon^2} \right\} - \frac{\rho_w}{\theta_i} \frac{B_i^2}{\sigma_\varepsilon^2} \right\} \right],$$

where  $H$  is the total number of observations. (For three levels,  $L = 1$  and  $\rho_w = 0$ .) Antwiler (2001) provides the first derivatives of the log likelihood function needed to maximize  $\ln L$ . However, he does suggest that the complexity of the results might make numerical differentiation attractive. On the other hand, he finds the second derivatives of the function intractable and resorts to numerical second derivatives in his application. The complex part of the Hessian is the cross derivatives between  $\beta$  and the variance parameters, and the lower right part for the variance parameters themselves. However, these are not needed. As in any generalized regression model, the variance estimators and the slope estimators are asymptotically uncorrelated. As such, one need only invert the part of the matrix with respect to  $\beta$  to get the appropriate asymptotic covariance matrix. The relevant block is

$$-\frac{\partial^2 \ln L}{\partial \beta \partial \beta'} = \frac{1}{\sigma_\varepsilon^2} \sum_{i=1}^L \sum_{j=1}^{M_i} \sum_{k=1}^{N_{ij}} \sum_{t=1}^{T_{ijk}} \mathbf{x}_{ijkl} \mathbf{x}'_{ijkl} - \frac{\rho_w}{\sigma_\varepsilon^2} \sum_{i=1}^L \sum_{j=1}^{M_i} \sum_{k=1}^{N_{ij}} \frac{1}{\theta_{ijk}} \left( \sum_{t=1}^{T_{ijk}} \mathbf{x}_{ijkl} \right) \left( \sum_{t=1}^{T_{ijk}} \mathbf{x}'_{ijkl} \right) \\ - \frac{\rho_v}{\sigma_\varepsilon^2} \sum_{i=1}^L \sum_{j=1}^{M_i} \frac{1}{\theta_{ij}} \left( \sum_{k=1}^{N_{ij}} \frac{1}{\theta_{ijk}} \left( \sum_{t=1}^{T_{ijk}} \mathbf{x}_{ijkl} \right) \right) \left( \sum_{k=1}^{N_{ij}} \frac{1}{\theta_{ijk}} \left( \sum_{t=1}^{T_{ijk}} \mathbf{x}'_{ijkl} \right) \right) \quad \text{--- (14-90)} \\ - \frac{\rho_u}{\sigma_\varepsilon^2} \sum_{i=1}^L \left( \sum_{j=1}^{M_i} \frac{1}{\theta_{ij}} \left( \sum_{k=1}^{N_{ij}} \frac{1}{\theta_{ijk}} \left( \sum_{t=1}^{T_{ijk}} \mathbf{x}_{ijkl} \right) \right) \right) \left( \sum_{j=1}^{M_i} \frac{1}{\theta_{ij}} \left( \sum_{k=1}^{N_{ij}} \frac{1}{\theta_{ijk}} \left( \sum_{t=1}^{T_{ijk}} \mathbf{x}'_{ijkl} \right) \right) \right).$$

The maximum likelihood estimator of  $\beta$  is FGLS based on the maximum likelihood estimators of the variance parameters. Thus, expression (14-90) provides the appropriate covariance matrix for the GLS or maximum likelihood estimator. The difference will be in how the variance components are computed. Baltagi et al. (2001) suggest a variety of methods for the three-level model. For more than three levels, the MLE becomes more attractive.

Given the complexity of the results, one might prefer simply to use OLS in spite of its inefficiency. As might be expected, the standard errors will be biased owing to the correlation across observations; there is evidence that the bias is downward. [See Moulton (1986).] In that event, the robust estimator in (11-4) would be the natural alternative. In the example given earlier, the nesting structure was obvious. In other cases, such as our application in Example 11.12, that might not be true. In Example 14.12 [and in the application in Baltagi (2005)], statewide observations are grouped into regions based on intuition. The impact of an incorrect grouping is unclear. Both OLS and FGLS would remain consistent—both are equivalent to GLS with the wrong weights, which we considered earlier. However, the impact on the asymptotic covariance matrix for the estimator remains to be analyzed.

## CHAPTER 14 ♦ Maximum Likelihood Estimation 579

**Example 14.12 Statewide Productivity**

Munnell (1990) analyzed the productivity of public capital at the state level using a Cobb-Douglas production function. We will use the data from that study to estimate a three-level log linear regression model,

$$\begin{aligned} \ln gsp_{jkt} = & \alpha + \beta_1 \ln pc_{jkt} + \beta_2 \ln hwy_{jkt} + \beta_3 \ln water_{jkt} \\ & + \beta_4 \ln util_{jkt} + \beta_5 \ln emp_{jkt} + \beta_6 unemp_{jkt} + \varepsilon_{jkt} + u_{jk} + v_j, \\ j = 1, \dots, 9; t = 1, \dots, 17, k = 1, \dots, N_j, \end{aligned}$$

where the variables in the model are

- $gsp$  = gross state product
- $p\_cap$  = public capital =  $hwy + water + util$
- $hwy$  = highway capital,
- $water$  = water utility capital,
- $util$  = utility capital,
- $pc$  = private capital,
- $emp$  = employment (labor),
- $unemp$  = unemployment rate,

and we have defined  $M = 9$  regions each consisting of a group of the 48 continental states:

- $Gulf$  = AL, FL, LA, MS,
- $Midwest$  = IL, IN, KY, MI, MN, OH, WI,
- $Mid\ Atlantic$  = DE, MD, NJ, NY, PA, VA,
- $Mountain$  = CO, ID, MT, ND, SD, WY,
- $New\ England$  = CT, ME, MA, NH, RI, VT,
- $South$  = GA, NC, SC, TN, WV,
- $Southwest$  = AZ, NV, NM, TX, UT,
- $Tornado\ Alley$  = AR, IA, KS, MO, NE, OK,
- $West\ Coast$  = CA, OR, WA.

For each state, we have 17 years of data, from 1970 to 1986.<sup>25</sup> The two- and three-level random effects models were estimated by maximum likelihood. The two-level model was also fit by FGLS using the methods developed in Section 11.5.3.

Table 14.10 presents the estimates of the production function using pooled OLS, OLS for the fixed effects model and both FGLS and maximum likelihood for the random effects models. Overall, the estimates are similar, though the OLS estimates do stand somewhat apart. This suggests, as one might suspect, that there are omitted effects in the pooled model. The  $F$  statistic for testing the significance of the fixed effects is 76.712 with 47 and 762 degrees of freedom. The critical value from the table is 1.379, so on this basis, one would reject the hypothesis of no common effects. Note, as well, the extremely large differences between the conventional OLS standard errors and the robust (cluster) corrected values. The three or four fold differences strongly suggest that there are latent effects at least at the state level. It remains to consider which approach, fixed or random effects is preferred. The Hausman test for fixed vs. random effects produces a chi-squared value of 18.987. The critical value is 12.592. This would imply that the fixed effects model would be the preferred specification. When we repeat the calculation of the Hausman statistic using the three-level estimates in the last column of Table 14.9, the statistic falls slightly to 15.327. Finally, note the similarity of all three sets of random effects estimates. In fact, under the hypothesis of mean independence, all three are consistent estimators. It is tempting at this point to carry out a likelihood ratio test

<sup>25</sup>The data were downloaded from the web site for Baltagi (2005) at <http://www.wiley.com/legacy/wileychi/baltagi3e/>. See Appendix Table F1.

**580 PART III ♦ Estimation Methodology**
**TABLE 14.10** Estimated Statewide Production Function

	<i>OLS</i>		<i>Fixed Effects</i>	<i>Random Effects FGLS</i>	<i>Random Effects ML</i>	<i>Nested Random Effects</i>
	<i>Estimate</i>	<i>Std. Err.<sup>a</sup></i>	<i>Estimate (Std. Err.)</i>	<i>Estimate (Std. Err.)</i>	<i>Estimate (Std. Err.)</i>	<i>Estimate (Std. Err.)</i>
	$\alpha$	1.9260 (0.2143)		2.1608 (0.1380)	2.1759 (0.1477)	2.1348 (0.1514)
$\beta_1$	0.3120 (0.04678)	0.01109 (0.04678)	0.2350 (0.02621)	0.2755 (0.01972)	0.2703 (0.02110)	0.2724 (0.02141)
$\beta_2$	0.05888 (0.05078)	0.01541 (0.03450)	0.07675 (0.03124)	0.06167 (0.02168)	0.06268 (0.02269)	0.06645 (0.02287)
$\beta_3$	0.1186 (0.03450)	0.01236 (0.01384)	0.0786 (0.0150)	0.07572 (0.01381)	0.07545 (0.01397)	0.07392 (0.01399)
$\beta_4$	0.00856 (0.04062)	0.01235 (0.002946)	-0.11478 (0.01814)	-0.09672 (0.01683)	-0.1004 (0.01730)	-0.1004 (0.01698)
$\beta_5$	0.5497 (0.06770)	0.01554 (0.00980)	0.8011 (0.02976)	0.7450 (0.02482)	0.7542 (0.02664)	0.7539 (0.02613)
$\beta_6$	-0.00727 (0.002946)		-0.005179 (0.000980)	-0.005963 (0.0008814)	-0.005809 (0.0009014)	-0.005878 (0.0009002)
$\sigma_\varepsilon$	0.085422		0.03676493	0.0367649	0.0366974	0.0366964
$\sigma_u$				0.0771064	0.0875682	0.0791243
$\sigma_v$						0.0386299
$\ln L$	853.1372		1565.501		1429.075	1430.30576

<sup>a</sup>Robust (cluster) standard errors in parentheses. The covariance matrix is multiplied by a degrees of freedom correction,  $nT/(nT - k) = 8161/7100$ .

of the hypothesis of the two-level model against the broader alternative three-level model. The test statistic would be twice the difference of the log likelihoods, which is 2.46. For one degree of freedom, the critical chi-squared with one degree of freedom is 3.84, so on this basis, we would not reject the hypothesis of the two-level model. We note, however, that there is a problem with this testing procedure. The hypothesis that a variance is zero is not well defined for the likelihood ratio test—the parameter under the null hypothesis is on the boundary of the parameter space ( $\sigma_v^2 \geq 0$ ). In this instance, the familiar distribution theory does not apply.

#### 14.9.6.c Random Effects in Nonlinear Models: MLE using Quadrature

Section 14.9.5.b describes a nonlinear model for panel data, the geometric regression model,

$$\text{Prob}[Y_{it} = y_{it} | \mathbf{x}_{it}] = \theta_{it}(1 - \theta_{it})^{y_{it}}, y_{it} = 0, 1, \dots; i = 1, \dots, n, t = 1, \dots, T_i,$$

$$\theta_{it} = 1/(1 + \lambda_{it}), \lambda_{it} = \exp(\mathbf{x}'_{it}\boldsymbol{\beta}).$$

As noted, this is a panel data model, although as stated, it has none of the features we have used for the panel data in the linear case. It is a regression model,

$$E[y_{it} | \mathbf{x}_{it}] = \lambda_{it},$$

which implies that

$$y_{it} = \lambda_{it} + \varepsilon_{it}.$$

This is simply a tautology that defines the deviation of  $y_{it}$  from its conditional mean. It might seem natural at this point to introduce a common fixed or random effect, as we

CHAPTER 14 ♦ Maximum Likelihood Estimation **581**

did earlier in the linear case, as in

$$y_{it} = \lambda_{it} + \varepsilon_{it} + c_i.$$

However, the difficulty in this specification is that whereas  $\varepsilon_{it}$  is defined residually just as the difference between  $y_{it}$  and its mean,  $c_i$  is a freely varying random variable. Without extremely complex constraints on how  $c_i$  varies, the model as stated cannot prevent  $y_{it}$  from being negative. When building the specification for a nonlinear model, greater care must be taken to preserve the internal consistency of the specification. A frequent approach in **index function models** such as this one is to introduce the common effect in the conditional mean function. The random effects geometric regression model, for example, might appear

$$\begin{aligned}\text{Prob}[Y_{it} = y_{it} | \mathbf{x}_{it}] &= \theta_{it}(1 - \theta_{it})^{y_{it}}, \quad y_{it} = 0, 1, \dots; i = 1, \dots, n, t = 1, \dots, T_i, \\ \theta_{it} &= 1/(1 + \lambda_{it}), \quad \lambda_{it} = \exp(\mathbf{x}'_{it} \boldsymbol{\beta} + u_i),\end{aligned}$$

$f(u_i)$  = the specification of the distribution of random effects over individuals.

By this specification, it is now appropriate to state the model specification as

$$\text{Prob}[Y_{it} = y_{it} | \mathbf{x}_{it}, u_i] = \theta_{it}(1 - \theta_{it})^{y_{it}}.$$

That is, our statement of the probability is now conditioned on both the observed data and the unobserved random effect. The random common effect can then vary freely and the inherent characteristics of the model are preserved.

Two questions now arise:

- How does one obtain maximum likelihood estimates of the parameters of the model? We will pursue that question now.
- If we ignore the individual heterogeneity and simply estimate the pooled model, will we obtain consistent estimators of the model parameters? The answer is sometimes, but usually not. The favorable cases are the simple loglinear models such as the geometric and Poisson models that we consider in this chapter. The unfavorable cases are most of the other common applications in the literature, including, notably, models for binary choice, censored regressions, sample selection, and, generally, nonlinear models that do not have simple exponential means. [Note that this is the crucial issue in the consideration of robust covariance matrix estimation in Sections 14.8.3 and 14.8.4. See, as well, Freedman (2006).]

We will now develop a maximum likelihood estimator for a nonlinear random effects model. To set up the methodology for applications later in the book, we will do this in a generic specification, then return to the specific application of the geometric regression model in Example 14.2. Assume, then, that the panel data model defines the probability distribution of a random variable,  $y_{it}$ , conditioned on a data vector,  $\mathbf{x}_{it}$ , and an unobserved common random effect,  $u_i$ . As always, there are  $T_i$  observations in the group, and the data on  $\mathbf{x}_{it}$  and now  $u_i$  are assumed to be strictly exogenously determined. Our model for one individual is, then,

$$p(y_{it} | \mathbf{x}_{it}, u_i) = f(y_{it} | \mathbf{x}_{it}, u_i, \boldsymbol{\theta}),$$

## 582 PART III ♦ Estimation Methodology

where  $p(y_{it} | \mathbf{x}_{it}, u_i)$  indicates that we are defining a conditional density while  $f(y_{it} | \mathbf{x}_{it}, u_i, \boldsymbol{\theta})$  defines the functional form and emphasizes the vector of parameters to be estimated. We are also going to assume that, but for the common  $u_i$ , observations within a group would be independent—the dependence of observations in the group arises through the presence of the common  $u_i$ . The joint density of the  $T_i$  observations on  $y_{it}$  given  $u_i$  under these assumptions would be

$$p(y_{i1}, y_{i2}, \dots, y_{iT_i} | \mathbf{X}_i, u_i) = \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, u_i, \boldsymbol{\theta}),$$

because conditioned on  $u_i$ , the observations are independent. But because  $u_i$  is part of the observation on the group, to construct the log-likelihood, we will require

$$p(y_{i1}, y_{i2}, \dots, y_{iT_i}, u_i | \mathbf{X}_i) = \left[ \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, u_i, \boldsymbol{\theta}) \right] f(u_i).$$

The likelihood function is the joint density for the observed random variables. Because  $u_i$  is an unobserved random effect, to construct the likelihood function, we will then have to integrate it out of the joint density. Thus,

$$p(y_{i1}, y_{i2}, \dots, y_{iT_i} | \mathbf{X}_i) = \int_{u_i} \left[ \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, u_i, \boldsymbol{\theta}) \right] f(u_i) du_i.$$

The contribution to the log-likelihood function of group  $i$  is, then,

$$\ln L_i = \ln \int_{u_i} \left[ \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, u_i, \boldsymbol{\theta}) \right] f(u_i) du_i.$$

There are two practical problems to be solved to implement this estimator. First, it will be rare that the integral will exist in closed form. (It does when the density of  $y_{it}$  is normal with linear conditional mean and the random effect is normal, because, as we have seen, this is the random effects linear model.) As such, the practical complication that arises is how the integrals are to be computed. Second, it remains to specify the distribution of  $u_i$  over which the integration is taken. The distribution of the common effect is part of the model specification. Several approaches for this model have now appeared in the literature. The one we will develop here extends the random effects model with normally distributed effects that we have analyzed in the previous section. The technique is **Butler and Moffitt's (1982) method**. It was originally proposed for extending the random effects model to a binary choice setting (see Chapter 17), but, as we shall see presently, it is straightforward to extend it to a wide range of other models. The computations center on a technique for approximating integrals known as **Gauss–Hermite quadrature**.

We assume that  $u_i$  is normally distributed with mean zero and variance  $\sigma_u^2$ . Thus,

$$f(u_i) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left(-\frac{u_i^2}{2\sigma_u^2}\right).$$

## CHAPTER 14 ♦ Maximum Likelihood Estimation 583

With this assumption, the  $i$ th term in the log-likelihood is

$$\ln L_i = \ln \int_{-\infty}^{\infty} \left[ \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, u_i, \boldsymbol{\theta}) \right] \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left(-\frac{u_i^2}{2\sigma_u^2}\right) du_i.$$

To put this function in a form that will be convenient for us later, we now let  $w_i = u_i / (\sigma_u \sqrt{2})$  so that  $u_i = \sigma_u \sqrt{2}w_i = \phi w_i$  and the Jacobian of the transformation from  $u_i$  to  $w_i$  is  $du_i = \phi dw_i$ . Now, we make the change of variable in the integral, to produce the function

$$\ln L_i = \ln \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \left[ \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi w_i, \boldsymbol{\theta}) \right] \exp(-w_i^2) dw_i.$$

For the moment, let

$$g(w_i) = \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi w_i, \boldsymbol{\theta}).$$

Then, the function we are manipulating is

$$\ln L_i = \ln \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} g(w_i) \exp(-w_i^2) dw_i.$$

The payoff to all this manipulation is that integrals of this form can be computed very accurately by Gauss–Hermite quadrature. Gauss–Hermite quadrature replaces the integration with a weighted sum of the functions evaluated at a specific set of points. For the general case, this is

$$\int_{-\infty}^{\infty} g(w_i) \exp(-w_i^2) dw_i \approx \sum_{h=1}^H z_h g(v_h)$$

where  $z_h$  is the weight and  $v_h$  is the node. Tables of the weights and nodes are found in popular sources such as Abramovitz and Stegun (1971). For example, the nodes and weights for a four-point quadrature are

$$v_h = \pm 0.52464762327529002 \text{ and } \pm 1.6506801238857849,$$

$$z_h = 0.80491409000549996 \text{ and } 0.081312835447250001.$$

In practice, it is common to use eight or more points, up to a practical limit of about 96. Assembling all of the parts, we obtain the approximation to the contribution to the log-likelihood,

$$\ln L_i = \ln \frac{1}{\sqrt{\pi}} \sum_{h=1}^H z_h \left[ \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi v_h, \boldsymbol{\theta}) \right].$$

The Hermite approximation to the log-likelihood function is

$$\ln L = \frac{1}{\sqrt{\pi}} \sum_{i=1}^n \ln \sum_{h=1}^H z_h \left[ \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi v_h, \boldsymbol{\theta}) \right].$$



This function is now to be maximized with respect to  $\boldsymbol{\theta}$  and  $\phi$ . Maximization is a complex problem. However, it has been automated in contemporary software for some models,

## 584 PART III ♦ Estimation Methodology

notably the binary choice models mentioned earlier, and is in fact quite straightforward to implement in many other models as well. The first and second derivatives of the log-likelihood function are correspondingly complex but still computable using quadrature. The estimate of  $\sigma_u$  and an appropriate standard error are obtained from  $\hat{\phi}$  using the result  $\phi = \sigma_u \sqrt{2}$ . The hypothesis of no cross-period correlation can be tested, in principle, using any of the three standard testing procedures.

### **Example 14.13 Random Effects Geometric Regression Model**

We will use the preceding to construct a random effects model for the *DocVis* count variable analyzed in Example 14.10. Using (14-90), the approximate log-likelihood function will be

$$\ln L_H = \frac{1}{\sqrt{\pi}} \sum_{i=1}^n \ln \sum_{h=1}^H z_h \left[ \prod_{t=1}^{T_i} \theta_{it}(1-\theta_{it})^{y_{it}} \right],$$

$$\theta_{it} = 1/(1+\lambda_{it}), \lambda_{it} = \exp(\mathbf{x}'_{it}\beta + \phi v_h).$$

The derivatives of the log-likelihood are approximated as well. The following is the general result—development is left as an exercise:

$$\begin{aligned} \frac{\partial \log L}{\partial (\beta)} &= \sum_{i=1}^n \frac{1}{L_i} \frac{\partial L_i}{\partial (\beta)} \\ &\approx \sum_{i=1}^n \frac{\left\{ \frac{1}{\sqrt{\pi}} \sum_{h=1}^H z_h \left[ \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi v_h, \beta) \right] \left[ \sum_{t=1}^{T_i} \frac{\partial \log f(y_{it} | \mathbf{x}_{it}, \phi v_h, \beta)}{\partial (\beta)} \right] \right\}}{\left\{ \frac{1}{\sqrt{\pi}} \sum_{h=1}^H z_h \left[ \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \phi v_h, \beta) \right] \right\}}. \end{aligned}$$

It remains only to specialize this to our geometric regression model. For this case, the density is given earlier. The missing components of the preceding derivatives are the partial derivatives with respect to  $\beta$  and  $\phi$  that were obtained in Section 14.9.5.b. The necessary result is

$$\frac{\partial \ln f(y_{it} | \mathbf{x}_{it}, \phi v_h, \beta)}{\partial (\beta)} = [\theta_{it}(1+y_{it}) - 1] \begin{pmatrix} \mathbf{x}_{it} \\ v_h \end{pmatrix}.$$

Maximum likelihood estimates of the parameters of the random effects geometric regression model are given in Example 14.13 with the fixed effects estimates for this model.

#### **14.9.6.d Fixed Effects in Nonlinear Models: Full MLE**

Using the same modeling framework that we used in the previous section, we now define a fixed effects model as an index function model with a group-specific constant term. As before, the “model” is the assumed density for a random variable,

$$p(y_{it} | d_{it}, \mathbf{x}_{it}) = f(y_{it} | \alpha_i d_{it} + \mathbf{x}'_{it}\beta),$$

where  $d_{it}$  is a dummy variable that takes the value one in every period for individual  $i$  and zero otherwise. (In more involved models, such as the censored regression model we examine in Chapter 15, there might be other parameters, such as a variance. For now, it is convenient to omit them—the development can be extended to add them later.) For convenience, we have redefined  $\mathbf{x}_{it}$  to be the nonconstant variables in the

## CHAPTER 14 ♦ Maximum Likelihood Estimation 585

model.<sup>26</sup> The parameters to be estimated are the  $K$  elements of  $\beta$  and the  $n$  individual constant terms. The log-likelihood function for the fixed effects model is

$$\ln L = \sum_{i=1}^n \sum_{t=1}^{T_i} \ln f(y_{it} | \alpha_i + \mathbf{x}'_{it}\beta),$$

where  $f(\cdot)$  is the probability density function of the observed outcome, for example, the geometric regression model that we used in our previous example. It will be convenient to let  $z_{it} = \alpha_i + \mathbf{x}'_{it}\beta$  so that  $p(y_{it} | d_{it}, \mathbf{x}_{it}) = f(y_{it} | z_{it})$ .

In the fixed effects linear regression case, we found that estimation of the parameters was made possible by a transformation of the data to deviations from group means that eliminated the person-specific constants from the equation. (See Section 11.4.1.) In a few cases of nonlinear models, it is also possible to eliminate the fixed effects from the likelihood function, although in general not by taking deviations from means. One example is the **exponential regression model** that is used for lifetimes of electronic components and electrical equipment such as light bulbs:

$$f(y_{it} | \alpha_i + \mathbf{x}'_{it}\beta) = \theta_{it} \exp(-\theta_{it} y_{it}), \theta_{it} = \exp(\alpha_i + \mathbf{x}'_{it}\beta), y_{it} \geq 0.$$

It will be convenient to write  $\theta_{it} = \gamma_i \exp(\mathbf{x}'_{it}\beta) = \gamma_i \Delta_{it}$ . We are exploiting the invariance property of the MLE—estimating  $\gamma_i = \exp(\alpha_i)$  is the same as estimating  $\alpha_i$ . The log-likelihood is

$$\begin{aligned} \ln L &= \sum_{i=1}^n \sum_{t=1}^{T_i} \ln \theta_{it} - \theta_{it} y_{it} \\ &= \sum_{i=1}^n \sum_{t=1}^{T_i} \ln(\gamma_i \Delta_{it}) - (\gamma_i \Delta_{it}) y_{it}. \end{aligned} \tag{14-91}$$

The MLE will be found by equating the  $n + K$  partial derivatives with respect to  $\gamma_i$  and  $\beta$  to zero. For each constant term,

$$\frac{\partial \ln L}{\partial \gamma_i} = \sum_{t=1}^{T_i} \left( \frac{1}{\gamma_i} - \Delta_{it} y_{it} \right).$$

Equating this to zero provides a solution for  $\gamma_i$  in terms of the data and  $\beta$ ,

$$\gamma_i = \frac{T_i}{\sum_{s=1}^{T_i} \Delta_{is} y_{is}}. \tag{14-92}$$

[Note the analogous result for the linear model in (11-25).] Inserting this solution back in the log-likelihood function in (14-91), we obtain the concentrated log-likelihood,

$$\ln L_C = \sum_{i=1}^n \sum_{t=1}^{T_i} \left[ \ln \left( \frac{T_i \Delta_{it}}{\sum_{s=1}^{T_i} \Delta_{is} y_{is}} \right) - \left( \frac{T_i \Delta_{it}}{\sum_{s=1}^{T_i} \Delta_{is} y_{is}} \right) y_{it} \right]$$

<sup>26</sup>In estimating a fixed effects linear regression model in Section 11.4, we found that it was not possible to analyze models with time-invariant variables. The same limitation applies in the nonlinear case, for essentially the same reasons. The time-invariant effects are absorbed in the constant term. In estimation, the columns of the data matrix with time-invariant variables will be transformed to columns of zeros when we compute derivatives of the log-likelihood function.

### 586 PART III ♦ Estimation Methodology

which is now only a function of  $\beta$ . This function can now be maximized with respect to  $\beta$  alone. The MLEs for  $\alpha_i$  are then found as the logs of the results of (14.29). Note, once again, we have eliminated the constants from the estimation problem, but not by computing deviations from group means. That is specific to the linear model.

The concentrated log-likelihood is only obtainable in only a small handful of cases, including the linear model, the exponential model (as just shown), the Poisson regression model, and a few others. Lancaster (2000) lists some of these and discusses the underlying methodological issues. In most cases, if one desires to estimate the parameters of a fixed effects model, it will be necessary to actually compute the possibly huge number of constant terms,  $\alpha_i$ , at the same time as the main parameters,  $\beta$ . This has widely been viewed as a practical obstacle to estimation of this model because of the need to invert a potentially large second derivatives matrix, but this is a misconception. [See, e.g., Maddala (1987), p. 317.] The likelihood equations for the fixed effects model are

$$\frac{\partial \ln L}{\partial \alpha_i} = \sum_{t=1}^{T_i} \frac{\partial \ln f(y_{it} | z_{it})}{\partial z_{it}} \frac{\partial z_{it}}{\partial \alpha_i} = \sum_{t=1}^{T_i} g_{it} = g_i = 0,$$

and

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^n \sum_{t=1}^{T_i} \frac{\partial \ln f(y_{it} | z_{it})}{\partial z_{it}} \frac{\partial z_{it}}{\partial \beta} = \sum_{i=1}^n \sum_{t=1}^{T_i} g_{it} \mathbf{x}_{it} = \mathbf{0}.$$

The second derivatives matrix is

$$\frac{\partial^2 \ln L}{\partial \alpha_i^2} = \sum_{t=1}^{T_i} \frac{\partial^2 \ln f(y_{it} | z_{it})}{\partial z_{it}^2} = \sum_{t=1}^{T_i} h_{it} = h_i < 0,$$

$$\frac{\partial^2 \ln L}{\partial \beta \partial \alpha_i} = \sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it},$$

$$\frac{\partial^2 \ln L}{\partial \beta \partial \beta'} = \sum_{i=1}^n \sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it} \mathbf{x}'_{it} = \mathbf{H}_{\beta \beta'},$$

where  $\mathbf{H}_{\beta \beta'}$  is a negative definite matrix. The likelihood equations are a large system, but the solution turns out to be surprisingly straightforward. [See Greene (2001).]

By using the formula for the partitioned inverse, we find that the  $K \times K$  submatrix of the inverse of the Hessian that corresponds to  $\beta$ , which would provide the asymptotic covariance matrix for the MLE, is

$$\begin{aligned} \mathbf{H}^{\beta \beta'} &= \left\{ \sum_{i=1}^n \left[ \sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it} \mathbf{x}'_{it} - \frac{1}{h_i} \left( \sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it} \right) \left( \sum_{t=1}^{T_i} h_{it} \mathbf{x}'_{it} \right) \right] \right\}^{-1}, \\ &= \left\{ \sum_{i=1}^n \left[ \sum_{t=1}^{T_i} h_{it} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \right] \right\}^{-1}, \quad \text{where } \bar{\mathbf{x}}_i = \frac{\sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it}}{h_i}. \end{aligned}$$

Note the striking similarity to the result we had in (9.28) for the fixed effects model in the linear case. [A similar result is noted briefly in Chamberlain (1984).] By assembling the Hessian as a partitioned matrix for  $\beta$  and the full vector of constant terms, then

## CHAPTER 14 ♦ Maximum Likelihood Estimation 587

using (A-66b) and the preceding definitions to isolate one diagonal element, we find

$$\mathbf{H}^{\alpha_i \alpha_i} = \frac{1}{h_i} + \bar{\mathbf{x}}_i' \mathbf{H}^{\beta \beta'} \bar{\mathbf{x}}_i.$$

Once again, the result has the same format as its counterpart in the linear model. [See (11.23).] In principle, the negatives of these would be the estimators of the asymptotic variances of the maximum likelihood estimators. (Asymptotic properties in this model are problematic, as we consider shortly.)

All of these can be computed quite easily once the parameter estimates are in hand, so that in fact, practical estimation of the model is not really the obstacle. [This must be qualified, however. Consider the likelihood equation for one of the constants in the geometric regression model. This would be

$$\sum_{t=1}^{T_i} [\theta_{it}(1 + y_{it}) - 1] = 0.$$

Suppose  $y_{it}$  equals zero in every period for individual  $i$ . Then, the solution occurs where  $\Sigma_i(\theta_{it} - 1) = 0$ . But  $\theta_{it}$  is between zero and one, so the sum must be negative and cannot equal zero. The likelihood equation has no solution with finite coefficients. Such groups would have to be removed from the sample to fit this model.]

It is shown in Greene (2001) in spite of the potentially large number of parameters in the model, Newton's method can be used with the following iteration, which uses only the  $K \times K$  matrix computed earlier and a few  $K \times 1$  vectors:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(s+1)} &= \hat{\boldsymbol{\beta}}^{(s)} - \left\{ \sum_{i=1}^n \left[ \sum_{t=1}^{T_i} h_{it} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i) (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \right] \right\}^{-1} \left\{ \sum_{i=1}^n \left[ \sum_{t=1}^{T_i} g_{it} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i) \right] \right\} \\ &= \hat{\boldsymbol{\beta}}^{(s)} + \Delta_{\boldsymbol{\beta}}^{(s)},\end{aligned}$$

and

$$\hat{\alpha}_l^{(s+1)} = \hat{\alpha}_l^{(s)} - [(g_{ll}/h_{ll}) + \bar{\mathbf{x}}_l' \Delta_{\boldsymbol{\beta}}^{(s)}].^{27}$$

This is a large amount of computation involving many summations, but it is linear in the number of parameters and does not involve any  $n \times n$  matrices.

In addition to the theoretical virtues and shortcomings of this model, we note the practical aspect of estimation of what are possibly a huge number of parameters,  $n + K$ . In the fixed effects case,  $n$  is not limited, and could be in the thousands in a typical application. [In Example 14.15,  $n$  is 7,293. As of this writing, the largest application of the method described here that we are aware of is Kingdon and Cassen's (2007) study in which they fit a fixed effects probit model with well over 140,000 dummy variable coefficients.] The problems with the fixed effects estimator are statistical, not practical.<sup>28</sup> The estimator relies on  $T_i$  increasing for the constant terms to be consistent—in essence, each  $\alpha_i$  is estimated with  $T_i$  observations. In this setting, not only is  $T_i$  fixed, it is also

<sup>27</sup>Similar results appear in Prentice and Gloeckler (1978) who attribute it to Rao (1973) and Chamberlain (1980, 1984).

<sup>28</sup>See Vytlacil, Aakvik, and Heckman (2005), Chamberlain (1980, 1984), Newey (1994), Bover and Arellano (1997), and Chen (1998) for some extensions of parametric and semiparametric forms of the binary choice models with fixed effects.

## 588 PART III ♦ Estimation Methodology

**TABLE 14.11** Panel Data Estimates of a Geometric Regression for DOCVIS

<i>Variable</i>	<i>Pooled</i>		<i>Random Effects<sup>a</sup></i>		<i>Fixed Effects</i>	
	<i>Estimate</i>	<i>St. Er.</i>	<i>Estimate</i>	<i>St. Er.</i>	<i>Estimate</i>	<i>St. Er.</i>
Constant	1.0918	0.1112	0.3998	0.09531		
Age	0.0180	0.0013	0.02208	0.001220	0.04845	0.003511
Education	-0.0473	0.0069	-0.04507	0.006262	-0.05437	0.03721
Income	-0.0468	0.0075	-0.1959	0.06103	-0.1892	0.09127
Kids	-0.1569	0.0319	-0.1242	0.02336	-0.002543	0.03687

<sup>a</sup>Estimated  $\sigma_u = 0.9542921$ .

likely to be quite small. As such, the estimators of the constant terms are not consistent (not because they converge to something other than what they are trying to estimate, but because they do not converge at all). There is, as well, a small sample (small  $T_i$ ) bias in the slope estimators. This is the **incidental parameters problem**. [See Neyman and Scott (1948) and Lancaster (2000).] We will examine the incidental parameters problem in a bit more detail with a Monte Carlo study in Section 15.3.

### Example 14.14 Fixed and Random Effects Geometric Regression

Example 14.10 presents pooled estimates for the geometric regression model

$$f(y_{it} | \mathbf{x}_{it}) = \theta_{it}(1 - \theta_{it})^{y_{it}}, \theta_{it} = 1/(1 + \lambda_{it}), \lambda_{it} = \exp(c_i + \mathbf{x}'_{it}\beta), y_{it} = 0, 1, \dots$$

We will now reestimate the model under the assumptions of the random and fixed effects specifications. The methods of the preceding two sections are applied directly—no modification of the procedures was required. Table 14.11 presents the three sets of maximum likelihood estimates. The estimates vary considerably. The average group size is about five. This implies that the fixed effects estimator may well be subject to a small sample bias. Save for the coefficient on *Kids*, the fixed effects and random effects estimates are quite similar. On the other hand, the two panel models give similar results to the pooled model except for the *Income* coefficient. On this basis, it is difficult to see, based solely on the results, which should be the preferred model. The model is nonlinear to begin with, so the pooled model, which might otherwise be preferred on the basis of computational ease, now has no redeeming virtues. None of the three models is robust to misspecification. Unlike the linear model, in this and other nonlinear models, the fixed effects estimator is inconsistent when  $T$  is small in both random and fixed effects models. The random effects estimator is consistent in the random effects model, but, as usual, not in the fixed effects model. The pooled estimator is inconsistent in both random and fixed effects cases (which calls into question the virtue of the robust covariance matrix). It might be tempting to use a Hausman specification test (see Section 11.5.5); however, the conditions that underlie the test are not met—unlike the linear model where the fixed effects is consistent in both cases, here it is inconsistent in both cases. For better or worse, that leaves the analyst with the need to choose the model based on the underlying theory.

## 14.10 LATENT CLASS AND FINITE MIXTURE MODELS

In this final application of maximum likelihood estimation, rather than explore a particular model, we will develop a technique that has been used in many different settings. The latent class modeling framework specifies that the distribution of the observed data

CHAPTER 14 ♦ Maximum Likelihood Estimation **589**

is a mixture of a finite number of underlying distributions. The model can be motivated in several ways:

- In the classic application of the technique, the observed data are drawn from a mix of distinct underlying populations. Consider, for example, a historical or fossilized record of the intersection (or collision) of two populations. The anthropological record consists of measurements on some variable that would differ imperfectly, but substantively, between the populations. However, the analyst has no definitive marker for which subpopulation an observation is drawn from. Given a sample of observations, they are interested in two statistical problems: (1) estimate the parameters of the underlying populations and (2) classify the observations in hand as having originated in which population. The technique has seen a number of recent applications in health econometrics. For example, in a study of obesity, Greene, Harris, Hollingsworth, and Maitra (2008) speculated that their ordered choice model (see Chapter 17) might systematically vary in a sample that contained (it was believed) some individuals who have a genetic predisposition toward obesity and most that did not. In another contemporary application, Lambert (1992) studied the number of defective outcomes in a production process. When a “zero defectives” condition is observed, it could indicate either regime 1, “the process is under control,” or regime 2, “the process is not under control but just happens to produce a zero observation.”
- In a narrower sense, one might view parameter heterogeneity in a population as a form of discrete mixing. We have modeled parameter heterogeneity using continuous distributions in ~~Chapter 11 and 15~~. The “finite mixture” approach takes the distribution of parameters across individuals to be discrete. (Of course, this is another way to interpret the first point.)
- The finite mixing approach is a means by which a distribution (model) can be constructed from a mixture of underlying distributions. Goldfeld and Quandt’s mixture of normals model in Example 13.4 is a case in which a nonnormal distribution is created by mixing two normal distributions with different parameters.

#### 14.10.1 A FINITE MIXTURE MODEL

To lay the foundation for the more fully developed model that follows, we revisit the mixture of normals model from Example 13.4. Consider a population that consists of a latent mixture of two underlying normal distributions. Neglecting for the moment that it is unknown which applies to a given individual, we have, for individual  $i$ ,

$$f(y_i | \text{class}_i = 1) = N[\mu_1, \sigma_1^2] = \frac{\exp[-\frac{1}{2}(y_i - \mu_1)^2 / \sigma_1^2]}{\sigma_1 \sqrt{2\pi}},$$

(14-93)

and

$$f(y_i | \text{class}_i = 2) = N[\mu_2, \sigma_2^2] = \frac{\exp[-\frac{1}{2}(y_i - \mu_2)^2 / \sigma_2^2]}{\sigma_2 \sqrt{2\pi}}.$$

The contribution to the likelihood function is  $f(y_i | \text{class}_i = 1)$  for an individual in class 1 and  $f(y_i | \text{class} = 2)$  for an individual in class 2. Assume that there is a true proportion  $\lambda = \text{Prob}(\text{class}_i = 1)$  of individuals in the population that are in class 1, and  $(1 - \lambda)$  in

## 590 PART III ♦ Estimation Methodology

class 2. Then the unconditional (marginal) density for individual  $i$  is

$$\begin{aligned} f(y_i) &= \lambda f(y_i | \text{class}_i = 1) + (1 - \lambda) f(y_i | \text{class}_i = 2) \\ &= E_{\text{classes}} f(y_i | \text{class}_i). \end{aligned} \quad (14-94)$$

The parameters to be estimated are  $\lambda$ ,  $\mu_1$ ,  $\mu_2$ ,  $\sigma_1$ , and  $\sigma_2$ . Combining terms, the log-likelihood for a sample of  $n$  individual observations would be

$$\ln L = \sum_{i=1}^n \ln \left( \frac{\lambda \exp \left[ -\frac{1}{2}(y_i - \mu_1)^2 / \sigma_1^2 \right]}{\sigma_1 \sqrt{2\pi}} + \frac{(1 - \lambda) \exp \left[ -\frac{1}{2}(y_i - \mu_2)^2 / \sigma_2^2 \right]}{\sigma_2 \sqrt{2\pi}} \right). \quad (14-95)$$

This is the mixture density that we saw in Example 13.4. We suggested the method of moments as an estimator of the five parameters in that example. However, this appears to be a straightforward problem in maximum likelihood estimation.

### **Example 14.15 Latent Class Model for Grade Point Averages**

Appendix Table F14.1 contains a data set of 32 observations used by Spector and Mazzeo (1980) to study whether a new method of teaching economics, the Personalized System of Instruction (PSI), significantly influenced performance in later economics courses. Variables in the data set include

- $GPA_i$  = the student's grade point average,
- $GRADE_i$  = dummy variable for whether the student's grade in intermediate macroeconomics was higher than in the principles course,
- $PSI_i$  = dummy variable for whether the individual participated in the PSI,
- $TUCE_i$  = the student's score on a pretest in economics.

We will use these data to develop a finite mixture normal model for the distribution of grade point averages.

We begin by computing maximum likelihood estimates of the parameters in (14-95). To estimate the parameters using an iterative method, it is necessary to devise a set of starting values. It might seem natural to use the simple values from a one-class model,  $\bar{y}$  and  $s_y$ , and a value such as 1/2 for  $\lambda$ . However, the optimizer will immediately stop on these values, as the derivatives will be zero at this point. Rather, it is common to use some value near these—perturbing them slightly (a few percent), just to get the iterations started. Table 14.12 contains the estimates for this two-class finite mixture model. The estimates for the one-class model are the sample mean and standard deviation of  $GPA$ . [Because these are the MLEs,  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (GPA_i - \bar{GPA})^2$ .] The means and standard deviations of the two classes are noticeably different—the model appears to be revealing a distinct splitting of the data into two classes. (Whether two is the appropriate number of classes is considered in Section 14.17.e). It is tempting at this point to identify the two classes with some other covariate, either in the data set or not, such as  $PSI$ . However, at this point, there is no basis for doing so—the classes are "latent." As the analysis continues, however, we will want to investigate whether any observed data help to predict the class membership.

**TABLE 14.12** Estimated Normal Mixture Model

<b>Parameter</b>	<b>One Class</b>		<b>Latent Class 1</b>		<b>Latent Class 2</b>	
	<b>Estimate</b>	<b>Std. Err.</b>	<b>Estimate</b>	<b>Std. Err.</b>	<b>Estimate</b>	<b>Std. Err.</b>
$\mu$	3.1172	0.08251	3.64187	0.3452	2.8894	0.2514
$\sigma$	0.4594	0.04070	0.2524	0.2625	0.3218	0.1095
<b>Probability</b>	1.0000	0.0000	0.3028	0.3497	0.6972	0.3497
<b>ln L</b>	-20.51274		-19.63654			

## CHAPTER 14 ♦ Maximum Likelihood Estimation 591

## 14.10.2 MEASURED AND UNMEASURED HETEROGENEITY

The development thus far has assumed that the analyst has no information about class membership. Estimation of the “prior” probabilities ( $\lambda$  in the preceding example) is part of the estimation problem. There may be some, albeit imperfect, information about class membership in the sample as well. For our earlier example of grade point averages, we also know the individual’s score on a test of economic literacy (*TUCE*). Use of this information might sharpen the estimates of the class probabilities. The mixture of normals problem, for example, might be formulated

$$f(y_i | \mathbf{z}_i) = \begin{pmatrix} \frac{\text{Prob}(class = 1 | \mathbf{z}_i) \exp[-\frac{1}{2}(y_i - \mu_1)^2 / \sigma_1^2]}{\sigma_1 \sqrt{2\pi}} \\ + \frac{[1 - \text{Prob}(class = 1 | \mathbf{z}_i)] \exp[-\frac{1}{2}(y_i - \mu_2)^2 / \sigma_2^2]}{\sigma_2 \sqrt{2\pi}} \end{pmatrix},$$

where  $\mathbf{z}_i$  is the vector of variables that help to explain the class probabilities. To make the mixture model amenable to estimation, it is necessary to parameterize the probabilities. The logit probability model is a common device. (See Section 14.10.6. For applications, see Greene (2007d, Section 2.3.3) and references cited.) For the two-class case, this might appear as follows:

$$\text{Prob}(class = 1 | \mathbf{z}_i) = \frac{\exp(\mathbf{z}'_i \boldsymbol{\theta})}{1 + \exp(\mathbf{z}'_i \boldsymbol{\theta})}, \text{Prob}(class = 2 | \mathbf{z}_i) = 1 - \text{Prob}(class = 1 | \mathbf{z}_i). \quad (14-96)$$

(The more general  $J$  class case is shown in Section 14.10.6.) The log-likelihood for our mixture of two normals example becomes

$$\begin{aligned} \ln L &= \sum_{i=1}^n \ln L_i \\ &= \sum_{i=1}^n \ln \left( \frac{\left( \frac{\exp(\mathbf{z}'_i \boldsymbol{\theta})}{1 + \exp(\mathbf{z}'_i \boldsymbol{\theta})} \right) \frac{\exp[-\frac{1}{2}(y_i - \mu_1)^2 / \sigma_1^2]}{\sigma_1 \sqrt{2\pi}}}{+ \left( \frac{1}{1 + \exp(\mathbf{z}'_i \boldsymbol{\theta})} \right) \frac{\exp[-\frac{1}{2}(y_i - \mu_2)^2 / \sigma_2^2]}{\sigma_2 \sqrt{2\pi}}} \right). \end{aligned} \quad (14-97)$$

The log-likelihood is now maximized with respect to  $\mu_1$ ,  $\sigma_1$ ,  $\mu_2$ ,  $\sigma_2$ , and  $\boldsymbol{\theta}$ . If  $\mathbf{z}_i$  contains a constant term and some other observed variables, then the earlier model returns if the coefficients on those other variables all equal zero. In this case, it follows that  $\lambda = \ln[\theta/(1-\theta)]$ . (This device is usually used to ensure that  $0 < \lambda < 1$  in the earlier model.)

## 14.10.3 PREDICTING CLASS MEMBERSHIP

The model in (14-97) now characterizes two random variables,  $y_i$ , the outcome variable of interest, and  $class_i$ , the indicator of which class the individual resides in. We have a joint distribution,  $f(y_i, class_i)$ , which we are modeling in terms of the conditional density,  $f(y_i | class_i)$  in (14-93), and the marginal density of  $class_i$  in (14-96). We have initially assumed the latter to be a simple Bernoulli distribution with  $\text{Prob}(class_i = 1) = \lambda$ , but then modified in the previous section to equal  $\text{Prob}(class_i = 1 | \mathbf{z}_i) = \Lambda(\mathbf{z}'_i \boldsymbol{\theta})$ . These can be viewed as the “prior” probabilities in a Bayesian sense. If we wish to make a prediction as to which class the individual came from, using all the information that we have on that individual, then the prior probability is going to waste some information.

## 592 PART III ♦ Estimation Methodology

The “posterior,” or conditional (on the remaining data) probability,

$$\text{Prob}(\text{class}_i = 1 | \mathbf{z}_i, y_i) = \frac{f(y_i, \text{class} = 1 | \mathbf{z}_i)}{f(y_i)}, \quad (14-98)$$

will be based on more information than the marginal probabilities. We have the elements that we need to compute this conditional probability. Use **Baye's theorem** to write this as

$$\begin{aligned} & \text{Prob}(\text{class}_i = 1 | \mathbf{z}_i, y_i) \\ &= \frac{f(y_i | \text{class}_i = 1, \mathbf{z}_i) \text{Prob}(\text{class}_i = 1 | \mathbf{z}_i)}{f(y_i | \text{class}_i = 1, \mathbf{z}_i) \text{Prob}(\text{class}_i = 1 | \mathbf{z}_i) + f(y_i | \text{class}_i = 2, \mathbf{z}_i) \text{Prob}(\text{class}_i = 2 | \mathbf{z}_i)}. \end{aligned} \quad (14-99)$$

The denominator is  $L_i$  (not  $\ln L_i$ ) from (14-97). The numerator is the first term in  $L_i$ . To continue our mixture of two normals example, the conditional (posterior) probability is

$$\text{Prob}(\text{class}_i = 1 | \mathbf{z}_i, y_i) = \frac{\left( \frac{\exp(\mathbf{z}'_i \boldsymbol{\theta})}{1 + \exp(\mathbf{z}'_i \boldsymbol{\theta})} \right) \frac{\exp[-\frac{1}{2}(y_i - \mu_1)^2 / \sigma_1^2]}{\sigma_1 \sqrt{2\pi}}}{L_i}, \quad (14-100)$$

while the unconditional probability is in (14-96). The conditional probability for the second class is computed using the other two marginal densities in the numerator (or by subtraction from one). Note that the conditional probabilities are functions of the data even if the unconditional ones are not. To come to the problem suggested at the outset, then, the natural predictor of  $\text{class}_i$  is the class associated with the largest estimated posterior probability.

### 14.10.4 A CONDITIONAL LATENT CLASS MODEL

To complete the construction of the latent class model, we note that the means (and, in principle, the variances) in the original model could be conditioned on observed data as well. For our normal mixture models, we might make the marginal mean,  $\mu_j$ , a conditional mean:

$$\mu_{ij} = \mathbf{x}'_i \boldsymbol{\beta}_j.$$

In the data of Example 14.4, we also observe an indicator of whether the individual has participated in a special program designed to enhance the economics program (PSI). We might modify the model,

$$f(y_i | \text{class}_i = 1, \text{PSI}_i) = N[\mu_{i1}, \sigma_1^2] = \frac{\exp[-\frac{1}{2}(y_i - \beta_{1,1} - \beta_{2,1}\text{PSI}_i)^2 / \sigma_1^2]}{\sigma_1 \sqrt{2\pi}},$$

and similarly for  $f(y_i | \text{class}_i = 2, \text{PSI}_i)$ . The model is now a **latent class linear regression** model.

More generally, as we will see shortly, the latent class, or **finite mixture model** for a variable  $y_i$  can be formulated as

$$f(y_i | \text{class}_i = j, \mathbf{x}_i) = h_j(y_i, \mathbf{x}_i, \boldsymbol{\gamma}_j),$$

where  $h_j$  denotes the density conditioned on class  $j$  – indexed by  $j$  to indicate, for example, the  $j$ th parameter vector  $\boldsymbol{\gamma}_j = (\boldsymbol{\beta}_j, \sigma_j)$  and so on. The marginal class probabilities are

$$\text{Prob}(\text{class}_i = j | \mathbf{z}_i) = p_j(j, \mathbf{z}_i, \boldsymbol{\theta}).$$

## CHAPTER 14 ♦ Maximum Likelihood Estimation 593

The methodology can be applied to any model for  $y_i$ . In the example in Section 16.7c.3, we will model a binary dependent variable with a probit model. The methodology has been applied in many other settings, such as stochastic frontier models [Orea and Kumbhakar (2004), Greene (2004)], Poisson regression models [Wedel et al. (1993)], and a wide variety of count, discrete choice, and limited dependent variable models [McLachlan and Peel (2000), Greene (2007b)].

**Example 14.16 Latent Class Regression Model for Grade Point Averages**

Combining 14.10.2 and 14.10.4, we have a latent class model for grade point averages,

$$f(GPA_i | \text{class}_i = j, PSI_i) = \frac{\exp\left[-\frac{1}{2}(y_i - \beta_{1j} - \beta_{2j}PSI_i)^2/\sigma_j^2\right]}{\sigma_j \sqrt{2\pi}}, \quad j = 1, 2,$$

$$\text{Prob}(\text{class}_i = 1 | TUCE_i) = \frac{\exp(\theta_1 + \theta_2 TUCE_i)}{1 + \exp(\theta_1 + \theta_2 TUCE_i)},$$

$$\text{Prob}(\text{class}_i = 2 | TUCE_i) = 1 - \text{Prob}(\text{class}_i = 1 | TUCE_i).$$

The log-likelihood is now

$$\ln L = \sum_{i=1}^n \ln \left( \left( \frac{\exp(\theta_1 + \theta_2 TUCE_i)}{1 + \exp(\theta_1 + \theta_2 TUCE_i)} \right) \frac{\exp\left[-\frac{1}{2}(y_i - \beta_{1,1} - \beta_{2,1}PSI_i)^2/\sigma_1^2\right]}{\sigma_1 \sqrt{2\pi}} + \left( \frac{1}{1 + \exp(\theta_1 + \theta_2 TUCE_i)} \right) \frac{\exp\left[-\frac{1}{2}(y_i - \beta_{1,2} - \beta_{2,2}PSI_i)^2/\sigma_2^2\right]}{\sigma_2 \sqrt{2\pi}} \right).$$

Maximum likelihood estimates of the parameters are given in Table 14.13.

Table 14.14 lists the observations sorted by GPA. The predictions of class membership reflect what one might guess from the coefficients in the table of coefficients. Class 2 members on average have lower GPAs than in class 1. The listing in Table 14.14 shows this clustering. It also suggests how the latent class model is using the sample information. If the results in Table 14.12—just estimating the means, constant class probabilities—are used to produce the same table, when sorted, the highest 10 GPAs are in class 1 and the remainder are in class 2. The more elaborate model is adding information on  $TUCE$  to the computation. A low  $TUCE$  score can push a high GPA individual into class 2. (Of course, this is largely what multiple linear regression does as well).

**TABLE 14.13** Estimated Latent Class Linear Regression Model for GPA

<b>Parameter</b>	<b>One Class</b>		<b>Latent Class 1</b>		<b>Latent Class 2</b>	
	<b>Estimate</b>	<b>Std. Err.</b>	<b>Estimate</b>	<b>Std. Err.</b>	<b>Estimate</b>	<b>Std. Err.</b>
$\beta_1$	3.1011	0.1117	3.3928	0.1733	2.7926	0.04988
$\beta_2$	0.03675	0.1689	-0.1074	0.2006	-0.5703	0.07553
$\sigma = \mathbf{e}'\mathbf{e}/n$	0.4443	0.0003086	0.3812	0.09337	0.1119	0.04487
$\theta_1$	0.0000	0.0000	-6.8392	3.07867	0.0000	0.0000
$\theta_2$	0.0000	0.0000	0.3518	0.1601	0.0000	0.0000
$\text{Prob}   TUCE$	1.0000		0.7063		0.2937	
$\ln L$	-20.48752				-13.39966	

**594 PART III ♦ Estimation Methodology**
**TABLE 14.14** Estimated Latent Class Probabilities

<b>GPA</b>	<b>TUCE</b>	<b>PSI</b>	<b>CLASS</b>	<b>P1</b>	<b>P1*</b>	<b>P2</b>	<b>P2*</b>
2.06	22	1	2	0.7109	0.0116	0.2891	0.9884
2.39	19	1	2	0.4612	0.0467	0.5388	0.9533
2.63	20	0	2	0.5489	0.1217	0.4511	0.8783
2.66	20	0	2	0.5489	0.1020	0.4511	0.8980
2.67	24	1	1	0.8325	0.9992	0.1675	0.0008
2.74	19	0	2	0.4612	0.0608	0.5388	0.9392
2.75	25	0	2	0.8760	0.3499	0.1240	0.6501
2.76	17	0	2	0.2975	0.0317	0.7025	0.9683
2.83	19	0	2	0.4612	0.0821	0.5388	0.9179
2.83	27	1	1	0.9345	1.0000	0.0655	0.0000
2.86	17	0	2	0.2975	0.0532	0.7025	0.9468
2.87	21	0	2	0.6336	0.2013	0.3664	0.7987
2.89	14	1	1	0.1285	1.0000	0.8715	0.0000
2.89	22	0	2	0.7109	0.3065	0.2891	0.6935
2.92	12	0	2	0.0680	0.0186	0.9320	0.9814
3.03	25	0	1	0.8760	0.9260	0.1240	0.0740
3.10	21	1	1	0.6336	1.0000	0.3664	0.0000
3.12	23	1	1	0.7775	1.0000	0.2225	0.0000
3.16	25	1	1	0.8760	1.0000	0.1240	0.0000
3.26	25	0	1	0.8760	0.9999	0.1240	0.0001
3.28	24	0	1	0.8325	0.9999	0.1675	0.0001
3.32	23	0	1	0.7775	1.0000	0.2225	0.0000
3.39	17	1	1	0.2975	1.0000	0.7025	0.0000
3.51	26	1	1	0.9094	1.0000	0.0906	0.0000
3.53	26	0	1	0.9094	1.0000	0.0906	0.0000
3.54	24	1	1	0.8325	1.0000	0.1675	0.0000
3.57	23	0	1	0.7775	1.0000	0.2225	0.0000
3.62	28	1	1	0.9530	1.0000	0.0470	0.0000
3.65	21	1	1	0.6336	1.0000	0.3664	0.0000
3.92	29	0	1	0.9665	1.0000	0.0335	0.0000
4.00	21	0	1	0.6336	1.0000	0.3664	0.0000
4.00	23	1	1	0.7775	1.0000	0.2225	0.0000

**14.10.5 DETERMINING THE NUMBER OF CLASSES**

There is an unsolved inference issue remaining in the specification of the model. The number of classes has been taken as a known parameter—two in our main example thus far, three in the following application. Ideally, one would like to determine the appropriate number of classes statistically. However,  $J$  is not a parameter in the model. A likelihood ratio test, for example, will not provide a valid result. Consider the original model in Example 14.  The model has two classes and five parameters in total. It would seem natural to test down to a one-class model that contains only the mean and variance using the LR test. However, the number of restrictions here is actually ambiguous. If  $\mu_1 = \mu_2$  and  $\sigma_1 = \sigma_2$ , then the mixing probability is irrelevant—the two class densities are the same, and it is a one-class model. Thus, the number of restrictions needed to get from the two-class model to the one-class model is ambiguous. It is neither two nor three. One strategy that has been suggested is to test upward, adding classes until the marginal class insignificantly changes the log-likelihood or one of the information criteria such as the AIC or BIC (see Section 14.6.5). Unfortunately, this approach is

## CHAPTER 14 ♦ Maximum Likelihood Estimation 595

likewise problematic because the estimates from any specification that is too short are inconsistent. The alternative would be to test down from a specification known to be too large. Heckman and Singer (1984b) discuss this possibility and note that when the number of classes  comes larger than appropriate, the estimator should break down. In our Example 14.14, if we expand to four classes, the optimizer breaks down, and it is no longer possible to compute the estimates. A five-class model does produce estimates, but some are nonsensical. This does provide at least the directions to seek a viable strategy. The authoritative treatise on finite mixture models by McLachlan and Peel (2000, Chapter 6) contains extensive discussion of this issue.

#### 14.10.6 A PANEL DATA APPLICATION

The latent class model is a useful framework for applications in panel data. The class probabilities partly play the role of common random effects, as we will now explore. The latent class model can be interpreted as a random parameters model, as suggested in Section 14.2, with a discrete distribution of the parameters. 

Suppose that  $\beta_j$  is generated from a discrete distribution with  $J$  outcomes, or classes, so that the distribution of  $\beta_j$  is over these classes. Thus, the model states that an individual belongs to one of the  $J$  latent classes, indexed by the parameter vector, but it is unknown from the sample data exactly which one. We will use the sample data to estimate the parameter vectors, the parameters of the underlying probability distribution and the probabilities of class membership. The corresponding model formulation is now

$$f(y_{it} | \mathbf{x}_{it}, \mathbf{z}_i, \Delta, \beta_1, \beta_2, \dots, \beta_J) = \sum_{j=1}^J p_{ij}(\mathbf{z}_i, \Delta) f(y_{it} | \text{class} = j, \mathbf{x}_{it}, \beta_j),$$

where it remains to parameterize the class probabilities,  $p_{ij}$ , and the structural model,  $f(y_{it} | \text{class} = j, \mathbf{x}_{it}, \beta_j)$ . The parameter matrix,  $\Delta$ , contains the parameters of the discrete probability distribution. It has  $J$  rows, one for each class, and  $M$  columns, for the  $M$  variables in  $\mathbf{z}_i$ . At a minimum,  $M = 1$  and  $\mathbf{z}_i$  contains a constant term if the class probabilities are fixed parameters as in Example 14.15. Finally, to accommodate the panel data nature of the sampling situation, we suppose that conditioned on  $\beta_j$ , that is, on membership in class  $j$ , which is fixed over time, the observations on  $y_{it}$  are independent. Therefore, for a group of  $T_i$  observations, the joint density is

$$f(y_{i1}, y_{i2}, \dots, y_{iT_i} | \text{class} = j, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT_i}, \beta_j) = \prod_{t=1}^{T_i} f(y_{it} | \text{class} = j, \mathbf{x}_{it}, \beta_j).$$

The log-likelihood function for a panel of data is

$$\ln L = \sum_{i=1}^n \ln \left[ \sum_{j=1}^J p_{ij}(\Delta, \mathbf{z}_i) \prod_{t=1}^{T_i} f(y_{it} | \text{class} = j, \mathbf{x}_{it}, \beta_j) \right].$$

The class probabilities must be constrained to sum to 1. The approach that is usually used is to reparameterize them as a set of logit probabilities, as we did in the preceding

## 596 PART III ♦ Estimation Methodology

examples. Then,

$$p_{ij}(\mathbf{z}_i, \Delta) = \frac{\exp(\theta_{ij})}{\sum_{j=1}^J \exp(\theta_{ij})}, J = 1, \dots, J, \theta_{ij} = \mathbf{z}'_i \boldsymbol{\delta}_j, \theta_{iJ} = 0 (\boldsymbol{\delta}_J = \mathbf{0}). \quad (14-101)$$

(See Section 17.21 for development of this model for the set of probabilities.) Note the restriction on  $\theta_{ij}$ . This is an identification restriction. Without it, the same set of probabilities will arise if an arbitrary vector is added to every  $\boldsymbol{\delta}_j$ . The resulting log likelihood is a continuous function of the parameters  $\beta_1, \dots, \beta_J$  and  $\delta_1, \dots, \delta_J$ . For all its apparent complexity, estimation of this model by direct maximization of the log-likelihood is not especially difficult. [See Section E.3 and Greene (2001, 2007b). The EM algorithm discussed in Section E.3.7 is especially well suited for estimating the parameters of latent class models. See McLachlan and Peel (2000).] The number of classes that can be identified is likely to be relatively small (on the order of 5 or 10 at most), however, which has been viewed as a drawback of the approach. In general, the more complex the model for  $y_{it}$ , the more difficult it becomes to expand the number of classes. Also, as might be expected, the less rich the data set in terms of cross-group variation, the more difficult it is to estimate latent class models.

Estimation produces values for the structural parameters,  $(\beta_j, \delta_j)$ ,  $j = 1, \dots, J$ . With these in hand, we can compute the prior class probabilities,  $p_{ij}$  using (14-101). For prediction purposes, we are also interested in the posterior (on the data) class probabilities, which we can compute using Bayes theorem [see (14-99)]. The conditional probability is

$$\begin{aligned} & \text{Prob(class} = j \mid \text{observation } i) \\ &= \frac{f(\text{observation } i \mid \text{class} = j) \text{Prob(class } j)}{\sum_{j=1}^J f(\text{observation } i \mid \text{class} = j) \text{Prob(class } j)} \\ &= \frac{f(y_{i1}, y_{i2}, \dots, y_{iT_i} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT_i}, \boldsymbol{\beta}_j) p_{ij}(\mathbf{z}_j, \Delta)}{\sum_{j=1}^J f(y_{i1}, y_{i2}, \dots, y_{iT_i} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT_i}, \boldsymbol{\beta}_j) p_{ij}(\mathbf{z}_j, \Delta)} \\ &= w_{ij}. \end{aligned} \quad (14-102)$$

The set of probabilities,  $\mathbf{w}_i = (w_{i1}, w_{i2}, \dots, w_{iJ})$  gives the posterior density over the distribution of values of  $\boldsymbol{\beta}$ , that is,  $[\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_J]$ .

### Example 17 Latent Class Model for Health Care Utilization

In Example 11.13, we proposed an exponential regression model,

$$y_{it} = \text{DocVis}_{it} = \exp(\mathbf{x}'_{it} \boldsymbol{\beta}) + \varepsilon_{it},$$

for the variable DocVis, the number of visits to the doctor, in the German health care data. (See Example 11.13 for details.) The regression results for the specification,

$$\mathbf{x}_{it} = (1, \text{Age}_{it}, \text{Education}_{it}, \text{Income}_{it}, \text{Kids}_{it})$$

are repeated (in parentheses) in Table 14.15 for convenience. The nonlinear least squares estimator is only semiparametric; it makes no assumption about the distribution of  $\text{DocVis}_{it}$  or about  $\varepsilon_{it}$ . We do see striking increases in the standard errors when the “” robust” asymptotic covariance matrix is used. (The estimates are given in Example 11.13.) The analysis at this point assumes that the nonlinear least squares estimator remains consistent in the presence of the cross-observation correlation. Given the way the model is specified, that is, only in terms of the conditional mean function, this is probably reasonable. The extension would imply a nonlinear generalized regression as opposed to a nonlinear ordinary regression.

## CHAPTER 14 ♦ Maximum Likelihood Estimation 597

**TABLE 14.15** Panel Data Estimates of a Geometric Regression for DocVis

Variable	Pooled MLE (Nonlinear Least Squares)		Random Effects <sup>a</sup>		Fixed Effects	
	Estimate	St. Er.	Estimate	St. Er.	Estimate	St. Er.
Constant	1.0918 (0.9801)	0.1082 (0.1813)	0.3998	0.09531		
Age	0.0180 (0.01873)	0.0013 (0.00198)	0.02208	0.001220	0.04845	0.003511
Education	-0.0473 (-0.03613)	0.0067 (0.01228)	-0.04507	0.006262	-0.05437	0.03721
Income	-0.4687 (-0.5911)	0.0726 (0.1282)	-0.1959	0.06103	-0.1982	0.09127
Kids	-0.1569 (-0.1692)	0.0306 (0.04882)	-0.1242	0.02336	-0.002543	0.03687

<sup>a</sup>Estimated  $\sigma_u = 0.9542921$ .

In Example 14.10, we narrowed this model by assuming that the observations on doctor visits were generated by a geometric distribution,

$$f(y_i | \mathbf{x}_i) = \theta_i(1 - \theta_i)^{y_i}, \theta_i = 1/(1 + \lambda_i), \lambda_i = \exp(\mathbf{x}'_i \boldsymbol{\beta}), y_i = 0, 1, \dots$$

The conditional mean is still  $\exp(\mathbf{x}'_i \boldsymbol{\beta})$ , but this specification adds the structure of a particular distribution for outcomes. The pooled model was estimated in Example 14.10. Example 14.14 added the panel data assumptions of random then fixed effects to the model. The model is now

$$f(y_{it} | \mathbf{x}_{it}) = \theta_{it}(1 - \theta_{it})^{y_{it}}, \theta_{it} = 1/(1 + \lambda_{it}), \lambda_{it} = \exp(c_i + \mathbf{x}'_{it} \boldsymbol{\beta}), y_{it} = 0, 1, \dots$$

The pooled, random effects and fixed effects estimates appear in Table 14.15. The pooled estimates, where the standard errors are corrected for the panel data grouping, are comparable to the nonlinear least squares estimates with the robust standard errors. The parameter estimates are similar—both are consistent and this is a very large sample. The smaller standard errors seen for the MLE are the product of the more detailed specification.

We will now relax the specification by assuming a two-class finite mixture model. We also specify that the class probabilities are functions of gender and marital status. For the latent class specification,

$$\text{Prob}(\text{class}_i = 1 | \mathbf{z}_i) = \Lambda(\theta_1 + \theta_2 \text{Female}_i + \theta_3 \text{Married}_i).$$

The model structure is the geometric regression as before. Estimates of the parameters of the latent class model are shown in Table 14.16. See Section E3.7 for discussion of estimation methods.

Deb and Trivedi (2002) suggested that a meaningful distinction between groups of health care system users would be between “infrequent” and “frequent” users. To investigate whether our latent class model is picking up this distinction in the data, we used (14-102) to predict the class memberships (class 1 or 2). We then linearly regressed  $\text{DocVis}_{it}$  on a constant and a dummy variable for class 2. The results are

$$\text{DocVis}_{it} = 5.8034 (0.0465) - 4.7801 (0.06282) \text{Class2}_i + e_{it},$$

where estimated standard errors are in parentheses. The linear regression suggests that the class membership dummy variable is strongly segregating the observations into frequent and infrequent users. The information in the regression is summarized in the descriptive statistics in Table 14.17.

## 598 PART III ♦ Estimation Methodology

**TABLE 14.16** Estimated Latent Class Linear Regression Model for GPA

<b>Parameter</b>	<b>One Class</b>		<b>Latent Class 1</b>		<b>Latent Class 2</b>	
	<b>Estimate</b>	<b>Std. Err.</b>	<b>Estimate</b>	<b>Std. Err.</b>	<b>Estimate</b>	<b>Std. Err.</b>
$\beta_1$	1.0918	0.1082	1.6423	0.05351	-0.3344	0.09288
$\beta_2$	0.0180	0.0013	0.01691	0.0007324	0.02649	0.001248
$\beta_3$	-0.0473	0.0067	-0.04473	0.003451	-0.06502	0.005739
$\beta_4$	-0.4687	0.0726	-0.4567	0.04688	0.01395	0.06964
$\beta_5$	-0.1569	0.0306	-0.1177	0.01611	-0.1388	0.02738
$\theta_1$	0.0000	0.0000	-0.4280	0.06938	0.0000	0.0000
$\theta_2$	0.0000	0.0000	0.8255	0.06322	0.0000	0.0000
$\theta_3$	0.0000	0.0000	-0.07829	0.07143	0.0000	0.0000
$Prob   \bar{z}$	1.0000		0.47697		0.52303	
$\ln L$	-61917.97				-58708.63	

**TABLE 14.17** Descriptive Statistics for Doctor Visits

<b>Class</b>	<b>Mean</b>	<b>Standard Deviation</b>
All, $n = 27,326$	3.18352	7.47579
Class 1, $n = 12,349$	5.80347	1.63076
Class 2, $n = 14,977$	1.02330	3.18352

## 14.11 SUMMARY AND CONCLUSIONS

This chapter has presented the theory and several applications of maximum likelihood estimation, which is the most frequently used estimation technique in econometrics after least squares. The maximum likelihood estimators are consistent, asymptotically normally distributed, and efficient among estimators that have these properties. The drawback to the technique is that it requires a fully parametric, detailed specification of the data generating process. As such, it is vulnerable to misspecification problems. The previous chapter considered GMM estimation techniques which are less parametric, but more robust to variation in the underlying data generating process. Together, ML and GMM estimation account for the large majority of empirical estimation in econometrics.

### Key Terms and Concepts

- AIC
- Asymptotic efficiency
- Asymptotic normality
- Asymptotic variance
- Autocorrelation
- Baye's theorem
- BHHH estimator
- BIC
- Butler and Moffitt's model
- Cluster estimator
- Concentrated log-likelihood
- Conditional likelihood
- Consistency
- Cramér–Rao lower bound
- Efficient score
- Estimable parameters
- Exclusion restriction
- Exponential regression model
- Finite mixture model
- Fixed effects
- Full information maximum likelihood (FIML)
- Gauss–Hermite quadrature
- Generalized sum of squares
- Geometric regression
- GMM estimator
- Identification
- Incidental parameters problem

## CHAPTER 14 ♦ Maximum Likelihood Estimation 599

- Index function model
- Information matrix
- Information matrix equality
- Invariance
- Jacobian
- Kullback–Leibler information criterion
- Latent regression
- Lagrange multiplier statistic
- Lagrange multiplier (LM) test
- Latent class model
- Latent class linear regression model
- Likelihood equation
- Likelihood function
- Likelihood inequality
- Likelihood ratio
- Likelihood ratio index
- Likelihood ratio statistic
- Likelihood ratio (LR) test
- Limited information maximum likelihood
- Logistic probability mode
- Logit model
- Loglinear conditional mean
- Maximum likelihood
- Maximum likelihood estimator
- $M$  estimator
- Method of scoring
- Murphy and Topel estimator
- Newton's method
- Noncentral chi-squared distribution
- Nonlinear least squares
- Nonnested models
- Normalization
- Oberhofer–Kmenta estimator
- Outer product of gradients estimator (OPG)
- Parameter space
- Precision parameter
- Pseudo-log likelihood function
- Pseudo MLE
- Pseudo  $R$  squared
- Quadrature
- Random effects
- Regularity conditions
- Sandwich estimator
- Score test
- Score vector
- Stochastic frontier
- Two-step maximum likelihood estimation
- Wald statistic
- Wald test
- Vuong test

### **Exercises**

1. Assume that the distribution of  $x$  is  $f(x) = 1/\theta$ ,  $0 \leq x \leq \theta$ . In random sampling from this distribution, prove that the sample maximum is a consistent estimator of  $\theta$ . Note that you can prove that the maximum is the maximum likelihood estimator of  $\theta$ . But the usual properties do not apply here. Why not? (Hint: Attempt to verify that the expected first derivative of the log-likelihood with respect to  $\theta$  is zero.)
2. In random sampling from the exponential distribution  $f(x) = (1/\theta)e^{-x/\theta}$ ,  $x \geq 0$ ,  $\theta > 0$ , find the maximum likelihood estimator of  $\theta$  and obtain the asymptotic distribution of this estimator.
3. *Mixture distribution.* Suppose that the joint distribution of the two random variables  $x$  and  $y$  is

$$f(x, y) = \frac{\theta e^{-(\beta+\theta)y}(\beta y)^x}{x!}, \quad \beta, \theta > 0, y \geq 0, x = 0, 1, 2, \dots$$

- a. Find the maximum likelihood estimators of  $\beta$  and  $\theta$  and their asymptotic joint distribution.
- b. Find the maximum likelihood estimator of  $\theta/(\beta + \theta)$  and its asymptotic distribution.
- c. Prove that  $f(x)$  is of the form

$$f(x) = \gamma(1 - \gamma)^x, x = 0, 1, 2, \dots,$$

and find the maximum likelihood estimator of  $\gamma$  and its asymptotic distribution.

- d. Prove that  $f(y|x)$  is of the form

$$f(y|x) = \frac{\lambda e^{-\lambda y}(\lambda y)^x}{x!}, \quad y \geq 0, \lambda > 0.$$

## 600 PART III ♦ Estimation Methodology

Prove that  $f(y|x)$  integrates to 1. Find the maximum likelihood estimator of  $\lambda$  and its asymptotic distribution. (Hint: In the conditional distribution, just carry the  $x$ 's along as constants.)

- e. Prove that

$$f(y) = \theta e^{-\theta y}, \quad y \geq 0, \quad \theta > 0.$$

Find the maximum likelihood estimator of  $\theta$  and its asymptotic variance.

- f. Prove that

$$f(x|y) = \frac{e^{-\beta y}(\beta y)^x}{x!}, \quad x = 0, 1, 2, \dots, \beta > 0.$$

Based on this distribution, what is the maximum likelihood estimator of  $\beta$ ?

4. Suppose that  $x$  has the Weibull distribution

$$f(x) = \alpha \beta x^{\beta-1} e^{-\alpha x^\beta}, \quad x \geq 0, \alpha, \beta > 0.$$

- a. Obtain the log-likelihood function for a random sample of  $n$  observations.
- b. Obtain the likelihood equations for maximum likelihood estimation of  $\alpha$  and  $\beta$ . Note that the first provides an explicit solution for  $\alpha$  in terms of the data and  $\beta$ . But, after inserting this in the second, we obtain only an implicit solution for  $\beta$ . How would you obtain the maximum likelihood estimators?
- c. Obtain the second derivatives matrix of the log-likelihood with respect to  $\alpha$  and  $\beta$ . The exact expectations of the elements involving  $\beta$  involve the derivatives of the gamma function and are quite messy analytically. Of course, your exact result provides an empirical estimator. How would you estimate the asymptotic covariance matrix for your estimators in part b?
- d. Prove that  $\alpha \beta \text{Cov}[\ln x, x^\beta] = 1$ . (Hint: The expected first derivatives of the log-likelihood function are zero.)

5. The following data were generated by the Weibull distribution of Exercise 4:

1.3043	0.49254	1.2742	1.4019	0.32556	0.29965	0.26423
1.0878	1.9461	0.47615	3.6454	0.15344	1.2357	0.96381
0.33453	1.1227	2.0296	1.2797	0.96080	2.0070	

- a. Obtain the maximum likelihood estimates of  $\alpha$  and  $\beta$ , and estimate the asymptotic covariance matrix for the estimates.

- b. Carry out a Wald test of the hypothesis that  $\beta = 1$ .
- c. Obtain the maximum likelihood estimate of  $\alpha$  under the hypothesis that  $\beta = 1$ .
- d. Using the results of parts a and c, carry out a likelihood ratio test of the hypothesis that  $\beta = 1$ .
- e. Carry out a Lagrange multiplier test of the hypothesis that  $\beta = 1$ .

6. **Limited Information Maximum Likelihood Estimation.** Consider a bivariate distribution for  $x$  and  $y$  that is a function of two parameters,  $\alpha$  and  $\beta$ . The joint density is  $f(x, y|\alpha, \beta)$ . We consider maximum likelihood estimation of the two parameters. The full information maximum likelihood estimator is the now familiar maximum likelihood estimator of the two parameters. Now, suppose that we can factor the joint distribution as done in Exercise 3, but in this case, we have

CHAPTER 14 ♦ Maximum Likelihood Estimation **601**

$f(x, y | \alpha, \beta) = f(y | x, \alpha, \beta) f(x | \alpha)$ . That is, the conditional density for  $y$  is a function of both parameters, but the marginal distribution for  $x$  involves only  $\alpha$ .

- a. Write down the general form for the log-likelihood function using the joint density.
- b. Because the joint density equals the product of the conditional times the marginal, the log-likelihood function can be written equivalently in terms of the factored density. Write this down, in general terms.
- c. The parameter  $\alpha$  can be estimated by itself using only the data on  $x$  and the log likelihood formed using the marginal density for  $x$ . It can also be estimated with  $\beta$  by using the full log-likelihood function and data on both  $y$  and  $x$ . Show this.
- d. Show that the first estimator in part c has a larger asymptotic variance than the second one. This is the difference between a limited information maximum likelihood estimator and a full information maximum likelihood estimator.
- e. Show that if  $\partial^2 \ln f(y | x, \alpha, \beta) / \partial \alpha \partial \beta = 0$ , then the result in part d is no longer true.
7. Show that the likelihood inequality in Theorem 14.3 holds for the Poisson distribution used in Section 14.3 by showing that  $E[(1/n) \ln L(\theta | y)]$  is uniquely maximized at  $\theta = \theta_0$ . (*Hint:* First show that the expectation is  $-\theta + \theta_0 \ln \theta - E_0[\ln y_i!]$ .)
8. Show that the likelihood inequality in Theorem 14.3 holds for the normal distribution.
9. For random sampling from the classical regression model in (14-3), reparameterize the likelihood function in terms of  $\eta = 1/\sigma$  and  $\delta = (1/\sigma)\beta$ . Find the maximum likelihood estimators of  $\eta$  and  $\delta$  and obtain the asymptotic covariance matrix of the estimators of these parameters.
10. Consider sampling from a multivariate normal distribution with mean vector  $\mu = (\mu_1, \mu_2, \dots, \mu_M)$  and covariance matrix  $\sigma^2 \mathbf{I}$ . The log-likelihood function is

$$\ln L = \frac{-nM}{2} \ln(2\pi) - \frac{nM}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu})' (\mathbf{y}_i - \boldsymbol{\mu}).$$

Show that the maximum likelihood estimates of the parameters are  $\hat{\mu} = \bar{y}_m$ , and

$$\hat{\sigma}_{ML}^2 = \frac{\sum_{i=1}^n \sum_{m=1}^M (y_{im} - \bar{y}_m)^2}{nM} = \frac{1}{M} \sum_{m=1}^M \frac{1}{n} \sum_{i=1}^n (y_{im} - \bar{y}_m)^2 = \frac{1}{M} \sum_{m=1}^M \hat{\sigma}_m^2.$$

Derive the second derivatives matrix and show that the asymptotic covariance matrix for the maximum likelihood estimators is

$$\left\{ -E \left[ \frac{\partial^2 \ln L}{\partial \theta \partial \theta'} \right] \right\}^{-1} = \begin{bmatrix} \sigma^2 \mathbf{I}/n & \mathbf{0} \\ \mathbf{0} & 2\sigma^4/(nM) \end{bmatrix}.$$

Suppose that we wished to test the hypothesis that the means of the  $M$  distributions were all equal to a particular value  $\mu^0$ . Show that the Wald statistic would be

$$\mathbf{W} = (\bar{\mathbf{y}} - \mu^0 \mathbf{i})' \left( \frac{\hat{\sigma}^2}{n} \mathbf{I} \right)^{-1} (\bar{\mathbf{y}} - \mu^0 \mathbf{i}) = \left( \frac{n}{\hat{\sigma}^2} \right) (\bar{\mathbf{y}} - \mu^0 \mathbf{i})' (\bar{\mathbf{y}} - \mu^0 \mathbf{i}),$$

where  $\bar{\mathbf{y}}$  is the vector of sample means.

11. Prove the result claimed in Example 4.7.

## 602 PART III ♦ Estimation Methodology

### Applications

1. **Binary Choice** This application will be based on the health care data analyzed in Example 11.1 and several others. Details on obtaining the data are given in Example 11.1. We consider analysis of a dependent variable,  $y_{it}$ , that takes values 0 and 1 and 0 with probabilities  $F(\mathbf{x}'_i \boldsymbol{\beta})$  and  $1 - F(\mathbf{x}'_i \boldsymbol{\beta})$ , where  $F$  is a function that defines a probability. The dependent variable,  $y_{it}$ , is constructed from the count variable  $DocVis$ , which is the number of visits to the doctor in the given year. Construct the binary variable

$$y_{it} = 1 \text{ if } DocVis_{it} > 0, 0 \text{ otherwise.}$$

We will build a model for the probability that  $y_{it}$  equals one. The independent variables of interest will be,

$$\mathbf{x}_{it} = (1, age_{it}, educ_{it}, female_{it}, married_{it}, hsat_{it}).$$

- a. According to the model, the theoretical density for  $y_{it}$  is

$$f(y_{it} | \mathbf{x}_{it}) = F(\mathbf{x}'_{it} \boldsymbol{\beta}) \text{ for } y_{it} = 1 \text{ and } 1 - F(\mathbf{x}'_{it} \boldsymbol{\beta}) \text{ for } y_{it} = 0.$$

We will assume that a “logit model” (see Section 11.1) is appropriate, so that

$$F(\mathbf{x}'_{it} \boldsymbol{\beta}) = \Lambda(\mathbf{x}'_{it} \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}'_{it} \boldsymbol{\beta})}{1 + \exp(\mathbf{x}'_{it} \boldsymbol{\beta})}.$$

Show that for the two outcomes, the probabilities may be combined into the density function

$$f(y_{it} | \mathbf{x}_{it}) = g(y_{it}, \mathbf{x}_{it}, \boldsymbol{\beta}) = \Lambda[(2y_{it} - 1)\mathbf{x}'_{it} \boldsymbol{\beta}].$$

Now, use this result to construct the log-likelihood function for a sample of data on  $(y_{it}, \mathbf{x}_{it})$ . (Note: We will be ignoring the panel aspect of the data set. Build the model as if this were a cross section.)

- b. Derive the likelihood equations for estimation of  $\boldsymbol{\beta}$ .
- c. Derive the second derivatives matrix of the log likelihood function. (Hint: The following will prove useful in the derivation:  $d\Lambda(t)/dt = \Lambda(t)[1 - \Lambda(t)]$ .)
- d. Show how to use Newton's method to estimate the parameters of the model.
- e. Does the method of scoring differ from Newton's method? Derive the negative of the expectation of the second derivatives matrix.
- f. Obtain maximum likelihood estimates of the parameters for the data and variables noted. Report your results: estimates, standard errors, etc., as well as the value of the log-likelihood.
- g. Test the hypothesis that the coefficients on female and marital status are zero. Show how to do the test using Wald, LM, and LR tests, and then carry out the tests.
- h. Test the hypothesis that all the coefficients in the model save for the constant term are equal to zero.

## 15

# SIMULATION-BASED ESTIMATION AND INFERENCE AND RANDOM PARAMETER MODELS

---

## 15.1 INTRODUCTION

Simulation-based methods have become increasingly popular in econometrics. They are extremely computer intensive, but steady improvements in recent years in computation hardware and software have reduced that cost enormously. The payoff has been in the form of methods for solving estimation and inference problems that have previously been unsolvable in analytic form. The methods are used for two main functions. First, simulation-based methods are used to infer the characteristics of random variables, including estimators, functions of estimators, test statistics, and so on, by sampling from their distributions. Second, simulation is used in constructing estimators that involve complicated integrals that do not exist in a closed form that can be evaluated. In such cases, when the integral can be written in the form of an expectation, simulation methods can be used to evaluate it to within acceptable degrees of approximation by estimating the expectation as the mean of a random sample. The technique of maximum simulated likelihood (MSL) is essentially a classical sampling theory counterpart to the hierarchical Bayesian estimator considered in Chapter 16. Since the celebrated paper of Berry, Levinsohn, and Pakes (1995) and a related literature advocated by McFadden and Train (2000), maximum simulated likelihood estimation has been used in a large and growing number of studies.

The following are three examples from earlier chapters that have relied on simulation methods.

***Example 15.1 Inferring the Sampling Distribution of the Least Squares Estimator***

In Example 4.1, we demonstrated the idea of a sampling distribution by drawing several thousand samples from a population and computing a least squares coefficient with each sample. We then examined the distribution of the sample of linear regression coefficients. A histogram suggested that the distribution appeared to be normal and centered over the true population value of the coefficient.

***Example 15.2 Bootstrapping the Variance of the LAD Estimator***

In Example 4.5, we compared the asymptotic variance of the least absolute deviations (LAD) estimator to that of the ordinary least squares (OLS) estimator. The form of the asymptotic variance of the LAD estimator is not known except in the special case of normally distributed disturbances. We relied, instead, on a random sampling method to approximate features of the sampling distribution of the LAD estimator. We used a device (bootstrapping) that allowed us to draw a sample of observations from the population that produces the estimator. With that random sample, by computing the corresponding sample statistics, we can infer

## 604 PART III ♦ Estimation Methodology

characteristics of the distribution such as its variance and its 2.5th and 97.5th percentiles which can be used to construct a confidence interval.

### **Example 15.3 Least Simulated Sum of Squares**

Familiar estimation and inference methods, such as least squares and maximum likelihood, rely on “closed form” expressions that can be evaluated exactly [at least in principle—likelihood equations such as (14-4)] may require an iterative solution. Model building and analysis often require evaluation of expressions that cannot be computed directly. Familiar examples include expectations that involve integrals with no d form such as the random effects nonlinear regression model presented in Section 14.9.2. The estimation problem posed there involved nonlinear least squares estimation of the parameters of

$$E[y_{it}|\mathbf{x}_{it}, u_i] = h(\mathbf{x}'_{it}\beta + u_i).$$

Minimizing the sum of squares,

$$S(\beta) = \sum_i \sum_t [y_{it} - h(\mathbf{x}'_{it}\beta + u_i)]^2,$$

is not feasible because  $u_i$  is not observed. In this formulation,

$$E[y|\mathbf{x}_{it}] = E_u E[y_{it}|\mathbf{x}_{it}, u_i] = \int_u E[y_{it}|\mathbf{x}_{it}, u_i] f(u_i) du_i,$$

so the feasible estimation problem would involve the sum of squares,

$$S^*(\beta) = \sum_j \sum_t \left[ y_{it} - \int_u h(\mathbf{x}'_{it}\beta + u_i) f(u_i) du_i \right]^2.$$

When the function is linear and  $u_i$  is normally distributed, this is a simple problem—it reduces to ordinary linear least squares. If either condition is not met, then the integral generally remains in the estimation problem. Although the integral,

$$E_u[h(\mathbf{x}'_{it}\beta + u_i)] = \int_u h(\mathbf{x}'_{it}\beta + u_i) f(u_i) du_i,$$

cannot be computed, if a large sample of  $R$  observations from the population of  $u_i$ , that is,  $u_{ir}, r = 1, \dots, R$ , were observed, then by virtue of the law of large numbers, we could rely on

$$\frac{1}{R} \sum_r h(\mathbf{x}'_{it}\beta + u_{ir}) = E_u E[y_{it}|\mathbf{x}_{it}, u_i] = \int_u h(\mathbf{x}'_{it}\beta + u_i) f(u_i) du_i. \quad (15-1)$$

We are suppressing the extra parameter,  $\sigma_u$ , which would become part of the estimation problem. A convenient way to formulate the problem is to write  $u_i = \sigma_u v_i$  where  $v_i$  has zero mean and variance one. By using this device, integrals can be replaced with sums that are feasible to compute. Our “simulated sum of squares” becomes

$$S_{\text{simulated}}(\beta) = \sum_i \sum_t \left[ y_{it} - \left( \frac{1}{R} \sum_r h(\mathbf{x}'_{it}\beta + \sigma_u v_{ir}) \right) \right]^2, \quad (15-2)$$

which can be minimized by conventional methods. As long as (15-1) holds, then

$$\frac{1}{nT} \sum_i \sum_t \left[ y_{it} - \left( \frac{1}{R} \sum_r h(\mathbf{x}'_{it}\beta + \sigma_u v_{ir}) \right) \right]^2 \xrightarrow{nT} \sum_i \sum_t \left[ y_{it} - \int_v h(\mathbf{x}'_{it}\beta + \sigma_u v_i) f(v_i) dv_i \right]^2 \quad (15-3)$$

and it follows that with sufficiently increasing  $R$ , the  $\beta$  that minimizes the left-hand side converges (in  $nT$ ) to the same parameter vector that minimizes the probability limit of the right-hand side. We are thus able to substitute a computer simulation for the intractable computation on the right-hand side of the expression.

This chapter will describe some of the (increasingly) more common applications of simulation methods in econometrics. We begin in Section 15.2 with the essential tool at the heart of all the computations, random number generation. Section 15.3 describes

**CHAPTER 15 ♦ Simulation-Based Estimation and Inference 605**

simulation-based inference using the method of Krinsky and Robb as an alternative to the delta method (see Section 4.4.4). The method of bootstrapping for inferring the features of the distribution of an estimator is described in Section 15.4. In Section 15.5, we will use a Monte Carlo study to learn about the behavior of a test statistic and the behavior of the fixed effects estimator in some nonlinear models. Sections 15.6 to 15.9 present simulation-based estimation methods. The essential ingredient of this entire set of results is the computation of integrals. Section 15.6.1 describes an application of a simulation-based estimator, a nonlinear random effects model. Section 15.6.2 discusses methods of integration. Then, the methods are applied to the estimation of the random effects model. Sections 15.7–15.9 describe several techniques and applications, including maximum simulated likelihood estimation for random parameter and hierarchical models. A third major (perhaps *the* major) application of simulation-based estimation in the current literature is Bayesian analysis using Markov Chain Monte Carlo (MCMC or MC<sup>2</sup>) methods. Bayesian methods are discussed separately in Chapter 16. Sections 15.10 and 15.11 consider two remaining aspects of modeling parameter heterogeneity, estimation of individual specific parameters, and a comparison of modeling with continuous distributions to modeling with discrete distributions using latent class models.

## 15.2 RANDOM NUMBER GENERATION

All the techniques we will consider here rely on samples of observations from an underlying population. We will sometimes call these “random samples,” though it will emerge shortly that they are never actually random. One of the important aspects of this entire body of research is the need to be able to replicate one’s computations. If the samples of draws used in any kind of simulation-based analysis were truly random, then this would be impossible. Although the methods we consider here will appear to be random, they are, in fact, deterministic—the “samples” can be replicated. For this reason, the sampling methods described in this section are more often labeled “pseudo-random number generators.” (This does raise an intriguing question: Is it possible to generate truly random draws from a population with a computer? The answer for practical purposes is no.) This section will begin with a description of some of the mechanical aspects of random number generation. We will then detail the methods of generating particular kinds of random samples. [See Train (2009, Chapter 3) for extensive further discussion.]

### 15.2.1 GENERATING PSEUDO-RANDOM NUMBERS

Data are generated internally in a computer using **pseudo-random number generators**. These computer programs generate sequences of values that appear to be strings of draws from a specified probability distribution. There are many types of random number generators, but most take advantage of the inherent inaccuracy of the digital representation of real numbers. The method of generation is usually by the following steps:

1. Set a **seed**.
2. Update the seed by  $\text{seed}_j = \text{seed}_{j-1} \times s$  value.
3.  $x_j = \text{seed}_j \times x$  value.
4. Transform  $x_j$  if necessary, and then move  $x_j$  to desired place in memory.
5. Return to step 2, or exit if no additional values are needed.

**606 PART III ♦ Estimation Methodology**

Random number generators produce sequences of values that resemble strings of random draws from the specified distribution. In fact, the sequence of values produced by the preceding method is not truly random at all; it is a deterministic **Markov chain** of values. The set of 32 bits in the random value only appear random when subjected to certain tests. [See Press et al. (1986).] Because the series is, in fact, deterministic, at any point that this type of generator produces a value it has produced before, it must thereafter replicate the entire sequence. Because modern digital computers typically use 32-bit double words to represent numbers, it follows that the longest string of values that this kind of generator can produce is  $2^{32} - 1$  (about 4.3 billion). This length is the **period** of a random number generator. (A generator with a shorter period than this would be inefficient, because it is possible to achieve this period with some fairly simple algorithms.) Some improvements in the periodicity of a generator can be achieved by the method of **shuffling**. By this method, a set of, say, 128 values is maintained in an array. The random draw is used to select one of these 128 positions from which the draw is taken and then the value in the array is replaced with a draw from the generator. The period of the generator can also be increased by combining several generators. [See L'Ecuyer (1998), Gentle (2002, 2003), and Greene (2007b).]

The deterministic nature of pseudo-random number generators is both a flaw and a virtue. Many Monte Carlo studies require billions of draws, so the finite period of any generator represents a nontrivial consideration. On the other hand, being able to reproduce a sequence of values just by resetting the seed to its initial value allows the researcher to replicate a study.<sup>1</sup> The seed itself can be a problem. It is known that certain seeds in particular generators will produce shorter series or series that do not pass randomness tests. For example, *congruential* generators of the sort just discussed should be started from odd seeds.

**15.2.2 SAMPLING FROM A STANDARD UNIFORM POPULATION**

The output of the generator described in Section 15.2.1 will be a pseudo-draw from the  $U[0, 1]$  population. (In principle, the draw should be from the closed interval  $[0, 1]$ . However, the actual draw produced by the generator will be strictly between zero and one with probability just slightly below one. In the application described, the draw will be constructed from the sequence of 32 bits in a double word. All but two of the  $2^{31}-1$  strings of bits will produce a value in  $(0, 1)$ . The practical result is consistent with the theoretical one, that the probabilities attached to the terminal points are zero also.) When sampling from a standard uniform,  $U[0, 1]$  population, the sequence is a kind of difference equation, because given the initial seed,  $x_j$  is ultimately a function of  $x_{j-1}$ . In most cases, the result at step 3 is a pseudo-draw from the continuous uniform distribution in the range zero to one, which can then be transformed to a draw from another distribution by using the fundamental probability transformation.

---

<sup>1</sup>Readers of empirical studies are often interested in replicating the computations. In Monte Carlo studies, at least in principle, data can be replicated efficiently merely by providing the random number generator and the seed.

## CHAPTER 15 ♦ Simulation-Based Estimation and Inference **607**

### 15.2.3 SAMPLING FROM CONTINUOUS DISTRIBUTIONS

One is usually interested in obtaining a sequence of draws,  $x_1, \dots, x_R$ , from some particular population such as the normal with mean  $\mu$  and variance  $\sigma^2$ . A sequence of draws from  $U[0, 1]$ ,  $u_1, \dots, u_R$ , produced by the random number generator is an intermediate step. These will be transformed into draws from the desired population. A common approach is to use the **fundamental probability transformation**. For continuous distributions, this is done by treating the draw,  $u_r = F_r$  as if  $F_r$  were  $F(x_r)$ , where  $F(\cdot)$  is the cdf of  $x$ . For example, if we desire draws from the exponential distribution with known  $\theta$ , then  $F(x) = 1 - \exp(-\theta x)$ . The inverse transform is  $x = (-1/\theta) \ln(1 - F)$ . For example, for a draw of  $u = 0.4$  with  $\theta = 5$ , the associated  $x$  would be  $(-1/5) \ln(1 - .4) = 0.1022$ . For the logistic population with cdf  $F(x) = \Lambda(x) = \exp(x)/[1 + \exp(x)]$ , the inverse transformation is  $x = \ln[F/(1 - F)]$ . There are many references, for example, Evans, Hastings, and Peacock (2000) and Gentle (2003), that contain tables of inverse transformations that can be used to construct random number generators.

One of the most common applications is the draws from the standard normal distribution. This is complicated because there is no closed form for  $\Phi^{-1}(F)$ . There are several ways to proceed. A well-known approximation to the inverse function is given in Abramovitz and Stegun (1971):

$$\Phi^{-1}(F) = x \approx T - \frac{c_0 + c_1 T + c_2 T^2}{1 + d_1 T + d_2 T^2 + d_3 T^3},$$

where  $T = [\ln(1/H^2)]^{1/2}$  and  $H = F$  if  $F > 0.5$  and  $1 - F$  otherwise. The sign is then reversed if  $F < 0.5$ . A second method is to transform the  $U[0, 1]$  values directly to a standard normal value. The Box–Muller (1958) method is  $z = (-2 \ln u_1)^{1/2} \cos(2\pi u_2)$ , where  $u_1$  and  $u_2$  are two independent  $U[0, 1]$  draws. A second  $N[0, 1]$  draw can be obtained from the same two values by replacing cos with sin in the transformation. The Marsaglia–Bray (1964) generator is  $z_i = x_i[-(2/v) \ln v]^{1/2}$ , where  $x_i = 2u_i - 1$ ,  $u_i$  is a random draw from  $U[0, 1]$  and  $v = u_1^2 + u_2^2$ ,  $i = 1, 2$ . The pair of draws is rejected and redrawn if  $v \geq 1$ .

Sequences of draws from the standard normal distribution can easily be transformed into draws from other distributions by making use of the results in Section B.4. For example, the square of a standard normal draw will be a draw from chi-squared[1], and the sum of  $K$  chi-squared[1]s is chi-squared [ $K$ ]. From this relationship, it is possible to produce samples from the chi-squared[ $K$ ],  $t[n]$ , and  $F[K, n]$  distributions.

A related problem is obtaining draws from the truncated normal distribution. The random variable with truncated normal distribution is obtained from one with a normal distribution by discarding the part of the range above a value  $U$  and below a value  $L$ . The density of the resulting random variable is that of a normal distribution restricted to the range  $[L, U]$ . The truncated normal density is

$$f(x|L \leq x \leq U) = \frac{f(x)}{\text{Prob}[L \leq x \leq U]} = \frac{(1/\sigma)\phi[(x - \mu)/\sigma]}{\Phi[(U - \mu)/\sigma] - \Phi[(L - \mu)/\sigma]},$$

where  $\phi(t) = (2\pi)^{-1/2} \exp(-t^2/2)$  and  $\Phi(t)$  is the cdf. An obviously inefficient (albeit effective) method of drawing values from the truncated normal  $[\mu, \sigma^2]$  distribution in the range  $[L, U]$  is simply to draw  $F$  from the  $U[0, 1]$  distribution and transform it first to a standard normal variate as discussed previously and then to the  $N[\mu, \sigma^2]$  variate by

## 608 PART III ♦ Estimation Methodology

using  $x = \mu + \sigma\Phi^{-1}(F)$ . Finally, the value  $x$  is retained if it falls in the range  $[L, U]$  and discarded otherwise. This rejection method will require, on average,  $1/\{\Phi[(U - \mu)/\sigma] - \Phi[(L - \mu)/\sigma]\}$  draws per observation, which could be substantial. A direct transformation that requires only one draw is as follows: Let  $P_j = \Phi[(j - \mu)/\sigma]$ ,  $j = L, U$ . Then

$$x = \mu + \sigma\Phi^{-1}[P_L + F \times (P_U - P_L)]. \quad (15-4)$$

### 15.2.4 SAMPLING FROM A MULTIVARIATE NORMAL POPULATION

A common application involves draws from a multivariate normal distribution with specified mean  $\mu$  and covariance matrix  $\Sigma$ . To sample from this  $K$ -variate distribution, we begin with a draw,  $\mathbf{z}$ , from the  $K$ -variate standard normal distribution. This is done by first computing  $K$  independent standard normal draws,  $z_1, \dots, z_K$  using the method of the previous section and stacking them in the vector  $\mathbf{z}$ . Let  $\mathbf{C}$  be a square root of  $\Sigma$  such that  $\mathbf{CC}' = \Sigma$ . The desired draw is then  $\mathbf{x} = \mu + \mathbf{C}\mathbf{z}$ , which will have covariance matrix  $E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)'] = \mathbf{CE}[\mathbf{zz}']\mathbf{C}' = \mathbf{C}\mathbf{I}\mathbf{C}' = \Sigma$ . For the square root matrix, the usual device is the **Cholesky decomposition**, in which  $\mathbf{C}$  is a lower triangular matrix. (See Section A.6.11.) For example, suppose we wish to sample from the bivariate normal distribution with mean vector  $\mu$ , unit variances and correlation coefficient  $\rho$ . Then,



$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \text{ and } \mathbf{C} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix}.$$

The transformation of two draws  $z_1$  and  $z_2$  is  $x_1 = \mu_1 + z_1$  and  $x_2 = \mu_2 + [\rho z_1 + (1 - \rho^2)^{1/2} z_2]$ . Section 15.3 and Example 15.4 following show a more involved application.

### 15.2.5 SAMPLING FROM DISCRETE POPULATIONS

There is generally no inverse transformation available for discrete distributions such as the **Poisson**. An inefficient, though usually unavoidable method for some distributions is to draw the  $F$  and then search sequentially for the smallest value that has cdf equal to or greater than  $F$ . For example, a generator for the Poisson distribution is constructed as follows. The pdf is  $\text{Prob}[x = j] = p_j = \exp(-\mu)\mu^j/j!$  where  $\mu$  is the mean of the random variable. The generator will use the recursion  $p_j = p_{j-1} \times \mu/j$ ,  $j = 1, \dots$  beginning with  $p_0 = \exp(-\mu)$ . An algorithm that requires only a single random draw is as follows:

Initialize	$c = \exp(-\mu); p = c; x = 0;$
Draw	$F$ from $U[0, 1];$
Deliver $x$	* exit with draw $x$ if $c > F$ ;
Iterate	$x = x + 1; p = p \times \mu/x; c = c + p;$ go to *.

This method is based explicitly on the pdf and cdf of the distribution. Other methods are suggested by Knuth (1969) and Press et al. (1986, pp. 203–209).

The most common application of random sampling from a discrete distribution is, fortunately, also the simplest. The method of bootstrapping, and countless other applications involve random samples of draws from the **discrete uniform distribution**,  $\text{Prob}(x = j) = 1/n$ ,  $j = 1, \dots, n$ . In the bootstrapping application, we are going to draw random samples of observations from the sequence of integers  $1, \dots, n$ , where each value must

CHAPTER 15 ♦ Simulation-Based Estimation and Inference **609**

be equally likely. In principle, the random draw could be obtained by partitioning the unit interval into  $n$  equal parts,  $[0, a_1], [a_1, a_2], \dots, [a_{n-2}, a_{n-1}], [a_{n-1}, 1]$ ;  $a_j = j/n$ ,  $j = 1, \dots, n - 1$ . Then, random draw  $F$  delivers  $x = j$  if  $F$  falls into interval  $j$ . This would entail a search, which could be time consuming. However, a simple method that will be much faster is simply to deliver  $x =$  the integer part of  $(n \times F + 1.0)$ . (Once again, we are making use of the practical result that  $F$  will equal exactly 1.0 (and  $x$  will equal  $n + 1$ ) with ignorable probability.)

### 15.3 SIMULATION-BASED STATISTICAL INFERENCE: THE METHOD OF KRINSKY AND ROBB

Most of the theoretical development in this text has concerned the statistical properties of estimators—that is, the characteristics of sampling distributions such as the mean (probability limits), variance (asymptotic variance), and quantiles (such as the boundaries for confidence intervals). In cases in which these properties cannot be derived explicitly, it is often possible to infer them by using random sampling methods to draw samples from the population that produced an estimator and deduce the characteristics from the features of such a random sample. In Example 4.4, we computed a set of least squares regression coefficients,  $b_1, \dots, b_K$ , and then examined the behavior of a nonlinear function  $c_k = b_k/(1 - b_m)$  using the **delta method**. In some cases, the asymptotic properties of nonlinear functions such as these are difficult to derive directly from the theoretical distribution of the parameters. The sampling methods described here can be used for that purpose. A second common application is learning about the behavior of test statistics. For example, at the end of Section 5.6 and in Section 14.9.1 [see (14-47)], we defined a Lagrange multiplier statistic for testing the hypothesis that certain coefficients are zero in a linear regression model. Under the assumption that the disturbances are normally distributed, the statistic has a limiting chi-squared distribution, which implies that the analyst knows what critical value to employ if they use this statistic. Whether the statistic has this distribution if the disturbances are not normally distributed is unknown. Monte Carlo methods can be helpful in determining if the guidance of the chi-squared result is useful in more general cases. Finally, in Section 14.7, we defined a two-step maximum likelihood estimator. Computation of the asymptotic variance of such an estimator can be challenging. Monte Carlo methods, in particular, bootstrapping methods, can be used as an effective substitute for the intractable derivation of the appropriate asymptotic distribution of an estimator. This and the next two sections will detail these three procedures and develop applications to illustrate their use.

The method of Krinsky and Robb is suggested as a way to estimate the asymptotic covariance matrix of  $\mathbf{c} = \mathbf{f}(\mathbf{b})$ , where  $\mathbf{b}$  is an estimated parameter vector with asymptotic covariance matrix  $\Sigma$  and  $\mathbf{f}(\mathbf{b})$  defines a set of possibly nonlinear functions of  $\mathbf{b}$ . We assume that  $\mathbf{f}(\mathbf{b})$  is a set of continuous and continuously differentiable functions that do not involve the sample size and whose derivatives do not equal zero at  $\boldsymbol{\beta} = \text{plim } \mathbf{b}$ . (These are the conditions underlying the Slutsky theorem in Section D.2.3.) In Section 4.4.4, we used the delta method to estimate the asymptotic covariance matrix of  $\mathbf{c}$ ;  $\text{Est. Asy. Var}[\mathbf{c}] = \mathbf{G}\mathbf{S}\mathbf{G}'$ , where  $\mathbf{S}$  is the estimate of  $\Sigma$  and  $\mathbf{G}$  is the matrix of partial derivatives,  $\mathbf{G} = \partial\mathbf{f}(\mathbf{b})/\partial\mathbf{b}'$ . The recent literature contains some occasional skepticism about the

## 610 PART III ♦ Estimation Methodology

accuracy of the delta method. The method of Krinsky and Robb (1986, 1990, 1991) is often suggested as an alternative. In a study of the behavior of estimated elasticities based on a translog model, the authors (1986) advocated an alternative approach based on Monte Carlo methods and the law of large numbers. We have consistently estimated  $\beta$  and  $(\sigma^2/n)\mathbf{Q}^{-1}$ , the mean and variance of the asymptotic normal distribution of the estimator  $\mathbf{b}$ , with  $\mathbf{b}$  and  $s^2(\mathbf{X}'\mathbf{X})^{-1}$ . It follows that we could estimate the mean and variance of the distribution of a function of  $\mathbf{b}$  by drawing a random sample of observations from the asymptotic normal population generating  $\mathbf{b}$ , and using the empirical mean and variance of the sample of functions to estimate the parameters of the distribution of the function. The quantiles of the sample of draws, for example, the .025th and .975th quantiles, can be used to estimate the boundaries of a confidence interval of the functions. The multivariate normal sample would be drawn using the method described in Section 15.2.4.

Krinsky and Robb (1986) reported huge differences in the standard errors produced by the delta method compared to the simulation-based estimator. In a subsequent paper (1990), they reported that the entire difference could be attributed to a bug in the software they used—upon redoing the computations, their estimates were essentially the same with the two methods. It is difficult to draw a conclusion about the effectiveness of the delta method based on the received results—it does seem at this juncture that the delta method remains an effective device that can often be employed with a hand calculator as opposed to the much more computation-intensive Krinsky and Robb (1986) technique. Unfortunately, the results of any comparison will depend on the data, the model, and the functions being computed. The amount of nonlinearity in the sense of the complexity of the functions seems not to be the answer. Krinsky and Robb's case was motivated by the extreme complexity of the elasticities in a translog model. In another study, Hole (2006) examines a similarly complex problem and finds that the delta method still appears to be the more accurate procedure.

### **Example 15.4 Long Run Elasticities**

A dynamic version of the demand for gasoline model is estimated in Example 4.4. The model is

$$\begin{aligned}\ln(G/\text{Pop})_t = & \beta_1 + \beta_2 \ln P_{G,t} + \beta_3 \ln(\text{Income}/\text{Pop})_t + \beta_4 \ln P_{nc,t} \\ & + \beta_5 \ln P_{uc,t} + \gamma \ln(G/\text{Pop})_{t-1} + \varepsilon_t.\end{aligned}$$

In this model, the short-run price and income elasticities are  $\beta_2$  and  $\beta_3$ . The long-run elasticities are  $\phi_2 = \beta_2/(1-\gamma)$  and  $\phi_3 = \beta_3/(1-\gamma)$ , respectively. To estimate the long-run elasticities, we estimated the parameters by least squares and then computed these two nonlinear functions of the estimates. Estimates of the full set of model parameters and the estimated asymptotic covariance matrix are given in Example 4.4. The delta method was used to estimate the asymptotic standard errors for the estimates of  $\phi_2$  and  $\phi_3$ . The three estimates of the specific parameters and the  $3 \times 3$  submatrix of the estimated asymptotic covariance matrix are

$$\begin{aligned}\text{Est.} \begin{pmatrix} \beta_2 \\ \beta_3 \\ \gamma \end{pmatrix} &= \begin{pmatrix} b_2 \\ b_3 \\ c \end{pmatrix} = \begin{pmatrix} -0.069532 \\ 0.164047 \\ 0.830971 \end{pmatrix}, \\ \text{Est. Asy. Var} \begin{pmatrix} b_2 \\ b_3 \\ c \end{pmatrix} &= \begin{pmatrix} 0.00021705 & 1.61265e-5 & -0.0001109 \\ 1.61265e-5 & 0.0030279 & -0.0021881 \\ -0.0001109 & -0.0021881 & 0.0020943 \end{pmatrix}.\end{aligned}$$

The method suggested by Krinsky and Robb would use a random number generator to draw a large trivariate sample,  $(b_2, b_3, c)_r, r = 1, \dots, R$ , from the normal distribution with this mean vector and covariance matrix, and then compute the sample of observations on  $f_2$  and  $f_3$  and obtain the empirical mean and variance and the .025 and .975 quantiles from the sample. The method of drawing such a sample is shown in Section 15.2.4. We will require the square

CHAPTER 15 ♦ Simulation-Based Estimation and Inference **611****TABLE 15.1** Simulation Results

	<i>Regression Estimate</i>		<i>Simulated Values</i>	
	<i>Estimate</i>	<i>Std.Error</i>	<i>Mean</i>	<i>Std.Dev.</i>
$\beta_2$	-0.069532	0.0147327	-0.068791	0.0138485
$\beta_3$	0.164047	0.0550265	0.162634	0.0558856
$\gamma$	0.830971	0.0457635	0.831083	0.0460514
$\phi_2$	-0.411358	0.152296	-0.453815	0.219110
$\phi_3$	0.970522	0.162386	0.950042	0.199458

**TABLE 15.2** Estimated Confidence Intervals

	$\phi_2$		$\phi_3$	
	<i>Lower</i>	<i>Upper</i>	<i>Lower</i>	<i>Upper</i>
Delta Method	-0.718098	-0.104618	0.643460	1.297585
Krinsky and Robb	-0.895125	-0.012505	0.548313	1.351772
Sample Quantiles	-0.983866	-0.209776	0.539668	1.321617

root of the covariance matrix. The Cholesky matrix is

$$\mathbf{C} = \begin{pmatrix} 0.0147326 & 0 & 0 \\ 0.00109461 & 0.0550155 & 0 \\ -0.0075275 & -0.0396227 & 0.0216259 \end{pmatrix}$$

The sample is drawn by obtaining vectors of three random draws from the standard normal population,  $\mathbf{v}_r = (v_1, v_2, v_3)', r = 1, \dots, R$ . The draws needed for the estimation are then obtained by computing  $\mathbf{b}_r = \mathbf{b} + \mathbf{C}\mathbf{v}_r$ , where  $\mathbf{b}$  is the set of least squares estimates. We then compute the sample of estimated long-run elasticities,  $f_{2r} = b_{2r}/(1 - c_r)$  and  $f_{3r} = b_{3r}/(1 - c_r)$ . The mean and variance of the sample observations constitute the estimates of the functions and asymptotic standard errors.

Table 15.1 shows the results of these computations based on 1,000 draws from the underlying distribution. The estimates from Example 4.4 using the delta method are shown as well. The two sets of estimates are in quite reasonable agreement. A 95 percent confidence interval for  $\phi_2$  based on the estimates, the  $t$  distribution with  $51 - 6 = 45$  degrees of freedom and the delta method would be  $-0.411358 \pm 2.014103(0.152296)$ . The result for  $\phi_3$  would be  $0.970522 \pm 2.014103(0.162386)$ . These are shown in Table 15.2 with the same computation using the Krinsky and Robb estimated standard errors. The table also shows the empirical estimates of these quantiles computed using the 26th and 975th values in the samples. There is reasonable agreement in the estimates, though there is also evident a considerable amount of sample variability, even in a sample as large as 1,000.

We note, finally, that it is generally not possible to replicate results such as these across software platforms, because they use different random number generators. Within a given platform, replicability can be obtained by setting the seed for the random number generator.

## 15.4 BOOTSTRAPPING STANDARD ERRORS AND CONFIDENCE INTERVALS

The technique of **bootstrapping** is used to obtain a description of the sampling properties of empirical estimators using the sample data themselves, rather than broad theoretical results.<sup>2</sup> Suppose that  $\hat{\theta}_n$  is an estimator of a parameter vector  $\theta$  based on a sample

<sup>2</sup>See Efron (1979), Efron and Tibshirani (1994), and Davidson and Hinkley (1997), Brownstone and Kazimi (1998), Horowitz (2001), and MacKinnon (2002).

## 612 PART III ♦ Estimation Methodology

$\mathbf{Z} = [(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)]$ . An approximation to the statistical properties of  $\hat{\theta}_n$  can be obtained by studying a sample of bootstrap estimators  $\hat{\theta}(b)_m, b=1, \dots, B$ , obtained by sampling  $m$  observations, *with replacement*, from  $\mathbf{Z}$  and recomputing  $\hat{\theta}$  with each sample. After a total of  $B$  times, the desired sampling characteristic is computed from

$$\hat{\Theta} = [\hat{\theta}(1)_m, \hat{\theta}(2)_m, \dots, \hat{\theta}(B)_m].$$

The most common application of bootstrapping for consistent estimators when  $n$  is reasonably large is approximating the asymptotic covariance matrix of the estimator  $\hat{\theta}_n$  with

$$Est.Asy.Var[\hat{\theta}_n] = \frac{1}{B-1} \sum_{b=1}^B [\hat{\theta}(b)_m - \bar{\hat{\theta}}_B] [\hat{\theta}(b)_m - \bar{\hat{\theta}}_B]', \quad (15-5)$$

where  $\bar{\hat{\theta}}_B$  is the average of the  $B$  bootstrapped estimates of  $\theta$ . There are few theoretical prescriptions for the number of replications,  $B$ . Andrews and Buchinsky (2000) and Cameron and Trivedi (2005, pp. 361–362) make some suggestions for particular applications; Davidson and MacKinnon (2000) recommend at least 399. Several hundred is the norm; we have used 1,000 in our application to follow. This technique was developed by Efron (1979) and has been appearing with increasing frequency in the applied econometrics literature. [See, for example, Veall (1987, 1992), Vinod (1993), and Vinod and Raj (1994). Extensive surveys of uses and methods in econometrics appear in Cameron and Trivedi (2005), Horowitz (2001), and Davidson and MacKinnon (2006).] An application of this technique to the least absolute deviations estimator in the linear model is shown in the following example and in Chapter 4.

The preceding is known as a **paired bootstrap**. The pairing is the joint sampling of  $y_i$  and  $\mathbf{x}_i$ . An alternative approach in a regression context would be to sample the observations on  $\mathbf{x}_i$  only and then with each  $\mathbf{x}_i$  sampled, generate the accompanying  $y_i$  by randomly generating the disturbance, then  $\hat{y}_i(b) = \mathbf{x}_i(b)' \hat{\theta}_n + \hat{\varepsilon}_i(b)$ . This would be a **parametric bootstrap** in that in order to simulate the disturbances, we need either to know (or assume) the data generating process that produces  $\varepsilon_i$ . In other contexts, such as in discrete choice modeling in Chapter 17, one would bootstrap sample the exogenous data in the model and then generate the dependent variable by this method using the appropriate underlying DGP. This is the approach used in Example 15.7 and in Greene (2004b) in a study of the incidental parameters problem in several limited dependent variable models. The obvious disadvantage of the parametric bootstrap is that one cannot learn of the influence of an unknown DGP for  $\varepsilon$  by assuming it is known. For example, if the bootstrap is being used to accommodate unknown heteroscedasticity in the model, a parametric bootstrap that assumes homoscedasticity would defeat the purpose. The more natural application would be a **nonparametric-bootstrap**, in which both  $\mathbf{x}_i$  and  $y_i$ , and, implicitly,  $\varepsilon_i$ , are sampled simultaneously.

### Example 15.5 Bootstrapping the Variance of the Median

There are few cases in which an exact expression for the sampling variance of the median is known. Example 15.7, examines the case of the median of a sample of 500 observations from the  $t$ -distribution with 10 degrees of freedom. This is one of those cases in which there is no exact formula for the asymptotic variance of the median. However, we can use the bootstrap technique to estimate one empirically. In one run of the experiment, we obtained

CHAPTER 15 ♦ Simulation-Based Estimation and Inference **613**

a sample of 500 observations for which we computed the median,  $-0.00786$ . We drew 100 samples of 500 with replacement from this sample of 500 and recomputed the median with each of these samples. The empirical square root of the mean squared deviation around this estimate of  $-0.00786$  was 0.056. In contrast, consider the same calculation for the mean. The sample mean is  $-0.07247$ . The sample standard deviation is 1.08469, so the standard error of the mean is 0.04657. (The bootstrap estimate of the standard error of the mean was 0.052.) This agrees with our expectation in that the sample mean should generally be a more efficient estimator of the mean of the distribution in a large sample. There is another approach we might take in this situation. Consider the regression model

$$y_i = \alpha + \varepsilon_i,$$

where  $\varepsilon_i$  has a symmetric distribution with finite variance. The least absolute deviations estimator of the coefficient in this model is an estimator of the median (which equals the mean) of the distribution. So, this presents another estimator. Once again, the bootstrap estimator must be used to estimate the asymptotic variance of the estimator. Using the same data, we fit this regression model using the LAD estimator. The coefficient estimate is  $-.05397$  with a bootstrap estimated standard error of 0.05872. The estimated standard error agrees with the earlier one. The difference in the estimated coefficient stems from the different computations—the regression estimate is the solution to a linear programming problem while the earlier estimate is the actual sample median.

The bootstrap estimation procedure has also been suggested as a method of reducing bias. In principle, we would compute  $\hat{\theta}_n - \text{bias}(\hat{\theta}_n) = \hat{\theta}_n - \{E[\hat{\theta}_n] - \theta\}$ . Since neither  $\theta$  nor the exact expectation of  $\hat{\theta}_n$  is known, we estimate the first with the mean of the bootstrap replications and the second with the estimator, itself. The revised estimator is

$$\hat{\theta}_{n,B} = \hat{\theta}_n - \left[ \frac{1}{B} \sum_{b=1}^B \hat{\theta}_n \right] = 2\hat{\theta}_n - \bar{\hat{\theta}}_B. \quad (15-6)$$

(Efron and Tibshirani (1994, p. 138) provide justification for what appears to be the wrong sign on the correction.) Davidson and MacKinnon (2006) argue that the smaller bias of the corrected estimator is offset by an increased variance compared to the uncorrected estimator. [See, as well, Cameron and Trivedi (2005).] The authors offer some other cautions for practitioners contemplating use of this technique. First, perhaps obviously, the extension of the method to samples with dependent observations presents some obstacles. For time-series data, the technique makes little sense—none of the bootstrapped samples will be a time series, so the properties of the resulting estimators will not satisfy the underlying assumptions needed to make the technique appropriate.

A second common application of bootstrapping methods is the computation of confidence intervals for parameters. This calculation will be useful when the underlying data generating process is unknown, and the bootstrap method is being used to obtain appropriate standard errors for estimated parameters. A natural approach to bootstrapping confidence intervals for parameters would be to compute the estimated asymptotic covariance matrix using (15-5) and then form confidence intervals in the usual fashion. An improvement in terms of the bias of the estimator is provided by the **percentile method** [Cameron and Trivedi (2005, p. 364)]. By this technique, during each bootstrap replication, we compute

$$t_k^*(b) = \frac{\hat{\theta}_k(b) - \hat{\theta}_{n,k}}{se.(\hat{\theta}_{n,k})}, \quad (15-7)$$

## 614 PART III ♦ Estimation Methodology

where “ $k$ ” indicates the  $k$ th parameter in the model, and  $\hat{\theta}_{n,k}$ ,  $s.e.(\hat{\theta}_{n,k})$  and  $\hat{\theta}_k(b)$  are the original estimator and estimated standard error from the full sample and the bootstrap replicate. Then, with all  $B$  replicates in hand, the bootstrap confidence interval is

$$\hat{\theta}_{n,k} + t_k^*[\alpha/2]s.e.(\hat{\theta}_{n,k}) \text{ to } \hat{\theta}_{n,k} + t_k^*[1 - \alpha/2]s.e.(\hat{\theta}_{n,k}). \quad (15-8)$$

(Note that  $t_k^*[\alpha/2]$  is negative, which explains the plus sign in left term.) For example, in our application, next, we compute the estimator and the asymptotic covariance matrix using the full sample. We compute 1,000 bootstrap replications, and compute the  $t$  ratio in (15-7) for the education coefficient in each of the 1,000 replicates. After the bootstrap samples are accumulated, we sorted the results from (15-7), and the 25th and 975th largest values provide the values of  $t^*$ .

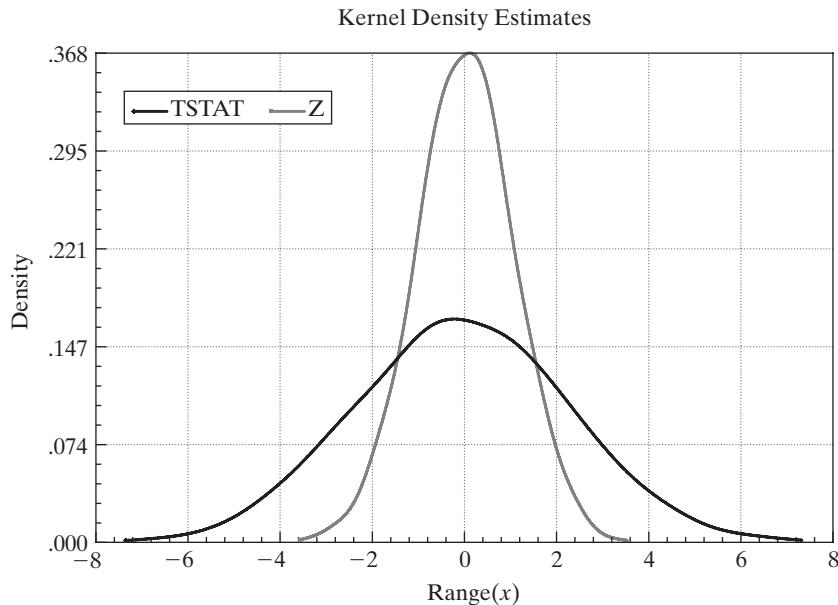
Example 15.6 demonstrates the computation of a confidence interval for a coefficient using the bootstrap. The application uses the Cornwell and Rupert panel data set used in Example 11.1 and several later applications. There are 595 groups of seven observations in the data set. Bootstrapping with panel data requires an additional element in the computations. The bootstrap replications are based on sampling over  $i$ , not  $t$ . Thus, the bootstrap sample consists of  $n$  blocks of  $T$  (or  $T_i$ ) observations—the  $i$ th group as a whole is sampled. This produces, then, a **block bootstrap** sample.

### Example 15.6 Bootstrapping Standard Errors and Confidence Intervals in a Panel

Example 11.1 presents least squares estimates and robust standard errors for the labor supply equation using Cornwell and Rupert's panel data set. There are 595 individuals and seven periods in the data set. As seen in the results in Table 11.1 (reproduced below), using a clustering correction in a robust covariance matrix for the least squares estimator produces substantial changes in the estimated standard errors. Table 15.3 reproduces the least squares coefficients and the standard errors associated with the conventional  $s^2(\mathbf{X}'\mathbf{X})^{-1}$  and the robust standard errors using the clustering correction, and presents the bootstrapped standard errors using 1,000 bootstrap replications. The resemblance between the original estimates in the leftmost column and the average of the bootstrap replications in the rightmost column is to be expected; the sample is quite large and the number of replications is large. What is striking (and reassuring) is the ability of the bootstrapping procedure to detect and mimic the effect of the clustering that is evident in the second and third columns of estimated standard errors.

**TABLE 15.3** Bootstrap Estimates of Standard Errors for a Wage Equation

Variable	Least Squares Estimate	Standard Error	Cluster Robust Std. Error	Bootstrap Std. Error	Bootstrap Coefficient
Constant	5.25112	0.07129	0.1233	0.12421	5.25907
Wks	0.00422	0.00108	0.001538	0.00159	0.00409
South	-0.05564	0.01253	0.02610	0.02557	-0.05417
SMSA	0.15167	0.01207	0.02405	0.02383	0.15140
MS	0.04845	0.02057	0.04085	0.04208	0.04676
Exp	0.04010	0.00216	0.004067	0.00418	0.04017
Exp <sup>2</sup>	-0.00067	0.00004744	0.00009111	0.00009235	-0.00067
Occ	-0.14001	0.01466	0.02718	0.02733	-0.13912
Ind	0.04679	0.01179	0.02361	0.02350	0.04728
Union	0.09263	0.01280	0.02362	0.02390	0.09126
Ed	0.05670	0.00261	0.005552	0.00576	0.05656
Fem	-0.36779	0.02510	0.04547	0.04562	-0.36855
Blk	-0.16694	0.02204	0.04423	0.04663	-0.16811

CHAPTER 15 ♦ Simulation-Based Estimation and Inference **615****FIGURE 15.1** Distributions of Test Statistics.

We also computed a confidence interval for the coefficient on  $Ed$  using the conventional, symmetric approach,  $b_{Ed} \pm 1.96s(b_{Ed})$ , and the percentile method in (15-7)–(15-8). The two intervals are

$$\begin{aligned} \text{Conventional: } & 0.051583 \text{ to } 0.061825 \\ \text{Percentile: } & 0.045560 \text{ to } 0.067909 \end{aligned}$$

Not surprisingly (given the larger standard errors), the percentile method gives a much wider interval. Figure 15.1 shows a kernel density estimator of the distribution of the  $t$  statistics computed using (15-7). It is substantially wider than the (approximate) standard normal density shown with it. This demonstrates the impact of the latent effect of the clustering on the standard errors, and ultimately on the test statistic used to compute the confidence intervals.

## 15.5 MONTE CARLO STUDIES

Simulated data generated by the methods of the preceding sections have various uses in econometrics. One of the more common applications is the analysis of the properties of estimators or in obtaining comparisons of the properties of estimators. For example, in time-series settings, most of the known results for characterizing the sampling distributions of estimators are asymptotic, large-sample results. But the typical time series is not very long, and descriptions that rely on  $T$ , the number of observations, going to infinity may not be very accurate. Exact finite-sample properties are usually intractable, however, which leaves the analyst with only the choice of learning about the behavior of the estimators experimentally.

In the typical application, one would either compare the properties of two or more estimators while holding the sampling conditions fixed or study how the properties of an estimator are affected by changing conditions such as the sample size or the value of an underlying parameter.

## 616 PART III ♦ Estimation Methodology

### **Example 15.7 Monte Carlo Study of the Mean Versus the Median**

In Example D.8, we compared the asymptotic distributions of the sample mean and the sample median in random sampling from the normal distribution. The basic result is that both estimators are consistent, but the mean is asymptotically more efficient by a factor of

$$\frac{\text{Asy. Var[Median]}}{\text{Asy. Var[Mean]}} = \frac{\pi}{2} = 1.5708.$$

This result is useful, but it does not tell which is the better estimator in small samples, nor does it suggest how the estimators would behave in some other distribution. It is known that the mean is affected by outlying observations whereas the median is not. The effect is averaged out in large samples, but the small-sample behavior might be very different. To investigate the issue, we constructed the following experiment: We sampled 500 observations from the  $t$  distribution with  $d$  degrees of freedom by sampling  $d + 1$  values from the standard normal distribution and then computing

$$t_{ir} = \frac{z_{ir,d+1}}{\sqrt{\frac{1}{d} \sum_{l=1}^d z_{ir,l}^2}}, \quad i = 1, \dots, 500, \quad r = 1, \dots, 100.$$

The  $t$  distribution with a low value of  $d$  was chosen because it has very thick tails and because large outlying values have high probability. For each value of  $d$ , we generated  $R = 100$  replications. For each of the 100 replications, we obtained the mean and median. Because both are unbiased, we compared the mean squared errors around the true expectations using

$$M_d = \frac{(1/R) \sum_{r=1}^R (\text{median}_r - 0)^2}{(1/R) \sum_{r=1}^R (\bar{x}_r - 0)^2}.$$

We obtained ratios of 0.6761, 1.2779, and 1.3765 for  $d = 3, 6$ , and 10, respectively. (You might want to repeat this experiment with different degrees of freedom.) These results agree with what intuition would suggest. As the degrees of freedom parameter increases, which brings the distribution closer to the normal distribution, the sample mean becomes more efficient—the ratio should approach its limiting value of 1.5708 as  $d$  increases. What might be surprising is the apparent overwhelming advantage of the median when the distribution is very nonnormal even in a sample as large as 500.

The preceding is a very small application of the technique. In a typical study, there are many more parameters to be varied and more dimensions upon which the results are to be studied. One of the practical problems in this setting is how to organize the results. There is a tendency in Monte Carlo work to proliferate tables indiscriminately. It is incumbent on the analyst to collect the results in a fashion that is useful to the reader. For example, this requires some judgment on how finely one should vary the parameters of interest. One useful possibility that will often mimic the thought process of the reader is to collect the results of bivariate tables in carefully designed contour plots.

There are any number of situations in which Monte Carlo simulation offers the only method of learning about finite-sample properties of estimators. Still, there are a number of problems with Monte Carlo studies. To achieve any level of generality, the number of parameters that must be varied and hence the amount of information that must be distilled can become enormous. Second, they are limited by the design of the experiments, so the results they produce are rarely generalizable. For our example, we may have learned something about the  $t$  distribution, but the results that would apply in other distributions remain to be described. And, unfortunately, real data will rarely conform to any specific distribution, so no matter how many other distributions we

## CHAPTER 15 ♦ Simulation-Based Estimation and Inference **617**

analyze, our results would still only be suggestive. In more general terms, this problem of **specificity** [Hendry (1984)] limits most Monte Carlo studies to quite narrow ranges of applicability. There are very few that have proved general enough to have provided a widely cited result.<sup>3</sup>

### 15.5.1 A MONTE CARLO STUDY: BEHAVIOR OF A TEST STATISTIC

Monte Carlo methods are often used to study the behavior of test statistics when their true properties are uncertain. This is often the case with Lagrange multiplier statistics. For example, Baltagi (2005) reports on the development of several new test statistics for panel data models such as a test for serial correlation. Examining the behavior of a test statistic is fairly straightforward. We are interested in two characteristics: the true **size of the test**—that is, the probability that it rejects the null hypothesis when that hypothesis is actually true (the probability of a type 1 error) and the **power of the test**—that is the probability that it will correctly reject a false null hypothesis (one minus the probability of a type 2 error). As we will see, the power of a test is a function of the alternative against which the null is tested.

To illustrate a Monte Carlo study of a test statistic, we consider how a familiar procedure behaves when the model assumptions are incorrect. Consider the linear regression model

$$y_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i, \quad \varepsilon_i | (x_i, z_i) \sim N[0, \sigma^2].$$

The Lagrange multiplier statistic for testing the null hypothesis that  $\gamma$  equals zero for this model is

$$LM = \mathbf{e}'_0 \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{e}_0 / (\mathbf{e}'_0 \mathbf{e}_0 / n)$$

where  $\mathbf{X} = (\mathbf{1}, \mathbf{x}, \mathbf{z})$  and  $\mathbf{e}_0$  is the vector of least squares residuals obtained from the regression of  $\mathbf{y}$  on the constant and  $\mathbf{x}$  (and not  $\mathbf{z}$ ). (See Section 14.6.3.) Under the assumptions of the preceding model, above, the large sample distribution of the LM statistic is chi-squared with one degree of freedom. Thus, our testing procedure is to compute LM and then reject the null hypothesis  $\gamma = 0$  if LM is greater than the critical value. We will use a nominal size of 0.05, so the critical value is 3.84. The theory for the statistic is well developed when the specification of the model is correct. [See, for example, Godfrey (1988).] We are interested in two specification errors. First, how does the statistic behave if the normality assumption is not met? Because the LM statistic is based on the likelihood function, if some distribution other than the normal governs  $\varepsilon_i$ , then the LM statistic would not be based on the OLS estimator. We will examine the behavior of the statistic under the true specification that  $\varepsilon_i$  comes from a  $t$  distribution with five degrees of freedom. Second, how does the statistic behave if the homoscedasticity assumption is not met? The statistic is entirely wrong if the disturbances are heteroscedastic. We will examine the case in which the conditional variance is  $\text{Var}[\varepsilon_i | (x_i, z_i)] = \sigma^2 [\exp(0.2x_i)]^2$ .

The design of the experiment is as follows: We will base the analysis on a sample of 50 observations. We draw 50 observations on  $x_i$  and  $z_i$  from independent  $N[0, 1]$  populations at the outset of each cycle. For each of 1,000 replications, we draw a sample of 50  $\varepsilon_i$ 's according to the assumed specification. The LM statistic is computed and the

---

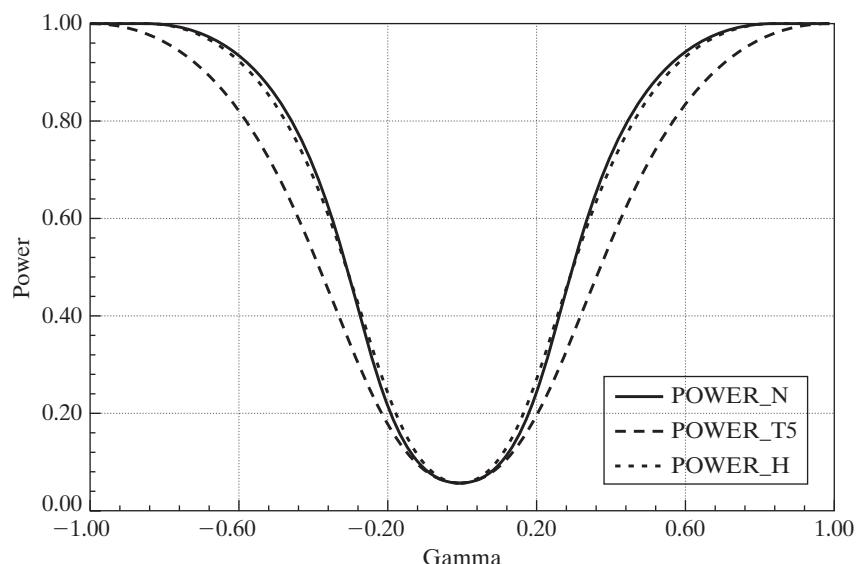
<sup>3</sup>Two that have withstood the test of time are Griliches and Rao (1969) and Kmenta and Gilbert (1968).

**618 PART III ♦ Estimation Methodology**
**TABLE 15.4** Size and Power Functions for LM Test

<i>Model</i>	0.5 -0.1	<i>Gamma</i>									
		0.1 -0.1	0.2 -0.2	0.3 -0.3	0.4 -0.4	0.5 -0.5	0.6 -0.6	0.7 -0.7	0.8 -0.8	0.9 -0.9	1.0 -1.0
Normal	0.059	0.090	0.235	0.464	0.691	0.859	0.957	0.989	0.998	1.000	1.000
		0.103	0.236	0.451	0.686	0.863	0.961	0.989	0.999	1.000	1.000
<i>t</i> (5)	0.052	0.083	0.169	0.320	0.508	0.680	0.816	0.911	0.956	0.976	0.994
		0.080	0.177	0.312	0.500	0.677	0.822	0.921	0.953	0.984	0.993
Het.	0.071	0.098	0.249	0.457	0.666	0.835	0.944	0.984	0.995	0.998	1.000
		0.107	0.239	0.442	0.651	0.832	0.940	0.985	0.996	1.000	1.000

proportion of the computed statistics that exceed 3.84 is recorded. The experiment is repeated for  $\gamma = 0$  to ascertain the true size of the test and for values of  $\gamma$  including  $-1, \dots, -0.2, -0.1, 0, 0.1, 0.2, \dots, 1.0$  to assess the power of the test. The cycle of tests is repeated for the two scenarios, the  $t(5)$  distribution and the model with heteroscedasticity.

Table 15.4 lists the results of the experiment. The first row shows the expected results for the LM statistic under the model assumptions for which it is appropriate. The size of the test appears to be in line with the theoretical results. Comparing the first and third rows, it appears that the presence of heteroscedasticity seems not to degrade the power of the statistic. But the different distributional assumption does. Figure 15.2 plots the values in the table, and displays the characteristic form of the power function for a test statistic.

**FIGURE 15.2** Power Functions.


CHAPTER 15 ♦ Simulation-Based Estimation and Inference **619****15.5.2 A MONTE CARLO STUDY: THE INCIDENTAL PARAMETERS PROBLEM**

Section 14.9.6.d examines the maximum likelihood estimator of a panel data model with fixed effects,

$$f(y_{it} | \mathbf{x}_{it}) = g(y_{it}, \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i, \boldsymbol{\theta})$$

where the individual effects may be correlated with  $x_{it}$ . The extra parameter vector  $\boldsymbol{\theta}$  represents  $M$  other parameters that might appear in the model, such as the disturbance variance,  $\sigma_e^2$ , in a linear regression model with normally distributed disturbance. The development there considers the mechanical problem of maximizing the log-likelihood

$$\ln L = \sum_{i=1}^n \sum_{t=1}^{T_i} \ln g(y_{it}, \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i, \boldsymbol{\theta})$$

with respect to the  $n + K + M$  parameters  $(\alpha_1, \dots, \alpha_n, \boldsymbol{\beta}, \boldsymbol{\theta})$ . A statistical problem with this estimator that was suggested there is a phenomenon labeled the **incidental parameters problem** [see Neyman and Scott (1948), Lancaster (2000)]. With the exception of a very small number of specific models (such as the Poisson regression model in Section 19.3.2), the “brute force,” unconditional maximum likelihood estimator of the parameters in this model is inconsistent. The result is straightforward to visualize with respect to the individual effects. Suppose that  $\boldsymbol{\beta}$  and  $\boldsymbol{\theta}$  were actually known. Then, each  $\alpha_i$  would be estimated with  $T_i$  observations. Because  $T_i$  is assumed to be fixed (and small), there is no asymptotic result to provide consistency for the MLE of  $\alpha_i$ . But,  $\boldsymbol{\beta}$  and  $\boldsymbol{\theta}$  are estimated with  $\sum_i T_i = N$  observations, so their large sample behavior is less transparent. One known result concerns the logit model for binary choice (see Section 17.2–17.5). Kalbfleisch and Sprott (1970), Andersen (1973), Hsiao (1996), and Abrevaya (1997) have established that in the binary logit model, if  $T_i = 2$ , then  $\text{plim } \hat{\boldsymbol{\beta}}_{\text{MLE}} = 2\boldsymbol{\beta}$ . Two other cases are known with certainty. In the linear regression model with fixed effects and normally distributed disturbances, the slope estimator,  $\mathbf{b}_{\text{LSDV}}$  is unbiased and consistent, however, the MLE of the variance,  $\sigma^2$  converges to  $(T - 1)\sigma^2 / T$ . (The degrees of freedom correction will adjust for this, but the MLE does not correct for degrees of freedom.) Finally, in the Poisson regression model (Section 19.3.2), the unconditional MLE is consistent [see Cameron and Trivedi (1988)]. Almost nothing else is known with certainty—that is, as a firm theoretical result—about the behavior of the maximum likelihood estimator in the presence of fixed effects. The literature appears to take as given the qualitative wisdom of Hsiao and Abrevaya, that the FE/MLE is inconsistent when  $T$  is small and fixed. (The implication that the severity of the inconsistency declines as  $T$  increases makes sense, but, again, remains to be shown analytically.)

The result for the two-period binary logit model is a standard result for discrete choice estimation. Several authors, all using Monte Carlo methods have pursued the result for the logit model for larger values of  $T$ . [See, for example, Katz (2001).] Greene (2004) analyzed the incidental parameters problem for other discrete choice models using Monte Carlo methods. We will examine part of that study.

The current studies are preceded by a small study in Heckman (1981) which examined the behavior of the fixed effects MLE in the following experiment:

$$\begin{aligned} z_{it} &= 0.1t + 0.5z_{i,t-1} + u_{it}, z_{i0} = 5 + 10.0u_{i0}, \\ u_{it} &\sim U[-0.5, 0.5], i = 1, \dots, 100, t = 0, \dots, 8, \\ Y_{it} &= \sigma_\tau \tau_i + \beta z_{it} + \varepsilon_{it}, \tau_i \sim N[0, 1], \varepsilon_{it} \sim N[0, 1], \\ y_{it} &= 1 \text{ if } Y_{it} > 0, 0 \text{ otherwise}. \end{aligned}$$

## 620 PART III ♦ Estimation Methodology

Heckman attempted to learn something about the behavior of the MLE for the probit model with  $T = 8$ . He used values of  $\beta = -1.0, -0.1$ , and  $1.0$  and  $\sigma_\tau = 0.5, 1.0$ , and  $3.0$ . The mean values of the maximum likelihood estimates of  $\beta$  for the nine cases are as follows:

	$\beta = -1.0$	$\beta = -0.1$	$\beta = 1.0$
$\sigma_\tau = 0.5$	-0.96	-0.10	0.93
$\sigma_\tau = 1.0$	-0.95	-0.09	0.91
$\sigma_\tau = 3.0$	-0.96	-0.10	0.90

The findings here disagree with the received wisdom. Where there appears to be a bias (that is, excluding the center column), it seems to be quite small, and toward, not away from zero.

The Heckman study used a very small sample and, moreover, analyzed the fixed effects estimator in a random effects model (note that  $\tau_i$  is independent of  $z_{it}$ ). Greene (2004a), using the same parameter values, number of replications, and sample design, found persistent biases away from zero on the order of 15–20 percent. Numerous authors have extended the logit result for  $T = 2$  with larger values of  $T$ , and likewise persistently found biases, away from zero that diminish with increases in  $T$ . Greene (2004a) redid the experiment for the logit model and then replicated it for the probit and ordered probit models. The experiment is designed as follows: All models are based on the same index function

$$\begin{aligned} w_{it} &= \alpha_i + \beta x_{it} + \delta d_{it}, \quad \text{where } \beta = \delta = 1, \\ x_{it} &\sim N[0, 1], d_{it} = \mathbf{1}[x_{it} + h_{it} > 0], \quad \text{where } h_{it} \sim N[0, 1], \\ \alpha_i &= \sqrt{T} \bar{x}_i + v_i, v_i \sim N[0, 1]. \end{aligned}$$

The regressors  $d_{it}$  and  $x_{it}$  are constructed to be correlated. The random term  $h_{it}$  is used to produce independent variation in  $d_{it}$ . There is, however, no within group correlation in  $x_{it}$  or  $d_{it}$  built into the data generator. (Other experiments suggested that the marginal distribution of  $x_{it}$  mattered little to the outcome of the experiment.) The correlations between the variables are approximately 0.7 between  $x_{it}$  and  $d_{it}$ , 0.4 between  $\alpha_i$  and  $x_{it}$ , and 0.2 between  $\alpha_i$  and  $d_{it}$ . The individual effect is produced from independent variation,  $v_i$  as well as the group mean of  $x_{it}$ . The latter is scaled by  $\sqrt{T}$  to maintain the unit variances of the two parts—without the scaling, the covariance between  $\alpha_i$  and  $x_{it}$  falls to zero as  $T$  increases and  $\bar{x}_i$  converges to its mean of zero). Thus, the data generator for the index function satisfies the assumptions of the fixed effects model. The sample used for the results below contains  $n = 1,000$  individuals. The data generating processes for the discrete dependent variables are as follows:

$$\begin{aligned} \text{probit:} \quad y_{it} &= \mathbf{1}[w_{it} + \varepsilon_{it} > 0], \varepsilon_{it} \sim N[0, 1], \\ \text{ordered probit:} \quad y_{it} &= \mathbf{1}[w_{it} + \varepsilon_{it} > 0] + \mathbf{1}[w_{it} + \varepsilon_{it} > 3], \varepsilon_{it} \sim N[0, 1], \\ \text{logit:} \quad y_{it} &= \mathbf{1}[w_{it} + v_{it} > 0], v_{it} = \log[u_{it}/(1 - u_{it})], u_{it} \sim U[0, 1]. \end{aligned}$$

(The three discrete dependent variables are described in Chapter 17.)

Table 15.5 reports the results of computing the MLE with 200 replications. Models were fit with  $T = 2, 3, 5, 8, 10$ , and  $20$ . (Note that this includes Heckman's experiment.) Each model specification and group size ( $T$ ) is fit 200 times with random draws for  $\varepsilon_{it}$  or  $u_{it}$ . The data on the regressors were drawn at the beginning of each experiment (that

CHAPTER 15 ♦ Simulation-Based Estimation and Inference **621****TABLE 15.5** Means of Empirical Sampling Distributions,  $N = 1,000$  Individuals Based on 200 Replications

	<b><math>T = 2</math></b>		<b><math>T = 3</math></b>		<b><math>T = 5</math></b>		<b><math>T = 8</math></b>		<b><math>T = 10</math></b>		<b><math>T = 20</math></b>	
	$\beta$	$\delta$	$\beta$	$\delta$	$\beta$	$\delta$	$\beta$	$\delta$	$\beta$	$\delta$	$\beta$	$\delta$
Logit Coeff	2.020	2.027	1.698	1.668	1.379	1.323	1.217	1.156	1.161	1.135	1.069	1.062
Logit M.E. <sup>a</sup>	1.676	1.660	1.523	1.477	1.319	1.254	1.191	1.128	1.140	1.111	1.034	1.052
Probit Coeff	2.083	1.938	1.821	1.777	1.589	1.407	1.328	1.243	1.247	1.169	1.108	1.068
Probit M.E. <sup>a</sup>	1.474	1.388	1.392	1.354	1.406	1.231	1.241	1.152	1.190	1.110	1.088	1.047
Ord. Probit	2.328	2.605	1.592	1.806	1.305	1.415	1.166	1.220	1.131	1.158	1.058	1.068

<sup>a</sup>Average ratio of estimated marginal effect to true marginal effect.

is, for each  $T$ ) and held constant for the replications. The table contains the average estimate of the coefficient and, for the binary choice models, the partial effects. The value at the extreme left corresponds to the received result, the 100 percent bias in the  $T = 2$  case. The remaining values show, as intuition would suggest, that the bias decreases with increasing  $T$ . The benchmark case of  $T = 8$ , appears to be less benign than Heckman's results suggested. One encouraging finding for the model builder is that the biases in the estimated marginal effects appears to be somewhat less than for the coefficients. Greene (2004b) extends this analysis to some other models, including the tobit and truncated regression models discussed in Chapter 14. The results there suggest that the conventional wisdom for the tobit model may not be correct—the incidental parameters problem seems to appear in the estimator of  $\sigma^2$  in the tobit model, not in the estimators of the slopes. This is consistent with the linear regression model, but not with the binary choice models.

## 15.6 SIMULATION-BASED ESTIMATION

Sections 15.3–15.5 developed a set of tools for inference about model parameters using simulation methods. This section will describe methods for using simulation as part of the estimation process. The modeling framework arises when integrals that cannot be computed directly appear in the estimation criterion function (sum of squares, log-likelihood, and so on). To illustrate, and begin the development, in Section 15.6.1, we will construct a nonlinear model with random effects. Section 15.6.2 will describe how simulation is used to evaluate integrals for maximum likelihood estimation. Section 15.6.3 will develop an application, the random effects regression model.

### 15.6.1 RANDOM EFFECTS IN A NONLINEAR MODEL

In Example 11.16, we considered a nonlinear regression model for the number of doctor visits in the German Socioeconomic Panel. The basic form of the nonlinear regression model is

$$E[y_{it} | \mathbf{x}_{it}] = \exp(\mathbf{x}'_{it}\boldsymbol{\beta}), t = 1, \dots, T_i, i = 1, \dots, n.$$

In order to accommodate unobserved heterogeneity in the panel data, we extended the model to include a random effect,

$$E[y_{it} | \mathbf{x}_{it}, u_i] = \exp(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i), \quad (15.9)$$

## 622 PART III ♦ Estimation Methodology

where  $u_i$  is an unobserved random effect with zero mean and constant variance, possibly normally distributed—we will turn to that shortly. We will now go a step further and specify a particular probability distribution for  $y_{it}$ . Since it is a count, the Poisson regression model would be a natural choice,

$$p(y_{it}|\mathbf{x}_{it}, u_i) = \frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!}, \mu_{it} = \exp(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i). \quad (15-10)$$

Conditioned on  $\mathbf{x}_{it}$ , and  $u_i$ , the  $T_i$  observations for individual  $i$  are independent. That is, by conditioning on  $u_i$ , we treat them as data, the same as  $\mathbf{x}_{it}$ . Thus, the  $T_i$  observations are independent when they are conditioned on  $\mathbf{x}_{it}$  and  $u_i$ . The joint density for the  $T_i$  observations for individual  $i$  is the product,

$$p(y_{i1}, y_{i2}, \dots, y_{iT_i}|\mathbf{X}_i, u_i) = \prod_{t=1}^{T_i} \frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!}, \mu_{it} = \exp(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i), t = 1, \dots, T_i. \quad (15-11)$$

In principle at this point, the log-likelihood function to be maximized would be

$$\ln L = \sum_{i=1}^n \ln \left[ \prod_{t=1}^{T_i} \frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!} \right], \mu_{it} = \exp(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i). \quad (15-12)$$

But, it is not possible to maximize this log-likelihood because the unobserved  $u_i, i = 1, \dots, n$ , appears in it. The joint distribution of  $(y_{i1}, y_{i2}, \dots, y_{iT_i}, u_i)$  is equal to the marginal distribution for  $u_i$  times the conditional distribution of  $\mathbf{y}_i = (y_{i1}, \dots, y_{iT_i})$  given  $u_i$ :

$$p(y_{i1}, y_{i2}, \dots, y_{iT_i}, u_i|\mathbf{X}_i) = p(y_{i1}, y_{i2}, \dots, y_{iT_i}|\mathbf{X}_i, u_i) f(u_i),$$

where  $f(u_i)$  is the marginal density for  $u_i$ . Now, we can obtain the marginal distribution of  $(y_{i1}, y_{i2}, \dots, y_{iT_i})$  without  $u_i$  by

$$p(y_{i1}, y_{i2}, \dots, y_{iT_i}|\mathbf{X}_i) = \int_{u_i} p(y_{i1}, y_{i2}, \dots, y_{iT_i}|\mathbf{X}_i, u_i) f(u_i) du_i.$$

For the specific application, with the Poisson conditional distributions for  $y_{it}|u_i$  and a normal distribution for the random effect,

$$p(y_{i1}, y_{i2}, \dots, y_{iT_i}|\mathbf{X}_i) = \int_{-\infty}^{\infty} \left[ \prod_{t=1}^{T_i} \frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!} \right] \frac{1}{\sigma} \phi\left(\frac{u_i}{\sigma}\right) du_i, \mu_{it} = \exp(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i).$$

The log-likelihood function will now be

$$\ln L = \sum_{i=1}^n \ln \left\{ \int_{-\infty}^{\infty} \left[ \prod_{t=1}^{T_i} \frac{\exp(-\mu_{it})\mu_{it}^{y_{it}}}{y_{it}!} \right] \frac{1}{\sigma} \phi\left(\frac{u_i}{\sigma}\right) du_i \right\}, \mu_{it} = \exp(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i). \quad (15-13)$$

The optimization problem is now free of the unobserved  $u_i$ , but that complication has been traded for another one, the integral that remains in the function.

To complete this part of the derivation, we will simplify the log-likelihood function slightly in a way that will make it fit more naturally into the derivations to follow. Make the change of variable  $u_i = \sigma w_i$  where  $w_i$  has mean zero and standard deviation one. Then, the Jacobian is  $du_i = \sigma dw_i$ , and the limits of integration for  $w_i$  are the same as for

CHAPTER 15 ♦ Simulation-Based Estimation and Inference **623**

$u_i$ . Making the substitution and multiplying by the Jacobian, the log-likelihood function becomes

$$\ln L = \sum_{i=1}^n \ln \left\{ \int_{-\infty}^{\infty} \left[ \prod_{t=1}^{T_i} \frac{\exp(-\mu_{it}) \mu_{it}^{y_{it}}}{y_{it}!} \right] \phi(w_i) dw_i \right\}, \quad \mu_{it} = \exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \sigma w_i). \quad (15-14)$$

The log-likelihood is then maximized over  $(\boldsymbol{\beta}, \sigma)$ . The purpose of the simplification is to parameterize the model so that the distribution of the variable that is being integrated out has no parameters of its own. Thus, in (15-14),  $w_i$  is normally distributed with mean zero and variance one.

In the next section, we will turn to how to compute the integrals. Section 14.9.6.c analyzes this model and suggests the **Gauss–Hermite quadrature** method for computing the integrals. In this section, we will derive a method based on simulation, **Monte Carlo integration**.<sup>4</sup>

### 15.6.2 MONTE CARLO INTEGRATION

Integrals often appear in econometric estimators in “open form,” that is, in a form for which there is no specific  $\text{form}$  function that is equivalent to them. (E.g., the integral,  $\int_0^t \theta \exp(-\theta w) dw = 1 - \exp(-\theta t)$ , is in closed form. The integral in (15-14) is in open form.) There are various devices available for approximating open form integrals—Gauss–Hermite and Gauss–Laguerre quadrature noted in Section 14.9.6.c and in Appendix E2.4 are two. The technique of Monte Carlo integration can often be used when the integral is in the form

$$h(y) = \int_w g(y|w) f(w) dw = E_w[g(y|w)]$$

where  $f(w)$  is the density of  $w$  and  $w$  is a random variable that can be simulated. [There are some necessary conditions on  $w$  and  $g(y|w)$  that will be met in the applications that interest us here. Some details appear in Cameron and Trivedi (2005) and Train (2003).]

If  $w_1, w_2, \dots, w_n$  are a random sample of observations on the random variable  $w$  and  $g(w)$  is a function of  $w$  with finite mean and variance, then by the law of large numbers [Theorem D.4 and the corollary in (D-5)],

$$\text{plim} \frac{1}{n} \sum_{i=1}^n g(w_i) = E[g(w)].$$

The function in (15-14) is in this form;

$$\begin{aligned} & \int_{-\infty}^{\infty} \left[ \prod_{t=1}^{T_i} \frac{\exp[-\exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \sigma w_i)][\exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \sigma w_i)]^{y_{it}}}{y_{it}!} \right] \phi(w_i) dw_i \\ &= E_{w_i}[g(y_{i1}, y_{i2}, \dots, y_{iT_i}|w_i, \mathbf{X}_i, \boldsymbol{\beta}, \sigma)] \end{aligned}$$

<sup>4</sup>The term “Monte Carlo” is in reference to the casino at Monte Carlo, where random number generation is a crucial element of the business.

## 624 PART III ♦ Estimation Methodology

where

$$g(y_{i1}, y_{i2}, \dots, y_{iT_i} | w_i, \mathbf{X}_i, \boldsymbol{\beta}, \sigma) = \prod_{t=1}^{T_i} \frac{\exp[-\exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \sigma w_i)][\exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \sigma w_i)]^{y_{it}}}{y_{it}!}$$

and  $w_i$  is a random variable with standard normal distribution. It follows, then, that

$$\begin{aligned} \text{plim} \frac{1}{R} \sum_{r=1}^R \prod_{t=1}^{T_i} & \frac{\exp[-\exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \sigma w_{ir})][\exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \sigma w_{ir})]^{y_{it}}}{y_{it}!} \\ &= \int_{-\infty}^{\infty} \left[ \prod_{t=1}^{T_i} \frac{\exp[-\exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \sigma w_i)][\exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \sigma w_i)]^{y_{it}}}{y_{it}!} \right] \phi(w_i) dw_i. \end{aligned} \quad (15-15)$$

This suggests the strategy for computing the integral. We can use the methods developed in Section 15.2 to produce the necessary set of random draws on  $w_i$  from the standard normal distribution and then compute the approximation to the integral according to (15-15).

### Example 15.8 Fractional Moments of the Truncated Normal Distribution

The following function appeared in Greene's (1990) study of the stochastic frontier model:

$$h(M, \varepsilon) = \frac{\int_0^\infty z^M \frac{1}{\sigma} \phi\left[\frac{z - (-\varepsilon - \theta\sigma^2)}{\sigma}\right] dz}{\int_0^\infty \frac{1}{\sigma} \phi\left[\frac{z - (-\varepsilon - \theta\sigma^2)}{\sigma}\right] dz}.$$

The integral only exists in closed form for integer values of  $M$ . However, the weighting function that appears in the integral is of the form

$$f(z|z > 0) = \frac{f(z)}{\text{Prob}[z > 0]} = \frac{\frac{1}{\sigma} \phi\left(\frac{z - \mu}{\sigma}\right)}{\int_0^\infty \frac{1}{\sigma} \phi\left(\frac{z - \mu}{\sigma}\right) dz}.$$

This is a truncated normal distribution. It is the distribution of a normally distributed variable  $z$  with mean  $\mu$  and standard deviation  $\sigma$ , conditioned on  $z$  being greater than zero. The integral is equal to the expected value of  $z^M$  given that  $z$  is greater than zero when  $z$  is normally distributed with mean  $\mu = -\varepsilon - \theta\sigma^2$  and variance  $\sigma^2$ .

The truncated normal distribution is examined in Section 15.2. The function  $h(M, \varepsilon)$  is the expected value of  $z^M$  when  $z$  is the truncation of a normal random variable with mean  $\mu$  and standard deviation  $\sigma$ . To evaluate the integral by Monte Carlo integration, we would require a sample  $z_1, \dots, z_R$  from this distribution. We have the results we need in (15-4) with  $L = 0$  so  $P_L = \Phi[0 - (-\varepsilon - \theta\sigma^2)/\sigma] = \Phi(\varepsilon/\sigma + \theta\sigma)$  and  $U = +\infty$  so  $P_U = 1$ . Then, a draw on  $z$  is obtained by

$$z = \mu + \sigma \Phi^{-1}[P_L + F(1 - P_L)],$$

where  $F$  is the primitive draw from  $U[0, 1]$ . Finally, the integral is approximated by the simple average of the draws,

$$h(M, \varepsilon) \approx \frac{1}{R} \sum_{r=1}^R z[\varepsilon, \theta, \sigma, F_r]^M.$$

This is an application of Monte Carlo integration. In certain cases, an integral can be approximated by computing the sample average of a set of function values. The approach taken here was to interpret the integral as an expected value. Our basic statistical result for the behavior of sample means implies that with a large enough

## CHAPTER 15 ♦ Simulation-Based Estimation and Inference **625**

sample, we can approximate the integral as closely as we like. The general approach is widely applicable in Bayesian econometrics and has begun to appear in classical statistics and econometrics as well.<sup>5</sup>

### **15.6.2.a Halton Sequences and Random Draws for Simulation-Based Integration**

Monte Carlo integration is used to evaluate the expectation

$$E[g(x)] = \int_x g(x) f(x) dx$$

where  $f(x)$  is the density of the random variable  $x$  and  $g(x)$  is a smooth function. The Monte Carlo approximation is

$$\widehat{E[g(x)]} = \frac{1}{R} \sum_{r=1}^R g(x_r).$$

Convergence of the approximation to the expectation is based on the law of large numbers—a random sample of draws on  $g(x)$  will converge in probability to its expectation. The standard approach to simulation-based integration is to use random draws from the specified distribution. Conventional simulation-based estimation uses a random number generator to produce the draws from a specified distribution. The central component of this approach is drawn from the standard continuous uniform distribution,  $U[0, 1]$ . Draws from other distributions are obtained from these draws by using transformations. In particular, for a draw from the normal distribution, where  $u_i$  is one draw from  $U[0, 1]$ ,  $v_i = \Phi^{-1}(u_i)$ . Given that the initial draws satisfy the necessary assumptions, the central issue for purposes of specifying the simulation is the number of draws. Good performance in this connection requires very large numbers of draws. Results differ on the number needed in a given application, but the general finding is that when simulation is done in this fashion, the number is large (hundreds or thousands). A consequence of this is that for large-scale problems, the amount of computation time in simulation-based estimation can be extremely large. Numerous methods have been devised for reducing the numbers of draws needed to obtain a satisfactory approximation. One such method is to introduce some autocorrelation into the draws—a small amount of negative correlation across the draws will reduce the variance of the simulation. **Antithetic draws**, whereby each draw in a sequence is included with its mirror image ( $w_i$  and  $-w_i$  for normally distributed draws,  $w_i$  and  $1 - w_i$  for uniform, for example) is one such method. [See Geweke (1988) and Train (2009, Chapter 9).]

Procedures have been devised in the numerical analysis literature for taking “intelligent” draws from the uniform distribution, rather than random ones. [See Train (1999, 2009) and Bhat (1999) for extensive discussion and further references.] An emerging literature has documented dramatic speed gains with no degradation in simulation performance through the use of a smaller number of **Halton draws** or other constructed, nonrandom sequences instead of a large number of random draws. These procedures appear to reduce vastly the number of draws needed for estimation (sometimes by a

---

<sup>5</sup>See Geweke (1986, 1988, 1989, 2005) for discussion and applications. A number of other references are given in Poirier (1995, p. 654) and Koop (2003).

## 626 PART III ♦ Estimation Methodology

factor of 90 percent or more) and reduce the simulation error associated with a given number of draws. In one application of the method to be discussed here, Bhat (1999) found that 100 Halton draws produced lower simulation error than 1,000 random numbers.

A sequence of Halton draws is generated as follows: Let  $r$  be a prime number. Expand the sequence of integers  $g = 1, 2, \dots$  in terms of the base  $r$  as

$$g = \sum_{i=0}^I b_i r^i \text{ where, by construction, } 0 \leq b_i \leq r - 1 \text{ and } r^I \leq g < r^{I+1}.$$

The Halton sequence of values that corresponds to this series is

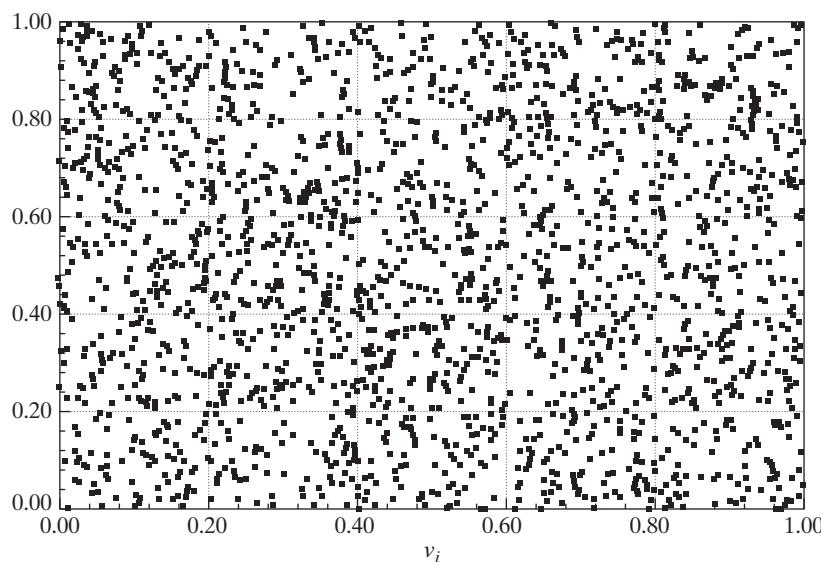
$$H(g) = \sum_{i=0}^I b_i r^{-i-1}.$$

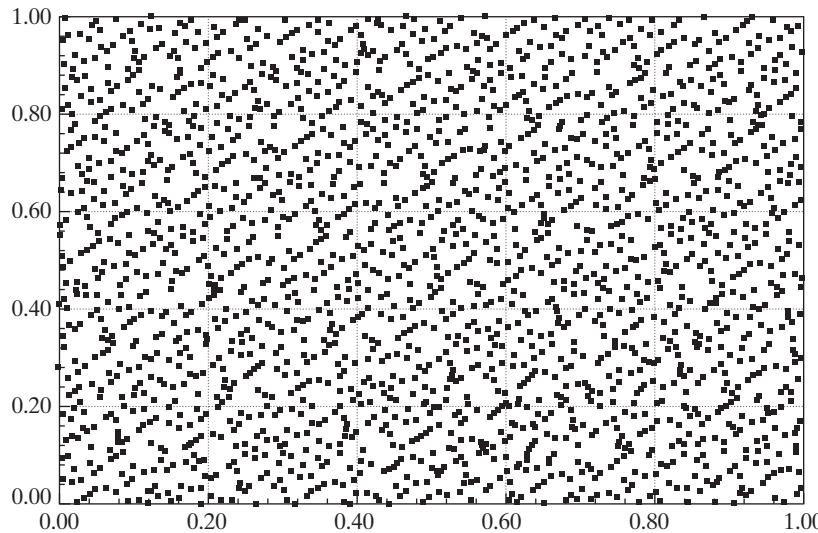
For example, using base 5, the integer 37 has  $b_0 = 2$ ,  $b_1 = 2$ , and  $b_3 = 1$ . Then

$$H_5(37) = 2 \times 5^{-1} + 2 \times 5^{-2} + 1 \times 5^{-3} = 0.488.$$

The sequence of Halton values is efficiently spread over the unit interval. The sequence is not random as the sequence of pseudo-random numbers is; it is a well-defined deterministic sequence. But, randomness is not the key to obtaining accurate approximations to integrals. Uniform coverage of the support of the random variable is the central requirement. The large numbers of random draws are required to obtain smooth and dense coverage of the unit interval. Figures 15.3 and 15.4 show two sequences of 1,000 Halton draws and two sequences of 1,000 pseudo-random draws. The Halton draws are based on  $r = 7$  and  $r = 9$ . The clumping evident in the first figure is the feature (among others) that mandates large samples for simulations.

**FIGURE 15.3** Bivariate Distribution of Random Uniform Draws.





**FIGURE 15.4** Bivariate Distribution of Halton (7) and Halton (9).

#### *Example 15.9 Estimating the Lognormal Mean*

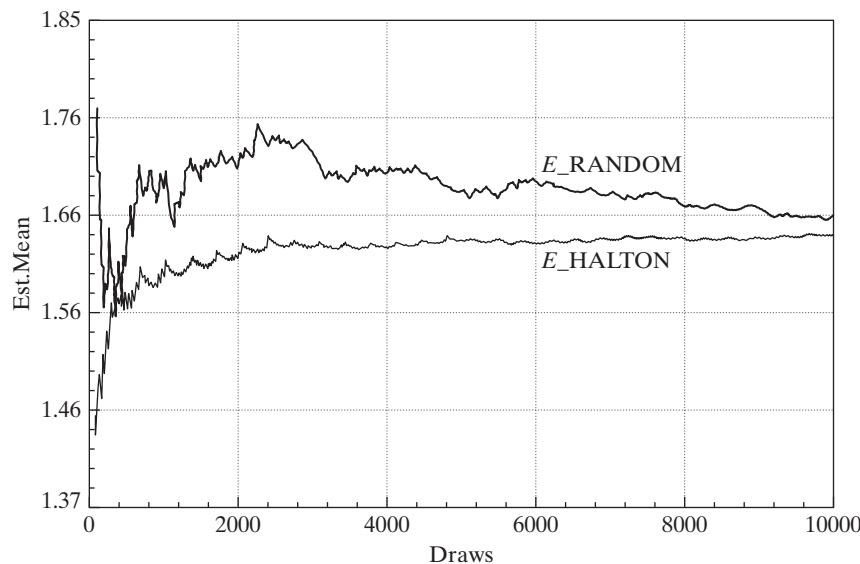
We are interested in estimating the mean of a standard lognormally distributed variable. Formally, this is

$$E[y] = \int_{-\infty}^{\infty} \exp(x) \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right] dx = 1.649.$$

To use simulation for the estimation, we will average  $n$  draws on  $y = \exp(x)$  where  $x$  is drawn from the standard normal distribution. To examine the behavior of the Halton sequence as compared to that of a set of random draws, we did the following experiment. Let  $x_{i,t}$  = the sequence of values for a standard normally distributed variable. We draw  $t = 1, \dots, 10,000$  draws. For  $i = 1$ , we used a random number generator. For  $i = 2$ , we used the sequence of the first 10,000 Halton draws using  $r = 7$ . The Halton draws were converted to standard normal using the inverse normal transformation. To finish preparation of the data, we transformed  $x_{i,t}$  to  $y_{i,t} = \exp(x_{i,t})$ . Then, for  $n = 100, 110, \dots, 10,000$ , we averaged the first  $n$  observations in the sample. Figure 15.5 plots the evolution of the sample means as a function of the sample size. The lower trace is the sequence of Halton-based means. The greater stability of the Halton estimator is clearly evident in the figure.

#### 15.6.2.b Computing Multivariate Normal Probabilities Using the GHK Simulator

The computation of bivariate normal probabilities is typically done using quadrature and requires a large amount of computing effort. Quadrature methods have been developed for trivariate probabilities as well, but the amount of computing effort needed at this level is enormous. For integrals of level greater than three, satisfactory (in terms of speed and accuracy) direct approximations remain to be developed. Our work thus far does suggest an alternative approach. Suppose that  $\mathbf{x}$  has a  $K$ -variate normal distribution with mean vector  $\mathbf{0}$  and covariance matrix  $\Sigma$ . (No generality is sacrificed by the assumption of a zero mean, because we could just subtract a nonzero mean from the random vector wherever it appears in any result.) We wish to compute the  $K$ -variate probability,  $\text{Prob}[a_1 < x_1 < b_1, a_2 < x_2 < b_2, \dots, a_K < x_K < b_K]$ . Our Monte Carlo

**628 PART III ♦ Estimation Methodology**


**FIGURE 15.5** Estimates of  $E[\exp(x)]$  Based on Random Draws and Halton Sequences, by Sample Size.

integration technique is well suited for this problem. As a first approach, consider sampling  $R$  observations,  $\mathbf{x}_r$ ,  $r = 1, \dots, R$ , from this multivariate normal distribution, using the method described in Section 15.2.4. Now, define

$$d_r = \mathbf{1}[a_1 < x_{r1} < b_1, a_2 < x_{r2} < b_2, \dots, a_K < x_{rK} < b_K].$$

(That is,  $d_r = 1$  if the condition is true and 0 otherwise.) Based on our earlier results, it follows that

$$\text{plim } \bar{d} = \text{plim} \frac{1}{R} \sum_{r=1}^R d_r = \text{Prob}[a_1 < x_1 < b_1, a_2 < x_2 < b_2, \dots, a_K < x_K < b_K].^6$$

This method is valid in principle, but in practice it has proved to be unsatisfactory for several reasons. For large-order problems, it requires an enormous number of draws from the distribution to give reasonable accuracy. Also, even with large numbers of draws, it appears to be problematic when the desired tail area is very small. Nonetheless, the idea is sound, and recent research has built on this idea to produce some quite accurate and efficient simulation methods for this computation. A survey of the methods is given in McFadden and Ruud (1994).<sup>7</sup>

Among the simulation methods examined in the survey, the **GHK smooth recursive simulator** appears to be the most accurate.<sup>8</sup> The method is surprisingly simple. The

<sup>6</sup>This method was suggested by Lerman and Manski (1981).

<sup>7</sup>A symposium on the topic of simulation methods appears in *Review of Economic Statistics*, Vol. 76, November 1994. See, especially, McFadden and Ruud (1994), Stern (1994), Geweke, Keane, and Runkle (1994), and Breslaw (1994). See, as well, Gourieroux and Monfort (1996).

<sup>8</sup>See Geweke (1989), Hajivassiliou (1990), and Keane (1994). Details on the properties of the simulator are given in Börsch-Supan and Hajivassiliou (1993).

CHAPTER 15 ♦ Simulation-Based Estimation and Inference **629**

general approach uses

$$\text{Prob}[a_1 < x_1 < b_1, a_2 < x_2 < b_2, \dots, a_K < x_K < b_K] \approx \frac{1}{R} \sum_{r=1}^R \prod_{k=1}^K Q_{rk},$$

where  $Q_{rk}$  are easily computed univariate probabilities. The probabilities  $Q_{rk}$  are computed according to the following recursion: We first factor  $\Sigma$  using the **Cholesky factorization**  $\Sigma = \mathbf{C}\mathbf{C}'$ , where  $\mathbf{C}$  is a lower triangular matrix (see Section A.6.11). The elements of  $\mathbf{C}$  are  $c_{km}$ , where  $c_{km} = 0$  if  $m > k$ . Then we begin the recursion with

$$Q_{r1} = \Phi(b_1/l_{11}) - \Phi(a_1/l_{11}).$$

Note that  $l_{11} = \sigma_{11}$ , so this is just the marginal probability,  $\text{Prob}[a_1 < x_1 < b_1]$ . Now, using (15-4), we generate a random observation  $\varepsilon_{r1}$  from the truncated standard normal distribution in the range

$$A_{r1} \text{ to } B_{r1} = a_1/l_{11} \text{ to } b_1/l_{11}.$$

(Note, again, that the range is standardized since  $l_{11} = \sigma_{11}$ .) For steps  $k = 2, \dots, K$ , compute

$$A_{rk} = \left[ a_k - \sum_{m=1}^{k-1} l_{km} \varepsilon_{rm} \right] / l_{kk},$$

$$B_{rk} = \left[ b_k - \sum_{m=1}^{k-1} l_{km} \varepsilon_{rm} \right] / l_{kk}.$$

Then 

$$Q_{rk} = \Phi(B_{rk}) - \Phi(A_{rk}).$$

Finally, in preparation for the next step in the recursion, we generate a random draw from the truncated standard normal distribution in the range  $A_{rk}$  to  $B_{rk}$ . This process is replicated  $R$  times, and the estimated probability is the sample average of the simulated probabilities.

The GHK simulator has been found to be impressively fast and accurate for fairly moderate numbers of replications. Its main usage has been in computing functions and derivatives for maximum likelihood estimation of models that involve multivariate normal integrals. We will revisit this in the context of the method of simulated moments when we examine the probit model in Chapter 17.

### 15.6.3 SIMULATION-BASED ESTIMATION OF RANDOM EFFECTS MODELS

In Section 15.4.2, (15-17), and (15-18), we developed a random effects specification for the Poisson regression model. For feasible estimation and inference, we replace the log-likelihood function,

$$\ln L = \sum_{i=1}^n \ln \left\{ \int_{-\infty}^{\infty} \left[ \prod_{t=1}^{T_i} \frac{\exp[-\exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \sigma w_i)][\exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \sigma w_i)]^{y_{it}}}{y_{it}!} \right] \phi(w_i) dw_i \right\},$$

### 630 PART III ♦ Estimation Methodology

with the simulated log-likelihood function,

$$\ln L_S = \sum_{i=1}^n \ln \left\{ \frac{1}{R} \sum_{r=1}^R \prod_{t=1}^{T_i} \frac{\exp[-\exp(\mathbf{x}'_{it}\beta + \sigma w_{ir})][\exp(\mathbf{x}'_{it}\beta + \sigma w_{ir})]^{y_{it}}]}{y_{it}!} \right\}. \quad (15-16)$$

We now consider how to estimate the estimate the parameters via maximum simulated likelihood. In spite of its complexity, the simulated log-likelihood will be treated in the same way that other log-likelihoods were handled in Chapter 14. That is, we treat  $\ln L_S$  as a function of the unknown parameters conditioned on the data,  $\ln L_S(\beta, \sigma)$  and maximize the function using the methods described in Appendix E, such as the DFP or BFGS gradient methods. What is needed here to complete the derivation are expressions for the derivatives of the function. We note that the function is a sum of  $n$  terms; asymptotic results will be obtained in  $n$ ; each observation can be viewed as one  $T_i$ -variate observation.

In order to develop a general set of results, it will be convenient to write each single density in the simulated function as

$$P_{itr}(\beta, \sigma) = f(y_{it} | \mathbf{x}_{it}, w_{ir}, \beta, \sigma) = P_{itr}(\theta) = P_{itr}.$$

For our specific application in (15-16),

$$P_{itr} = \frac{\exp[-\exp(\mathbf{x}'_{it}\beta + \sigma w_{ir})][\exp(\mathbf{x}'_{it}\beta + \sigma w_{ir})]^{y_{it}}}{y_{it}!}.$$

The simulated log-likelihod is, then,

$$\ln L_S = \sum_{i=1}^n \ln \left\{ \frac{1}{R} \sum_{r=1}^R \prod_{t=1}^{T_i} P_{itr}(\theta) \right\}. \quad (15-17)$$

Continuing this shorthand, then, we will also define

$$P_{ir} = P_{ir}(\theta) = \prod_{t=1}^{T_i} P_{itr}(\theta) \quad \text{[Note: } T_i \text{ is the number of observations for individual } i\text{]}$$

so that

$$\ln L_S = \sum_{i=1}^n \ln \left\{ \frac{1}{R} \sum_{r=1}^R P_{ir}(\theta) \right\}.$$

And, finally,

$$P_i = P_i(\theta) = \frac{1}{R} \sum_{r=1}^R P_{ir} \quad \text{[Note: } R \text{ is the number of individuals]}$$

so that

$$\ln L_S = \sum_{i=1}^n \ln P_i(\theta). \quad (15-18)$$

With this general template, we will be able to accommodate richer specifications of the index function, now  $\mathbf{x}'_{it}\beta + \sigma w_i$ , and other models such as the linear regression, binary choice models, and so on, simply by changing the specification of  $P_{itr}$ .

The algorithm will use the usual procedure,

$$\hat{\theta}^{(k)} = \hat{\theta}^{(k-1)} + \text{update vector},$$

CHAPTER 15 ♦ Simulation-Based Estimation and Inference **631**

starting from an initial value,  $\hat{\theta}^{(0)}$ , and will exit when the update vector is sufficiently small. A natural initial value would be from a model with no random effects; that is, the pooled estimator for the linear or Poisson or other model with  $\sigma = 0$ . Thus, at entry to the iteration (update), we will compute

$$\ln \hat{L}_S^{(k-1)} = \sum_{i=1}^n \ln \left\{ \frac{1}{R} \sum_{r=1}^R \prod_{t=1}^{T_i} \frac{\exp[-\exp(\mathbf{x}'_{it}\hat{\beta}^{(k-1)} + \hat{\sigma}^{(k-1)}w_{ir})] [\exp(\mathbf{x}'_{it}\hat{\beta}^{(k-1)} + \hat{\sigma}^{(k-1)}w_{ir})]^{y_{it}}}{y_{it}!} \right\}.$$

To use a gradient method for the update, we will need the first derivatives of the function. Computation of an asymptotic covariance matrix may require the Hessian, so we will obtain this as well.

Before proceeding, we note two important aspects of the computation. First, a question remains about the number of draws,  $R$ , required for the maximum simulated likelihood estimator to be consistent. The approximated function,

$$\hat{E}_w[f(y|\mathbf{x}, w)] = \frac{1}{R} \sum_{r=1}^R f(y|\mathbf{x}, w_r)$$

is an unbiased estimator of  $E_w[f(y|\mathbf{x}, w)]$ . However, what appears in the simulated log-likelihood is  $\ln E_w[f(y|\mathbf{x}, w)]$ , and the log of the estimator is a biased estimator of the log of its expectation. To maintain the asymptotic equivalence of the MSL estimator of  $\theta$  and the true MLE (if  $w$  were observed), it is necessary for the estimators of these terms in the log-likelihood to converge to their expectations faster than the expectation of  $\ln L$  converges to its expectation. The requirement [see Gourieroux and Monfort (1996)] is that  $n^{1/2}/R \rightarrow 0$ . The estimator remains consistent if  $n^{1/2}$  and  $R$  increase at the same rate; however, the asymptotic covariance matrix of the MSL estimator will then be larger than that of the true MLE. In practical terms, this suggests that the number of draws be on the order of  $n^{5+\delta}$  for some positive  $\delta$ . [This does not state, however, what  $R$  should be for a given  $n$ ; it only establishes the properties of the MSL estimator as  $n$  increases. For better or worse, researchers who have one sample of  $n$  observations often rely on the numerical stability of the estimator with respect to changes in  $R$  as their guide. Hajivassiliou (2000) gives some suggestions.] Note, as well, that the use of Halton sequences or any other autocorrelated sequences for the simulation, which is becoming more prevalent, interrupts this result. The appropriate counterpart to the Gourieroux and Monfort result for random sampling remains to be derived. One might suspect that the convergence result would persist, however. The usual standard is several hundred.

Second, it is essential that the same (pseudo- or Halton) draws be used every time the function or derivatives or any function involving these is computed for observation  $i$ . This can be achieved by creating the pool of draws for the entire sample before the optimization begins, and simply dipping into the same point in the pool each time a computation is required for observation  $i$ . Alternatively, if computer memory is an issue and the draws are re-created for each individual each time, the same practical result can be achieved by setting a preassigned seed for individual  $i$ ,  $seed(i) = s(i)$  for some simple monotonic function of  $i$ , and resetting the seed when draws for individual  $i$  are needed.

To obtain the derivatives, we begin with

$$\frac{\partial \ln L}{\partial \theta} \sum_{i=1}^n \frac{(1/R) \sum_{r=1}^R \partial \left( \prod_{t=1}^{T_i} P_{itr}(\theta) \right) / \partial \theta}{(1/R) \sum_{r=1}^R \prod_{t=1}^{T_i} P_{itr}(\theta)}. \quad (15-19)$$

### 632 PART III ♦ Estimation Methodology

For the derivative term,

$$\begin{aligned}
 \partial \prod_{t=1}^{T_i} P_{itr}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} &= \left( \prod_{t=1}^{T_i} P_{itr}(\boldsymbol{\theta}) \right) \partial \left( \ln \prod_{t=1}^{T_i} P_{itr}(\boldsymbol{\theta}) \right) / \partial \boldsymbol{\theta} \\
 &= \left( \prod_{t=1}^{T_i} P_{itr}(\boldsymbol{\theta}) \right) \sum_{t=1}^{T_i} \partial \ln P_{itr}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \\
 &= P_{ir}(\boldsymbol{\theta}) \left( \sum_{t=1}^{T_i} \partial \ln P_{itr}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \right) = P_{ir}(\boldsymbol{\theta}) \sum_{t=1}^{T_i} \mathbf{g}_{itr}(\boldsymbol{\theta}) \\
 &= P_{ir}(\boldsymbol{\theta}) \mathbf{g}_{ir}(\boldsymbol{\theta}).
 \end{aligned} \tag{15-20}$$

Now, insert the result of (15-20) in (15-19) to obtain

$$\frac{\partial \ln L_S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^n \frac{\sum_{r=1}^R P_{ir}(\boldsymbol{\theta}) \mathbf{g}_{ir}(\boldsymbol{\theta})}{\sum_{r=1}^R P_{ir}(\boldsymbol{\theta})}. \tag{15-21}$$

Define the weight  $Q_{ir}(\boldsymbol{\theta}) = P_{ir}(\boldsymbol{\theta}) / \sum_{r=1}^R P_{ir}(\boldsymbol{\theta})$  so that  $0 < Q_{ir}(\boldsymbol{\theta}) < 1$  and  $\sum_{r=1}^R Q_{ir}(\boldsymbol{\theta}) = 1$ . Then,

$$\frac{\partial \ln L_S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^n \sum_{r=1}^R Q_{ir}(\boldsymbol{\theta}) \mathbf{g}_{ir}(\boldsymbol{\theta}) = \sum_{i=1}^n \bar{\mathbf{g}}_i(\boldsymbol{\theta}). \tag{15-22}$$

To obtain the second derivatives, define  $\mathbf{H}_{itr}(\boldsymbol{\theta}) = \partial^2 \ln P_{itr}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$  and let

$$\mathbf{H}_{ir}(\boldsymbol{\theta}) = \sum_{t=1}^{T_i} \mathbf{H}_{itr}(\boldsymbol{\theta})$$

and

$$\bar{\mathbf{H}}_i(\boldsymbol{\theta}) = \sum_{r=1}^R Q_{ir}(\boldsymbol{\theta}) \mathbf{H}_{ir}(\boldsymbol{\theta}). \tag{15-23}$$

Then, working from (15-21), the second derivatives matrix breaks into three parts as follows:

$$\frac{\partial^2 \ln L_S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{i=1}^n \left[ \begin{array}{c} \frac{\sum_{r=1}^R P_{ir}(\boldsymbol{\theta}) \mathbf{H}_{ir}(\boldsymbol{\theta})}{\sum_{r=1}^R P_{ir}(\boldsymbol{\theta})} + \frac{\sum_{r=1}^R P_{ir}(\boldsymbol{\theta}) \mathbf{g}_{ir}(\boldsymbol{\theta}) \mathbf{g}_{ir}(\boldsymbol{\theta})'}{\sum_{r=1}^R P_{ir}(\boldsymbol{\theta})} \\ - \frac{\left[ \sum_{r=1}^R P_{ir}(\boldsymbol{\theta}) \mathbf{g}_{ir}(\boldsymbol{\theta}) \right] \left[ \sum_{r=1}^R P_{ir}(\boldsymbol{\theta}) \mathbf{g}_{ir}(\boldsymbol{\theta}) \right]'}{\left[ \sum_{r=1}^R P_{ir}(\boldsymbol{\theta}) \right]^2} \end{array} \right].$$

We can now use (15-20)–(15-23) to combine these terms;

$$\frac{\partial^2 \ln L_S}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{i=1}^n \left\{ \bar{\mathbf{H}}_i(\boldsymbol{\theta}) + \sum_{r=1}^R Q_{ir}(\boldsymbol{\theta}) [\mathbf{g}_{ir}(\boldsymbol{\theta}) - \bar{\mathbf{g}}_i(\boldsymbol{\theta})] [\mathbf{g}_{ir}(\boldsymbol{\theta}) - \bar{\mathbf{g}}_i(\boldsymbol{\theta})]' \right\}. \tag{15-24}$$

**CHAPTER 15 ♦ Simulation-Based Estimation and Inference 633**

An estimator of the asymptotic covariance matrix for the MSLE can be obtained by computing the negative inverse of this matrix.

**Example 15.10 Poisson Regression Model with Random Effects**

For the Poisson regression model,  $\theta = (\beta', \sigma^2)$  and

$$\begin{aligned} P_{itr}(\theta) &= \frac{\exp[-\exp(\mathbf{x}'_it\beta + \sigma w_{ir})][\exp(\mathbf{x}'_it\beta + \sigma w_{ir})]^{y_{it}}}{y_{it}!} = \frac{\exp[-\mu_{itr}(\theta)]\mu_{itr}(\theta)^{y_{it}}}{y_{it}!} \\ \mathbf{g}_{itr}(\theta) &= [y_{it} - \mu_{itr}(\theta)] \begin{pmatrix} \mathbf{x}'_it \\ w_{ir} \end{pmatrix} \\ \mathbf{H}_{itr}(\theta) &= -\mu_{itr}(\theta) \begin{pmatrix} \mathbf{x}'_it \\ w_{ir} \end{pmatrix} \begin{pmatrix} \mathbf{x}'_it \\ w_{ir} \end{pmatrix}' . \end{aligned} \quad (15-25)$$

Estimates of the random effects model parameters would be obtained by using these expressions in the preceding general template. We will apply these results in an application in Chapter 19 where the Poisson regression model is developed in greater detail.

**Example 15.11 Maximum Simulated Likelihood Estimation of the Random Effects Linear Regression Model**

The preceding method can also be used to estimate a linear regression model with random effects. We have already seen two ways to estimate this model, using two-step FGLS in Section 11.5.3 and by (closed form) maximum likelihood in Section 14.9.6.a. It might seem redundant to construct yet a third estimator for the model. However, this third approach will be the only feasible method when we generalize the model to have other random parameters in the next section. To use the simulation estimator, we define  $\theta = (\beta, \sigma_u, \sigma_e)$ . We will require

$$\begin{aligned} P_{itr}(\theta) &= \frac{1}{\sigma_e \sqrt{2\pi}} \exp \left[ -\frac{(y_{it} - \mathbf{x}'_it\beta - \sigma_u w_{ir})^2}{2\sigma_e^2} \right], \\ \mathbf{g}_{itr}(\theta) &= \begin{bmatrix} \left( \frac{(y_{it} - \mathbf{x}'_it\beta - \sigma_u w_{ir})}{\sigma_e^2} \right) \begin{pmatrix} \mathbf{x}'_it \\ w_{ir} \end{pmatrix} \\ \frac{(y_{it} - \mathbf{x}'_it\beta - \sigma_u w_{ir})^2}{\sigma_e^3} - \frac{1}{\sigma_e} \end{bmatrix} = \begin{bmatrix} (\varepsilon_{itr}/\sigma_e^2) \begin{pmatrix} \mathbf{x}'_it \\ w_{ir} \end{pmatrix} \\ (1/\sigma_e)[(\varepsilon_{itr}^2/\sigma_e^2) - 1] \end{bmatrix} \quad (15-26) \\ \mathbf{H}_{itr}(\theta) &= \begin{bmatrix} -(1/\sigma_e^2) \begin{pmatrix} \mathbf{x}'_it \\ w_{ir} \end{pmatrix} \begin{pmatrix} \mathbf{x}'_it \\ w_{ir} \end{pmatrix}' & -(2\varepsilon_{itr}/\sigma_e^3) \begin{pmatrix} \mathbf{x}'_it \\ w_{ir} \end{pmatrix} \\ -(2\varepsilon_{itr}^3/\sigma_e^3) (\mathbf{x}'_it w_{ir}) & -(3\varepsilon_{itr}^2/\sigma_e^4) + (1/\sigma_e^2) \end{bmatrix}. \end{aligned}$$

Note in the computation of the disturbance variance,  $\sigma_e^2$ , we are using the sum of squared simulated residuals. However, the estimator of the variance of the heterogeneity,  $\sigma_u$ , is not being computed as a mean square. It is essentially the regression coefficient on  $w_{ir}$ . One surprising implication is that the actual estimate of  $\sigma_u$  can be negative. This is the same result that we have encountered in other situations. In no case is there a natural estimator of  $\sigma_u^2$  that is based on a sum of squares. However, in this context, there is yet another surprising aspect of this calculation. In the simulated log-likelihood function, if every  $w_{ir}$  for every individual were changed to  $-w_{ir}$  and  $\sigma_e$  changed to  $-\sigma_e$ , the exact same value of the function and all derivatives results. The implication is that the sign of  $\sigma_e$  is not identified in this setting. With no loss of generality, it is normalized to positive (+) to be consistent with the underlying theory that it is a standard deviation.

## 634 PART III ♦ Estimation Methodology

### 15.7 A RANDOM PARAMETERS LINEAR REGRESSION MODEL

We will slightly reinterpret the random effects model as

$$\begin{aligned} y_{it} &= \beta_{0i} + \mathbf{x}'_{it1}\boldsymbol{\beta}_1 + \varepsilon_{it}, \\ \beta_{0i} &= \beta_0 + u_i. \end{aligned} \tag{15-27}$$

This is equivalent to the random effects model, though in (15-27), we reinterpret it as a regression model with a randomly distributed constant term. In Section 11.11.1, we built a linear regression model that provided for parameter heterogeneity across individuals,

$$\begin{aligned} y_{it} &= \mathbf{x}'_{it}\boldsymbol{\beta}_i + \varepsilon_{it}, \\ \boldsymbol{\beta}_i &\Rightarrow \boldsymbol{\beta} + \mathbf{u}_i, \end{aligned} \tag{15-28}$$

where  $\mathbf{u}_i$  has mean vector  $\mathbf{0}$  and covariance matrix  $\boldsymbol{\Gamma}$ . In that development, we took a fixed effects approach in that no restriction was placed on the covariance between  $\mathbf{u}_i$  and  $\mathbf{x}_{it}$ . Consistent with these assumptions, we constructed an estimator that involved  $n$  regressions of  $\mathbf{y}_i$  on  $\mathbf{X}_i$  to estimate  $\boldsymbol{\beta}$  one unit at a time. Each estimator is consistent in  $T_i$ . (This is precisely the approach taken in the fixed effects model, where there are  $n$  unit specific constants and a common  $\boldsymbol{\beta}$ . The approach there is to estimate  $\boldsymbol{\beta}$  first and then to regress  $\mathbf{y}_i - \mathbf{X}_i\mathbf{b}_{LSDV}$  on  $\mathbf{d}_i$  to estimate  $\alpha_i$ .) In the same way that assuming that  $u_i$  is uncorrelated with  $\mathbf{x}_{it}$  in the fixed effects model provided a way to use FGLS to estimate the parameters of the random effects model, if we assume in (15-28) that  $\mathbf{u}_i$  is uncorrelated with  $\mathbf{X}_i$ , we can extend the random effects model in Section 15.6 to a model in which some or all of the other coefficients in the regression model, not just the constant term, are randomly distributed. The theoretical proposition is that the model is now extended to allow individual heterogeneity in all coefficients.

To implement the extended model, we will begin with a simple formulation in which  $\mathbf{u}_i$  has diagonal covariance matrix—this specification is quite common in the literature. The implication is that the random parameters are uncorrelated;  $\beta_{i,k}$  has mean  $\beta_k$  and variance  $\gamma_k^2$ . The model in (15-26) can be modified to allow this case with a few minor changes in notation. Write

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \Lambda \mathbf{w}_i \tag{15-29}$$

where  $\Lambda$  is a diagonal matrix with the standard deviations ( $\gamma_1, \gamma_2, \dots, \gamma_K$ ) of  $(u_{i1}, \dots, u_{iK})$  on the diagonal and  $\mathbf{w}_i$  is now a random vector with zero means and unit standard deviations. The parameter vector in the model is now

$$\boldsymbol{\theta} = (\beta_1, \dots, \beta_K, \lambda_1, \dots, \lambda_K, \sigma_\varepsilon).$$

(In an application, some of the  $\gamma$ 's might be fixed at zero to make the corresponding parameters nonrandom.) In order to extend the model, the disturbance in (15-20),  $\varepsilon_{itr} = (y_{it} - \mathbf{x}'_{it}\boldsymbol{\beta} - \sigma_w \mathbf{w}_{ir})$ , becomes

$$\varepsilon_{itr} = y_{it} - \mathbf{x}'_{it}(\boldsymbol{\beta} + \Lambda \mathbf{w}_{ir}). \tag{15-30}$$

Now, combine (15-17) and (15-29) with (15-30) to produce

$$\ln L_S = \sum_{i=1}^n \ln \left\{ \frac{1}{R} \sum_{r=1}^R \prod_{t=1}^{T_i} \frac{1}{\sigma_\varepsilon \sqrt{2\pi}} \exp \left[ \frac{(y_{it} - \mathbf{x}'_{it}(\boldsymbol{\beta} + \Lambda \mathbf{w}_{it}))^2}{2\sigma_\varepsilon^2} \right] \right\}. \tag{15-31}$$

CHAPTER 15 ♦ Simulation-Based Estimation and Inference **635**

In the derivatives in (15-26), the only change needed to accommodate this extended model is that the scalar  $w_{ir}$  becomes the vector  $(w_{ir,1}x_{it,1}, w_{ir,2}x_{it,2}, \dots, w_{ir,K}x_{it,K})$ . This is the element-by-element product of the regressors,  $\mathbf{x}_{it}$ , and the vector of random draws,  which is the **Hadamard product**, **direct product**, or **Schur product** of the two vectors, denoted  $\mathbf{x}_{it} \bullet \mathbf{w}_{ir}$ .

Although only a minor change in notation in the random effects template in (15-26), this formulation brings a substantial change in the formulation of the model. The integral in (15-20) is now a  $K$  dimensional integral. Maximum simulated likelihood estimation proceeds as before, with potentially much more computation as each “draw” now requires a  $K$ -variate vector of pseudo-random draws.

The random parameters model can now be extended to one with a full covariance matrix,  $\boldsymbol{\Gamma}$  as we did with the fixed effects case. We will now let  $\boldsymbol{\Lambda}$  in (15-29) be the Cholesky factorization of  $\boldsymbol{\Gamma}$ , so  $\boldsymbol{\Gamma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ . (This was already the case for the simpler model with diagonal  $\boldsymbol{\Gamma}$ .) The implementation in (15-26) will be a bit complicated. The derivatives with respect to  $\boldsymbol{\beta}$  are unchanged. For the derivatives with respect to  $\boldsymbol{\Lambda}$ , it is useful to assume for the moment that  $\boldsymbol{\Lambda}$  is a full matrix, not a lower triangular one. Then, the scalar  $w_{ir}$  in the derivative expression becomes a  $K^2 \times 1$  vector in which the  $(k-1) \times K + l^{\text{th}}$  element is  $x_{it,k} \times w_{ir,l}$ . The full set of these is the **Kronecker product** of  $\mathbf{x}_{it}$  and  $\mathbf{w}_{ir}$ ,  $\mathbf{x}_{it} \otimes \mathbf{w}_{ir}$ . The necessary elements for maximization of the log-likelihood function are then obtained by discarding the elements for which  $\boldsymbol{\Lambda}_{kl}$  are known to be zero—these correspond to  $l > k$ .

In (15-26), for the full model, for computing the MSL estimators, the derivatives with respect to  $(\boldsymbol{\beta}, \boldsymbol{\Lambda})$  are equated to zero. The result after some manipulation is

$$\frac{\partial \ln L_S}{\partial (\boldsymbol{\beta}, \boldsymbol{\Lambda})} = \sum_{i=1}^n \frac{1}{R} \sum_{r=1}^R \sum_{t=1}^{T_i} \frac{(y_{it} - \mathbf{x}'_{it}(\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{w}_{it}))}{\sigma_\varepsilon^2} \begin{bmatrix} \mathbf{x}_{it} \\ \mathbf{x}_{it} \otimes \mathbf{w}_{ir} \end{bmatrix} = \mathbf{0}.$$

By multiplying this by  $\sigma_\varepsilon^2$ , we find, as usual, that  $\sigma_\varepsilon^2$  is not needed for computation of the estimates of  $(\boldsymbol{\beta}, \boldsymbol{\Lambda})$ . Thus, we can view the solution as the counterpart to least squares, which might call, instead, the minimum simulated sum of squares estimator. Once the simulated sum of squares is minimized with respect to  $\boldsymbol{\beta}$  and  $\boldsymbol{\Lambda}$ , then the solution for  $\sigma_\varepsilon^2$  can be obtained via the likelihood equation,

$$\frac{\partial \ln L_S}{\partial \sigma_\varepsilon^2} = \sum_{i=1}^n \left\{ \frac{1}{R} \sum_{r=1}^R \left[ \frac{-T_i}{2\sigma_\varepsilon^2} + \frac{\sum_{t=1}^{T_i} (y_{it} - \mathbf{x}'_{it}(\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{w}_{it}))^2}{2\sigma_\varepsilon^4} \right] \right\} = 0.$$

Multiply both sides of this equation by  $-2\sigma_\varepsilon^4$  to obtain the equivalent condition

$$\frac{\partial \ln L_S}{\partial \sigma_\varepsilon^2} = \sum_{i=1}^n \left\{ \frac{1}{R} \sum_{r=1}^R T_i \left[ -\sigma_\varepsilon^2 + \frac{\sum_{t=1}^{T_i} (y_{it} - \mathbf{x}'_{it}(\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{w}_{it}))^2}{T_i} \right] \right\} = 0.$$

By expanding this expression and manipulating it a bit, we find the solution for  $\sigma_\varepsilon^2$  is

$$\hat{\sigma}_\varepsilon^2 = \sum_{i=1}^n Q_i \frac{1}{R} \sum_{r=1}^R \hat{\sigma}_{\varepsilon,ir}^2, \text{ where } \hat{\sigma}_{\varepsilon,ir}^2 = \frac{\sum_{t=1}^{T_i} (y_{it} - \mathbf{x}'_{it}(\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{w}_{it}))^2}{T_i}$$

and  $Q_i = T_i / \Sigma_i T_i$  is a weight for each group that equals  $1/n$  if  $T_i$  is the same for all  $i$ .

### 636 PART III ♦ Estimation Methodology

#### Example 15.12 Random Parameters Wage Equation

Estimates of the random effects log wage equation from the Cornwell and Rupert study in Examples 15.1 and 15.6 are shown in Table 15.6. The table presents estimates based on several assumptions. The encompassing model is

$$\ln \text{Wage}_{it} = \beta_{1,i} + \beta_{2,i} \text{Wks}_{i,t} + \dots + \beta_{12,i} \text{Fem}_i + \beta_{13,i} \text{Blk}_i + \varepsilon_{it}, \quad (15-32)$$

$$\beta_{k,i} = \beta_k + \lambda_k w_{ik}, w_{ik} \sim N[0, 1], k = 1, \dots, 13. \quad (15-33)$$

Under the assumption of homogeneity, that is,  $\lambda_k = 0$ , the pooled OLS estimator is consistent and efficient. As we saw in Chapter 11, under the random effects assumption, that is  $\lambda_k = 0$  for  $k = 2, \dots, 13$  but  $\lambda_1 \neq 0$ , the OLS estimator is consistent, as are the next three estimators that explicitly account for the heterogeneity. To consider the full specification, write the model in the equivalent form

$$\begin{aligned} \ln \text{Wage}_{it} &= \mathbf{x}'_{it} \boldsymbol{\beta} + \left( \lambda_1 w_{i,1} + \sum_{k=2}^{13} \lambda_k w_{i,k} x_{it,k} \right) + \varepsilon_{it} \\ &= \mathbf{x}'_{it} \boldsymbol{\beta} + W_{it} + \varepsilon_{it}. \end{aligned}$$

**TABLE 15.6** Estimated Wage Equations (Standard Errors in Parentheses)

Variable	Pooled OLS	Feasible Two Step GLS	Maximum Likelihood	Maximum Simulated Likelihood <sup>a</sup>	Random Parameters Max. Simulated Likelihood <sup>a</sup>	
				$\beta$	$\lambda$	
Wks	.00422 (.00108)	.00096 (.00059)	.00084 (.00060)	.00086 (.00099)	-.00029 (.00082)	.00614 (.00042)
South	-.05564 (.01253)	-.00825 (.02246)	.00577 (.03159)	.00935 (.03106)	.04941 (.02002)	.20997 (.01702)
SMSA	.15167 (.01207)	-.02840 (.01616)	-.04748 (.01896)	-.04913 (.03710)	-.05486 (.01747)	.01165 (.02738)
MS	.04845 (.02057)	-.07090 (.01793)	-.04138 (.01899)	-.04142 (.02176)	-.06358* (.01896)	.02524 (.03190)
Exp	.04010 (.00216)	.08748 (.00225)	.10721 (.00248)	.10668 (.00290)	.09291 (.00216)	.01803 (.00092)
Exp <sup>2</sup>	-.00067 (.0000474)	-.00076 (.0000496)	-.00051 (.0000545)	-.00050 (.0000661)	-.00019 (.0000732)	.0000812 (.00002)
Occ	-.14001 (.01466)	-.04322 (.01299)	-.02512 (.01378)	-.02437 (.02485)	-.00963 (.01331)	.02565 (.01019)
Ind	.04679 (.01179)	.00378 (.01373)	.01380 (.01529)	.01610 (.03670)	.00207 (.01357)	.02575 (.02420)
Union	.09263 (.01280)	.05835 (.01350)	.03873 (.01481)	.03724 (.02814)	.05749 (.01469)	.15260 (.02022)
Ed	.05670 (.00261)	.10707 (.00511)	.13562 (.01267)	.13952 (.03746)	.09356 (.00359)	.00409 (.00160)
Fem	-.36779 (.02510)	-.30938 (.04554)	-.17562 (.11310)	-.11694 (.10784)	-.03864 (.02467)	.28310 (.00760)
Blk	-.16694 (.02204)	-.21950 (.05252)	-.26121 (.13747)	-.15184 (.08356)	-.26864 (.03156)	.02930 (.03841)
Constant	5.25112 (.07129)	4.04144 (.08330)	3.12622 (.17761)	3.08362 (.48917)	3.81680 (.06905)	.26347 (.01628)
$\sigma_u$	.00000	.31453	.15334	.21164 (.03070)		
$\sigma_\varepsilon$	.34936	.15206	.83949	.15326 (.00217)	.14354 (.00208)	
ln L	-1523.254		307.873	568.446	668.630	

<sup>a</sup> Based on 500 Halton draws

CHAPTER 15 ♦ Simulation-Based Estimation and Inference **637**

This is still a regression:  $E[W_{it} + \varepsilon_{it} | \mathbf{X}] = 0$ . (For the product terms,  $E[\lambda_k w_{i,k} x_{it,k} | \mathbf{X}] = \lambda_k x_{it,k} E[w_{i,k} | \mathbf{X}] = 0$ .) Therefore, even OLS remains consistent. The heterogeneity induces heteroscedasticity in  $W_{it}$  so the OLS estimator is inefficient and the conventional covariance matrix will be inappropriate. The random effects estimators of  $\beta$  in the center three columns of Table 15.6 are also consistent, by a similar logic. However, they likewise are inefficient. The result at work, which is specific to the linear regression model, is that we are estimating the mean parameters,  $\beta_k$ , and the variance parameters,  $\lambda_k$  and  $\sigma_\varepsilon$ , separately. Certainly, if  $\lambda_k$  is nonzero for  $k = 2, \dots, 13$ , then the pooled and RE estimators that assume they are zero are all inconsistent. With  $\beta$  estimated consistently in an otherwise misspecified model, we would call the MLE and MSLE **pseudo maximum likelihood estimators**.

Comparing the ML and MSL estimators of the random effects model, we find the estimates are similar, though in a few cases, noticeably different nonetheless. The estimates tend to differ most when the estimates themselves have large standard errors (small  $t$  ratios). This is partly due to the different methods of estimation in a finite sample of 595 observations. We could attribute at least some of the difference to the approximation error in the simulation compared to the exact evaluation of the (closed form) integral in the MLE. The difference in the log-likelihood functions would be attributable to this as well. Note, however, that the difference is smaller than it first appears—the comparison of 586.446 to 307.883 is misleading; the comparison should be of the difference of the two values from the log-likelihood from the pooled model of  $-1523.254$ . This produces a difference of about 14 percent.

The full random parameters model is shown in the last two columns. Based on the likelihood ratio statistic of  $2(668.630 - 568.446) = 200.368$  with 12 degrees of freedom, we would reject the hypothesis that  $\lambda_2 = \lambda_3 = \dots = \lambda_{13} = 0$ . The 95 percent critical value with 12 degrees of freedom is 21.03. This random parameters formulation of the model suggests a need to reconsider the notion of “statistical significance” of the estimated parameters. In view of (15-33), it may be the case that the mean parameter might well be significantly different from zero while the corresponding standard deviation,  $\lambda$ , might be large as well, suggesting that a large proportion of the population remains statistically close to zero. Consider the estimate of  $\beta_{12,i}$ , the coefficient on  $Fem_i$ . The estimate of the mean,  $\beta_{12}$ , is  $-0.03864$  with an estimated standard error of 0.02467. This implies a confidence interval for this parameter of  $-0.03864 \pm 1.96(0.02467) = [-0.086993, 0.009713]$  But, this is only the location of the center of the distribution. With an estimate of  $\lambda_k$  of 0.2831, the random parameters model suggests that in the population, 95 percent of individuals have an effect of  $Fem_i$  within  $-0.03864 \pm 1.96(0.2831) = [-0.5935, 0.5163]$ . This is still centered near zero but has a different interpretation from the simple confidence interval for  $\beta$  itself. This analysis suggests that it might be an interesting exercise to estimate  $\beta_i$  rather than just the parameters of the distribution. We will consider that estimation problem in Section 15.10.

The next example examines a random parameters model in which the covariance matrix of the random parameters is allowed to be a free, positive definite matrix. That is

$$\begin{aligned} y_{it} &= \mathbf{x}'_{it} \boldsymbol{\beta}_i + \varepsilon_{it} \\ \boldsymbol{\beta}_i &= \boldsymbol{\beta} + \mathbf{u}_i, E[\mathbf{u}_i | \mathbf{X}] = \mathbf{0}, \text{Var}[\mathbf{u}_i | \mathbf{X}] = \sum. \end{aligned} \tag{15-34}$$

This is the counterpart to the fixed effects model in Section 11.4. Note that the difference in the specifications is the random effects assumption,  $E[\mathbf{u}_i | \mathbf{X}] = \mathbf{0}$ . We continue to use the Cholesky decomposition of  $\sum$  in the reparameterized model

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \Lambda \mathbf{w}_i, E[\mathbf{w}_i | \mathbf{X}] = \mathbf{0}, \text{Var}[\mathbf{w}_i | \mathbf{X}] = \mathbf{I}.$$

### 638 PART III ♦ Estimation Methodology

#### **Example 15.13 Least Simulated Sum of Squares Estimates of a Production Function Model**

In Example 11.19, we examined Munell's production model for gross state product,

$$\ln gsp_{it} = \beta_1 + \beta_2 \ln pc_{it} + \beta_3 \ln hwy_{it} + \beta_4 \ln water_{it} + \beta_5 \ln util_{it} + \beta_6 \ln emp_{it} + \beta_7 unemp_{it} + \varepsilon_{it}, i = 1, \dots, 48; t = 1, \dots, 17.$$

The panel consists of state-level data for 17 years. The model in Example 11.19 (and Munell's) provide no means for parameter heterogeneity save for the constant term. We have reestimated the model using the Hildreth and Houck approach. The OLS, feasible GLS and maximum likelihood estimates are given in Table 15.7. The chi-squared statistic for testing the null hypothesis of parameter homogeneity is 25,556.26, with  $7(47) = 329$  degrees of freedom. The critical value from the table is 372.299, so the hypothesis would be rejected. Unlike the other cases we have examined in this chapter, the FGLS estimates are very different from OLS in these estimates, in spite of the fact that both estimators are consistent and the sample is fairly large. The underlying standard deviations are computed using  $\mathbf{G}$  as the covariance matrix. [For these data, subtracting the second matrix rendered  $\mathbf{G}$  not positive definite so, in the table, the standard deviations are based on the estimates using only the first term in (11.97).] The increase in the standard errors is striking. This suggests that there is considerable variation in the parameters across states. We have used (11.97) to compute the estimates of the state-specific coefficients.

The rightmost columns of Table 15.7 present the maximum simulated likelihood estimates of the random parameters production function model. They somewhat resemble the OLS estimates, more so than the FGLS estimates, which are computed by an entirely different method. The values in parentheses under the parameter estimates are the estimates of the standard deviations of the distribution of the square roots of the diagonal elements of  $\Sigma$ . These are obtained by computing the square roots of the diagonal elements of  $\Lambda\Lambda'$ . The

**TABLE 15.7** Estimated Random Coefficients Models

<i>Variable</i>	<i>Least Squares</i>		<i>Feasible GLS</i>			<i>Maximum Simulated Likelihood</i>	
	<i>Estimate</i>	<i>Standard Error</i>	<i>Estimate</i>	<i>Standard Error</i>	<i>Popn. Std. Deviation</i>	<i>Estimate</i>	<i>Std. Error</i>
Constant	1.9260	0.05250	1.6533	1.08331	7.0782	1.9463 (0.0411)	0.03569
ln pc	0.3120	0.01109	0.09409	0.05152	0.3036	0.2962 (0.0730)	0.00882
ln hwy	0.05888	0.01541	0.1050	0.1736	1.1112	0.09515 (0.146)	0.01157
ln water	0.1186	0.01236	0.07672	0.06743	0.4340	0.2434 (0.343)	0.01929
ln util	0.00856	0.01235	-0.01489	0.09886	0.6322	-0.1855 (0.281)	0.02713
ln emp	0.5497	0.01554	0.9190	0.1044	0.6595	0.6795 (0.121)	0.02274
unemp	-0.00727	0.001384	-0.004706	0.002067	0.01266	-0.02318 (0.0308)	0.002712
$\sigma_\varepsilon$		0.08542		0.2129		0.02748	
ln L		853.1372				1567.233	

CHAPTER 15 ♦ Simulation-Based Estimation and Inference **639**

estimate of  $\Lambda$  is shown here.

$$\hat{\Lambda} = \begin{matrix} 0.04114 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.00715 & 0.07266 & 0 & 0 & 0 & 0 & 0 \\ -0.02446 & 0.12392 & 0.07247 & 0 & 0 & 0 & 0 \\ 0.09972 & -0.00644 & 0.31916 & 0.07614 & 0 & 0 & 0 \\ -0.08928 & 0.02143 & -0.25105 & 0.07583 & 0.04053 & 0 & 0 \\ 0.03842 & -0.06321 & -0.03992 & -0.06693 & -0.05490 & 0.00857 & 0 \\ -0.00833 & -0.00257 & -0.02478 & 0.01594 & 0.00102 & -0.00185 & 0.0018. \end{matrix}$$

An estimate of the correlation matrix for the parameters might also be informative. This is also derived from  $\hat{\Lambda}$  by computing  $\hat{\Sigma} = \hat{\Lambda}\hat{\Lambda}'$  and then transforming the covariances to correlations by dividing by the products of the respective standard deviations (the values in parentheses in Table 15.7). The result is

$$\mathbf{R} = \begin{matrix} 1 & & & & & & \\ 0.0979 & 1 & & & & & \\ -0.1680 & 0.83040 & 1 & & & & \\ 0.2907 & 0.00980 & 0.3983 & 1 & & & \\ -0.3180 & 0.04481 & -0.3266 & -0.8659 & 1 & & \\ 0.3176 & -0.48890 & -0.6622 & -0.3277 & -0.06073 & 1 & \\ -0.2700 & -0.10940 & -0.4253 & -0.7097 & 0.94190 & -0.08228 & 1. \end{matrix}$$

## 15.8 HIERARCHICAL LINEAR MODELS

Example 11.20 examined an application of a “two-level model,” or “hierarchical model,” for mortgage rates,

$$RM_{it} = \beta_{1i} + \beta_{2,i} J_{it} + \text{various terms relating to the mortgate} + \varepsilon_{it}.$$

The second level equation is

$$\begin{aligned} \beta_{2,i} = & \alpha_1 + \alpha_2 \text{GFA}_i + \alpha_3 \text{one-year treasury rate} + \alpha_4 \text{ten-year treasure rate} \\ & + \alpha_5 \text{credit risk} + \alpha_6 \text{prepayment risk} + \dots + u_i. \end{aligned}$$

Recent research in many fields has extended the idea of hierarchical modeling to the full set of parameters in the model. (Depending on the field studied, the reader may find these labeled “hierarchical models,” **mixed models**, “random parameters models,” or “random effects models.” The last of these generalizes our notion of random effects.) A two-level formulation of the model in (11.20) might appear as

$$\begin{aligned} y_{it} &= \mathbf{x}'_{it} \boldsymbol{\beta}_i + \varepsilon_{it}, \\ \boldsymbol{\beta}_i &= \boldsymbol{\beta} + \Delta \mathbf{z}_i + \mathbf{u}_i. \end{aligned}$$

(A three-level model is shown in Example 15.14.) This model retains the earlier stochastic specification but adds the measurement equation to the generation of the random parameters. In principle, this is actually only a minor extension of the model used thus far. The model of the previous section now becomes

$$y_{it} = \mathbf{x}'_{it} (\boldsymbol{\beta} + \Delta \mathbf{z}_i + \Lambda \mathbf{w}_i) + \varepsilon_{it},$$

which is essentially the same as our earlier model in (15.28)–(15.31) with the addition of product (interaction) terms of the form  $\delta_{kl} x_{itk} z_{il}$ , which suggests how it might be

## 640 PART III ♦ Estimation Methodology

estimated (simply by adding the interaction terms to the previous formulation.) In the template in (15-26), the term  $\sigma_u w_{ir}$  becomes  $\mathbf{x}'_i(\Delta \mathbf{z}_i + \Lambda \mathbf{w}_i)$ ,  $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_v, \lambda', \sigma_\varepsilon)'$  where  $\boldsymbol{\beta}'$  is a row vector composed of the rows of  $\Delta$ , and  $\lambda'$  is a row vector composed of the rows of  $\Lambda$ . The scalar term  $w_{ir}$  in the derivatives is replaced by a column vector of terms contained in  $(\mathbf{x}'_i \otimes \mathbf{z}_i, \mathbf{x}'_i \otimes \mathbf{w}_i)$ .

The hierarchical model can be extended in several useful directions. Recent analyses have expanded the model to accommodate multilevel stratification in data sets such as those we considered in the treatment of nested random effects in Section 14.9.6.b. A three-level model would appear as in the next example that relates to home sales,

$$\begin{aligned} y_{ijt} &= \mathbf{x}'_{ijt} \boldsymbol{\beta}_{ij} + \varepsilon_{it}, t = site, j = neighborhood, i = community, \\ \boldsymbol{\beta}_{ij} &= \boldsymbol{\beta}_i + \Delta \mathbf{z}_{ij} + \mathbf{u}_{ij} \\ \boldsymbol{\beta}_i &= \boldsymbol{\pi} + \Phi \mathbf{r}_i + \mathbf{v}_i. \end{aligned} \tag{15-35}$$

### Example 15.14 Hierarchical Linear Model of Home Prices

Beron, Murdoch, and Thayer (1999) used a hedonic pricing model to analyze the sale prices of 76,343 homes in four California counties: Los Angeles, San Bernardino, Riverside, and Orange. The data set is stratified into 2,185 census tracts and 131 school districts. Home prices are modeled using a three-level random parameters pricing model. (We will change their notation somewhat to make roles of the components of the model more obvious.) Let *site* denote the specific location (sale), *nei* denote the neighborhood, and *com* denote the community, the highest level of aggregation. The pricing equation is

$$\begin{aligned} \ln Price_{site,nei,com} &= \pi_{nei,com}^0 + \sum_{k=1}^K \pi_{nei,com}^k X_{k,site,nei,com} + \varepsilon_{site,nei,com}, \\ \pi_{nei,com}^k &= \beta_{com}^{0,k} + \sum_{l=1}^L \beta_{com}^{l,k} z_{k,nei,com} + r_{nei,com}^k, k = 0, \dots, K, \\ \beta_{com}^{l,k} &= \gamma^{0,l,k} + \sum_{m=1}^M \gamma^{m,l,k} e_{m,com} + u_{com}^{l,k}, l = 1, \dots, L. \end{aligned}$$

There are  $K$  level-one variables,  $x_k$ , and a constant in the main equation,  $L$  level-two variables,  $z_l$ , and a constant in the second-level equations, and  $M$  level-three variables,  $e_m$ , and a constant in the third-level equations. The variables in the model are as follows. The level-one variables define the hedonic pricing model,

**x** = house size, number of bathrooms, lot size, presence of central heating, presence of air conditioning, presence of a pool, quality of the view, age of the house, distance to the nearest beach.

Levels two and three are measured at the neighborhood and community levels

**z** = percentage of the neighborhood below the poverty line, racial makeup of the neighborhood, percentage of residents over 65, average time to travel to work

and

**e** = FBI crime index, average achievement test score in school district, air quality measure, visibility index.

The model is estimated by maximum simulated likelihood.

CHAPTER 15 ♦ Simulation-Based Estimation and Inference **641**

The **hierarchical linear model** analyzed in this section is also called a “mixed model” and “random parameters” model. Although the three terms are usually used interchangeably, each highlights a different aspect of the structural model in (15-35). The “hierarchical” aspect of the model refers to the layering of coefficients that is built into stratified and panel data structures, such as in Example (15-5). The random parameters feature is a signature feature of the model that relates to the modeling of heterogeneity across units in the sample. Note that the model in (15-35) [1] Ceron et al.’s application could be formulated without the random terms in the lower-level equations. This would then provide a convenient way to introduce interactions of variables in the linear regression model. The addition of the random component is motivated on precisely the basis that  $u_i$  appears in the familiar random effects model in Section 11.5 and (15-35). It is important to bear in mind, in all these structures, strict mean independence is maintained between  $u_i$  and all other variables in the model. In most treatments, we go yet a step further and assume a particular distribution for  $u_i$ , typically joint normal. Finally, the “mixed” model aspect of the specification relates to the underlying integration that removes the heterogeneity, for example, in (15-13). The unconditional estimated model is a mixture of the underlying models, where the weights in the mixture are provided by the underlying density of the random component.

## 15.9 NONLINEAR RANDOM PARAMETER MODELS

Most of the preceding applications have used the linear regression model to illustrate and demonstrate the procedures. However, the template used to build the model has no intrinsic features that limit it to the linear regression. The initial description of the model and the first example were applied to a nonlinear model, the Poisson regression. We will examine a random parameters binary choice model in the next section as well. This random parameters model has been used in a wide variety of settings. One [2] the most common is the multinomial choice models that we will discuss in Chapter 17.

The simulation-based random parameters estimator/model is extremely flexible. [See Train and McFadden (2000) for discussion.] The simulation method, in addition to extending the reach of a wide variety of model classes, also allows great flexibility in terms of the model itself. For example, constraining a parameter to have only one sign is a perennial issue. Use of a lognormal specification of the parameter,  $\beta_i = \exp(\beta + \sigma w_i)$  provides one method of restricting a random parameter to be consistent with a theoretical restriction. Researchers often find that the lognormal distribution produces unrealistically large values of the parameter. A model with parameters that vary in a restricted range that has found use is the random variable with symmetric about zero triangular distribution,

$$f(w) = \mathbf{1}[-a \leq w \leq 0](a + w)/a^2 + \mathbf{1}[0 < w \leq a](a - w)/a^2.$$

A draw from this distribution with  $a = 1$  can be computed as

$$w = \mathbf{1}[u \leq .5][(2u)^{1/2} - 1] + \mathbf{1}[u > .5][1 - (2(1 - u))^{1/2}],$$

where  $u$  is the  $U[0, 1]$  draw. Then, the parameter restricted to the range  $\beta \pm \lambda$  is obtained as  $\beta + \lambda w$ . A further refinement to restrict the sign of the random coefficient is to force  $\lambda = \beta$ , so that  $\beta_i$  ranges from 0 to  $2\lambda$ . [Discussion of this sort of model construction is

## 642 PART III ♦ Estimation Methodology

given in Train and Sonnier (2003) and Train (2009).] There is a large variety of methods for simulation that allow the model to be extended beyond the linear model and beyond the simple normal distribution for the random parameters.

Random parameters models have been implemented in several contemporary computer packages. The PROC MIXED package of routines in SAS uses a kind of generalized least squares for linear, Poisson, and binary choice models. The GLAMM program [Rabe-Hesketh, Skrondal, and Pickles (2005)] written for Stata uses quadrature methods for several models including linear, Poisson, and binary choice. The RPM and RPL procedures in LIMDEP/NLOGIT use the methods described here for linear, binary choice, censored data, multinomial, ordered choice, and several others. Finally, the MLWin package (<http://cmm.bristol.ac.uk/MLwiN/>) is a large implementation of some of the models discussed here. MLwin uses MCMC methods with noninformative priors to carry out maximum simulated likelihood estimation.

### 15.10 INDIVIDUAL PARAMETER ESTIMATES

In our analysis of the various random parameters specifications, we have focused on estimation of the population parameters,  $\beta$ ,  $\Delta$  and  $\Lambda$  in the model,

$$\beta_i = \beta + \Delta z_i + \Lambda w_i,$$

for example, in Example 15.13, where we estimated the parameters of the normal distribution of  $\beta_{\text{rem},i}$ . At a few points, it is noted that it might be useful to estimate the individual specific  $\beta_i$ . We did a similar exercise in analyzing the Hildreth/Houck/Swamy model in Section 11.11.1. The model is

$$\begin{aligned} y_i &= \mathbf{X}_i \beta_i + \epsilon_i \\ \beta_i &= \beta + u_i, \end{aligned}$$

where no restriction is placed on the correlation between  $u_i$  and  $\mathbf{X}_i$ . In this “fixed effects” case, we obtained a feasible GLS estimator for the population mean,  $\beta$ ,

$$\hat{\beta} = \sum_{i=1}^n \hat{\mathbf{W}}_i \mathbf{b}_i,$$

where

$$\hat{\mathbf{W}}_i = \left\{ \sum_{i=1}^n [\hat{\Gamma} + \hat{\sigma}_e^2 (\mathbf{X}'_i \mathbf{X}_i)^{-1}]^{-1} \right\}^{-1} [\hat{\Gamma} + \hat{\sigma}_e^2 (\mathbf{X}'_i \mathbf{X}_i)^{-1}]^{-1}$$

and

$$\mathbf{b}_i = (\mathbf{X}'_i \mathbf{X}_i)^{-1} \mathbf{X}'_i \mathbf{y}_i.$$

For each group, we then proposed an estimator of  $E[\beta_i]$  information in hand about group  $i$ ] as

$$\text{Est. } E[\beta_i | \mathbf{y}_i, \mathbf{X}_i] = \hat{\beta} + \hat{\mathbf{Q}}_i (\mathbf{b}_i - \hat{\beta})$$

where

$$\hat{\mathbf{Q}}_i = \left\{ [s_i^2 (\mathbf{X}'_i \mathbf{X}_i)]^{-1} + \hat{\Gamma}^{-1} \right\}^{-1} \hat{\Gamma}^{-1}. \quad (15-36)$$

CHAPTER 15 ♦ Simulation-Based Estimation and Inference **643**

The estimator of  $E[\beta_i | \mathbf{y}_i, \mathbf{X}_i]$  is equal to the estimator of the population mean plus a proportion of the difference between  $\hat{\beta}$  and  $\mathbf{b}_i$ . (The matrix  $\hat{\mathbf{Q}}_i$  is between  $\mathbf{0}$  and  $\mathbf{I}$ . If there were a single column in  $\mathbf{X}_i$ , then  $\hat{q}_i$  would equal  $(1/\hat{\gamma})/(1/\hat{\gamma}) + [1/(s_i^2/\mathbf{x}'_i \mathbf{x}_i)]$ .)

We obtain an analogous result for the mixed models we have examined in this chapter. From the initial model assumption, we have

$$f(y_{it} | \mathbf{x}_{it}, \beta_i, \theta)$$

where

$$\beta_i = \beta + \Delta \mathbf{z}_i + \Lambda \mathbf{w}_i \quad (15-37)$$

and  $\theta$  is any other parameters in the model, such as  $\sigma_e$  in the linear regression model. For a panel, since we are conditioning on  $\beta_i$ , that is, on  $\mathbf{w}_i$ , the  $T_i$  observations are independent, and it follows that

$$f(y_{i1}, y_{i2}, \dots, y_{iT_i} | \mathbf{X}_i, \beta_i, \theta) = f(\mathbf{y}_i | \mathbf{X}_i, \beta_i, \theta) = \prod_t f(y_{it} | \mathbf{x}_{it}, \beta_i, \theta). \quad (15-38)$$

This is the contribution of group  $i$  to the likelihood function (not its log) for the sample, given  $\beta_i$ ; that is, note that the log of this term is what appears in the simulated log likelihood function in (15-31) for the normal linear model and in (15-16) for the Poisson model. The marginal density for  $\beta_i$  is induced by the density of  $\mathbf{w}_i$  in (15-37). For example, if  $\mathbf{w}_i$  is joint normally distributed, then  $f(\beta_i) = N[\beta + \Delta \mathbf{z}_i, \Lambda \Lambda']$ . As we noted earlier in Section 15.9, some other distribution might apply. Write this generically as the marginal density of  $\beta_i$ ,  $f(\beta_i | \mathbf{z}_i, \Omega)$ , where  $\Omega$  is the parameters of the underlying distribution of  $\beta_i$ , for example  $(\beta, \Delta, \Lambda)$  in (15-37). Then, the joint distribution of  $\mathbf{y}_i$  and  $\beta_i$  is

$$f(\mathbf{y}_i, \beta_i | \mathbf{X}_i, \mathbf{z}_i, \theta, \Omega) = f(\mathbf{y}_i | \mathbf{X}_i, \beta_i, \theta) f(\beta_i | \mathbf{z}_i, \Omega).$$

We will now use Bayes's theorem to obtain  $f(\beta_i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \theta, \Omega)$ :

$$\begin{aligned} f(\beta_i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \theta, \Omega) &= \frac{f(\mathbf{y}_i | \mathbf{X}_i, \beta_i, \theta) f(\beta_i | \mathbf{z}_i, \Omega)}{f(\mathbf{y}_i | \mathbf{X}_i, \beta_i, \theta) f(\beta_i | \mathbf{z}_i, \Omega)} \\ &= \frac{\int_{\beta_i} f(\mathbf{y}_i | \mathbf{X}_i, \beta_i | \mathbf{z}_i, \theta, \Omega) d\beta_i}{\int_{\beta_i} f(\mathbf{y}_i | \mathbf{X}_i, \beta_i, \theta) f(\beta_i | \mathbf{z}_i, \Omega) d\beta_i}. \end{aligned}$$

The denominator of this ratio is the integral of the term that appears in the log-likelihood conditional on  $\beta_i$ . We will return momentarily to computation of the integral. We now have the conditional distribution of  $\beta_i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \theta, \Omega$ . The conditional expectation of  $\beta_i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \theta, \Omega$  is

$$E[\beta_i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \theta, \Omega] = \frac{\int_{\beta_i} f(\mathbf{y}_i | \mathbf{X}_i, \beta_i, \theta) f(\beta_i | \mathbf{z}_i, \Omega)}{\int_{\beta_i} f(\mathbf{y}_i | \mathbf{X}_i, \beta_i, \theta) f(\beta_i | \mathbf{z}_i, \Omega) d\beta_i}.$$

## 644 PART III ♦ Estimation Methodology

Neither of these integrals will exist in closed form. However, using the methods already developed in this chapter, we can compute them by simulation. The simulation estimator will be

$$\begin{aligned} \text{Est. } E[\beta_i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}] &= \frac{(1/R) \sum_{r=1}^R \hat{\beta}_{ir} \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \hat{\beta}_{ir}, \hat{\theta})}{(1/R) \sum_{r=1}^R \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \hat{\beta}_{ir}, \hat{\theta})} \\ &= \sum_{r=1}^R \hat{Q}_{ir} \hat{\beta}_{ir} \end{aligned} \quad (15-39)$$

where  $\hat{Q}_{ir}$  is defined in (15-20)–(15-21) and

$$\hat{\beta}_{ir} = \hat{\beta} + \hat{\Delta} \mathbf{z}_i + \Lambda \mathbf{w}_{ir}.$$

This can be computed after the estimation of the population parameters. (It may be more efficient to do this computation during the iterations, since everything needed to do the calculation will be in place and available while the iterations are proceeding.) For example, for the random parameters linear model, we will use

$$f(y_{it} | \mathbf{x}_{it}, \hat{\beta}_{ir}, \hat{\theta}) = \frac{1}{\hat{\sigma}_e \sqrt{2\pi}} \exp \left[ -\frac{(y_{it} - \mathbf{x}'_{it}(\hat{\beta} + \hat{\Delta} \mathbf{z}_i + \hat{\Lambda} \mathbf{w}_{ir}))^2}{2\hat{\sigma}_e^2} \right]. \quad (15-40)$$

We can also estimate the conditional variance of  $\beta_i$  by estimating first, one element at a time,  $E[\beta_{i,k}^2 | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}]$ , then, again one element at a time

$$\text{Est. } \text{Var}[\beta_{i,k} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}] = \frac{\{ \text{Est. } E[\beta_{i,k}^2 | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}] \} - \{ \text{Est. } E[\beta_{i,k} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}] \}^2}{\{ \text{Est. } E[\beta_{i,k} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\Omega}] \}^2}. \quad (15-41)$$

With the estimates of the conditional mean and conditional variance in hand, we can then compute the limits of an interval that resembles a confidence interval as the mean plus and minus two estimated standard deviations. This will construct an interval that contains at least 95 percent of the conditional distribution of  $\beta_i$ .

Some aspects worth noting about this computation are as follows:

- The preceding interval suggest is a classical (sampling-theory-based) counterpart to the highest posterior density interval that would be computed for  $\beta_i$  for a hierarchical Bayesian estimator.
- The conditional distribution from which  $\beta_i$  is drawn might not be symmetric or normal, so a symmetric interval of the mean plus and minus two standard deviations may pick up more or less than 95 percent of the actual distribution. This is likely to be a small effect. In any event, in any population, whether symmetric or not, the mean plus and minus two standard deviations will typically encompass at least 95 percent of the mass of the distribution.
- It has been suggested that this classical interval is too narrow because it does not account for the sampling variability of the parameter estimators used to construct it. But, the suggested computation should be viewed as a “point” estimate of the interval, not an interval estimate as such. Accounting for the sampling variability of the estimators might well suggest that the endpoints of the interval should be somewhat farther apart. The Bayesian interval that produces the same estimation

**CHAPTER 15 ♦ Simulation-Based Estimation and Inference 645**

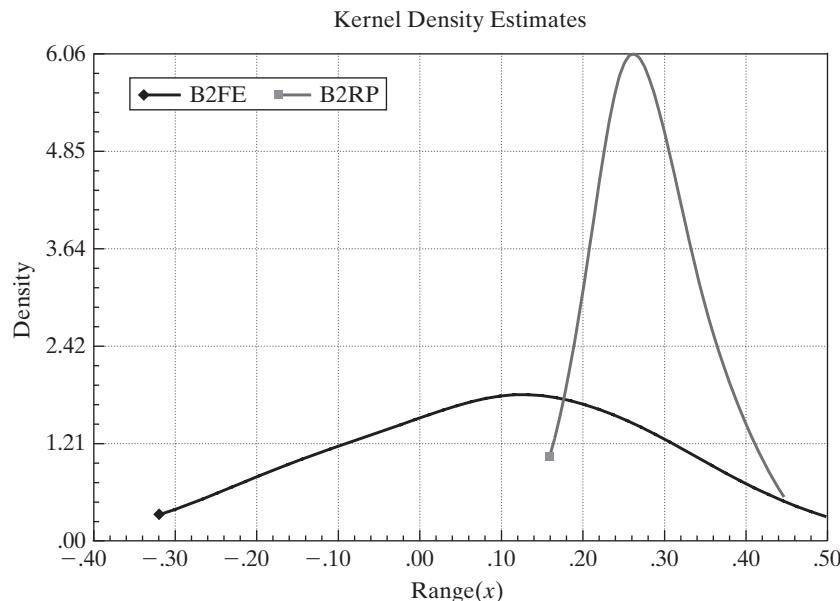
would be narrower because the estimator is posterior to, that is, applies only to the sample data.

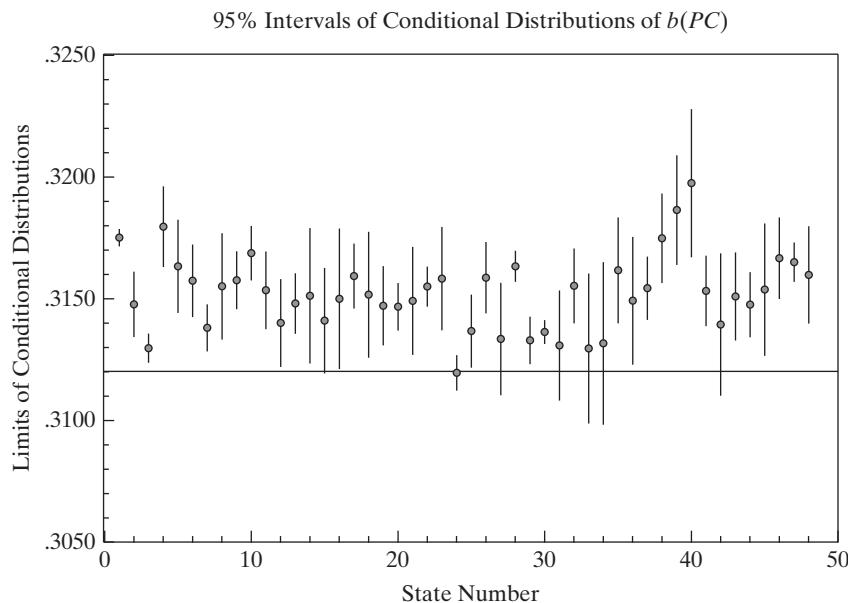
- Perhaps surprisingly so, even if the analysis departs from normal marginal distributions  $\beta_i$ , the sample distribution of the  $n$  estimated conditional means is  normal. Kernel estimators based on the  $n$  estimators, for example, can have variety of shapes.
- A common misperception found in the Bayesian and classical literatures alike is that the preceding produces an estimator of  $\beta_i$ . In fact, it is an estimator of conditional mean of the distribution from which  $\beta_i$  is an observation. By construction, for example, every individual with the same  $(y_i, X_i, z_i)$  has the same prediction even though the  $w_i$  and any other stochastic elements of the model, such as  $\varepsilon_i$ , will differ across individuals.

**Example 15.15 Individual State Estimates of Private Capital Coefficient**

Example 15.13 presents feasible GLS and maximum simulated likelihood estimates of Munnell's state production model. We have computed the estimates of  $E[\beta_{2j} | y_j, X_j]$  for the 48 states in the sample using (15-36) for the fixed effects estimates and (15-39) for the random effects estimates. Figures 15.6 and 15.7 examine the estimated coefficients for private capital. Figure 15.6 displays kernel density estimates for the population distributions based on the fixed and random effects estimates computed using (15-36) and (15-39). The much narrower distribution corresponds to the random effects estimates. The substantial overall difference of the distributions is presumably due in large part to the difference between the fixed effects and random effects assumptions. One might suspect on this basis that the random effects assumption is restrictive. Figure 15.7 shows the results based on the random parameters model, using (15-39) and (15-41) to compute the estimates. As expected, the range of variation of the estimators in the conditional distributions is much smaller than the overall range of variation shown in Figure 15.6.

**FIGURE 15.6** Kernel Density Estimates of Parameter Distributions.



**646 PART III ♦ Estimation Methodology**


**FIGURE 15.7** Estimates of Conditional Distributions for Private Capital Coefficient.

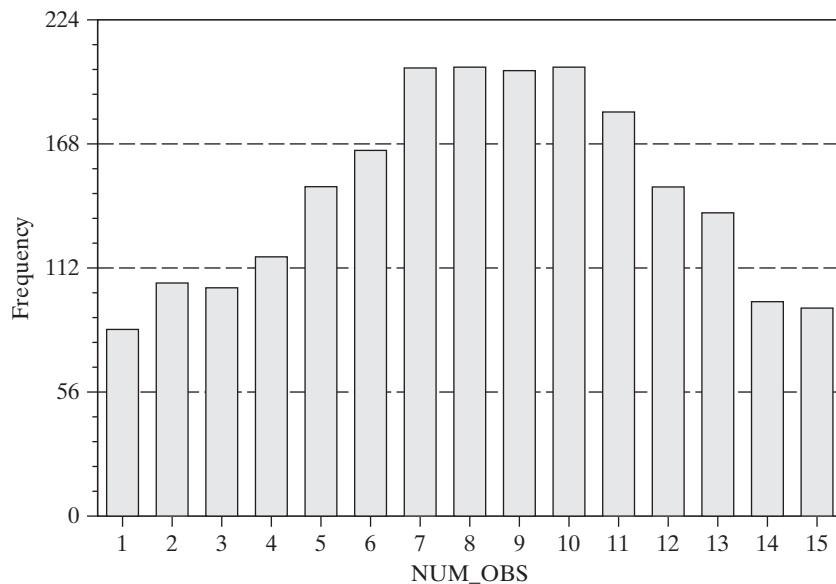
**Example 15.16 Mixed Linear Model for Wages**

Koop and Tobias (2004) analyzed a panel of 17,919 observations in their study of the relationship between wages and education, ability and family characteristics. (See the end of chapter applications in Chapters 3 and 5 and Appendix Table F3.2 for details on the location of the data.) The variables used in the analysis are

Person id	
Education	(time varying)
Log of hourly wage	(time varying)
Potential experience	(time varying)
Time trend	(time varying)
Ability	(time invariant)
Mother's education	(time invariant)
Father's education	(time invariant)
Dummy variable for residence in a broken home	(time invariant)
Number of siblings	(time invariant)

This is an unbalanced panel of 2,178 individuals; Figure 15.8 shows a frequency count of the numbers of observations in the sample. We will estimate the following hierarchical wage model

$$\begin{aligned} \ln \text{Wage}_{it} &= \beta_{1,i} + \beta_{2,i} \text{Education}_{it} + \beta_3 \text{Experience}_{it} + \beta_4 \text{Experience}_{it}^2 \\ &\quad + \beta_5 \text{Broken Home}_i + \beta_6 \text{Siblings}_i + \varepsilon_{it}, \\ \beta_{1,i} &= \alpha_{1,1} + \alpha_{1,2} \text{Ability}_i + \alpha_{1,3} \text{Mother's education}_i + \alpha_{1,4} \text{Father's education}_i + u_{1,i}, \\ \beta_{2,i} &= \alpha_{2,1} + \alpha_{2,2} \text{Ability}_i + \alpha_{2,3} \text{Mother's education}_i + \alpha_{2,4} \text{Father's education}_i + u_{2,i}. \end{aligned}$$

CHAPTER 15 ♦ Simulation-Based Estimation and Inference **647****FIGURE 15.8** Group Sizes for Wage Data Panel.

Estimates are computed using the maximum simulated likelihood method described in Sections 15.6.3 and 15.7. Estimates of the model parameters appear in Table 15.8. The four models in Table 15.8 are the pooled OLS estimates, the random effects model, and the random parameters models, first assuming that the random parameters are uncorrelated ( $\Gamma_{21} = 0$ ) and then allowing free correlation ( $\Gamma_{21} = \text{nonzero}$ ). The differences between the conventional and the robust standard errors in the pooled model are fairly large, which suggests the presence of latent common effects. The formal estimates of the random effects model confirm this. There are only minor differences between the FGLS and the ML estimates of the random effects model. But, the hypothesis of the pooled model is soundly rejected by the likelihood ratio test. The LM statistic [Section 11.5.4 and  $(11-3\sigma_{ij})^2 / (11-\sigma_{ij})$ ] is 11,709.7, which is far larger than the critical value of 3.84. So, the hypothesis of the pooled model is firmly rejected. The likelihood ratio statistic based on the MLEs is  $2(10,840.18 - (-885.674)) = 23,451.71$ , which produces the same conclusion. An alternative approach would be to test the hypothesis that  $\sigma_u^2 = 0$  using a Wald statistic—the standard  $t$  test. The software used for this exercise reparameterizes the log-likelihood in terms of  $\theta_1 = \sigma_u^2/\sigma_\varepsilon^2$  and  $\theta_2 = 1/\sigma_\varepsilon^2$ . One approach, based on the delta method (see Section 4.4.4), would be to estimate  $\sigma_u^2$  with the MLE of  $\theta_1/\theta_2$ . The asymptotic variance of this estimator would be estimated using Theorem 4.5. Alternatively, we might note that  $\sigma_\varepsilon^2$  must be positive in this model, so it is sufficient simply to test the hypothesis that  $\theta_1 = 0$ . Our MLE of  $\theta_1$  is 0.999206 and the estimated asymptotic standard error is 0.03934. Following this logic, then, the test statistic is  $0.999206/0.03934 = 25.397$ . This is far larger than the critical value of 1.96, so, once again, the hypothesis is rejected. We do note a problem with the LR and Wald tests. The hypothesis that  $\sigma_u^2 = 0$  produces a nonstandard test under the null hypothesis, because  $\sigma_u^2 = 0$  is on the boundary of the parameter space. Our standard theory for likelihood ratio testing (see Chapter 14) requires the restricted parameters to be in the interior of the parameter space, not on the edge. The distribution of the test statistic under the null hypothesis is not the familiar chi squared. This issue is confronted in Breusch and Pagan (1980) and Godfrey (1988) and analyzed at (great) length by Andrews (1998, 1999, 2000, 2001, 2002) and Andrews and Ploberger (1994, 1995). The simple expedient in this complex situation is to use the LM statistic, which remains consistent with the earlier conclusion.

**648 PART III ♦ Estimation Methodology****TABLE 15.8** Estimated Random Parameter Models

<i>Variable</i>	<i>Pooled OLS</i>		<i>Random Effects FGLS</i> [ <i>Random Effects MLE</i> ]		<i>Random</i> <i>Parameters</i>	<i>Random</i> <i>Parameters</i>
	<i>Estimate</i>	<i>Std.Err.</i> ( <i>Robust</i> )	<i>Estimate</i> [ <i>MLE</i> ]	<i>Std.Err.</i> [ <i>MLE</i> ]	<i>Estimate</i> ( <i>Std.Err.</i> )	<i>Estimate</i> ( <i>Std.Err.</i> )
Exp	0.04157	0.001819 (0.002242)	0.04698 [0.04715]	0.001468 [0.001481]	0.04758 (0.001108)	0.04802 (0.001118)
Exp <sup>2</sup>	-0.00144	0.0001002 (0.000126)	-0.00172 [-0.00172]	0.0000805 [0.000081]	-0.001750 (0.000063)	-0.001761 (0.0000631)
Broken	-0.02781	0.005296 (0.01074)	-0.03185 [-0.03224]	0.01089 [0.01172]	-0.01236 (0.003669)	-0.01980 (0.003534)
Sibs	-0.00120	0.0009143 (0.001975)	-0.002999 [-0.00310]	0.001925 [0.002071]	0.0000496 (0.000662)	-0.001953 (0.0006599)
Constant	0.09728	0.01589 (0.02783)	0.03281 [0.03306]	0.02438 [0.02566]	0.3277 (0.03803)	0.3935 (0.03778)
Ability					0.04232 (0.01064)	0.1107 (0.01077)
MEd					-0.01393 (0.0040)	-0.02887 (0.003990)
FEd					-0.007548 (0.003252)	0.002657 (0.003299)
$\sigma_{u1}$			0.172278 [0.18767]		0.004187 (0.001320)	0.5026
Educ	0.03854	0.001040 (0.002013)	0.04072 [0.04061]	0.001758 [0.001853]	0.01253 (0.003015)	0.007607 (0.002973)
Ability					-0.0002560 (0.000869)	-0.005316 (0.0008751)
MEd					0.001054 (0.000321)	0.002142 (0.0003165)
Fed					0.0007754 (0.000255)	0.00006752 (0.00001354)
$\sigma_{u2}$					0.01622 (0.000114)	0.03365
$\sigma_{u,12}$					0.0000	-0.01560
					0.0000	-0.92259
$\sigma_\varepsilon$	0.2542736		0.187017 [0.187742]		0.192741	0.1919182
$\Lambda_{11}$					0.004187 (0.001320)	0.5026 (0.008775)
$\Lambda_{21}$					0.0000 (0)	-0.03104 (0.0001114)
$\Lambda_{22}$					0.01622 (0.000113)	0.01298 (0.0006841)
ln L	-885.6740		[10480.18]		3550.594	3587.611

**CHAPTER 15 ♦ Simulation-Based Estimation and Inference 649**

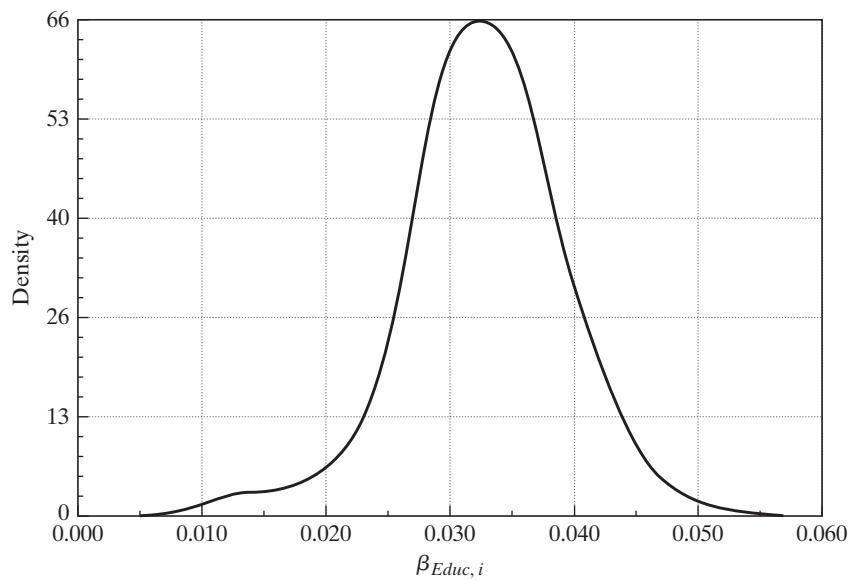
The third and fourth models in Table 15.8 present the mixed model estimates. The first of them imposes the restriction that  $\Gamma_{21} = 0$ , or that the two random parameters are uncorrelated. The second mixed model allows  $\Lambda_{21}$  to be a free parameter. The implied estimators for  $\sigma_{u1}$ ,  $\sigma_{u2}$  and  $\sigma_{u,21}$  are the elements of  $\Lambda\Lambda'$ , or

$$\begin{aligned}\sigma_{u1}^2 &= \Lambda_{11}^2, \\ \sigma_{u,21} &= \Lambda_{11}\Lambda_{21}, \\ \sigma_{u2}^2 &= \Lambda_{21}^2 + \Lambda_{22}^2.\end{aligned}$$

These estimates are shown separately in the table. Note that in all three random parameters models (including the random effects model which is equivalent to the mixed model with all  $\alpha_{1m} = 0$  save for  $\alpha_{1,1}$  and  $\alpha_{2,1}$  as well as  $\Lambda_{21} = \Lambda_{22} = 0.0$ ), the estimate of  $\sigma_e$  is relatively unchanged. The three models decompose the variation across groups in the parameters differently, but the overall variation of the dependent variable is largely the same.

The interesting coefficient in the model is  $\beta_{2,i}$ . Reading across the row for *Educ*, one might suspect that the random parameters model has washed out the impact of education, since the “coefficient” declines from 0.04072 to 0.007607. However, in the mixed models, the “mean” parameter,  $\alpha_{2,1}$ , is not the coefficient of interest. The coefficient on education in the model is  $\beta_{2,i} = \alpha_{2,1} + \alpha_{2,2} \text{Ability} + \beta_{2,3} \text{Mother's education} + \beta_{2,4} \text{Father's education} + u_{2,i}$ . A rough indication of the magnitude of this result can be seen by inserting the sample means for these variables, 0.052374, 11.4719, and 11.7092, respectively. With these values, the mean value for the education coefficient is approximately 0.0327. This is comparable, though somewhat smaller, than the estimates for the pooled and random effects model. Of course, variation in this parameter across the sample individuals was the objective of this specification. Figure 15.9 plots a kernel density estimate for the estimated conditional means for the 2,178 sample individuals. The figure shows the very wide range of variation in the sample estimates.

**FIGURE 15.9** Kernel Density Estimate for Education Coefficient.



## 650 PART III ♦ Estimation Methodology

### 15.11 MIXED MODELS AND LATENT CLASS MODELS

Sections 15.7–15.10 examined different approaches to modeling parameter heterogeneity. The fixed effects approach begun in Section 11.4 is extended to include the full set of regression coefficients in Section 11.11.1, where

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \\ \boldsymbol{\beta}_i &= \boldsymbol{\beta} + \mathbf{u}_i \end{aligned}$$

and no restriction is placed on  $E[\mathbf{u}_i | \mathbf{X}_i]$ . Estimation produces a feasible GLS estimate of  $\boldsymbol{\beta}$ . Estimation of  $\boldsymbol{\beta}$  begins with separate least squares estimation with each group,  $i$ —because of the correlation between  $\mathbf{u}_i$  and  $\mathbf{x}_{it}$ , the pooled estimator is not consistent. The efficient estimator of  $\boldsymbol{\beta}$  is then a mixture of the  $\mathbf{b}_i$ 's. We also examined an estimator of  $\boldsymbol{\beta}_i$ , using the optimal predictor from the conditional distributions, (15-39). The crucial assumption underlying the analysis is the possible correlation between  $\mathbf{X}_i$  and  $\mathbf{u}_i$ . We also considered two modifications of this random coefficients model. First, a restriction of the model in which some coefficients are nonrandom provides a useful simplification. The familiar fixed effects model of Section 11.4 is such a case, in which only the constant term varies across individuals. Second, we considered a hierarchical form of the model

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \Delta \mathbf{z}_i + \mathbf{u}_i. \quad (15-42)$$

This approach is applied to an analysis of mortgage rates in Example 11.20. [Plümper and Troeger's (2007) FEVD estimator examined in Section 11.4.5 is essentially this model as well.]

A second approach to random parameters modeling builds from the crucial assumption added to (15-42) that  $\mathbf{u}_i$  and  $\mathbf{X}_i$  are uncorrelated. The general model is defined in terms of the conditional density of the random variable,  $f(y_{it} | \mathbf{x}_{it}, \boldsymbol{\beta}_i, \boldsymbol{\theta})$  and the marginal density of the random coefficients,  $f(\boldsymbol{\beta}_i | \mathbf{z}_i, \boldsymbol{\Omega})$  in which  $\boldsymbol{\Omega}$  is the separate parameters of this distribution. This leads to the mixed models examined in this chapter. The random effects model that we examined in Section 11.5 and several other points is a special case in which only the constant term is random (like the fixed effects model). We also considered the specific case in which  $u_i$  is distributed normally with variance  $\sigma_u^2$ .

A third approach to modeling heterogeneity in parametric models is to use a discrete distribution, either as an approximation to an underlying continuous distribution, or as the model of the data generating process in its own right. (See Section 14.10.) This model adds to the preceding a nonparametric specification of the variation in  $\boldsymbol{\beta}_i$ ,

$$\text{Prob}(\boldsymbol{\beta}_i = \boldsymbol{\beta}_j | \mathbf{z}_i) = \pi_j, \quad j = 1, \dots, J.$$

A somewhat richer, semiparametric form that mimics (15-42) is

$$\text{Prob}(\boldsymbol{\beta}_i = \boldsymbol{\beta}_j | \mathbf{z}_i) = \pi_j(\mathbf{z}_i, \boldsymbol{\Omega}), \quad j = 1, \dots, J.$$

We continue to assume that the process generating variation in  $\boldsymbol{\beta}_i$  across individuals is independent of the process that produces  $\mathbf{X}_i$ —that is, in a broad sense, we retain the random effects approach. This latent class model is gaining popularity in the current

CHAPTER 15 ♦ Simulation-Based Estimation and Inference **651****TABLE 15.9** Estimated Random Parameters Model

	<i>Probit</i>	<i>RP Mean</i>	<i>RP Std. Dev.</i>	<i>Empirical Distn.</i>
Constant	-1.96 (0.23)	-3.91 (0.20)	2.70	-3.27 (0.57)
In Sales	0.18 (0.022)	0.36 (0.019)	0.28	0.32 (0.15)
Relative Size	1.07 (0.14)	6.01 (0.22)	5.99	3.33 (2.25)
Import	1.13 (0.15)	1.51 (0.13)	0.84	2.01 (0.58)
FDI	2.85 (0.40)	3.81 (0.33)	6.51	3.76 (1.69)
Productivity	-2.34 (0.72)	-5.10 (0.73)	13.03	-8.15 (8.29)
Raw materials	-0.28 (0.081)	-0.31 (0.075)	1.65	-0.18 (0.57)
Investment	0.19 (0.039)	0.27 (0.032)	1.42	0.27 (0.38)
ln L	-4114.05		-3498.654	

literature. In the last example of this chapter, we will examine a comparison of mixed and finite mixture models for a nonlinear model.

**Example 15.17 Maximum Simulated Likelihood Estimation of a Binary Choice Model**

Bertschek and Lechner (1998) analyzed the product innovations of a sample of German manufacturing firms. They used a probit model (Sections 17.2–17.4) to study firm innovations. The model is for  $\text{Prob}[y_{it} = 1 | \mathbf{x}_{it}, \boldsymbol{\beta}_i]$  where

$$y_{it} = 1 \text{ if firm } i \text{ realized a product innovation in year } t \text{ and } 0 \text{ if not.}$$

The independent variables in the model are

$X_{it,1}$  = constant

$X_{it,2}$  = log of sales

$X_{it,3}$  = relative size = ratio of employment in business unit to employment in the industry

$X_{it,4}$  = ratio of industry imports to (industry sales + imports)

$X_{it,5}$  = ratio of industry foreign direct investment to (industry sales + imports)

$X_{it,6}$  = productivity = ratio of industry value added to industry employment

$X_{it,7}$  = dummy variable indicating firm is in the raw materials sector

$X_{it,8}$  = dummy variable indicating firm is in the investment goods sector

The sample consists of 1,270 German firms observed for five years, 1984–1988. (See Appendix Table F15.1.) The density that enters the log-likelihood is

$$f(y_{it} | \mathbf{x}_{it}, \boldsymbol{\beta}_i) = \text{Prob}[y_{it} | \mathbf{x}'_{it} \boldsymbol{\beta}_i] = \Phi[(2y_{it} - 1)\mathbf{x}'_{it} \boldsymbol{\beta}_i], y_{it} = 0, 1,$$

where

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{v}_i, \mathbf{v}_i \sim N[\mathbf{0}, \Sigma].$$

To be consistent with Bertschek and Lechner (1998) we did not fit any firm specific time-invariant components in the main equation for  $\boldsymbol{\beta}_i$ .<sup>9</sup> Table 15.9 presents the estimated

<sup>9</sup> Apparently they did not use the second derivatives to compute the standard errors—we could not replicate these. Those shown in the Table 15.9 are our results.

## 652 PART III ♦ Estimation Methodology

**TABLE 15.10** Estimated Latent Class Model

	<i>Class 1</i>	<i>Class 2</i>	<i>Class 3</i>	<i>Posterior</i>
Constant	-2.32 (0.59)	-2.71 (0.69)	-8.97 (2.20)	-3.38 (2.14)
In Sales	0.32 (0.061)	0.23 (0.072)	0.57 (0.18)	0.34 (0.09)
Relative Size	4.38 (0.89)	0.72 (0.37)	1.42 (0.76)	2.58 (1.30)
Import	0.94 (0.37)	2.26 (0.53)	3.12 (1.38)	1.81 (0.74)
FDI	2.20 (1.16)	2.81 (1.11)	8.37 (1.93)	3.63 (1.98)
Productivity	-5.86 (2.70)	-7.70 (4.69)	-0.91 (6.76)	-5.48 (1.78)
Raw Materials	-0.11 (0.24)	-0.60 (0.42)	0.86 (0.70)	-0.08 (0.37)
Investment	0.13 (0.11)	0.41 (0.12)	0.47 (0.26)	0.29 (0.13)
ln <i>L</i>		-3503.55		
Class Prob (Prior)	0.469 (0.0352)	0.331 (0.0333)	0.200 (0.0246)	
Class Prob (Posterior)	0.469 (0.394)	0.331 (0.289)	0.200 (0.325)	
Pred. Count	649	366	255	

coefficients for the basic probit model in the first column. These are the values reported in the 1998 study. The estimates of the means,  $\beta$ , are shown in the second column. There appear to be large differences in the parameter estimates, although this can be misleading as there is large variation across the firms in the posterior estimates. The third column presents the square roots of the implied diagonal elements of  $\Sigma$  computed as the diagonal elements of  $\mathbf{C}\mathbf{C}'$ . These estimated standard deviations are for the underlying distribution of the parameter in the model—they are not estimates of the standard deviation of the sampling distribution of the estimator. That is shown for the mean parameter in the second column. The fourth column presents the sample means and standard deviations of the 1,270 estimated conditional estimates of the coefficients.

The latent class formulation developed in Section 14.10 provides an alternative approach for modeling latent parameter heterogeneity.<sup>10</sup> To illustrate the specification, we will reestimate the random parameters innovation model using a three-class latent class model. Estimates of the model parameters are presented in Table 15.10. The estimated conditionally unobserved means, which is comparable to the empirical means in the rightmost column in Table 17.4 for the random parameters model, are the sample average and standard deviation of the 1,270 firm-specific posterior mean parameter vectors. They are computed using  $\hat{\beta}_i = \sum_{j=1}^3 \hat{\pi}_{ij} \hat{\beta}_j$  where  $\hat{\pi}_{ij}$  is the conditional estimator of the class probabilities in (14-102). These estimates differ considerably from the probit model, but they are quite similar to the empirical means in Table 15.9. In each case, a confidence interval around the posterior mean contains the one-class pooled probit estimator. Finally, the (identical) prior and average of the sample posterior class probabilities are shown at the bottom of the table. The much larger empirical standard deviations reflect that the posterior estimates are based on aggregating the sample data and involve, as well, complicated functions of all the model parameters. The estimated numbers of class members are computed by assigning to each firm the predicted class associated with the highest posterior class probability.

<sup>10</sup> See Greene (2001) for a survey. For two examples, Nagin and Land (1993) employed the model to study age transitions through stages of criminal careers and Wang et al. (1998) and Wedel et al. (1993) used the Poisson regression model to study counts of patents.

## 15.12 SUMMARY AND CONCLUSIONS

This chapter has outlined several applications of simulation-assisted estimation and inference. The essential ingredient in any of these applications is a random number generator. We examined the most common method of generating what appear to be samples of random draws from a population—in fact, they are deterministic Markov chains that only appear to be random. Random number generators are used directly to obtain draws from the standard uniform distribution. The inverse probability transformation is then used to transform these to draws from other distributions. We examined several major applications involving random sampling:

- Random sampling, in the form of bootstrapping, allows us to infer the characteristics of the sampling distribution of an estimator, in particular its asymptotic variance. We used this result to examine the sampling variance of the median in random sampling from a nonnormal population. Bootstrapping is also a useful, robust method of constructing confidence intervals for parameters.
- Monte Carlo studies are used to examine the behavior of statistics when the precise sampling distribution of the statistic cannot be derived. We examined the behavior of a certain test statistic and of the maximum likelihood estimator in a fixed effects model.
- Many integrals that do not have closed forms can be transformed into expectations of random variables that can be sampled with a random number generator. This produces the technique of Monte Carlo integration. The technique of maximum simulated likelihood estimation allows the researcher to formulate likelihood functions (and other criteria such as moment equations) that involve expectations that can be integrated out of the function using Monte Carlo techniques. We used the method to fit random parameters models.

The techniques suggested here open up a vast range of applications of Bayesian statistics and econometrics in which the characteristics of a posterior distribution are deduced from random samples from the distribution, rather than brute force derivation of the analytic form. Bayesian methods based on this principle are discussed in the next chapter.

### **Key Terms and Concepts**

- Antithetic draws
- Block bootstrap
- Bootstrapping
- Cholesky decomposition
- Cholesky factorization
- Delta method
- Direct product
- Discrete uniform distribution
- Fundamental probability transformation
- Gauss–Hermite quadrature
- GHK smooth recursive stimulator
- Hadamard product
- Halton draws
- Hierarchical linear model
- Incidental parameters problem
- Kronecker product
- Markov chain
- Maximum stimulated likelihood
- Mixed model
- Monte Carlo integration
- Monte Carlo study
- Nonparametric bootstrap
- Paired bootstrap
- Parametric bootstrap
- Percentile method
- Period
- Poisson
- Power of a test
- Pseudo maximum likelihood estimator

## 654 PART III ♦ Estimation Methodology

- Pseudo-random number generator
- Random parameters
- Schur product
- Seed
- Simulation
- Size of a test
- Specificity
- Shuffling

### Exercises

1. The exponential distribution has density  $f(x) = \theta \exp(-\theta x)$ . How would you obtain a random sample of observations from an exponential population?
2. The Weibull population has survival function  $S(x) = \lambda p \exp(-(\lambda x)^p)$ . How would you obtain a random sample of observations from a Weibull population? (The survival function equals one minus the cdf.)
3. Suppose  $x$  and  $y$  are bivariate normally distributed with zero means, variances equal to one and correlation equal to  $\rho$ . Show how to use a Gibbs sample to estimate  $E[x^2 \exp(y) + y^2 \exp(x)]$ .

 Derive the first order conditions for nonlinear least squares estimation of the parameters in (15-2). How would you estimate the asymptotic covariance matrix for your estimator of  $\theta = (\beta, \sigma)$ ?

### Applications

1. Does the Wald statistic reject the null often? Construct a Monte Carlo study of the behavior of the Wald statistic for testing the hypothesis that  $\gamma$  equals zero in the model of Section 15.1. Recall, the Wald statistic is the square of the  $t$  ratio on the parameter in question. The procedure of the test is to reject the null hypothesis if the Wald statistic is greater than 3.84, the critical value from the chi-squared distribution with one degree of freedom. Replicate the study in Section 15.1 that is for all three assumptions about the underlying data.
2. A regression model that describes income as a function of experience is

$$\ln Income_i = \beta_1 + \beta_2 Experience_i + \beta_3 Experience_i^2 + \varepsilon_i.$$

The model implies that  $\ln Income$  is largest when  $\partial \ln Income / \partial Experience$  equals zero. The value of  $Experience$  at which this occurs is where  $\beta_2 + 2\beta_3 Experience = 0$ , or  $Experience^* = -\beta_2/\beta_3$ . Describe how to use the delta method to obtain a confidence interval for  $Experience^*$ . Now, describe how to use bootstrapping for this computation. A model of this sort using the Cornwell and Rupert data appears in Example 15.6. Using your proposals here, carry out the computations for that model using the Cornwell and Rupert data.

# 16

## BAYESIAN ESTIMATION AND INFERENCE

---

### 16.1 INTRODUCTION

The preceding chapters (and those that follow this one) are focused primarily on parametric specifications and classical estimation methods. These elements of the econometric method present a bit of a methodological dilemma for the researcher. They appear to straightjacket the analyst into a fixed and immutable specification of the model. But in any analysis, there is uncertainty as to the magnitudes, sometimes the signs and, at the extreme, even the meaning of parameters. It is rare that the presentation of a set of empirical results has not been preceded by at least some exploratory analysis. Proponents of the Bayesian methodology argue that the process of “estimation” is not one of deducing the values of fixed parameters, but rather, in accordance with the scientific method, one of continually updating and sharpening our subjective beliefs about the state of the world. Of course, this adherence to a subjective approach to model building is not necessarily a virtue. If one holds that “models” and “parameters” represent objective truths that the analyst seeks to discover, then the subjectivity of Bayesian methods may be less than perfectly comfortable.

Contemporary applications of Bayesian methods typically advance little of this theological debate. The modern practice of Bayesian econometrics is much more pragmatic. As we will see in several of the following examples, Bayesian methods have produced some remarkably efficient solutions to difficult estimation problems. Researchers often choose the techniques on practical grounds, rather than in adherence to their philosophical basis; indeed, for some, the Bayesian estimator is merely an algorithm.<sup>1</sup>

Bayesian methods have been employed by econometricians since well before Zellner’s classic (1971) presentation of the methodology to economists, but until fairly recently, were more or less at the margin of the field. With recent advances in technique (notably the Gibbs sampler) and the advance of computer software and hardware that has made simulation-based estimation routine, Bayesian methods that rely heavily on both have become widespread throughout the social sciences. There are libraries of work on Bayesian econometrics a rapidly expanding applied

---

<sup>1</sup>For example, from the home web site of MLWin, a widely used program for multilevel (random parameters) modeling, <http://www.cmm.bris.ac.uk/MLwiN/features/mcmc.shtml>, we find “Markov Chain Monte Carlo (MCMC) methods allow Bayesian models to be fitted, where prior distributions for the model parameters are specified. By default *MLwiN* sets diffuse priors which can be used to approximate maximum likelihood estimation.” Train (2001) is an interesting application that compares Bayesian and classical estimators of a random parameters model.

## 656 PART III ♦ Estimation Methodology

literature.<sup>2</sup> This chapter will introduce the vocabulary and techniques of Bayesian econometrics. Section 16.2 lays out the essential foundation for the method. The canonical application, the linear regression model, is developed in Section 16.3. Section 16.4 continues the methodological development. The fundamental tool of contemporary Bayesian econometrics, the Gibbs sampler, is presented in Section 16.5. Three applications and several more limited examples are presented in Sections 16.6, 16.7, and 16.8. Section 16.6 shows how to use the Gibbs sampler to estimate the parameters of a probit model without maximizing the likelihood function. This application also introduces the technique of data augmentation. Bayesian counterparts to the panel data random and fixed effects models are presented in Section 16.7. A hierarchical Bayesian treatment of the random parameters model is presented in Section 16.8 with a comparison to the classical treatment of the same model. Some conclusions are drawn in Section 16.9. The presentation here is nontechnical. A much more extensive entry level presentation is given by Lancaster (2004). Intermediate-level presentations appear in Cameron and Trivedi (2005, Chapter 13), and Koop (2003). A more challenging treatment is offered in Geweke (2005). The other sources listed in footnote 2 are oriented to applications.

### 16.2 BAYES THEOREM AND THE POSTERIOR DENSITY

The centerpiece of the Bayesian methodology is the **Bayes's theorem**: for events  $A$  and  $B$ , the conditional probability of event  $A$  given that  $B$  has occurred is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \quad (16-1)$$

Paraphrased for our applications here, we would write

$$P(\text{parameters} | \text{data}) = \frac{P(\text{data} | \text{parameters})P(\text{parameters})}{P(\text{data})}.$$

In this setting, the data are viewed as constants whose distributions do not involve the parameters of interest. For the purpose of the study, we treat the data as only a fixed set of additional information to be used in updating our beliefs about the parameters. Note the similarity to (12-1). Thus, we write

$$\begin{aligned} P(\text{parameters} | \text{data}) &\propto P(\text{data} | \text{parameters})P(\text{parameters}) \\ &= \text{Likelihood function} \times \text{Prior density}. \end{aligned} \quad (16-2)$$

The symbol  $\propto$  means “is proportional to.” In the preceding equation, we have dropped the marginal density of the data, so what remains is not a proper density until it is scaled by what will be an inessential proportionality constant. The first term on the right is the joint distribution of the observed random variables  $\mathbf{y}$ , given the parameters. As we

<sup>2</sup>Recent additions to the dozens of books on the subject include Gelman et al. (2004), Geweke (2005), Gill (2002), Koop (2003), Lancaster (2004), Congdon (2005), and Rossi et al. (2005). Readers with a historical bent will find Zellner (1971) and Leamer (1978) worthwhile reading. There are also many methodological surveys. Poirier and Tobias (2006) as well as Poirier (1988, 1995) sharply focus the nature of the methodological distinctions between the classical (frequentist) and Bayesian approaches.

## CHAPTER 16 ♦ Bayesian Estimation and Inference 657

shall analyze it here, this distribution is the normal distribution we have used in our previous analysis—see (12-1). The second term is the **prior beliefs** of the analyst. The left-hand side is the **posterior density** of the parameters, given the current body of data, or our *revised* beliefs about the distribution of the parameters after “seeing” the data. The posterior is a mixture of the prior information and the “current information,” that is, the data. Once obtained, this posterior density is available to be the prior density function when the next body of data or other usable information becomes available. The principle involved, which appears nowhere in the classical analysis, is one of continual accretion of knowledge about the parameters.

Traditional Bayesian estimation is heavily parameterized. The prior density and the likelihood function are crucial elements of the analysis, and both must be fully specified for estimation to proceed. The Bayesian “estimator” is the mean of the posterior density of the parameters, a quantity that is usually obtained either by integration (when closed forms exist), approximation of integrals [numerical techniques, or by Monte Carlo methods, which are discussed in Section 15.3.

**Example 16.1 Bayesian Estimation of a Probability**

Consider estimation of the probability that a production process will produce a defective product. In case 1, suppose the sampling design is to choose  $N = 25$  items from the production line and count the number of defectives. If the probability that any item is defective is a constant  $\theta$  between zero and one, then the likelihood for the sample of data is

$$L(\theta | \text{data}) = \theta^D(1 - \theta)^{25-D},$$

where  $D$  is the number of defectives, say, 8. The maximum likelihood estimator of  $\theta$  will be  $p = D/25 = 0.32$ , and the asymptotic variance of the maximum likelihood estimator is estimated by  $p(1 - p)/25 = 0.008704$ .

Now, consider a Bayesian approach to the same analysis. The posterior density is obtained by the following reasoning:

$$\begin{aligned} p(\theta | \text{data}) &= \frac{p(\theta, \text{data})}{p(\text{data})} = \frac{p(\theta, \text{data})}{\int_{\theta} p(\theta, \text{data}) d\theta} = \frac{p(\text{data} | \theta) p(\theta)}{p(\text{data})} \\ &= \frac{\text{Likelihood}(\text{data} | \theta) \times p(\theta)}{p(\text{data})} \end{aligned}$$

where  $p(\theta)$  is the prior density assumed for  $\theta$ . [We have taken some license with the terminology, since the likelihood function is conventionally defined as  $L(\theta | \text{data})$ .] Inserting the results of the sample first drawn, we have the posterior density:

$$p(\theta | \text{data}) = \frac{\theta^D(1 - \theta)^{N-D} p(\theta)}{\int_{\theta} \theta^D(1 - \theta)^{N-D} p(\theta) d\theta}.$$

What follows depends on the assumed prior for  $\theta$ . Suppose we begin with a “noninformative” prior that treats all *allowable* values of  $\theta$  as equally likely. This would imply a uniform distribution over  $(0,1)$ . Thus,  $p(\theta) = 1$ ,  $0 \leq \theta \leq 1$ . The denominator with this assumption is a beta integral (see Section E2.3) with parameters  $a = D + 1$  and  $b = N - D + 1$ , so the posterior density is

$$p(\theta | \text{data}) = \frac{\theta^D(1 - \theta)^{N-D}}{\left( \frac{\Gamma(D+1)\Gamma(N-D+1)}{\Gamma(D+1+N-D+1)} \right)} = \frac{\Gamma(N+2)\theta^D(1-\theta)^{N-D}}{\Gamma(D+1)\Gamma(N-D+1)}.$$

This is the density of a random variable with a beta distribution with parameters  $(\alpha, \beta) = (D+1, N-D+1)$ . (See Section B.4.6.) The mean of this random variable is  $(D+1)/(N+2) = 9/27 = 0.3333$  (as opposed to 0.32, the MLE). The posterior variance is  $[(D+1)/(N-D+1)]/[(N+3)(N+2)^2] = 0.007936$ .

## 658 PART III ♦ Estimation Methodology

There is a loose end in this example. If the uniform prior were noninformative, that would mean that the only information we had was in the likelihood function. Why didn't the Bayesian estimator and the MLE coincide? The reason is that the uniform prior over  $[0,1]$  is not really noninformative. It did introduce the information that  $\theta$  must fall in the unit interval. The prior mean is 0.5 and the prior variance is  $1/12$ . The posterior mean is an average of the MLE and the prior mean. Another less than obvious aspect of this result is the smaller variance of the Bayesian estimator. The principle that lies behind this (aside from the fact that the prior did in fact introduce some certainty in the estimator) is that the Bayesian estimator is conditioned on the specific sample data. The theory behind the classical MLE implies that it averages over the entire population that generates the data. This will always introduce a greater degree of "uncertainty" in the classical estimator compared to its Bayesian counterpart.

### 16.3 BAYESIAN ANALYSIS OF THE CLASSICAL REGRESSION MODEL

The complexity of the algebra involved in Bayesian analysis is often extremely burdensome. For the linear regression model, however, many fairly straightforward results have been obtained. To provide some of the flavor of the techniques, we present the full derivation only for some simple cases. In the interest of brevity, and to avoid the burden of excessive algebra, we refer the reader to one of the several sources that present the full derivation of the more complex cases.<sup>3</sup>

The classical normal regression model we have analyzed thus far is constructed around the conditional multivariate normal distribution  $N[\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}]$ . The interpretation is different here. In the sampling theory setting, this distribution embodies the information about the observed sample data *given* the assumed distribution and the fixed, albeit unknown, parameters of the model. In the Bayesian setting, this function summarizes the information that a particular realization of the data provides about the assumed distribution of the model parameters. To underscore that idea, we rename this joint density the *likelihood for  $\boldsymbol{\beta}$  and  $\sigma^2$  given the data*, so

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) = [2\pi\sigma^2]^{-n/2} e^{-[(1/(2\sigma^2))(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]}. \quad (16-3)$$

For purposes of the following results, some reformulation is useful. Let  $d = n - K$  (the degrees of freedom parameter), and substitute

$$\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y} - \mathbf{X}\mathbf{b} - \mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) = \mathbf{e} - \mathbf{X}(\boldsymbol{\beta} - \mathbf{b})$$

in the exponent. Expanding this produces

$$\left(-\frac{1}{2\sigma^2}\right)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \left(-\frac{1}{2}ds^2\right)\left(\frac{1}{\sigma^2}\right) - \frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\left(\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}\right)(\boldsymbol{\beta} - \mathbf{b}).$$

After a bit of manipulation (note that  $n/2 = d/2 + K/2$ ), the likelihood may be written

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) \\ = [2\pi]^{-d/2} [\sigma^2]^{-d/2} e^{-(d/2)(s^2/\sigma^2)} [2\pi]^{-K/2} [\sigma^2]^{-K/2} e^{-(1/2)(\boldsymbol{\beta}-\mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\mathbf{b})}. \end{aligned}$$

<sup>3</sup>These sources include Judge et al. (1982, 1985), Maddala (1977a), Mittelhammer et al. (2000), and the canonical reference for econometricians, Zellner (1971). A remarkable feature of the current literature is the degree to which the analytical components have become ever simpler while the applications have become progressively more complex. This will become evident in Sections 16.5–16.7.

## CHAPTER 16 ♦ Bayesian Estimation and Inference 659

This density embodies all that we have to learn about the parameters from the observed data. Because the data are taken to be constants in the joint density, we may multiply this joint density by the (very carefully chosen), inessential (because it does not involve  $\beta$  or  $\sigma^2$ ) constant function of the observations,

$$A = \frac{\left(\frac{d}{2}s^2\right)^{(d/2)+1}}{\Gamma\left(\frac{d}{2} + 1\right)} [2\pi]^{(d/2)} |\mathbf{X}'\mathbf{X}|^{-1/2}.$$

For convenience, let  $v = d/2$ . Then, multiplying  $L(\beta, \sigma^2 | \mathbf{y}, \mathbf{X})$  by  $A$  gives

$$\begin{aligned} L(\beta, \sigma^2 | \mathbf{y}, \mathbf{X}) &\propto \frac{[vs^2]^{v+1}}{\Gamma(v+1)} \left(\frac{1}{\sigma^2}\right)^v e^{-vs^2(1/\sigma^2)} [2\pi]^{-K/2} |\sigma^2(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2} \\ &\quad \times e^{-(1/2)(\beta-\mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\beta-\mathbf{b})}. \end{aligned} \quad (16-4)$$

The likelihood function is proportional to the product of a gamma density for  $z = 1/\sigma^2$  with parameters  $\lambda = vs^2$  and  $P = v + 1$  [see (B-39); this is an **inverted gamma distribution**] and a  $K$ -variate normal density for  $\beta | \sigma^2$  with mean vector  $\mathbf{b}$  and covariance matrix  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ . The reason will be clear shortly.

### 16.3.1 ANALYSIS WITH A NONINFORMATIVE PRIOR

The departure point for the Bayesian analysis of the model is the specification of a **prior distribution**. This distribution gives the analyst's prior beliefs about the parameters of the model. One of two approaches is generally taken. If no prior information is known about the parameters, then we can specify a **noninformative prior** that reflects that. We do this by specifying a "flat" prior for the parameter in question:<sup>4</sup>

$$g(\text{parameter}) \propto \text{constant}.$$

There are different ways that one might characterize the lack of prior information. The implication of a flat prior is that within the range of valid values for the parameter, all intervals of equal length—hence, in principle, all values—are equally likely. The second possibility, an **informative prior**, is treated in the next section. The posterior density is the result of combining the likelihood function with the prior density. Because it pools the full set of information available to the analyst, *once the data have been drawn*, the posterior density would be interpreted the same way the prior density was before the data were obtained.

To begin, we analyze the case in which  $\sigma^2$  is assumed to be known. This assumption is obviously unrealistic, and we do so only to establish a point of departure. Using Bayes's theorem, we construct the posterior density,

$$f(\beta | \mathbf{y}, \mathbf{X}, \sigma^2) = \frac{L(\beta | \sigma^2, \mathbf{y}, \mathbf{X})g(\beta | \sigma^2)}{f(\mathbf{y})} \propto L(\beta | \sigma^2, \mathbf{y}, \mathbf{X})g(\beta | \sigma^2),$$

---

<sup>4</sup>That this "improper" density might not integrate to one is only a minor difficulty. Any constant of integration would ultimately drop out of the final result. See Zellner (1971, pp. 41–53) for a discussion of noninformative priors.

## 660 PART III ♦ Estimation Methodology

assuming that the distribution of  $\mathbf{X}$  does not depend on  $\boldsymbol{\beta}$  or  $\sigma^2$ . Because  $g(\boldsymbol{\beta} | \sigma^2) \propto$  a constant, this density is the one in (16-4). For now, write

$$f(\boldsymbol{\beta} | \sigma^2, \mathbf{y}, \mathbf{X}) \propto h(\sigma^2) [2\pi]^{-K/2} |\sigma^2(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2} e^{-(1/2)(\boldsymbol{\beta}-\mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\mathbf{b})}, \quad (16-5)$$

where

$$h(\sigma^2) = \frac{[vs^2]^{v+1}}{\Gamma(v+1)} \left[ \frac{1}{\sigma^2} \right]^v e^{-vs^2(1/\sigma^2)}. \quad (16-6)$$

For the present, we treat  $h(\sigma^2)$  simply as a constant that involves  $\sigma^2$ , not as a probability density; (16-5) is *conditional* on  $\sigma^2$ . Thus, the posterior density  $f(\boldsymbol{\beta} | \sigma^2, \mathbf{y}, \mathbf{X})$  is proportional to a multivariate normal distribution with mean  $\mathbf{b}$  and covariance matrix  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ .

This result is familiar, but it is interpreted differently in this setting. First, we have combined our prior information about  $\boldsymbol{\beta}$  (in this case, no information) and the sample information to obtain a *posterior distribution*. Thus, on the basis of the sample data in hand, we obtain a distribution for  $\boldsymbol{\beta}$  with mean  $\mathbf{b}$  and covariance matrix  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ . The result is dominated by the sample information, as it should be if there is no prior information. In the absence of any prior information, the mean of the posterior distribution, which is a type of Bayesian point estimate, is the sampling theory estimator.

To generalize the preceding to an unknown  $\sigma^2$ , we specify a noninformative prior distribution for  $\ln \sigma$  over the entire real line.<sup>5</sup> By the change of variable formula, if  $g(\ln \sigma)$  is constant, then  $g(\sigma^2)$  is proportional to  $1/\sigma^2$ .<sup>6</sup> Assuming that  $\boldsymbol{\beta}$  and  $\sigma^2$  are independent, we now have the noninformative joint prior distribution:

$$g(\boldsymbol{\beta}, \sigma^2) = g_{\boldsymbol{\beta}}(\boldsymbol{\beta}) g_{\sigma^2}(\sigma^2) \propto \frac{1}{\sigma^2}.$$

We can obtain the **joint posterior distribution** for  $\boldsymbol{\beta}$  and  $\sigma^2$  by using

$$f(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) = L(\boldsymbol{\beta} | \sigma^2, \mathbf{y}, \mathbf{X}) g_{\sigma^2}(\sigma^2) \propto L(\boldsymbol{\beta} | \sigma^2, \mathbf{y}, \mathbf{X}) \times \frac{1}{\sigma^2}. \quad (16-7)$$

For the same reason as before, we multiply  $g_{\sigma^2}(\sigma^2)$  by a well-chosen constant, this time  $vs^2 \Gamma(v+1) / \Gamma(v+2) = vs^2/(v+1)$ . Multiplying (16-5) by this constant times  $g_{\sigma^2}(\sigma^2)$  and inserting  $h(\sigma^2)$  gives the joint posterior for  $\boldsymbol{\beta}$  and  $\sigma^2$ , given  $\mathbf{y}$  and  $\mathbf{X}$ :

$$\begin{aligned} f(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) &\propto \frac{[vs^2]^{v+2}}{\Gamma(v+2)} \left[ \frac{1}{\sigma^2} \right]^{v+1} e^{-vs^2(1/\sigma^2)} [2\pi]^{-K/2} |\sigma^2(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2} \\ &\quad \times e^{-(1/2)(\boldsymbol{\beta}-\mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\mathbf{b})}. \end{aligned}$$

To obtain the marginal posterior distribution for  $\boldsymbol{\beta}$ , it is now necessary to integrate  $\sigma^2$  out of the joint distribution (and vice versa to obtain the marginal distribution for  $\sigma^2$ ). By collecting the terms,  $f(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X})$  can be written as

$$f(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) \propto A \times \left( \frac{1}{\sigma^2} \right)^{P-1} e^{-\lambda(1/\sigma^2)},$$

<sup>5</sup>See Zellner (1971) for justification of this prior distribution.

<sup>6</sup>Many treatments of this model use  $\sigma$  rather than  $\sigma^2$  as the parameter of interest. The end results are identical. We have chosen this parameterization because it makes manipulation of the likelihood function with a gamma prior distribution especially convenient. See Zellner (1971, pp. 44–45) for discussion.

CHAPTER 16 ♦ Bayesian Estimation and Inference **661**

where

$$A = \frac{[vs^2]^{v+2}}{\Gamma(v+2)} [2\pi]^{-K/2} |(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2},$$

$$P = v + 2 + K/2 = (n - K)/2 + 2 + K/2 = (n + 4)/2,$$

and

$$\lambda = vs^2 + \frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})' \mathbf{X}' \mathbf{X} (\boldsymbol{\beta} - \mathbf{b}),$$

so the marginal posterior distribution for  $\boldsymbol{\beta}$  is

$$\int_0^\infty f(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) d\sigma^2 \propto A \int_0^\infty \left( \frac{1}{\sigma^2} \right)^{P-1} e^{-\lambda(1/\sigma^2)} d\sigma^2.$$

To do the integration, we have to make a change of variable;  $d(1/\sigma^2) = -(1/\sigma^2)^2 d\sigma^2$ , so  $d\sigma^2 = -(1/\sigma^2)^{-2} d(1/\sigma^2)$ . Making the substitution—the sign of the integral changes twice, once for the Jacobian and back again because the integral from  $\sigma^2 = 0$  to  $\infty$  is the negative of the integral from  $(1/\sigma^2) = 0$  to  $\infty$ —we obtain

$$\begin{aligned} \int_0^\infty f(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) d\sigma^2 &\propto A \int_0^\infty \left( \frac{1}{\sigma^2} \right)^{P-3} e^{-\lambda(1/\sigma^2)} d\left( \frac{1}{\sigma^2} \right) \\ &= A \times \frac{\Gamma(P-2)}{\lambda^{P-2}}. \end{aligned}$$

Reinserting the expressions for  $A$ ,  $P$ , and  $\lambda$  produces

$$f(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) \propto \frac{[vs^2]^{v+2} \Gamma(v+K/2)}{\Gamma(v+2)} [2\pi]^{-K/2} |\mathbf{X}'\mathbf{X}|^{-1/2} \frac{1}{[vs^2 + \frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})' \mathbf{X}' \mathbf{X} (\boldsymbol{\beta} - \mathbf{b})]^{v+K/2}}. \quad (16-8)$$

This density is proportional to a **multivariate *t* distribution**<sup>7</sup> and is a generalization of the familiar univariate distribution we have used at various points. This distribution has a degrees of freedom parameter,  $d = n - K$ , mean  $\mathbf{b}$ , and covariance matrix  $(d/(d-2)) \times [s^2(\mathbf{X}'\mathbf{X})^{-1}]$ . Each element of the  $K$ -element vector  $\boldsymbol{\beta}$  has a marginal distribution that is the univariate *t* distribution with degrees of freedom  $n - K$ , mean  $b_k$ , and variance equal to the  $k$ th diagonal element of the covariance matrix given earlier. Once again, this is the same as our sampling theory result. The difference is a matter of interpretation. In the current context, the estimated distribution is for  $\boldsymbol{\beta}$  and is centered at  $\mathbf{b}$ .

### 16.3.2 ESTIMATION WITH AN INFORMATIVE PRIOR DENSITY

Once we leave the simple case of noninformative priors, matters become quite complicated, both at a practical level and, methodologically, in terms of just where the prior comes from. The integration of  $\sigma^2$  out of the posterior in (16-7) is complicated by itself. It is made much more so if the prior distributions of  $\boldsymbol{\beta}$  and  $\sigma^2$  are at all involved. Partly to offset these difficulties, researchers usually use what is called a **conjugate prior**, which

<sup>7</sup>See, for example, Judge et al. (1985) for details. The expression appears in Zellner (1971, p. 67). Note that the exponent in the denominator is  $v + K/2 = n/2$ .

## 662 PART III ♦ Estimation Methodology

is one that has the same form as the conditional density and is therefore amenable to the integration needed to obtain the marginal distributions.<sup>8</sup>

### **Example 16.2 Estimation with a Conjugate Prior**

We continue Example 16.1, but we now assume a conjugate prior. For likelihood functions involving proportions, the beta prior is a common device, for reasons that will emerge shortly. The beta prior is

$$p(\theta) = \frac{\Gamma(\alpha + \beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}.$$

Then, the posterior density becomes

$$\frac{\theta^D(1-\theta)^{N-D} \frac{\Gamma(\alpha + \beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}}{\int_0^1 \theta^D(1-\theta)^{N-D} \frac{\Gamma(\alpha + \beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} d\theta} = \frac{\theta^{D+\alpha-1}(1-\theta)^{N-D+\beta-1}}{\int_0^1 \theta^{D+\alpha-1}(1-\theta)^{N-D+\beta-1} d\theta}.$$

The posterior density is, once again, a beta distribution, with parameters  $(D + \alpha, N - D + \beta)$ . The posterior mean is

$$E[\theta | \text{data}] = \frac{D + \alpha}{N + \alpha + \beta}.$$

(Our previous choice of the uniform density was equivalent to  $\alpha = \beta = 1$ .) Suppose we choose a prior that conforms to a prior mean of 0.5, but with less mass near zero and one than in the center, such as  $\alpha = \beta = 2$ . Then, the posterior mean would be  $(8 + 2)/(25 + 3) = 0.33571$ . (This is yet larger than the previous estimator. The reason is that the prior variance is now smaller than  $1/12$ , so the prior mean, still 0.5, receives yet greater weight than it did in the previous example.)

Suppose that we assume that the prior beliefs about  $\beta$  may be summarized in a  $K$ -variate normal distribution with mean  $\beta_0$  and variance matrix  $\Sigma_0$ . Once again, it is illuminating to begin with the case in which  $\sigma^2$  is assumed to be known. Proceeding in exactly the same fashion as before, we would obtain the following result: The posterior density of  $\beta$  conditioned on  $\sigma^2$  and the data will be normal with

$$\begin{aligned} E[\beta | \sigma^2, \mathbf{y}, \mathbf{X}] &= \{\Sigma_0^{-1} + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\}^{-1} \{\Sigma_0^{-1}\beta_0 + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{b}\} \\ &= \mathbf{F}\beta_0 + (\mathbf{I} - \mathbf{F})\mathbf{b}, \end{aligned} \quad (16-9)$$

where

$$\begin{aligned} \mathbf{F} &= \{\Sigma_0^{-1} + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\}^{-1} \Sigma_0^{-1} \\ &= \{[\text{prior variance}]^{-1} + [\text{conditional variance}]^{-1}\}^{-1} [\text{prior variance}]^{-1}. \end{aligned} \quad (16-10)$$

This vector is a matrix weighted average of the prior and the least squares (sample) coefficient estimates, where the weights are the inverses of the prior and the conditional

<sup>8</sup>Our choice of noninformative prior for  $\ln \sigma$  led to a convenient prior for  $\sigma^2$  in our derivation of the posterior for  $\beta$ . The idea that the prior can be specified arbitrarily in whatever form is mathematically convenient is very troubling; it is supposed to represent the accumulated prior belief about the parameter. On the other hand, it could be argued that the conjugate prior is the posterior of a previous analysis, which could justify its form. The issue of how priors should be specified is one of the focal points of the methodological debate. “Non-Bayesians” argue that it is disingenuous to claim the methodological high ground and then base the crucial prior density in a model purely on the basis of mathematical convenience. In a small sample, this assumed prior is going to dominate the results, whereas in a large one, the sampling theory estimates will dominate anyway.

## CHAPTER 16 ♦ Bayesian Estimation and Inference 663

covariance matrices.<sup>9</sup> The smaller the variance of the estimator, the larger its weight, which makes sense. Also, still taking  $\sigma^2$  as known, we can write the variance of the posterior normal distribution as

$$\text{Var}[\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}, \sigma^2] = \{ \boldsymbol{\Sigma}_0^{-1} + [\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1} \}^{-1}. \quad (16-11)$$

Notice that the posterior variance combines the prior and conditional variances on the basis of their inverses.<sup>10</sup> We may interpret the noninformative prior as having infinite elements in  $\boldsymbol{\Sigma}_0$ . This assumption would reduce this case to the earlier one.

Once again, it is necessary to account for the unknown  $\sigma^2$ . If our prior over  $\sigma^2$  is to be informative as well, then the resulting distribution can be extremely cumbersome. A conjugate prior for  $\boldsymbol{\beta}$  and  $\sigma^2$  that can be used is

$$g(\boldsymbol{\beta}, \sigma^2) = g_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta} | \sigma^2)g_{\sigma^2}(\sigma^2), \quad (16-12)$$

where  $g_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta} | \sigma^2)$  is normal, with mean  $\boldsymbol{\beta}^0$  and variance  $\sigma^2 \mathbf{A}$  and

$$g_{\sigma^2}(\sigma^2) = \frac{[m\sigma_0^2]^{m+1}}{\Gamma(m+1)} \left( \frac{1}{\sigma^2} \right)^m e^{-m\sigma_0^2(1/\sigma^2)}. \quad (16-13)$$

This distribution is an inverted gamma distribution. It implies that  $1/\sigma^2$  has a gamma distribution. The prior mean for  $\sigma^2$  is  $\sigma_0^2$  and the prior variance is  $\sigma_0^4/(m-1)$ .<sup>11</sup> The product in (16-12) produces what is called a **normal-gamma prior**, which is the natural conjugate prior for this form of the model. By integrating out  $\sigma^2$ , we would obtain the prior marginal for  $\boldsymbol{\beta}$  alone, which would be a multivariate *t* distribution.<sup>12</sup> Combining (16-12) with (16-13) produces the joint posterior distribution for  $\boldsymbol{\beta}$  and  $\sigma^2$ . Finally, the marginal posterior distribution for  $\boldsymbol{\beta}$  is obtained by integrating out  $\sigma^2$ . It has been shown that this posterior distribution is multivariate *t* with

$$E[\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}] = \{ [\bar{\sigma}^2 \mathbf{A}]^{-1} + [\bar{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}]^{-1} \}^{-1} \{ [\bar{\sigma}^2 \mathbf{A}]^{-1} \boldsymbol{\beta}_0 + [\bar{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}]^{-1} \mathbf{b} \} \quad (16-14)$$

and

$$\text{Var}[\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}] = \left( \frac{j}{j-2} \right) \{ [\bar{\sigma}^2 \mathbf{A}]^{-1} + [\bar{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}]^{-1} \}^{-1}, \quad (16-15)$$

where  $j$  is a degrees of freedom parameter and  $\bar{\sigma}^2$  is the Bayesian estimate of  $\sigma^2$ . The prior degrees of freedom  $m$  is a parameter of the prior distribution for  $\sigma^2$  that would have been determined at the outset. (See the following example.) Once again, it is clear that as the amount of data increases, the posterior density, and the estimates thereof, converge to the sampling theory results.

<sup>9</sup>Note that it will not follow that individual elements of the posterior mean vector lie between those of  $\boldsymbol{\beta}_0$  and  $\mathbf{b}$ . See Judge et al. (1985, pp. 109–110) and Chamberlain and Leamer (1976).

<sup>10</sup>Precisely this estimator was proposed by Theil and Goldberger (1961) as a way of combining a previously obtained estimate of a parameter and a current body of new data. They called their result a “mixed estimator.” The term “mixed estimation” takes an entirely different meaning in the current literature, as we saw in Chapter 15.

<sup>11</sup>You can show this result by using gamma integrals. Note that the density is a function of  $1/\sigma^2 = 1/x$  in the formula of (B-39), so to obtain  $E[\sigma^2]$ , we use the analog of  $E[1/x] = \lambda/(P-1)$  and  $E[(1/x)^2] = \lambda^2/[(P-1)(P-2)]$ . In the density for  $(1/\sigma^2)$ , the counterparts to  $\lambda$  and  $P$  are  $m\sigma_0^2$  and  $m+1$ .

<sup>12</sup>Full details of this (lengthy) derivation appear in Judge et al. (1985, pp. 106–110) and Zellner (1971).

## 664 PART III ♦ Estimation Methodology

**TABLE 16.1** Estimates of the MPC

Years	Estimated MPC	Variance of $\hat{b}$	Degrees of Freedom	Estimated $\sigma$
1940–1950	0.6848014	0.061878	9	24.954
1950–2000	0.92481	0.000065865	49	92.244

**Example 16.3 Bayesian Estimate of the Marginal Propensity to Consume**

In Example 3.2 an estimate of the marginal propensity to consume is obtained using 11 observations from 1940 to 1950, with the results shown in the top row of Table 16.1. A classical 95 percent confidence interval for  $\beta$  based on these estimates is (0.1221, 1.2475). (The very wide interval probably results from the obviously poor specification of the model.) Based on noninformative priors for  $\beta$  and  $\sigma^2$ , we would estimate the posterior density for  $\beta$  to be univariate  $t$  with nine degrees of freedom, with mean 0.6848014 and variance  $(11/9)0.061878 = 0.075628$ . An HPD interval for  $\beta$  would coincide with the confidence interval. Using the fourth quarter (yearly) values of the 1950–2000 data used in Example 5.3, we obtain the new estimates that appear in the second row of the table.

We take the first estimate and its estimated distribution as our prior for  $\beta$  and obtain a posterior density for  $\beta$  based on an informative prior instead. We assume for this exercise that  $\sigma^2$  may be taken as known at the sample value of 24.954. Then,

$$\hat{b} = \left[ \frac{1}{0.000065865} + \frac{1}{0.061878} \right]^{-1} \left[ \frac{0.92481}{0.000065865} + \frac{0.6848014}{0.061878} \right] = 0.92455$$

The weighted average is overwhelmingly dominated by the far more precise sample estimate from the larger sample. The posterior variance is the inverse in brackets, which is 0.000065795. This is close to the variance of the latter estimate. An HPD interval can be formed in the familiar fashion. It will be slightly narrower than the confidence interval, because the variance of the posterior distribution is slightly smaller than the variance of the sampling estimator. This reduction is the value of the prior information. (As we see here, the prior is not particularly informative.)

## 16.4 BAYESIAN INFERENCE

The posterior density is the Bayesian counterpart to the likelihood function. It embodies the information that is available to make inference about the econometric model. As we have seen, the mean and variance of the posterior distribution correspond to the classical (sampling theory) point estimator and asymptotic variance, although they are interpreted differently. Before we examine more intricate applications of Bayesian inference, it is useful to formalize some other components of the method, point and interval estimation and the Bayesian equivalent of testing a hypothesis.<sup>13</sup>

### 16.4.1 POINT ESTIMATION

The posterior density function embodies the prior and the likelihood and therefore contains all the researcher's information about the parameters. But for purposes of presenting results, the density is somewhat imprecise, and one normally prefers a point

<sup>13</sup>We do not include prediction in this list. The Bayesian approach would treat the prediction problem as one of estimation in the same fashion as “parameter” estimation. The value to be forecasted is among the unknown elements of the model that would be characterized by a prior and would enter the posterior density in a symmetric fashion along with the other parameters.

**CHAPTER 16 ♦ Bayesian Estimation and Inference 665**

or interval estimate. The natural approach would be to use the mean of the posterior distribution as the estimator. For the noninformative prior, we use  $\mathbf{b}$ , the **sampling theory** estimator.

One might ask at this point, why bother? These Bayesian point estimates are identical to the sampling theory estimates. All that has changed is our interpretation of the results. This situation is, however, exactly the way it should be. Remember that we entered the analysis with noninformative priors for  $\beta$  and  $\sigma^2$ . Therefore, the only information brought to bear on estimation is the sample data, and it would be peculiar if anything other than the sampling theory estimates emerged at the end. The results do change when our prior brings out of sample information into the estimates, as we shall see later.

The results will also change if we change our motivation for estimating  $\beta$ . The parameter estimates have been treated thus far as if they were an end in themselves. But in some settings, parameter estimates are obtained so as to enable the analyst to make a decision. Consider then, a **loss function**,  $H(\hat{\beta}, \beta)$ , which quantifies the cost of basing a decision on an estimate  $\hat{\beta}$  when the parameter is  $\beta$ . The expected, or average loss is

$$E_{\beta}[H(\hat{\beta}, \beta)] = \int_{\beta} H(\hat{\beta}, \beta) f(\beta | \mathbf{y}, \mathbf{X}) d\beta, \quad (16-16)$$

where the weighting function is the marginal posterior density. (The joint density for  $\beta$  and  $\sigma^2$  would be used if the loss were defined over both.) The Bayesian point estimate is the parameter vector that minimizes the expected loss. If the loss function is a quadratic form in  $(\hat{\beta} - \beta)$ , then the mean of the posterior distribution is the “minimum expected loss” (MELO) estimator. The proof is simple. For this case,

$$E[H(\hat{\beta}, \beta) | \mathbf{y}, \mathbf{X}] = E\left[\frac{1}{2}(\hat{\beta} - \beta)' \mathbf{W}(\hat{\beta} - \beta) | \mathbf{y}, \mathbf{X}\right].$$

To minimize this, we can use the result that

$$\begin{aligned} \partial E[H(\hat{\beta}, \beta) | \mathbf{y}, \mathbf{X}] / \partial \hat{\beta} &= E[\partial H(\hat{\beta}, \beta) / \partial \hat{\beta} | \mathbf{y}, \mathbf{X}] \\ &= E[-\mathbf{W}(\hat{\beta} - \beta) | \mathbf{y}, \mathbf{X}]. \end{aligned}$$

The minimum is found by equating this derivative to  $\mathbf{0}$ , whence, because  $-\mathbf{W}$  is irrelevant,  $\hat{\beta} = E[\beta | \mathbf{y}, \mathbf{X}]$ . This kind of loss function would state that errors in the positive and negative direction are equally bad, and large errors are much worse than small errors. If the loss function were a linear function instead, then the MELO estimator would be the median of the posterior distribution. These results are the same in the case of the noninformative prior that we have just examined.

#### 16.4.2 INTERVAL ESTIMATION

The counterpart to a confidence interval in this setting is an interval of the posterior distribution that contains a specified probability. Clearly, it is desirable to have this interval be as narrow as possible. For a unimodal density, this corresponds to an interval within which the density function is higher than any points outside it, which justifies the term **highest posterior density (HPD) interval**. For the case we have analyzed, which involves a symmetric distribution, we would form the HPD interval for  $\beta$  around the least squares estimate  $\mathbf{b}$ , with terminal values taken from the standard  $t$  tables.

## 666 PART III ♦ Estimation Methodology

### 16.4.3 HYPOTHESIS TESTING

The Bayesian methodology treats the classical approach to hypothesis testing with a large amount of skepticism. Two issues are especially problematic. First, a close examination of only the work we have done in Chapter 5 will show that because we are using consistent estimators, with a large enough sample, we will ultimately reject any (nested) hypothesis unless we adjust the significance level of the test downward as the sample size increases. Second, the all-or-nothing approach of either rejecting or not rejecting a hypothesis provides no method of simply sharpening our beliefs. Even the most committed of analysts might be reluctant to discard a strongly held prior based on a single sample of data, yet this is what the sampling methodolo<sup>14</sup> mandates. (Note, for example, the uncomfortable dilemma this creates in footnote <sup>14</sup> in Chapter 10.) The Bayesian approach to hypothesis testing is much more appealing in this regard. Indeed, the approach might be more appropriately called “comparing hypotheses,” because it essentially involves only making an assessment of which of two hypotheses has a higher probability of being correct.

The Bayesian approach to hypothesis testing bears large similarity to Bayesian estimation.<sup>14</sup> We have formulated two hypotheses, a “null,” denoted  $H_0$ , and an alternative, denoted  $H_1$ . These need not be complementary, as in  $H_0$ : “statement  $A$  is true” versus  $H_1$ : “statement  $A$  is not true,” since the intent of the procedure is not to reject one hypothesis in favor of the other. For simplicity, however, we will confine our attention to hypotheses about the parameters in the regression model, which often are complementary. Assume that before we begin our experimentation (data gathering, statistical analysis) we are able to assign **prior probabilities**  $P(H_0)$  and  $P(H_1)$  to the two hypotheses. The **prior odds ratio** is simply the ratio

$$\text{Odds}_{\text{prior}} = \frac{P(H_0)}{P(H_1)}. \quad (16-17)$$

For example, one’s uncertainty about the sign of a parameter might be summarized in a prior odds over  $H_0: \beta \geq 0$  versus  $H_1: \beta < 0$  of  $0.5/0.5 = 1$ . After the sample evidence is gathered, the prior will be modified, so the posterior is, in general,

$$\text{Odds}_{\text{posterior}} = B_{01} \times \text{Odds}_{\text{prior}}.$$

The value  $B_{01}$  is called the **Bayes factor** for comparing the two hypotheses. It summarizes the effect of the sample data on the prior odds. The end result,  $\text{Odds}_{\text{posterior}}$ , is a new odds ratio that can be carried forward as the prior in a subsequent analysis.

The Bayes factor is computed by assessing the likelihoods of the data observed under the two hypotheses. We return to our first departure point, the likelihood of the data, given the parameters:

$$f(\mathbf{y} | \boldsymbol{\beta}, \sigma^2, \mathbf{X}) = [2\pi\sigma^2]^{-n/2} e^{(-1/(2\sigma^2))(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}. \quad (16-18)$$

Based on our priors for the parameters, the expected, or average likelihood, assuming that hypothesis  $j$  is true ( $j = 0, 1$ ), is

$$f(\mathbf{y} | \mathbf{X}, H_j) = E_{\boldsymbol{\beta}, \sigma^2}[f(\mathbf{y} | \boldsymbol{\beta}, \sigma^2, \mathbf{X}, H_j)] = \int_{\sigma^2} \int_{\boldsymbol{\beta}} f(\mathbf{y} | \boldsymbol{\beta}, \sigma^2, \mathbf{X}, H_j) g(\boldsymbol{\beta}, \sigma^2) d\boldsymbol{\beta} d\sigma^2.$$

<sup>14</sup>For extensive discussion, see Zellner and Siow (1980) and Zellner (1985, pp. 275–305).

## CHAPTER 16 ♦ Bayesian Estimation and Inference 667

(This conditional density is also the **predictive density** for  $\mathbf{y}$ .) Therefore, based on the observed data, we use Bayes's theorem to reassess the probability of  $H_j$ ; the posterior probability is

$$P(H_j | \mathbf{y}, \mathbf{X}) = \frac{f(\mathbf{y} | \mathbf{X}, H_j) P(H_j)}{f(\mathbf{y})}.$$

The posterior odds ratio is  $P(H_0 | \mathbf{y}, \mathbf{X})/P(H_1 | \mathbf{y}, \mathbf{X})$ , so the Bayes factor is

$$B_{01} = \frac{f(\mathbf{y} | \mathbf{X}, H_0)}{f(\mathbf{y} | \mathbf{X}, H_1)}.$$

**Example 16.4 Posterior Odds for the Classical Regression Model**

Zellner (1971) analyzes the setting in which there are two possible explanations for the variation in a dependent variable  $y$ :

$$\text{Model 0: } y = \mathbf{x}'_0 \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_0$$

and

$$\text{Model 1: } y = \mathbf{x}'_1 \boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1.$$

We will briefly sketch his results. We form *informative priors* for  $[\boldsymbol{\beta}, \sigma^2]_j$ ,  $j = 0, 1$ , as specified in (16-12) and (16-13), that is, multivariate normal and inverted gamma, respectively. Zellner then derives the Bayes factor for the posterior odds ratio. The derivation is lengthy and complicated, but for large  $n$ , with some simplifying assumptions, a useful formulation emerges. First, assume that the priors for  $\sigma_0^2$  and  $\sigma_1^2$  are the same. Second, assume that  $[|\mathbf{A}_0^{-1}|/|\mathbf{A}_0^{-1} + \mathbf{X}'_0 \mathbf{X}_0|]/[|\mathbf{A}_1^{-1}|/|\mathbf{A}_1^{-1} + \mathbf{X}'_1 \mathbf{X}_1|] \rightarrow 1$ . The first of these would be the usual situation, in which the uncertainty concerns the covariation between  $y_i$  and  $\mathbf{x}_i$ , not the amount of residual variation (lack of fit). The second concerns the relative amounts of information in the prior ( $\mathbf{A}$ ) versus the likelihood ( $\mathbf{X}'\mathbf{X}$ ). These matrices are the inverses of the covariance matrices, or the **precision matrices**. [Note how these two matrices form the matrix weights in the computation of the posterior mean in (16-9).] Zellner (p. 310) discusses this assumption at some length. With these two assumptions, he shows that as  $n$  grows large,<sup>15</sup>

$$B_{01} \approx \left( \frac{s_0^2}{s_1^2} \right)^{-(n+m)/2} = \left( \frac{1 - R_0^2}{1 - R_1^2} \right)^{-(n+m)/2}.$$

Therefore, the result favors the model that provides the better fit using  $R^2$  as the fit measure. If we stretch Zellner's analysis a bit by interpreting model 1 as "the model" and model 0 as "no model" (that is, the relevant part of  $\boldsymbol{\beta}_0 = \mathbf{0}$ , so  $R_0^2 = 0$ ), then the ratio simplifies to

$$B_{01} = (1 - R_1^2)^{(n+m)/2}.$$

Thus, the better the fit of the regression, the lower the Bayes factor in favor of model 0 (no model), which makes intuitive sense.

Zellner and Siow (1980) have continued this analysis with noninformative priors for  $\boldsymbol{\beta}$  and  $\sigma_j^2$ . Specifically, they use the flat prior for  $\ln \sigma$  [see (16-7)] and a multivariate Cauchy prior (which has infinite variances) for  $\boldsymbol{\beta}$ . Their main result (3.10) is

$$B_{01} = \frac{\frac{1}{2}\sqrt{\pi}}{\Gamma[(k+1)/2]} \left( \frac{n-K}{2} \right)^{k/2} (1 - R^2)^{(n-K-1)/2}.$$

This result is very much like the previous one, with some slight differences due to degrees of freedom corrections and the several approximations used to reach the first one.

<sup>15</sup>A ratio of exponentials that appears in Zellner's result (his equation 10.50) is omitted. To the order of approximation in the result, this ratio vanishes from the final result. (Personal correspondence from A. Zellner to the author.)

## 668 PART III ♦ Estimation Methodology

### 16.4.4 LARGE-SAMPLE RESULTS

Although all statistical results for Bayesian estimators are necessarily “finite sample” (they are conditioned on the sample data), it remains of interest to consider how the estimators behave in large samples.<sup>16</sup> Do Bayesian estimators “converge” to something? To do this exercise, it is useful to envision having a sample that is the entire population. Then, the posterior distribution would characterize this entire population, not a sample from it. It stands to reason in this case, at least intuitively, that the posterior distribution should coincide with the likelihood function. It will (as usual) save for the influence of the prior. But as the sample size grows, one should expect the likelihood function to overwhelm the prior. It will, unless the strength of the prior grows with the sample size (that is, for example, if the prior variance is of order  $1/n$ ). An informative prior will still fade in its influence on the posterior unless it becomes *more* informative as the sample size grows.

The preceding suggests that the posterior mean will converge to the maximum likelihood estimator. The MLE is the parameter vector that is at the mode of the likelihood function. The Bayesian estimator is the **posterior mean**, not the mode, so a remaining question concerns the relationship between these two features. The **Bernstein-von Mises “theorem”** [See Cameron and Trivedi (2005, p. 433) and Train (2003, Chapter 12)] states that the posterior mean and the maximum likelihood estimator will converge to the same probability limit and have the same limiting normal distribution. A form of **central limit theorem** is at work.

But for remaining philosophical questions, the results suggest that for large samples, the choice between Bayesian and frequentist methods can be one of computational efficiency. (This is the thrust of the application in Section 16.8. Note, as well, footnote 1 at the beginning of this chapter. In an infinite sample, the maintained “uncertainty” of the Bayesian estimation framework would have to arise from deeper questions about the model. For example, the mean of the entire population is its mean; there is no uncertainty about the “parameter.”)

## 16.5 POSTERIOR DISTRIBUTIONS AND THE GIBBS SAMPLER

The preceding analysis has proceeded along a set of steps that includes formulating the likelihood function (the model), the prior density over the objects of estimation, and the posterior density. To complete the inference step, we then analytically derived the characteristics of the posterior density of interest, such as the mean or mode, and the variance. The complicated element of any of this analysis is determining the moments of the posterior density, for example, the mean:

$$\hat{\theta} = E[\theta | \text{data}] = \int_{\theta} \theta p(\theta | \text{data}) d\theta. \quad (16-19)$$

---

<sup>16</sup>The standard preamble in econometric studies, that the analysis to follow is “exact” as opposed to approximate or “large sample,” refers to this aspect—the analysis is conditioned on and, by implication, applies only to the sample data in hand. Any inference outside the sample, for example, to hypothesized random samples is, like the sampling theory counterpart, approximate.

## CHAPTER 16 ♦ Bayesian Estimation and Inference 669

There are relatively few applications for which integrals such as this can be derived in closed form. (This is one motivation for conjugate priors.) The modern approach to Bayesian inference takes a different strategy. The result in (16-19) is an expectation. Suppose it were possible to obtain a random sample, as large as desired, from the population defined by  $p(\theta | \text{data})$ . Then, using the same strategy we used throughout Chapter 15 for simulation-based estimation, we could use that sample's characteristics, such as mean, variance, quantiles, and so on, to infer the characteristics of the posterior distribution. Indeed, with an (essentially) infinite sample, we would be freed from having to limit our attention to a few simple features such as the mean and variance and we could view any features of the posterior distribution that we like. The (much less) complicated part of the analysis is the formulation of the posterior density.

It remains to determine how the sample is to be drawn from the posterior density. This element of the strategy is provided by a remarkable (and remarkably useful) result known as the **Gibbs sampler**. [See Casella and George (1992).] The central result of the Gibbs sampler is as follows: We wish to draw a random sample from the joint population  $(x, y)$ . The joint distribution of  $x$  and  $y$  is either unknown or intractable and it is not possible to sample from the joint distribution. However, assume that the conditional distributions  $f(x | y)$  and  $f(y | x)$  are known and simple enough that it is possible to draw univariate random samples from both of them. The following iteration will produce a bivariate random sample from the joint distribution:

### Gibbs Sampler

1. Begin the cycle with a value of  $x_0$  that is in the right range of  $x | y$ ,
2. Draw an observation  $y_0 | x_0$ ,
3. Draw an observation  $x_t | y_{t-1}$ ,
4. Draw an observation  $y_t | x_t$ .

Iteration of steps 3 and 4 for several thousand cycles will eventually produce a random sample from the joint distribution. (The first several thousand draws are discarded to avoid the influence of the initial conditions—this is called the **burn in**.) [Some technical details on the procedure appear in Cameron and Trivedi (Chapter Section 13.5).]

### **Example 16.5 Gibbs Sampling from the Normal Distribution**

To illustrate the mechanical aspects of the Gibbs sampler, consider random sampling from the joint normal distribution. We consider the bivariate normal distribution first. Suppose we wished to draw a random sample from the population

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right].$$

As we have seen in Chapter 15, a direct approach is to use the fact that linear functions of normally distributed variables are normally distributed. [See (B-80).] Thus, we might transform a series of independent normal draws  $(u_1, u_2)'$  by the Cholesky decomposition of the covariance matrix

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}_i = \begin{bmatrix} 1 & 0 \\ \theta_1 & \theta_2 \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_i = \mathbf{L}\mathbf{u}_i,$$

## 670 PART III ♦ Estimation Methodology

where  $\theta_1 = \rho$  and  $\theta_2 = \sqrt{1 - \rho^2}$ . The Gibbs sampler would take advantage of the result

$$x_1 | x_2 \sim N[\rho x_2, (1 - \rho^2)],$$

and

$$x_2 | x_1 \sim N[\rho x_1, (1 - \rho^2)].$$

To sample from a trivariate, or multivariate population, we can expand the Gibbs sequence in the natural fashion. For example, to sample from a trivariate population, we would use the Gibbs sequence

$$\begin{aligned} x_1 | x_2, x_3 &\sim N[\beta_{1,2}x_2 + \beta_{1,3}x_3, \Sigma_{1|2,3}], \\ x_2 | x_1, x_3 &\sim N[\beta_{2,1}x_1 + \beta_{2,3}x_3, \Sigma_{2|1,3}], \\ x_3 | x_1, x_2 &\sim N[\beta_{3,1}x_1 + \beta_{3,2}x_2, \Sigma_{3|1,2}], \end{aligned}$$

where the conditional means and variances are given in Theorem B.7. This defines a three-step cycle.

The availability of the Gibbs sampler frees the researcher from the necessity of deriving the analytical properties of the full, joint posterior distribution. Because the formulation of conditional priors is straightforward, and the derivation of the *conditional* posteriors is only slightly less so, this tool has facilitated a vast range of applications that previously were intractable. For an example, consider, once again, the classical normal regression model. From (16-7), the joint posterior for  $(\boldsymbol{\beta}, \sigma^2)$  is

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) &\propto \frac{[vs^2]^{v+2}}{\Gamma(v+2)} \left[ \frac{1}{\sigma^2} \right]^{v+1} \exp(-vs^2/\sigma^2) [2\pi]^{-K/2} |\sigma^2(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2} \\ &\quad \times \exp(-(1/2)(\boldsymbol{\beta} - \mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta} - \mathbf{b})). \end{aligned}$$

If we wished to use a simulation approach to characterizing the posterior distribution, we would need to draw a  $K + 1$  variate sample of observations from this intractable distribution. However, with the assumed priors, we found the conditional posterior for  $\boldsymbol{\beta}$  in (16-5):

$$p(\boldsymbol{\beta} | \sigma^2, \mathbf{y}, \mathbf{X}) = N[\mathbf{b}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}].$$

From (16-6), we can deduce that the conditional posterior for  $\sigma^2 | \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}$  is an inverted gamma distribution with parameters  $m\sigma_0^2 = v\hat{\sigma}^2$  and  $m = v$  in (16-13):

$$p(\sigma^2 | \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}) = \frac{[v\hat{\sigma}^2]^{v+1}}{\Gamma(v+1)} \left[ \frac{1}{\sigma^2} \right]^v \exp(-v\hat{\sigma}^2/\sigma^2), \quad \hat{\sigma}^2 = \frac{\sum_{i=1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{n-K}.$$

This sets up a Gibbs sampler for sampling from the joint posterior of  $\boldsymbol{\beta}$  and  $\sigma^2$ . We would cycle between random draws from the multivariate normal for  $\boldsymbol{\beta}$  and the inverted gamma distribution for  $\sigma^2$  to obtain a  $K + 1$  variate sample on  $(\boldsymbol{\beta}, \sigma^2)$ . [Of course, for this application, we do know the marginal posterior distribution for  $\boldsymbol{\beta}$ —see (16-8).]

The Gibbs sampler is not truly a random sampler; it is a Markov chain—each “draw” from the distribution is a function of the draw that precedes it. The random input at each cycle provides the randomness, which leads to the popular name for this strategy, **Markov–Chain Monte Carlo** or **MCMC** or **MC<sup>2</sup>** (pick one) estimation. In its simplest

CHAPTER 16 ♦ Bayesian Estimation and Inference **671**

form, it provides a remarkably efficient tool for studying the posterior distributions in very complicated models. The example in the next section shows a striking example of how to locate the MLE for a probit model without computing the likelihood function or its derivatives. In Section 16.8, we will examine an extension and refinement of the strategy, the Metropolis–Hastings algorithm.

In the next several sections, we will present some applications of Bayesian inference. In Section 16.9, we will return to some general issues in classical and Bayesian estimation and inference.

## 16.6 APPLICATION: BINOMIAL PROBIT MODEL

Consider inference about the binomial probit model for a dependent variable that is generated as follows (see Sections 17.2–17.4):

$$y_i^* = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N[0, 1], \quad (16-20)$$

$$y_i = 1 \quad \text{if } y_i^* > 0, \text{ otherwise } y_i = 0. \quad (16-21)$$

(Theoretical motivation for the model appears in Section 17.3.) The data consist of  $(\mathbf{y}, \mathbf{X}) = (y_i, \mathbf{x}_i), i = 1, \dots, n$ . The random variable  $y_i$  has a Bernoulli distribution with probabilities

$$\text{Prob}[y_i = 1 | \mathbf{x}_i] = \Phi(\mathbf{x}'_i \boldsymbol{\beta}),$$

$$\text{Prob}[y_i = 0 | \mathbf{x}_i] = 1 - \Phi(\mathbf{x}'_i \boldsymbol{\beta}).$$

The likelihood function for the observed data is

$$L(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^n [\Phi(\mathbf{x}'_i \boldsymbol{\beta})]^{y_i} [1 - \Phi(\mathbf{x}'_i \boldsymbol{\beta})]^{1-y_i}.$$

(Once again, we cheat a bit on the notation—the likelihood function is actually the joint density for the data, given  $\mathbf{X}$  and  $\boldsymbol{\beta}$ .) Classical maximum likelihood estimation of  $\boldsymbol{\beta}$  is developed in Section 17.4. To obtain the posterior mean (Bayesian estimator), we assume a noninformative, flat (improper) prior for  $\boldsymbol{\beta}$ ,

$$p(\boldsymbol{\beta}) \propto 1.$$

The posterior density would be

$$p(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) = \frac{\prod_{i=1}^n [\Phi(\mathbf{x}'_i \boldsymbol{\beta})]^{y_i} [1 - \Phi(\mathbf{x}'_i \boldsymbol{\beta})]^{1-y_i} (1)}{\int_{\boldsymbol{\beta}} \prod_{i=1}^n [\Phi(\mathbf{x}'_i \boldsymbol{\beta})]^{y_i} [1 - \Phi(\mathbf{x}'_i \boldsymbol{\beta})]^{1-y_i} (1) d\boldsymbol{\beta}},$$

and the estimator would be the posterior mean,

$$\hat{\boldsymbol{\beta}} = E[\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}] = \frac{\int_{\boldsymbol{\beta}} \boldsymbol{\beta} \prod_{i=1}^n [\Phi(\mathbf{x}'_i \boldsymbol{\beta})]^{y_i} [1 - \Phi(\mathbf{x}'_i \boldsymbol{\beta})]^{1-y_i} d\boldsymbol{\beta}}{\int_{\boldsymbol{\beta}} \prod_{i=1}^n [\Phi(\mathbf{x}'_i \boldsymbol{\beta})]^{y_i} [1 - \Phi(\mathbf{x}'_i \boldsymbol{\beta})]^{1-y_i} d\boldsymbol{\beta}}. \quad (16-22)$$

Evaluation of the integrals in (16-22) is hopelessly complicated, but a solution using the Gibbs sampler and a technique known as **data augmentation**, pioneered by Albert

## 672 PART III ♦ Estimation Methodology

and Chib (1993a) is surprisingly simple. We begin by treating the unobserved  $y_i^*$ 's as unknowns to be estimated, along with  $\beta$ . Thus, the  $(K + n) \times 1$  parameter vector is  $\theta = (\beta, \mathbf{y}^*)$ . We now construct a Gibbs sampler. Consider, first,  $p(\beta | \mathbf{y}^*, \mathbf{y}, \mathbf{X})$ . If  $y_i^*$  is known, then  $y_i$  is known [see (16-21)]. It follows that

$$p(\beta | \mathbf{y}^*, \mathbf{y}, \mathbf{X}) = p(\beta | \mathbf{y}^*, \mathbf{X}).$$

This posterior defines a linear regression model with normally distributed disturbances and known  $\sigma^2 = 1$ . It is precisely the model we saw in Section 16.3.1, and the posterior we need is in (16-5), with  $\sigma^2 = 1$ . So, based on our earlier results, it follows that

$$p(\beta | \mathbf{y}^*, \mathbf{y}, \mathbf{X}) = N[\mathbf{b}^*, (\mathbf{X}'\mathbf{X})^{-1}], \quad (16-23)$$

where

$$\mathbf{b}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*.$$

For  $y_i^*$ , ignoring  $y_i$  for the moment, it would follow immediately from (16-20) that

$$p(y_i^* | \beta, \mathbf{X}) = N[\mathbf{x}_i'\beta, 1].$$

However,  $y_i$  is informative about  $y_i^*$ . If  $y_i$  equals one, we know that  $y_i^* > 0$  and if  $y_i$  equals zero, then  $y_i^* \leq 0$ . The implication is that conditioned on  $\beta$ ,  $\mathbf{X}$ , and  $\mathbf{y}$ ,  $y_i^*$  has the truncated normal distribution that is developed in Sections 16.2.1 and 16.2.2. The standard notation for this is

$$\begin{aligned} p(y_i^* | y_i = 1, \beta, \mathbf{x}_i) &= N^+[\mathbf{x}_i'\beta, 1], \\ p(y_i^* | y_i = 0, \beta, \mathbf{x}_i) &= N^-[\mathbf{x}_i'\beta, 1]. \end{aligned} \quad (16-24)$$

Results (16-23) and (16-24) set up the components for a Gibbs sampler that we can use to estimate the posterior means  $E[\beta | \mathbf{y}, \mathbf{X}]$  and  $E[\mathbf{y}^* | \mathbf{y}, \mathbf{X}]$ . The following is our algorithm:

### Gibbs Sampler for the Binomial Probit Model

1. Compute  $\mathbf{X}'\mathbf{X}$  once at the outset and obtain  $\mathbf{L}$  such that  $\mathbf{LL}' = (\mathbf{X}'\mathbf{X})^{-1}$ .
2. Start  $\beta$  at any value such as  $\mathbf{0}$ .
3. Result (15.2.4) shows how to transform a draw from  $U[0, 1]$  to a draw from the truncated normal with underlying mean  $\mu$  and standard deviation  $\sigma$ . For this application, the draw is

$$\begin{aligned} y_{i,r}^*(r) &= \mathbf{x}_i'\beta_{r-1} + \Phi^{-1}[1 - (1 - U)\Phi(\mathbf{x}_i'\beta_{r-1})] && \text{if } y_i = 1, \\ y_{i,r}^*(r) &= \mathbf{x}_i'\beta_{r-1} + \Phi^{-1}[U\Phi(-\mathbf{x}_i'\beta_{r-1})] && \text{if } y_i = 0. \end{aligned}$$

This step is used to draw the  $n$  observations on  $y_{i,r}^*(r)$ .

4. Section 15.2.4 shows how to draw an observation from the multivariate normal population. For this application, we use the results at step 3 to compute  $\mathbf{b}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*(r)$ . We obtain a vector,  $\mathbf{v}$ , of  $K$  draws from the  $N[0, 1]$  population, then  $\beta(r) = \mathbf{b}^* + \mathbf{L}\mathbf{v}$ .

The iteration cycles between steps 3 and 4. This should be repeated several thousand times, discarding the burn-in draws, then the estimator of  $\beta$  is the sample mean of the retained draws. The posterior variance is computed with the variance of the retained draws. Posterior estimates of  $y_i^*$  would typically not be useful.

CHAPTER 16 ♦ Bayesian Estimation and Inference **673****TABLE 16.2** Probit Estimates for Grade Equation

<i>Variable</i>	<i>Maximum Likelihood</i>		<i>Posterior Means and Std. Devs</i>	
	<i>Estimate</i>	<i>Standard Error</i>	<i>Posterior Mean</i>	<i>Posterior S.D.</i>
Constant	-7.4523	2.5425	-8.6286	2.7995
GPA	1.6258	0.6939	1.8754	0.7668
TUCE	0.05173	0.08389	0.06277	0.08695
PSI	1.4263	0.5950	1.6072	0.6257

**Example 16.6**  Gibbs Sampler for a Probit Model

In Examples 14.14 and 14.15, we examined Spector and Mazzeo's (1980) widely traveled data on a binary choice outcome. (The example used the data for a different model.) The binary probit model studied in the paper was

$$\text{Prob}(GRADE_i = 1 | \beta, \mathbf{x}_i) = \Phi(\beta_0 + \beta_1 \text{GPA}_i + \beta_2 \text{TUCE}_i + \beta_3 \text{PSI}_i).$$

The variables are defined in Example 14.14. Their probit model is studied in Example 17.3. The sample contains 32 observations. Table 16.2 presents the maximum likelihood estimates and the posterior means and standard deviations for the probit model. For the Gibbs sampler, we used 5,000 draws, and discarded the first 1,000.

The results in Table 16.2 suggest the similarity of the posterior mean estimated with the Gibbs sampler to the maximum likelihood estimate. However, the sample is quite small, and the differences between the coefficients are still fairly substantial. For a striking example of the behavior of this procedure, we now revisit the German health care data examined in Example 14.15, and several other examples throughout the book. The probit model to be estimated is

$$\begin{aligned} \text{Prob}(\text{Doctor visits}_{it} > 0) = \Phi(\beta_1 + \beta_2 \text{Age}_{it} + \beta_3 \text{Education}_{it} + \beta_4 \text{Income}_{it} \\ + \beta_5 \text{Kids}_{it} + \beta_6 \text{Married}_{it} + \beta_7 \text{Female}_{it}). \end{aligned}$$

The sample contains data on 7,293 families and a total of 27,326 observations. We are pooling the data for this application. Table 16.3 presents the probit results for this model using the same procedure as before. (We used only 500 draws, and discarded the first 100.)

The similarity is what one would expect given the large sample size. We note before proceeding to other applications, notwithstanding the striking similarity of the Gibbs sampler to the MLE, that this is not an efficient method of estimating the parameters of a probit model. The estimator requires generation of thousands of samples of potentially thousands of observations. We used only 500 replications to produce Table 16.3. The computations took about five minutes. Using Newton's method to maximize the log-likelihood directly took less than five seconds. Unless one is wedded to the Bayesian paradigm, on strictly practical grounds, the MLE would be the preferred estimator.

**TABLE 16.3** Probit Estimates for Doctor Visits Equation

<i>Variable</i>	<i>Maximum Likelihood</i>		<i>Posterior Means and Std. Devs</i>	
	<i>Estimate</i>	<i>Standard Error</i>	<i>Posterior Mean</i>	<i>Posterior S.D.</i>
Constant	-0.12433	0.058146	-0.12628	0.054759
Age	0.011892	0.00079568	0.011979	0.00080073
Education	-0.014966	0.0035747	-0.015142	0.0036246
Income	-0.13242	0.046552	-0.12669	0.047979
Kids	-0.15212	0.018327	-0.15149	0.018400
Married	0.073522	0.020644	0.071977	0.020852
Female	0.35591	0.016017	0.35582	0.015913

## 674 PART III ♦ Estimation Methodology

This application of the Gibbs sampler demonstrates in an uncomplicated case how the algorithm can provide an alternative to actually maximizing the log-likelihood. We do note that the similarity of the method to the EM algorithm in Section E.3.7 is not coincidental. Both procedures use an estimate of the unobserved, censored data, and both estimate  $\beta$  by using OLS using the predicted data.

### 16.7 PANEL DATA APPLICATION: INDIVIDUAL EFFECTS MODELS

We consider a panel data model with common individual effects,

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + \varepsilon_{it}, \quad \varepsilon_{it} \sim N[0, \sigma^2_\varepsilon].$$

In the Bayesian framework, there is no need to distinguish between fixed and random effects. The classical distinction results from an asymmetric treatment of the data and the parameters. So, we will leave that unspecified for the moment. The implications will emerge later when we specify the prior densities over the model parameters.

The likelihood function for the sample under normality of  $\varepsilon_{it}$  is

$$p(\mathbf{y} | \alpha_1, \dots, \alpha_n, \beta, \sigma^2_\varepsilon, \mathbf{X}) = \prod_{i=1}^n \prod_{t=1}^{T_i} \frac{1}{\sigma_\varepsilon \sqrt{2\pi}} \exp\left(-\frac{(y_{it} - \alpha_i - \mathbf{x}'_{it}\beta)^2}{2\sigma^2_\varepsilon}\right).$$

The remaining analysis hinges on the specification of the prior distributions. We will consider three cases. Each illustrates an aspect of the methodology.

First, group the full set of location (regression) parameters in one  $(n + K) \times 1$  slope vector,  $\gamma$ . Then, with the disturbance variance,  $\theta = (\alpha, \beta, \sigma^2_\varepsilon) = (\gamma, \sigma^2_\varepsilon)$ . Define a conformable data matrix,  $\mathbf{Z} = (\mathbf{D}, \mathbf{X})$ , where  $\mathbf{D}$  contains the  $n$  dummy variables so that we may write the model,

$$\mathbf{y} = \mathbf{Z}\gamma + \varepsilon$$

in the familiar fashion for our common effects linear regression. (See Chapter 11.) We now assume the **uniform-inverse gamma prior** that we used in our earlier treatment of the linear model,

$$p(\gamma, \sigma^2_\varepsilon) \propto 1/\sigma^2_\varepsilon.$$

The resulting (marginal) posterior density for  $\gamma$  is precisely that in (16-8) (where now the slope vector includes the elements of  $\alpha$ ). The density is an  $(n + K)$  variate  $t$  with mean equal to the OLS estimator and covariance matrix  $[(\sum_i T_i - n - K)/(\sum_i T_i - n - K - 2)]s^2(\mathbf{Z}'\mathbf{Z})^{-1}$ . Because OLS in this model as stated means the within estimator, the implication is that with this noninformative prior over  $(\alpha, \beta)$ , the model is equivalent to the fixed effects model. Note, again, this is not a consequence of any assumption about correlation between effects and included variables. That has remained unstated; though, by implication, we would allow correlation between  $\mathbf{D}$  and  $\mathbf{X}$ .

Some observers are uncomfortable with the idea of a **uniform prior** over the entire real line. [See, for example, Koop (2003, pp. 22–23).] Others, for example, Zellner (1971, p. 20), are less concerned. Cameron and Trivedi (2005, pp. 425–427) suggest a middle ground.] Formally, our assumption of a uniform prior over the entire real line is an

## CHAPTER 16 ♦ Bayesian Estimation and Inference 675

**improper prior**, because it cannot have a positive density and integrate to one over the entire real line. As such, the posterior appears to be ill defined. However, note that the “improper” uniform prior will, in fact, fall out of the posterior, because it appears in both numerator and denominator. [Zellner (1971, p. 20) offers some more methodological commentary.] The practical solution for location parameters, such as a vector of regression slopes, is to assume a nearly flat, “almost uninformative” prior. The usual choice is a conjugate normal prior with an arbitrarily large variance. (It should be noted, of course, that as long as that variance is finite, even if it is large, the prior is informative. We return to this point in Section 16.9.)

Consider, then, the conventional **normal-gamma prior** over  $(\boldsymbol{\gamma}, \sigma_\varepsilon^2)$  where the conditional (on  $\sigma_\varepsilon^2$ ) prior normal density for the slope parameters has mean  $\boldsymbol{\gamma}_0$  and covariance matrix  $\sigma_\varepsilon^2 \mathbf{A}$ , where the  $(n + K) \times (n + K)$  matrix,  $\mathbf{A}$ , is yet to be specified. [See the discussion after (16-13).] The marginal posterior mean and variance for  $\boldsymbol{\gamma}$  for this set of assumptions are given in (16-14) and (16-15). We reach a point that presents two rather serious dilemmas for the researcher. The posterior was simple with our uniform, non-informative prior. Now, it is necessary actually to specify  $\mathbf{A}$ , which is potentially large. (In one of our main applications in this text, we are analyzing models with  $n = 7,293$  constant terms and about  $K = 7$  regressors.) It is hopelessly optimistic to expect to be able to specify all the variances and covariances in a matrix this large, unless we actually have the results of an earlier study (in which case we would also have a prior estimate of  $\boldsymbol{\gamma}$ ). A practical solution that is frequently chosen is to specify  $\mathbf{A}$  to be a diagonal matrix with extremely large diagonal elements, thus emulating a uniform prior without having to commit to one. The second practical issue then becomes dealing with the actual computation of the order  $(n + K)$  inverse matrix in (16-14) and (16-15). Under the strategy chosen, to make  $\mathbf{A}$  a multiple of the identity matrix, however, there are forms of partitioned inverse matrices that will allow solution to the actual computation.

Thus far, we have assumed that each  $\alpha_i$  is generated by a different normal distribution,  $-\boldsymbol{\gamma}_0$  and  $\mathbf{A}$ , however specified, have (potentially) different means and variances for the elements of  $\alpha$ . The third specification we consider is one in which all  $\alpha_i$ ’s in the model are assumed to be draws from the same population. To produce this specification, we use a **hierarchical prior** for the individual effects. The full model will be

$$\begin{aligned} y_{it} &= \alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it}, \quad \varepsilon_{it} \sim N[0, \sigma_\varepsilon^2], \\ p(\boldsymbol{\beta} | \sigma_\varepsilon^2) &= N[\boldsymbol{\beta}_0, \sigma_\varepsilon^2 \mathbf{A}], \\ p(\sigma_\varepsilon^2) &= \text{Gamma}(\sigma_0^2, m), \\ p(\alpha_i) &= N[\mu_\alpha, \tau_\alpha^2], \\ p(\mu_\alpha) &= N[a, Q], \\ p(\tau_\alpha^2) &= \text{Gamma}(\tau_0^2, v). \end{aligned}$$

We will not be able to derive the posterior density (joint or marginal) for the parameters of this model. However, it is possible to set up a Gibbs sampler that can be used to infer the characteristics of the posterior densities statistically. The sampler will be driven by conditional normal posteriors for the location parameters,  $[\boldsymbol{\beta} | \alpha, \sigma_\varepsilon^2, \mu_\alpha, \tau_\alpha^2]$ ,  $[\alpha_i | \boldsymbol{\beta}, \sigma_\varepsilon^2, \mu_\alpha, \tau_\alpha^2]$ , and  $[\mu_\alpha | \boldsymbol{\beta}, \alpha, \sigma_\varepsilon^2, \tau_\alpha^2]$  and conditional gamma densities for the scale (variance) parameters,  $[\sigma_\varepsilon^2 | \alpha, \boldsymbol{\beta}, \mu_\alpha, \tau_\alpha^2]$  and  $[\tau_\alpha^2 | \alpha, \boldsymbol{\beta}, \sigma_\varepsilon^2, \mu_\alpha]$ . [The procedure is

## 676 PART III ♦ Estimation Methodology

developed at length by Koop (2003, pp. 152–153).] The assumption of a common distribution for the individual effects and an independent prior for  $\beta$  produces a Bayesian counterpart to the random effects model.

### 16.8 HIERARCHICAL BAYES ESTIMATION OF A RANDOM PARAMETERS MODEL

We now consider a Bayesian approach to estimation of the random parameters model.<sup>17</sup> For an individual  $i$ , the conditional density for the dependent variable in period  $t$  is  $f(y_{it} | \mathbf{x}_{it}, \beta_i)$  where  $\beta_i$  is the individual specific  $K \times 1$  parameter vector and  $\mathbf{x}_{it}$  is individual specific data that enter the probability density.<sup>18</sup> For the sequence of  $T$  observations, assuming conditional (on  $\beta_i$ ) independence, person  $i$ 's contribution to the likelihood for the sample is

$$f(\mathbf{y}_i | \mathbf{X}_i, \beta_i) = \prod_{t=1}^T f(y_{it} | \mathbf{x}_{it}, \beta_i). \quad (16-25)$$

where  $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})$  and  $\mathbf{X}_i = [\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}]$ . We will suppose that  $\beta_i$  is distributed normally with mean  $\beta$  and covariance matrix  $\Sigma$ . (This is the “hierarchical” aspect of the model.) The unconditional density would be the expected value over the possible values of  $\beta_i$ ;

$$f(\mathbf{y}_i | \mathbf{X}_i, \beta, \Sigma) = \int_{\beta_i} \prod_{t=1}^T f(y_{it} | \mathbf{x}_{it}, \beta_i) \phi_K[\beta_i | \beta, \Sigma] d\beta_i, \quad (16-26)$$

where  $\phi_K[\beta_i | \beta, \Sigma]$  denotes the  $K$  variate normal prior density for  $\beta_i$  given  $\beta$  and  $\Sigma$ . Maximum likelihood estimation of this model, which entails estimation of the “deep” parameters,  $\beta$ , then estimation of the individual specific parameters,  $\beta_i$  is considered in Section 15.7. We now consider the Bayesian approach to estimation of the parameters of this model.

To approach this from a Bayesian viewpoint, we will assign noninformative prior densities to  $\beta$  and  $\Sigma$ . As is conventional, we assign a flat (noninformative) prior to  $\beta$ . The variance parameters are more involved. If it is assumed that the elements of  $\beta_i$  are conditionally independent, then each element of the (now) diagonal matrix  $\Sigma$  may be assigned the inverted gamma prior that we used in (16-13). A full matrix  $\Sigma$  is handled by assigning to  $\Sigma$  an **inverted Wishart** prior density with parameters scalar  $K$  and matrix  $K \times \mathbf{I}$ . [The Wishart density is a multivariate counterpart to the chi-squared

<sup>17</sup>Note that, there is occasional confusion as to what is meant by “random parameters” in a random parameters (RP) model. In the Bayesian framework we discuss in this chapter, the “randomness” of the random parameters in the model arises from the “uncertainty” of the analyst. As developed at several points in this book (and in the literature), the randomness of the parameters in the RP model is a characterization of the heterogeneity of parameters across individuals. Consider, for example, in the Bayesian framework of this section, in the RP model, each vector  $\beta_i$  is a random vector with a distribution (defined hierarchically). In the classical framework, each  $\beta_i$  represents a single draw from a parent population.

<sup>18</sup>To avoid a layer of complication, we will embed the time-invariant effect  $\Delta z_i$  in  $\mathbf{x}_{it}'\beta$ . A full treatment in the same fashion as the latent class model would be substantially more complicated in this setting, though it is quite straightforward in the maximum simulated likelihood approach discussed in Section 15.7.4.

## CHAPTER 16 ♦ Bayesian Estimation and Inference 677

distribution. Discussion may be found in Zellner (1971, pp. 389–394).] This produces the joint posterior density,

$$\Lambda(\beta_1, \dots, \beta_n, \boldsymbol{\beta}, \boldsymbol{\Sigma} | \text{all data}) = \left\{ \prod_{i=1}^n \prod_{t=1}^T f(y_{it} | \mathbf{x}_{it}, \beta_i) \phi_K[\beta_i | \boldsymbol{\beta}, \boldsymbol{\Sigma}] \right\} \times p(\boldsymbol{\beta}, \boldsymbol{\Sigma}). \quad (16-27)$$

This gives the joint density of all the unknown parameters conditioned on the observed data. Our Bayesian estimators of the parameters will be the posterior means for these  $(n+1)K + K(K+1)/2$  parameters. In principle, this requires integration of (16-27) with respect to the components. As one might guess at this point, that integration is hopelessly complex and not remotely feasible.

However, the techniques of Markov–Chain Monte Carlo (MCMC) simulation estimation (the Gibbs sampler) and the **Metropolis–Hastings algorithm** enable us to sample from the (hopelessly complex) joint density  $\Lambda(\beta_1, \dots, \beta_n, \boldsymbol{\beta}, \boldsymbol{\Sigma} | \text{all data})$  in a remarkably simple fashion. Train (2001 and 2002, Chapter 12) describe how to use these results for this random parameters model.<sup>19</sup> The usefulness of this result for our current problem is that it is, indeed, possible to partition the joint distribution, and we can easily sample from the conditional distributions. We begin by partitioning the parameters into  $\gamma = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$  and  $\delta = (\beta_1, \dots, \beta_n)$ . Train proposes the following strategy: To obtain a draw from  $\gamma | \delta$ , we will use the Gibbs sampler to obtain a draw from the distribution of  $(\boldsymbol{\beta} | \boldsymbol{\Sigma}, \delta)$  and then one from the distribution of  $(\boldsymbol{\Sigma} | \boldsymbol{\beta}, \delta)$ . We will lay out this first, then turn to sampling from  $\delta | \boldsymbol{\beta}, \boldsymbol{\Sigma}$ .

Conditioned on  $\delta$  and  $\boldsymbol{\Sigma}$ ,  $\boldsymbol{\beta}$  has a  $K$ -variate normal distribution with mean  $\bar{\boldsymbol{\beta}} = (1/n) \sum_{i=1}^n \beta_i$  and covariance matrix  $(1/n)\boldsymbol{\Sigma}$ . To sample from this distribution we will first obtain the Cholesky factorization of  $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}'$  where  $\mathbf{L}$  is a lower triangular matrix. [See Section A.6.11.] Let  $\mathbf{v}$  be a vector of  $K$  draws from the standard normal distribution. Then,  $\bar{\boldsymbol{\beta}} + \mathbf{L}\mathbf{v}$  has mean vector  $\bar{\boldsymbol{\beta}} + \mathbf{L} \times \mathbf{0} = \bar{\boldsymbol{\beta}}$  and covariance matrix  $\mathbf{L}\mathbf{L}' = \boldsymbol{\Sigma}$ , which is exactly what we need. So, this shows how to sample a draw from the conditional distribution  $\boldsymbol{\beta}$ .

To obtain a random draw from the distribution of  $\boldsymbol{\Sigma} | \boldsymbol{\beta}, \delta$ , we will require a random draw from the inverted Wishart distribution. The marginal posterior distribution of  $\boldsymbol{\Sigma} | \boldsymbol{\beta}, \delta$  is inverted Wishart with parameters scalar  $K+n$  and matrix  $\mathbf{W} = (K\mathbf{I} + n\mathbf{V})$ , where  $\mathbf{V} = (1/n) \sum_{i=1}^n (\beta_i - \bar{\boldsymbol{\beta}})(\beta_i - \bar{\boldsymbol{\beta}})'$ . Train (2001) suggests the following strategy for sampling a matrix from this distribution: Let  $\mathbf{M}$  be the lower triangular Cholesky factor of  $\mathbf{W}^{-1}$ , so  $\mathbf{M}\mathbf{M}' = \mathbf{W}^{-1}$ . Obtain  $K+n$  draws of  $\mathbf{v}_k = K$  standard normal variates. Then, obtain  $\mathbf{S} = \mathbf{M}(\sum_{k=1}^{K+n} \mathbf{v}_k \mathbf{v}_k') \mathbf{M}'$ . Then,  $\boldsymbol{\Sigma}^j = \mathbf{S}^{-1}$  is a draw from the inverted Wishart distribution. [This is fairly straightforward, as it involves only random sampling from the standard normal distribution. For a diagonal  $\boldsymbol{\Sigma}$  matrix, that is, uncorrelated parameters in  $\beta_i$ , it simplifies a bit further. A draw for the nonzero  $k$ th diagonal element can be obtained using  $(1 + n\mathbf{V}_{kk}) / \sum_{r=1}^{K+n} v_{rk}^2$ .]

<sup>19</sup>Train describes use of this method for “mixed (random parameters) multinomial logit” models. By writing the densities in generic form, we have extended his result to any general setting that yes a parameter vector in the fashion described at The classical version of this appears in Section 15.6.1 for the binomial probit model and in Section 17.17 for the mixed logit model.

## 678 PART III ♦ Estimation Methodology

The difficult step is sampling  $\beta_i$ . For this step, we use the Metropolis–Hastings (M–H) algorithm suggested by Chib and Greenberg (1995, 1996) and Gelman et al. (2004). The procedure involves the following steps:

1. Given  $\beta$  and  $\Sigma$  and “tuning constant”  $\tau$  (to be described next), compute  $\mathbf{d} = \tau \mathbf{L}\mathbf{v}$  where  $\mathbf{L}$  is the Cholesky factorization of  $\Sigma$  and  $\mathbf{v}$  is a vector of  $K$  independent standard normal draws.
2. Create a trial value  $\beta_{i1} = \beta_{i0} + \mathbf{d}$  where  $\beta_{i0}$  is the previous value.
3. The posterior distribution for  $\beta_i$  is the likelihood that appears in (16-26) times the joint normal prior density,  $\phi_K[\beta_i | \beta, \Sigma]$ . Evaluate this posterior density at the trial value  $\beta_{i1}$  and the previous value  $\beta_{i0}$ . Let

$$R_{i0} = \frac{f(\mathbf{y}_i | \mathbf{X}_i, \beta_{i1})\phi_K(\beta_{i1} | \beta, \Sigma)}{f(\mathbf{y}_i | \mathbf{X}_i, \beta_{i0})\phi_K(\beta_{i0} | \beta, \Sigma)}.$$

4. Draw one observation,  $u$ , from the standard uniform distribution,  $U[0, 1]$ .
5. If  $u < R_{i0}$ , then accept the trial (new) draw. Otherwise, reuse the old one.

This M–H iteration converges to a sequence of draws from the desired density. Overall, then, the algorithm uses the Gibbs sampler and the Metropolis–Hastings algorithm to produce the sequence of draws for all the parameters in the model. The sequence is repeated a large number of times to produce each draw from the joint posterior distribution. The entire sequence must then be repeated  $N$  times to produce the sample of  $N$  draws, which can then be analyzed, for example, by computing the posterior mean.

Some practical details remain. The tuning constant,  $\tau$  is used to control the iteration. A smaller  $\tau$  increases the acceptance rate. But at the same time, a smaller  $\tau$  makes new draws look more like old draws so this slows down the process. Gelman et al. (2004) suggest  $\tau = 0.4$  for  $K = 1$  and smaller values down to about 0.23 for higher dimensions, as will be typical. Each multivariate draw takes many runs of the MCMC sampler. The process must be started somewhere, though it does not matter much where. Nonetheless, a “burn-in” period is required to eliminate the influence of the starting value. Typical applications use several draws for this burn in period for each run of the sampler. How many sample observations are needed for accurate estimation is not certain, though several hundred would be a minimum. This means that there is a huge amount of computation done by this estimator. However, the computations are fairly simple. The only complicated step is computation of the acceptance criterion at step 3 of the M–H iteration. Depending on the model, this may, like the rest of the calculations, be quite simple.

### 16.9 SUMMARY AND CONCLUSIONS

This chapter has introduced the major elements of the Bayesian approach to estimation and inference. The contrast between Bayesian and classical, or frequentist, approaches to the analysis has been the subject of a decades-long dialogue among practitioners and philosophers. As the frequency of applications of Bayesian methods have grown dramatically in the modern literature, however, the approach to the body of techniques has typically become more pragmatic. The Gibbs sampler and related techniques including the Metropolis–Hastings algorithm have enabled some remarkable simplifications of heretofore intractable problems. For example, recent developments in commercial

## CHAPTER 16 ♦ Bayesian Estimation and Inference 679

software have produced a wide choice of “mixed” estimators which are various implementations of the maximum likelihood procedures and hierarchical Bayes procedures (such as the Sawtooth and MLWin programs). Unless one is dealing with a small sample, the choice between these can be based on convenience. There is little methodological difference. This returns us to the practical point noted earlier. The choice between the Bayesian approach and the sampling theory method in this application would not be based on a fundamental methodological criterion, but on purely practical considerations—the end result is the same.

This chapter concludes our survey of estimation and inference methods in econometrics. We will now turn to two major areas of applications, time series and (broadly) macroeconometrics, and microeconometrics which is primarily oriented to cross-section and panel data applications.

### **Key Terms and Concepts**

- Bayes factor
- Bayes's theorem
- Bernstein–von Mises theorem
- Burn in
- Central limit theorem
- Conjugate prior
- Data augmentation
- Gibbs sampler
- Hierarchical Bayes
- Hierarchical prior
- Highest posterior density (HPD) interval
- Improper prior
- Informative prior
- Inverted gamma distribution
- Inverted Wishart
- Joint posterior distribution
- Likelihood function
- Loss function
- Markov–Chain Monte Carlo (MCMC)
- Metropolis–Hastings algorithm
- Multivariate  $t$  distribution
- Noninformative prior
- Normal-gamma prior
- Posterior density
- Posterior mean
- Precision matrix
- Predictive density
- Prior beliefs
- Prior density
- Prior distribution
- Prior odds ratio
- Prior probabilities
- Sampling theory
- Uniform prior
- Uniform-inverse gamma prior

### **Exercise**

1. Suppose the distribution of  $y_i | \lambda$  is Poisson,

$$f(y_i | \lambda) = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!} = \frac{\exp(-\lambda)\lambda^{y_i}}{\Gamma(y_i + 1)}, \quad y_i = 0, 1, \dots, \lambda > 0.$$

We will obtain a sample of observations,  $y_1, \dots, y_n$ . Suppose our prior for  $\lambda$  is the inverted gamma, which will imply

$$p(\lambda) \propto \frac{1}{\lambda}.$$

- a. Construct the likelihood function,  $p(y_1, \dots, y_n | \lambda)$ .
- b. Construct the posterior density

$$p(\lambda | y_1, \dots, y_n) = \frac{p(y_1, \dots, y_n | \lambda)p(\lambda)}{\int_0^{\infty} p(y_1, \dots, y_n | \lambda)p(\lambda)d\lambda}.$$

- c. Prove that the Bayesian estimator of  $\lambda$  is the posterior mean,  $E[\lambda | y_1, \dots, y_n] = \bar{y}$ .
- d. Prove that the posterior variance is  $\text{Var}[\lambda | y_1, \dots, y_n] = \bar{y}/n$ .

## 680 PART III ♦ Estimation Methodology

(Hint: You will make heavy use of gamma integrals in solving this problem. Also, you will find it convenient to use  $\Sigma_i y_i = n\bar{y}$ .)

### Application

- Consider a model for the mix of male and female children in families. Let  $K_i$  denote the family size (number of children),  $K_i = 1, \dots$ . Let  $F_i$  denote the number of female children,  $F_i = 0, \dots, K_i$ . Suppose the density for the number of female children in a family with  $K_i$  children is binomial with constant success probability  $\theta$ :

$$p(F_i|K_i, \theta) = \binom{K_i}{F_i} \theta^{F_i} (1 - \theta)^{K_i - F_i}.$$

We are interested in analyzing the “probability,”  $\theta$ . Suppose the (conjugate) prior over  $\theta$  is a beta distribution with parameters  $a$  and  $b$ :

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}.$$

Your sample of 25 observations is given here:

$K_i$	2	1	1	5	5	4	4	5	1	2	4	4	2	4	3	2	3	2	3	5	3	2	5	4	1
$F_i$	1	1	1	3	2	3	2	4	0	2	3	1	1	3	2	1	3	1	2	4	2	1	1	4	1

- Compute the classical maximum likelihood estimate of  $\theta$ .
- Form the posterior density for  $\theta$  given  $(K_i, F_i), i = 1, \dots, 25$  conditioned on  $a$  and  $b$ .
- Using your sample of data, compute the posterior mean assuming  $a = b = 1$ .
- Using your sample of data, compute the posterior mean assuming  $a = b = 2$ .
- Using your sample of data, compute the posterior mean assuming  $a = 1$  and  $b = 2$ .

# 17

## DISCRETE CHOICE

---

### 17.1 INTRODUCTION

This is the first of three chapters that will survey models used in **microeconomics**. The analysis of individual choice that is the focus of this field is fundamentally about modeling discrete outcomes such as purchase decisions, for example whether or not to buy insurance, voting behavior, choice among a set of alternative brands, travel modes or places to live, and responses to survey questions about the strength of preferences or about self-assessed health or well-being. In these and any number of other cases, the “dependent variable” is not a quantitative measure of some economic outcome, but rather an indicator of whether or not some outcome occurred. It follows that the regression methods we have used up to this point are largely inappropriate. We turn, instead, to modeling probabilities and using econometric tools to make probabilistic statements about the occurrence of these events. We will also examine models for counts of occurrences. These are closer to familiar regression models, but are, once again, about discrete outcomes of behavioral choices. As such, in this setting as well, we will be modeling probabilities of events, rather than conditional mean functions.

The models that are analyzed in this and the next chapter are built on a platform of preferences of decision makers. We take a **random utility** view of the choices that are observed. The decision maker is faced with a situation or set of alternatives and reveals something about their underlying preferences by the choice that he or she makes. The choice(s) made will be affected by observable influences—this is, of course, the ultimate objective of advertising—and by unobservable characteristics of the chooser. The blend of these fundamental bases for individual choice is at the core of the broad range of models that we will examine here.<sup>1</sup>

This chapter and Chapter 18 will describe four broad frameworks for analysis:

**Binary Choice:** The individual faces a pair of choices and makes that choice between the two that provides the greater utility. Many such settings involve the choice between taking an action and not taking that action, for example the decision whether or not to purchase health insurance. In other cases, the decision might be between two distinctly different choices, such as the decision whether to travel to and from work via public or private transportation. In the binary choice case, the 0/1 outcome is merely a label for “no/yes”—the numerical values are a mere convenience.

**Multinomial Choice:** The individual chooses among more than two choices, once again, making the choice that provides the greatest utility. In the previous example, private travel might involve a choice of being a driver or passenger while public

---

<sup>1</sup>See Greene and Hensher (2010, Chapter 4) for an historical perspective on this approach to model specification.

**682 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

transport might involve a choice between bus and train. At one level, this is a minor variation of the binary choice case—the latter is, of course, a special case of the former. But, more elaborate models of multinomial choice allow a rich specification of consumer preferences. In the multinomial case, the observed response is simply a label for the selected choice; it might be a brand, the name of a place, or the type of travel mode. Numerical assignments are not meaningful in this setting.

**Ordered Choice:** The individual reveals the strength of his or her preferences with respect to a single outcome. Familiar cases involve survey questions about strength of feelings about a particular commodity such as a movie, or self-assessments of social outcomes such as health in general or self-assessed well-being. In the ordered choice setting, opinions are given meaningful numeric values, usually  $0, 1, \dots, J$  for some upper limit,  $J$ . For example, opinions might be labelled  $0, 1, 2, 3, 4$  to indicate the strength of preferences, for example, for a product, a movie, a candidate or a piece of legislation. But, in this context, the numerical values are only a ranking, not a quantitative measure. Thus a “1” is greater than a “0” in a qualitative sense, but not by one unit, and the difference between a “2” and a “1” is not the same as that between a “1” and a “0.”

In these three cases, although the numerical outcomes are merely labels of some nonquantitative outcome, the analysis will nonetheless have a regression-style motivation. Throughout, the models will be based on the idea that observed “covariates” are relevant in explaining the observed choices. For example, in the binary outcome “did or did not purchase health insurance,” a conditioning model suggests that covariates such as age, income, and family situation will help to explain the choice. This chapter will describe a range of models that have been developed around these considerations. We will also be interested in a fourth application of discrete outcome models:

**Event Counts:** The observed outcome is a count of the number of occurrences. In many cases, this is similar to the preceding three settings in that the “dependent variable” measures an individual choice, such as the number of visits to the physician or the hospital, the number of derogatory reports in one’s credit history, or the number of visits to a particular recreation site. In other cases, the event count might be the outcome of some natural process, such as incidence of a disease in a population or the number of defects per unit of time in a production process. In this setting, we will be doing a more familiar sort of regression modeling. However, the models will still be constructed specifically to accommodate the discrete nature of the observed response variable.

We will consider these four cases in turn. The four broad areas have many elements in common; however, there are also substantive differences between the particular models and analysis techniques used in each. This chapter will develop the first topic, models for binary choices. In each section, we will begin with an overview of applications and then present the single basic model that is the centerpiece of the methodology, and, finally, examine some recently developed extensions of the model. This chapter contains a very lengthy discussion of models for binary choices. This analysis is as long as it is because, first, the models discussed are used throughout microeconomics—the central model of binary choice in this area is as ubiquitous as linear regression. Second, all the econometric issues and features that are encountered in the other areas will appear in the analysis of binary choice, where we can examine them in a fairly straightforward fashion.

It will emerge that, at least in econometric terms, the models for multinomial and ordered choice considered in Chapter 18 can be built from the two fundamental building blocks, the model of random utility and the translation of that model into a description of binary choices. There are relatively few new econometric issues that arise here. Chapter 18 will be largely devoted to suggesting different approaches to modeling choices among multiple alternatives and models for ordered choices. Once again, models of preference scales, such as movie or product ratings, or self-assessments of health or well-being, can be naturally built up from the fundamental model of random utility. Finally, Chapter 18 will develop the well-known Poisson regression model for counts of events. We will then extend the model to demonstrate some recent applications and innovations.

Chapters 17 and 18 are a lengthy but far from complete survey of topics in estimating **qualitative response (QR)** models. None of these models can consistently be estimated with linear regression methods. In most cases, the method of estimation is **maximum likelihood**. Therefore, readers interested in the mechanics of estimation may want to review the material in Appendices D and E before continuing. The various properties of maximum likelihood estimators are discussed in Chapter 14. We shall assume throughout these chapters that the necessary conditions behind the optimality properties of maximum likelihood estimators are met and, therefore, we will not derive or establish these properties specifically for the QR models. Detailed proofs for most of these models can be found in surveys by Amemiya (1981), McFadden (1984), Maddala (1983), and Dhrymes (1984). Additional commentary on some of the issues of interest in the contemporary literature is given by Manski and McFadden (1981) and Maddala and Flores-Lagunes (2001). Agresti (2002) and Cameron and Trivedi (2005) contain numerous theoretical developments and applications. Greene (2008) and Hensher and Greene (2010) provide, among many others, general surveys of discrete choice models and methods.<sup>2</sup>

## 17.2 MODELS FOR BINARY OUTCOMES

For purposes of studying individual behavior, we will construct models that link the decision or outcome to a set of factors, at least in the spirit of regression. Our approach will be to analyze each of them in the general framework of probability models:

$$\text{Prob}(\text{event } j \text{ occurs}) = \text{Prob}(Y = j) = F[\text{relevant effects, parameters}]. \quad (17-1)$$

The study of qualitative choice focuses on appropriate specification, estimation, and use of models for the probabilities of events, where in most cases, the “event” is an individual’s choice among a set of two or more alternatives.

### **Example 17.1 Labor Force Participation Model**

In Example 5.2 we estimated an earnings equation for the subsample of 428 married women who participated in the formal labor market taken from a full sample of 753 observations. The semilog earnings equation is of the form

$$\ln \text{earnings} = \beta_1 + \beta_2 \text{age} + \beta_3 \text{age}^2 + \beta_4 \text{education} + \beta_5 \text{kids} + \varepsilon,$$

---

<sup>2</sup>There are dozens of book length surveys of discrete choice models. Two others that are heavily oriented to application of the methods are Train (2003) and Hensher, Rose, and Greene (2005).

## 684 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

where *earnings* is hourly wage times *hours worked*, *education* is measured in years of schooling, and *kids* is a binary variable which equals one if there are children under 18 in the household. What of the other 325 individuals? The underlying labor supply model described a market in which labor force participation was the outcome of a market process whereby the demanders of labor services were willing to offer a wage based on expected marginal product and individuals themselves made a decision whether or not to accept the offer depending on whether it exceeded their own reservation wage. The first of these depends on, among other things, education, while the second (we assume) depends on such variables as age, the presence of children in the household, other sources of income (husband's), and marginal tax rates on labor income. The sample we used to fit the earnings equation contains data on all these other variables. The models considered in this chapter would be appropriate for modeling the outcome  $y = 1$  if in the labor force, and 0 if not.

Models for explaining a binary (0/1) dependent variable are typically motivated in two contexts. The labor force participation model in Example 17.1 describes a process of individual choice between two alternatives in which the choice is influenced by observable effects (children, tax rates) and unobservable aspects of the preferences of the individual. The relationship between voting behavior and income is another example. In other cases, the **binary choice model** arises in a setting in which the nature of the observed data dictate the special treatment of a binary dependent variable model. In these cases, the analyst is essentially interested in a regression-like model of the sort considered in Chapters 2 through 7. With data on the variable of interest and a set of covariates, they are interested in specifying a relationship between the former and the latter, more or less along the lines of the models we have already studied. For example, in a model of the demand for tickets for sporting events, in which the variable of interest is number of tickets, it could happen that the observation consists only of whether the sports facility was filled to capacity (demand greater than or equal to capacity so  $Y = 1$ ) or not ( $Y = 0$ ). It will generally turn out that the models and techniques used in both cases are the same. Nonetheless, it is useful to examine both of them.

### 17.2.1 RANDOM UTILITY MODELS FOR INDIVIDUAL CHOICE

An interpretation of data on individual choices is provided by the random utility model. Let  $U_a$  and  $U_b$  represent an individual's utility of two choices. For example,  $U_a$  might be the utility of rental housing and  $U_b$  that of home ownership. The observed choice between the two reveals which one provides the greater utility, but not the unobservable utilities. Hence, the observed indicator equals 1 if  $U_a > U_b$  and 0 if  $U_a \leq U_b$ . A common formulation is the linear random utility model,

$$U_a = \mathbf{w}'\boldsymbol{\beta}_a + \mathbf{z}_a'\boldsymbol{\gamma}_a + \varepsilon_a \quad \text{and} \quad U_b = \mathbf{w}'\boldsymbol{\beta}_b + \mathbf{z}_b'\boldsymbol{\gamma}_b + \varepsilon_b. \quad (17-2)$$

In (17-2), the observable (measurable) vector of **characteristics** of the individual is denoted  $\mathbf{w}$ ; this might include gender, age, income, and other demographics. The vectors  $\mathbf{z}_a$  and  $\mathbf{z}_b$  denote features (**attributes**) of the two choices that might be choice specific. In a voting context, for example, the attributes might be indicators of the competing candidates' positions on important issues. The random terms,  $\varepsilon_a$  and  $\varepsilon_b$  represent the stochastic elements that are specific to and known only by the individual, but not by the observer (analyst). To continue our voting example,  $\varepsilon_a$  might represent an intangible, general "preference" for candidate  $a$ .

The completion of the model for the determination of the observed outcome (choice) is the revelation of the ranking of the preferences by the choice the individual makes. Thus, if we denote by  $Y = 1$  the consumer's choice of alternative  $a$ , we infer from  $Y = 1$  that  $U_a > U_b$ . Since the outcome is ultimately driven by the random elements in the utility functions, we have

$$\begin{aligned}\text{Prob}[Y = 1 | \mathbf{w}, \mathbf{z}_a, \mathbf{z}_b] &= \text{Prob}[U_a > U_b] \\ &= \text{Prob}[(\mathbf{w}'\boldsymbol{\beta}_a + \mathbf{z}_a'\boldsymbol{\gamma}_a + \varepsilon_a) - (\mathbf{x}'\boldsymbol{\beta}_b + \mathbf{z}_b'\boldsymbol{\gamma}_b + \varepsilon_b) > 0 | \mathbf{w}, \mathbf{z}_a, \mathbf{z}_b] \\ &= \text{Prob}[(\mathbf{w}'(\boldsymbol{\beta}_a - \boldsymbol{\beta}_b) + \mathbf{z}_a'\boldsymbol{\gamma}_a - \mathbf{z}_b'\boldsymbol{\gamma}_b + \varepsilon_a - \varepsilon_b) > 0 | \mathbf{w}, \mathbf{z}_a, \mathbf{z}_b] \\ &= \text{Prob}[\mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0 | \mathbf{x}],\end{aligned}$$

where  $\mathbf{x}'\boldsymbol{\beta}$  collects all the observable elements of the difference of the two utility functions and  $\varepsilon$  denotes the difference between the two random elements.

**Example 17.2 Structural Equations for a Binary Choice Model**

Nakosteen and Zimmer (1980) analyzed a model of migration based on the following structure:<sup>3</sup> For a given individual, the market wage that can be earned at the present location is

$$y_p^* = \mathbf{w}'_p \boldsymbol{\beta}_p + \varepsilon_p.$$

Variables in the equation include age, sex, race, growth in employment, and growth in per capita income. If the individual migrates to a new location, then his or her market wage would be

$$y_m^* = \mathbf{w}'_m \boldsymbol{\beta}_m + \varepsilon_m.$$

Migration entails costs that are related both to the individual and to the labor market:

$$C^* = \mathbf{z}'\boldsymbol{\alpha} + u.$$

Costs of moving are related to whether the individual is self-employed and whether that person recently changed his or her industry of employment. They migrate if the benefit  $y_m^* - y_p^*$  is greater than the cost,  $C$ . The net benefit of moving is

$$\begin{aligned}M^* &= y_m^* - y_p^* - C^* \\ &= \mathbf{w}'_m \boldsymbol{\beta}_m - \mathbf{w}'_p \boldsymbol{\beta}_p - \mathbf{z}'\boldsymbol{\alpha} + (\varepsilon_m - \varepsilon_p - u) \\ &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon.\end{aligned}$$

Because  $M^*$  is unobservable, we cannot treat this equation as an ordinary regression. The individual either moves or does not. After the fact, we observe only  $y_m^*$  if the individual has moved or  $y_p^*$  if he or she has not. But we do observe that  $M = 1$  for a move and  $M = 0$  for no move.

<sup>3</sup>A number of other studies have also used variants of this basic formulation. Some important examples are Willis and Rosen (1979) and Robinson and Tomes (1982). The study by Tunali (1986) examined in Example 17.6 is another application. The now standard approach, in which "participation" equals one if wage offer  $(\mathbf{x}'_w \boldsymbol{\beta}_w + \varepsilon_w)$  minus reservation wage  $(\mathbf{x}'_r \boldsymbol{\beta}_r + \varepsilon_r)$  is positive, is also used in Fernandez and Rodriguez-Poo (1997). Brock and Durlauf (2000) describe a number of models and situations involving individual behavior that give rise to binary choice models.

## 686 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

### 17.2.2 A LATENT REGRESSION MODEL

Discrete dependent-variable models are often cast in the form of **index function models**. We view the outcome of a discrete choice as a reflection of an underlying regression. As an often-cited example, consider the decision to make a large purchase. The theory states that the consumer makes a marginal benefit/marginal cost calculation based on the utilities achieved by making the purchase and by not making the purchase and by using the money for something else. We model the difference between benefit and cost as an unobserved variable  $y^*$  such that

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon.$$

Note that this is the result of the “net utility” calculation in the previous section and in Example 17.2. We assume that  $\varepsilon$  has mean zero and has either a standardized logistic with variance  $\pi^2/3$  or a standard normal distribution with variance one or some other specific distribution with known variance. We do not observe the net benefit of the purchase (i.e., net utility), only whether it is made or not. Therefore, our observation is

$$\begin{aligned} y &= 1 && \text{if } y^* > 0, \\ y &= 0 && \text{if } y^* \leq 0. \end{aligned} \tag{17-3}$$

In this formulation,  $\mathbf{x}'\boldsymbol{\beta}$  is called the index function. The assumption of known variance of  $\varepsilon$  is an innocent normalization. Suppose the variance of  $\varepsilon$  is scaled by an unrestricted parameter  $\sigma^2$ . The **latent regression** will be  $y^* = \mathbf{x}'\boldsymbol{\beta} + \sigma\varepsilon$ . But,  $(y^*/\sigma) = \mathbf{x}'(\boldsymbol{\beta}/\sigma) + \varepsilon$  is the same model with the same data. The observed data will be unchanged;  $y$  is still 0 or 1, depending only on the sign of  $y^*$  not on its scale. This means that there is no information about  $\sigma$  in the sample data so  $\sigma$  cannot be estimated. The parameter vector  $\boldsymbol{\beta}$  in this model is only “identified up to scale.” The assumption of zero for the threshold in (17-3) is likewise innocent if the model contains a constant term (and not if it does not).<sup>4</sup> Let  $a$  be the supposed nonzero threshold and  $\alpha$  be the unknown constant term and, for the present,  $\mathbf{x}$  and  $\boldsymbol{\beta}$  contain the rest of the index not including the constant term. Then, the probability that  $y$  equals one is

$$\text{Prob}(y^* > a | \mathbf{x}) = \text{Prob}(\alpha + \mathbf{x}'\boldsymbol{\beta} + \varepsilon > a | \mathbf{x}) = \text{Prob}[(\alpha - a) + \mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0 | \mathbf{x}].$$

Because  $\alpha$  is unknown, the difference  $(\alpha - a)$  remains an unknown parameter. The end result is that if the model contains a constant term, it is unchanged by the choice of the threshold in (17-3). The choice of zero is a normalization with no significance. With the two normalizations, then,

$$\text{Prob}(y^* > 0 | \mathbf{x}) = \text{Prob}(\varepsilon > -\mathbf{x}'\boldsymbol{\beta} | \mathbf{x}).$$

A remaining detail in the model is the choice of the specific distribution for  $\varepsilon$ . We will consider several. The overwhelming majority of applications are based either on the normal or the logistic distribution. If the distribution is symmetric, as are the normal and logistic, then

$$\text{Prob}(y^* > 0 | \mathbf{x}) = \text{Prob}(\varepsilon < \mathbf{x}'\boldsymbol{\beta} | \mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta}), \tag{17-4}$$

---

<sup>4</sup>Unless there is some compelling reason, binomial probability models should not be estimated without constant terms.

where  $F(t)$  is the cdf of the random variable,  $\varepsilon$ . This provides an underlying structural model for the probability.

### 17.2.3 FUNCTIONAL FORM AND REGRESSION

Consider the model of labor force participation suggested in Example 17.1. The respondent either works or seeks work ( $Y = 1$ ) or does not ( $Y = 0$ ) in the period in which our survey is taken. We believe that a set of factors, such as age, marital status, education, and work history, gathered in a vector  $\mathbf{x}$ , explain the decision, so that

$$\begin{aligned} \text{Prob}(Y = 1 | \mathbf{x}) &= F(\mathbf{x}, \boldsymbol{\beta}) \\ \text{Prob}(Y = 0 | \mathbf{x}) &= 1 - F(\mathbf{x}, \boldsymbol{\beta}). \end{aligned} \quad (17-5)$$

The set of parameters  $\boldsymbol{\beta}$  reflects the impact of changes in  $\mathbf{x}$  on the probability. For example, among the factors that might interest us is the marginal effect of marital status on the probability of labor force participation. The problem at this point is to devise a suitable model for the right-hand side of the equation. One possibility is to retain the familiar linear regression,

$$F(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}.$$

Because  $E[y | \mathbf{x}] = 0[1 - F(\mathbf{x}, \boldsymbol{\beta})] + 1[F(\mathbf{x}, \boldsymbol{\beta})] = F(\mathbf{x}, \boldsymbol{\beta})$ , we can construct the regression model,

$$\begin{aligned} y &= E[y | \mathbf{x}] + y - E[y | \mathbf{x}] \\ &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon. \end{aligned} \quad (17-6)$$

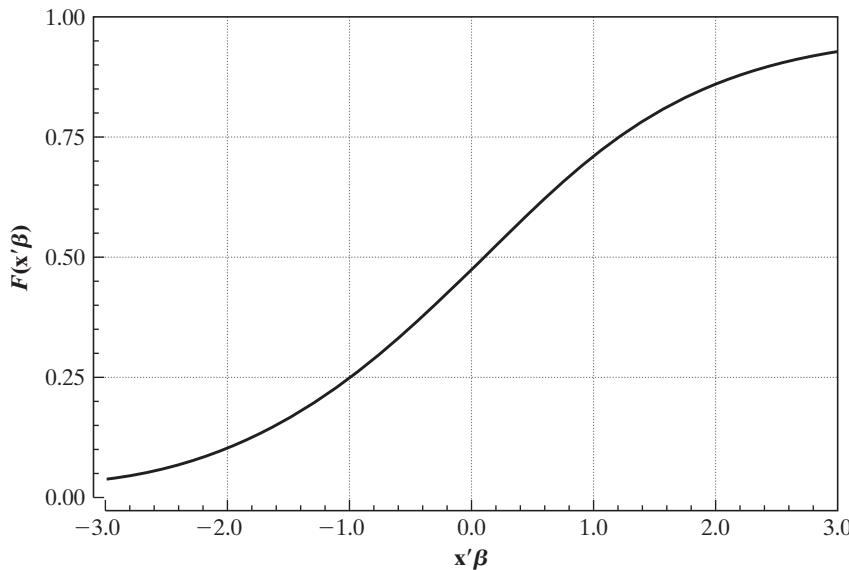
The **linear probability model** has a number of shortcomings. A minor complication arises because  $\varepsilon$  is heteroscedastic in a way that depends on  $\boldsymbol{\beta}$ . Because  $\mathbf{x}'\boldsymbol{\beta} + \varepsilon$  must equal 0 or 1,  $\varepsilon$  equals either  $-\mathbf{x}'\boldsymbol{\beta}$  or  $1 - \mathbf{x}'\boldsymbol{\beta}$ , with probabilities  $1 - F$  and  $F$ , respectively. Thus, you can easily show that in this model,

$$\text{Var}[\varepsilon | \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}(1 - \mathbf{x}'\boldsymbol{\beta}). \quad (17-7)$$

We could manage this complication with an FGLS estimator in the fashion of Chapter 9, though this only solves the estimation problem, not the theoretical one. A more serious flaw is that without some ad hoc tinkering with the disturbances, we cannot be assured that the predictions from this model will truly look like probabilities. We cannot constrain  $\mathbf{x}'\boldsymbol{\beta}$  to the 0–1 interval. Such a model produces both nonsense probabilities and negative variances. For these reasons, the linear probability model is becoming less frequently used except as a basis for comparison to some other more appropriate models.<sup>5</sup>

---

<sup>5</sup>The linear model is not beyond redemption. Aldrich and Nelson (1984) analyze the properties of the model at length. Judge et al. (1985) and Fomby, Hill, and Johnson (1984) give interesting discussions of the ways we may modify the model to force internal consistency. But the fixes are sample dependent, and the resulting estimator, such as it is, may have no known sampling properties. Additional discussion of weighted least squares appears in Amemiya (1977) and Mullahy (1990). Finally, its shortcomings notwithstanding, the linear probability model is applied by Caudill (1988), Heckman, and MaCurdy (1985), and Heckman and Snyder (1997). An exchange on the usefulness of the approach is Angrist (2001) and Moffitt (2001). See Angrist and Pischke (2009) for some applications.

**688 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**


**FIGURE 17.1** Model for a Probability.

Our requirement, then, is a model that will produce predictions consistent with the underlying theory in (17-4). For a given regressor vector, we would expect

$$\begin{aligned} \lim_{\mathbf{x}'\boldsymbol{\beta} \rightarrow +\infty} \text{Prob}(Y = 1 | \mathbf{x}) &= 1 \\ \lim_{\mathbf{x}'\boldsymbol{\beta} \rightarrow -\infty} \text{Prob}(Y = 1 | \mathbf{x}) &= 0. \end{aligned} \quad (17-8)$$

See Figure 17.1. In principle, any proper, continuous probability distribution defined over the real line will suffice. The normal distribution has been used in many analyses, giving rise to the **probit** model,

$$\text{Prob}(Y = 1 | \mathbf{x}) = \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \phi(t) dt = \Phi(\mathbf{x}'\boldsymbol{\beta}). \quad (17-9)$$

The function  $\Phi(t)$  is a commonly used notation for the standard normal distribution function. Partly because of its mathematical convenience, the logistic distribution,

$$\text{Prob}(Y = 1 | \mathbf{x}) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})} = \Lambda(\mathbf{x}'\boldsymbol{\beta}). \quad (17-10)$$

has also been used in many applications. We shall use the notation  $\Lambda(\cdot)$  to indicate the logistic cumulative distribution function. This model is called the **logit** model for reasons we shall discuss in the next section. Both of these distributions have the familiar bell shape of symmetric distributions. Other models which do not assume symmetry, such as the **Gumbel model**,

$$\text{Prob}(Y = 1 | \mathbf{x}) = \exp[-\exp(-\mathbf{x}'\boldsymbol{\beta})],$$

and **complementary log log model**,

$$\text{Prob}(Y = 1 | \mathbf{x}) = 1 - \exp[-\exp(\mathbf{x}'\boldsymbol{\beta})],$$

have also been employed. Still other distributions have been suggested,<sup>6</sup> but the probit and logit models are still the most common frameworks used in econometric applications.

The question of which distribution to use is a natural one. The logistic distribution is similar to the normal except in the tails, which are considerably heavier. (It more closely resembles a *t* distribution with seven degrees of freedom.) Therefore, for intermediate values of  $\mathbf{x}'\boldsymbol{\beta}$  (say, between  $-1.2$  and  $+1.2$ ), the two distributions tend to give similar probabilities. The logistic distribution tends to give larger probabilities to  $Y = 1$  when  $\mathbf{x}'\boldsymbol{\beta}$  is extremely small (and smaller probabilities to  $Y = 1$  when  $\mathbf{x}'\boldsymbol{\beta}$  is very large) than the normal distribution. It is difficult to provide practical generalities on this basis, however, as they would require knowledge of  $\boldsymbol{\beta}$ . We should expect different predictions from the two models, however, if the sample contains (1) very few “responses” ( $Y$ 's equal to 1) or very few “nonresponses” ( $Y$ 's equal to 0) and (2) very wide variation in an important independent variable, particularly if (1) is also true. There are practical reasons for favoring one or the other in some cases for mathematical convenience, but it is difficult to justify the choice of one distribution or another on theoretical grounds. Amemiya (1981) discusses a number of related issues, but as a general proposition, the question is unresolved. In most applications, the choice between these two seems not to make much difference. However, as seen in the following example, the symmetric and asymmetric distributions can give substantively different results, and here, the guidance on how to choose is unfortunately sparse.

The probability model is a regression:

$$E[y | \mathbf{x}] = F(\mathbf{x}'\boldsymbol{\beta}).$$

Whatever distribution is used, it is important to note that the parameters of the model, like those of any nonlinear regression model, are not necessarily the marginal effects we are accustomed to analyzing. In general,

$$\frac{\partial E[y | \mathbf{x}]}{\partial \mathbf{x}} = \left[ \frac{dF(\mathbf{x}'\boldsymbol{\beta})}{d(\mathbf{x}'\boldsymbol{\beta})} \right] \times \boldsymbol{\beta} = f(\mathbf{x}'\boldsymbol{\beta}) \times \boldsymbol{\beta}, \quad (17-11)$$

where  $f(\cdot)$  is the density function that corresponds to the cumulative distribution,  $F(\cdot)$ . For the normal distribution, this result is

$$\frac{\partial E[y | \mathbf{x}]}{\partial \mathbf{x}} = \phi(\mathbf{x}'\boldsymbol{\beta}) \times \boldsymbol{\beta}, \quad (17-12)$$

where  $\phi(t)$  is the standard normal density. For the logistic distribution,

$$\frac{d\Lambda(\mathbf{x}'\boldsymbol{\beta})}{d(\mathbf{x}'\boldsymbol{\beta})} = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{[1 + \exp(\mathbf{x}'\boldsymbol{\beta})]^2} = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})],$$

so, in the logit model,

$$\frac{\partial E[y | \mathbf{x}]}{\partial \mathbf{x}} = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})]\boldsymbol{\beta}. \quad (17-13)$$

---

<sup>6</sup>See, for example, Maddala (1983, pp. 27–32), Aldrich and Nelson (1984), and Greene (2001).

## 690 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

It is obvious that these values will vary with the values of  $\mathbf{x}$ . In interpreting the estimated model, it will be useful to calculate this value at, say, the means of the regressors and, where necessary, other pertinent values. For convenience, it is worth noting that the same scale factor applies to all the slopes in the model.

For computing **marginal effects**, one can evaluate the expressions at the sample means of the data or evaluate the marginal effects at every observation and use the sample average of the individual marginal effects—this produces the **average partial effects**. In large samples these generally give roughly the same answer (see Section 17.3.2). But that is not so in small- or moderate-sized samples. Current practice favors averaging the individual marginal effects when it is possible to do so.

Another complication for computing marginal effects in a binary choice model arises because  $\mathbf{x}$  will often include dummy variables—for example, a labor force participation equation will often contain a dummy variable for marital status. Because the derivative is with respect to a small change, it is not appropriate to apply (15) for the effect of a change in a dummy variable, or a change of state. The appropriate marginal effect for a binary independent variable, say,  $d$ , would be

$$\text{Marginal effect} = \text{Prob}[Y = 1 | \bar{\mathbf{x}}_{(d)}, d = 1] - \text{Prob}[Y = 1 | \bar{\mathbf{x}}_{(d)}, d = 0], \quad (17-14)$$

where  $\bar{\mathbf{x}}_{(d)}$  denotes the means of all the other variables in the model. Simply taking the derivative with respect to the binary variable as if it were continuous provides an approximation that is often surprisingly accurate. In Example 17.3, for the binary variable  $PSI$ , the difference in the two probabilities for the probit model is  $(0.5702 - 0.1057) = 0.4645$ , whereas the derivative approximation reported in Table 17.1 is 0.468. Nonetheless, it might be optimistic to rely on this outcome. We will revisit this computation in the examples and discussion to follow.

### 17.3 ESTIMATION AND INFERENCE IN BINARY CHOICE MODELS

With the exception of the linear probability model, estimation of binary choice models is usually based on the method of maximum likelihood. Each observation is treated as a single draw from a Bernoulli distribution (binomial with one draw). The model with success probability  $F(\mathbf{x}'\boldsymbol{\beta})$  and independent observations leads to the joint probability, or likelihood function,

$$\text{Prob}(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n | \mathbf{X}) = \prod_{y_i=0} [1 - F(\mathbf{x}'_i \boldsymbol{\beta})] \prod_{y_i=1} F(\mathbf{x}'_i \boldsymbol{\beta}).$$

where  $\mathbf{X}$  denotes  $[\mathbf{x}_i]_{i=1,\dots,n}$ . The likelihood function for a sample of  $n$  observations can be conveniently written as

$$L(\boldsymbol{\beta} | \text{data}) = \prod_{i=1}^n [F(\mathbf{x}'_i \boldsymbol{\beta})]^{y_i} [1 - F(\mathbf{x}'_i \boldsymbol{\beta})]^{1-y_i}. \quad (17-15)$$

Taking logs, we obtain

$$\ln L = \sum_{i=1}^n \{y_i \ln F(\mathbf{x}'_i \boldsymbol{\beta}) + (1 - y_i) \ln[1 - F(\mathbf{x}'_i \boldsymbol{\beta})]\}.^7 \quad (17-16)$$

The **likelihood equations** are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \left[ \frac{y_i f_i}{F_i} + (1 - y_i) \frac{-f_i}{(1 - F_i)} \right] \mathbf{x}_i = \mathbf{0}, \quad (17-17)$$

where  $f_i$  is the density,  $dF_i/d(\mathbf{x}'_i \boldsymbol{\beta})$ . [In (17-17) and later, we will use the subscript  $i$  to indicate that the function has an argument  $\mathbf{x}'_i \boldsymbol{\beta}$ .] The choice of a particular form for  $F_i$  leads to the empirical model.

Unless we are using the linear probability model, the likelihood equations in (17-17) will be nonlinear and require an iterative solution. All of the models we have seen thus far are relatively straightforward to analyze. For the logit model, by inserting (17-7) and (17-8) in (17-17), we get, after a bit of manipulation, the likelihood equations

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n (y_i - \Lambda_i) \mathbf{x}_i = \mathbf{0}. \quad (17-18)$$

Note that if  $\mathbf{x}_i$  contains a constant term, the first-order conditions imply that the average of the predicted probabilities must equal the proportion of ones in the sample.<sup>8</sup> This implication also bears some similarity to the least squares normal equations if we view the term  $y_i - \Lambda_i$  as a residual.<sup>9</sup> For the normal distribution, the log-likelihood is

$$\ln L = \sum_{y_i=0} \ln[1 - \Phi(\mathbf{x}'_i \boldsymbol{\beta})] + \sum_{y_i=1} \ln \Phi(\mathbf{x}'_i \boldsymbol{\beta}). \quad (17-19)$$

The first-order conditions for maximizing  $\ln L$  are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{y_i=0} \frac{-\phi(\mathbf{x}'_i \boldsymbol{\beta})}{1 - \Phi(\mathbf{x}'_i \boldsymbol{\beta})} + \sum_{y_i=1} \frac{\phi_i}{\Phi_i} \mathbf{x}_i = \sum_{y_i=0} \lambda_{0i} \mathbf{x}_i + \sum_{y_i=1} \lambda_{1i} \mathbf{x}_i.$$

Using the device suggested in footnote 7, we can reduce this to

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \left[ \frac{q_i \mathbf{x}'_i \boldsymbol{\beta}}{\Phi(q_i \mathbf{x}'_i \boldsymbol{\beta})} \right] \mathbf{x}_i = \sum_{i=1}^n \lambda_i \mathbf{x}_i = \mathbf{0}, \quad (17-20)$$

where  $q_i = 2y_i - 1$ .

The actual second derivatives for the logit model are quite simple:

$$\mathbf{H} = \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = - \sum_i \Lambda_i (1 - \Lambda_i) \mathbf{x}_i \mathbf{x}'_i. \quad (17-21)$$

<sup>7</sup>If the distribution is symmetric, as the normal and logistic are, then  $1 - F(-\mathbf{x}' \boldsymbol{\beta}) = F(-\mathbf{x}' \boldsymbol{\beta})$ . There is a further simplification. Let  $q = 2y - 1$ . Then  $\ln L = \sum_i \ln F(q_i \mathbf{x}'_i \boldsymbol{\beta})$ . See (17-21).

<sup>8</sup>The same result holds for the linear probability model. Although regularly observed in practice, the result has not been verified for the probit model.

<sup>9</sup>This sort of construction arises in many models. The first derivative of the log-likelihood with respect to the constant term produces the **generalized residual** in many settings. See, for example, Chesher, Lancaster, and Irish (1985) and the equivalent result for the tobit model in Section 19.3.4.d.

## 692 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

The second derivatives do not involve the random variable  $y_i$ , so Newton's method is also the **method of scoring** for the logit model. Note that the Hessian is always negative definite, so the log-likelihood is globally concave. Newton's method will usually converge to the maximum of the log-likelihood in just a few iterations unless the data are especially badly conditioned. The computation is slightly more involved for the probit model. A useful simplification is obtained by using the variable  $\lambda(y_i, \beta' \mathbf{x}_i) = \lambda_i$  that is defined in (17-20). The second derivatives can be obtained using the result that for any  $z$ ,  $d\phi(z)/dz = -z\phi(z)$ . Then, for the probit model,

$$\mathbf{H} = \frac{\partial^2 \ln L}{\partial \beta \partial \beta'} = \sum_{i=1}^n -\lambda_i (\lambda_i + \mathbf{x}'_i \beta) \mathbf{x}_i \mathbf{x}'_i. \quad (17-22)$$

This matrix is also negative definite for all values of  $\beta$ . The proof is less obvious than for the logit model.<sup>10</sup> It suffices to note that the scalar part in the summation is  $\text{Var}[\varepsilon | \varepsilon \leq \beta' \mathbf{x}] - 1$  when  $y = 1$  and  $\text{Var}[\varepsilon | \varepsilon \geq -\beta' \mathbf{x}] - 1$  when  $y = 0$ . The unconditional variance is one. Because truncation always reduces variance—see Theorem 18.2—in both cases, the variance is between zero and one, so the value is negative.<sup>11</sup>

The asymptotic covariance matrix for the maximum likelihood estimator can be estimated by using the inverse of the Hessian evaluated at the maximum likelihood estimates. There are also two other estimators available. The Berndt, Hall, Hall, and Hausman estimator [see (14-18) and Example 14.4] would be

$$\mathbf{B} = \sum_{i=1}^n g_i^2 \mathbf{x}_i \mathbf{x}'_i,$$

where  $g_i = (y_i - \Lambda_i)$  for the logit model [see (17-18)] and  $g_i = \lambda_i$  for the probit model [see (17-20)]. The third estimator would be based on the expected value of the Hessian. As we saw earlier, the Hessian for the logit model does not involve  $y_i$ , so  $\mathbf{H} = E[\mathbf{H}]$ . But because  $\lambda_i$  is a function of  $y_i$  [see (17-20)], this result is not true for the probit model. Amemiya (1981) showed that for the probit model,

$$E \left[ \frac{\partial^2 \ln L}{\partial \beta \partial \beta'} \right]_{\text{probit}} = \sum_{i=1}^n \lambda_{0i} \lambda_{1i} \mathbf{x}_i \mathbf{x}'_i. \quad (17-23)$$

Once again, the scalar part of the expression is always negative [see (17-20)] and note that  $\lambda_{0i}$  is always negative and  $\lambda_{1i}$  is always positive]. The estimator of the asymptotic covariance matrix for the maximum likelihood estimator is then the negative inverse of whatever matrix is used to estimate the expected Hessian. Since the actual Hessian is generally used for the iterations, this option is the usual choice. As we shall see later, though, for certain hypothesis tests, the BHHH estimator is a more convenient choice.

### 17.3.1 ROBUST COVARIANCE MATRIX ESTIMATION

The probit maximum likelihood estimator is often labeled a **quasi-maximum likelihood estimator (QMLE)** in view of the possibility that the normal probability model might be misspecified. White's (1982a) robust "sandwich" estimator for the asymptotic

<sup>10</sup>See, for example, Amemiya (1985, pp. 273–274) and Maddala (1983, p. 63).

<sup>11</sup>See Johnson and Kotz (1993) and Heckman (1979). We will make repeated use of this result in Chapter 19.

covariance matrix of the QMLE (see Section 14.8 for discussion),

$$\text{Est. Asy. Var}[\hat{\beta}] = [\hat{\mathbf{H}}]^{-1} \hat{\mathbf{B}} [\hat{\mathbf{H}}]^{-1},$$

has been used in a number of recent studies based on the probit model [e.g., Fernandez and Rodriguez-Poo (1997), Horowitz (1993), and Blundell, Laisney, and Lechner (1993)]. If the probit model is correctly specified, then  $\text{plim}(1/n)\hat{\mathbf{B}} = \mathbb{E}(1/n)(-\hat{\mathbf{H}})$  and either single matrix will suffice, so the robustness issue is moot (of course). On the other hand, the probit ( $Q$ -) maximum likelihood estimator is *not* consistent in the presence of any form of heteroscedasticity, unmeasured heterogeneity, omitted variables (even if they are orthogonal to the included ones), nonlinearity of the functional form of the index, or an error in the distributional assumption [with some narrow exceptions as described by Ruud (1986)]. Thus, in almost any case, the sandwich estimator provides an appropriate asymptotic covariance matrix for an estimator that is biased in an unknown direction. [See Section 14.8 and Freedman (2006).] White raises this issue explicitly, although it seems to receive little attention in the literature: "It is the consistency of the QMLE for the parameters of interest in a wide range of situations which insures its usefulness as the basis for robust estimation techniques" (1982a, p. 4). His very useful result is that if the quasi-maximum likelihood estimator converges to a probability limit, then the sandwich estimator can, under certain circumstances, be used to estimate the asymptotic covariance matrix of that estimator. But there is no guarantee that the QMLE *will* converge to anything interesting or useful. Simply computing a robust covariance matrix for an otherwise inconsistent estimator does not give it redemption. Consequently, the virtue of a robust covariance matrix in this setting is unclear.

### 17.3.2 MARGINAL EFFECTS AND AVERAGE PARTIAL EFFECTS

The predicted probabilities,  $F(\mathbf{x}'\hat{\beta}) = \hat{F}$  and the estimated partial effects  $f(\mathbf{x}'\hat{\beta}) \times \hat{\beta} = \hat{f}\hat{\beta}$  are nonlinear functions of the parameter estimates. To compute standard errors, we can use the linear approximation approach (delta method) discussed in Section 4.4.4. For the predicted probabilities,

$$\text{Asy. Var}[\hat{F}] = [\partial \hat{F} / \partial \hat{\beta}]' \mathbf{V} [\partial \hat{F} / \partial \hat{\beta}],$$

where

$$\mathbf{V} = \text{Asy. Var}[\hat{\beta}].$$

The estimated asymptotic covariance matrix of  $\hat{\beta}$  can be any of the three described earlier. Let  $z = \mathbf{x}'\hat{\beta}$ . Then the derivative vector is

$$[\partial \hat{F} / \partial \hat{\beta}] = [d\hat{F} / dz][\partial z / \partial \hat{\beta}] = \hat{f}\mathbf{x}.$$

Combining terms gives

$$\text{Asy. Var}[\hat{F}] = \hat{f}^2 \mathbf{x}' \mathbf{V} \mathbf{x},$$

which depends, of course, on the particular  $\mathbf{x}$  vector used. This result is useful when a marginal effect is computed for a dummy variable. In that case, the estimated effect is

$$\Delta \hat{F} = \hat{F}|_{(d=1)} - \hat{F}|_{(d=0)}. \quad (17-24)$$

## 694 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

The asymptotic variance would be

$$\text{Asy. Var}[\Delta \hat{F}] = [\partial \Delta \hat{F} / \partial \hat{\beta}]' \mathbf{V} [\partial \Delta \hat{F} / \partial \hat{\beta}], \quad (17-25)$$

where

$$[\partial \Delta \hat{F} / \partial \hat{\beta}] = \hat{f}_1 \begin{pmatrix} \bar{\mathbf{x}}_{(d)} \\ 1 \end{pmatrix} - \hat{f}_0 \begin{pmatrix} \bar{\mathbf{x}}_{(d)} \\ 0 \end{pmatrix}.$$

For the other marginal effects, let  $\hat{y} = \hat{f}\hat{\beta}$ . Then

$$\text{Asy. Var}[\hat{y}] = \left[ \frac{\partial \hat{y}}{\partial \hat{\beta}'} \right] \mathbf{V} \left[ \frac{\partial \hat{y}}{\partial \hat{\beta}'} \right]'.$$

The matrix of derivatives is

$$\hat{f} \left( \frac{\partial \hat{\beta}}{\partial \hat{\beta}'} \right) + \hat{\beta} \left( \frac{d\hat{f}}{dz} \right) \left( \frac{\partial z}{\partial \hat{\beta}'} \right) = \hat{f} \mathbf{I} + \left( \frac{d\hat{f}}{dz} \right) \hat{\beta} \mathbf{x}'.$$

For the probit model,  $df/dz = -z\phi$ , so

$$\text{Asy. Var}[\hat{y}] = \phi^2 [\mathbf{I} - (\mathbf{x}' \beta) \beta \mathbf{x}'] \mathbf{V} [\mathbf{I} - (\mathbf{x}' \beta) \beta \mathbf{x}']'.$$

For the logit model,  $\hat{f} = \hat{\Lambda}(1 - \hat{\Lambda})$ , so

$$\frac{d\hat{f}}{dz} = (1 - 2\hat{\Lambda}) \left( \frac{d\hat{\Lambda}}{dz} \right) = (1 - 2\hat{\Lambda})\hat{\Lambda}(1 - \hat{\Lambda}).$$

Collecting terms, we obtain

$$\text{Asy. Var}[\hat{y}] = [\Lambda(1 - \Lambda)]^2 [\mathbf{I} + (1 - 2\Lambda)\beta \mathbf{x}'] \mathbf{V} [\mathbf{I} + (1 - 2\Lambda)\mathbf{x} \beta'].$$

As before, the value obtained will depend on the  $\mathbf{x}$  vector used.

### Example 17.3 Probability Models

The data listed in Appendix Table F14.1 were taken from a study by Spector and Mazzeo (1980), which examined whether a new method of teaching economics, the Personalized System of Instruction (*PSI*), significantly influenced performance in later economics courses. The “dependent variable” used in our application is *GRADE*, which indicates the whether a student’s grade in an intermediate macroeconomics course was higher than that in the principles course. The other variables are *GPA*, their grade point average; *TUCE*, the score on a pretest that indicates entering knowledge of the material; and *PSI*, the binary variable indicator of whether the student was exposed to the new teaching method. (Spector and Mazzeo’s specific equation was somewhat different from the one estimated here.)

Table 17.1 presents four sets of parameter estimates. The slope parameters and derivatives were computed for four probability models: linear, probit, logit, and complementary log log. The last three sets of estimates are computed by maximizing the appropriate log-likelihood function. Inference is discussed in the next section, so standard errors are not presented here. The scale factor given in the last row is the density function evaluated at the means of the variables. Also, note that the slope given for *PSI* is the derivative, not the change in the function with *PSI* changed from zero to one with other variables held constant.

If one looked only at the coefficient estimates, then it would be natural to conclude that the four models had produced radically different estimates. But a comparison of the columns of slopes shows that this conclusion is clearly wrong. The models are very similar; in fact, the logit and probit models results are nearly identical.

The data used in this example are only moderately unbalanced between 0s and 1s for the dependent variable (21 and 11). As such, we might expect similar results for the probit

**TABLE 17.1** Estimated Probability Models

Variable	Linear		Logistic		Probit		Complementary log log	
	Coefficient	Slope	Coefficient	Slope	Coefficient	Slope	Coefficient	Slope
Constant	-1.498	—	-13.021	—	-7.452	—	-10.631	—
GPA	0.464	0.464	2.826	0.534	1.626	0.533	2.293	0.477
TUCE	0.010	0.010	0.095	0.018	0.052	0.017	0.041	0.009
PSI	0.379	0.379	2.379	0.450	1.426	0.468	1.562	0.325
$f(\bar{x}'\hat{\beta})$	1.000		0.189		0.328		0.208	

and logit models.<sup>12</sup> One indicator is a comparison of the coefficients. In view of the different variances of the distributions, one for the normal and  $\pi^2/3$  for the logistic, we might expect to obtain comparable estimates by multiplying the probit coefficients by  $\pi/\sqrt{3} \approx 1.8$ . Amemiya (1981) found, through trial and error, that scaling by 1.6 instead produced better results. This proportionality result is frequently cited. The result in (17-11) may help to explain the finding. The index  $\mathbf{x}'\beta$  is not the random variable. The marginal effect in the probit model for, say,  $x_k$  is  $\phi(\mathbf{x}'\beta_p)\beta_{pk}$ , whereas that for the logit is  $\Lambda(1 - \Lambda)\beta_{lk}$ . (The subscripts  $p$  and  $l$  are for probit and logit.) Amemiya suggests that his approximation works best at the center of the distribution, where  $F = 0.5$ , or  $\mathbf{x}'\beta = 0$  for either distribution. Suppose it is. Then  $\phi(0) = 0.3989$  and  $\Lambda(0)[1 - \Lambda(0)] = 0.25$ . If the marginal effects are to be the same, then  $0.3989 \beta_{pk} = 0.25 \beta_{lk}$ , or  $\beta_{lk} = 1.6 \beta_{pk}$ , which is the regularity observed by Amemiya. Note, though, that as we depart from the center of the distribution, the relationship will move away from 1.6. Because the logistic density descends more slowly than the normal, for unbalanced samples such as ours, the ratio of the logit coefficients to the probit coefficients will tend to be larger than 1.6. The ratios for the ones in Table 17.1 are closer to 1.7 than 1.6.

The computation of the derivatives of the conditional mean function is useful when the variable in question is continuous and often produces a reasonable approximation for a dummy variable. Another way to analyze the effect of a dummy variable on the whole distribution is to compute  $\text{Prob}(Y = 1)$  over the range of  $\mathbf{x}'\beta$  (using the sample estimates) and with the two values of the binary variable. Using the coefficients from the probit model in Table 17.1, we have the following probabilities as a function of GPA, at the mean of TUCE:

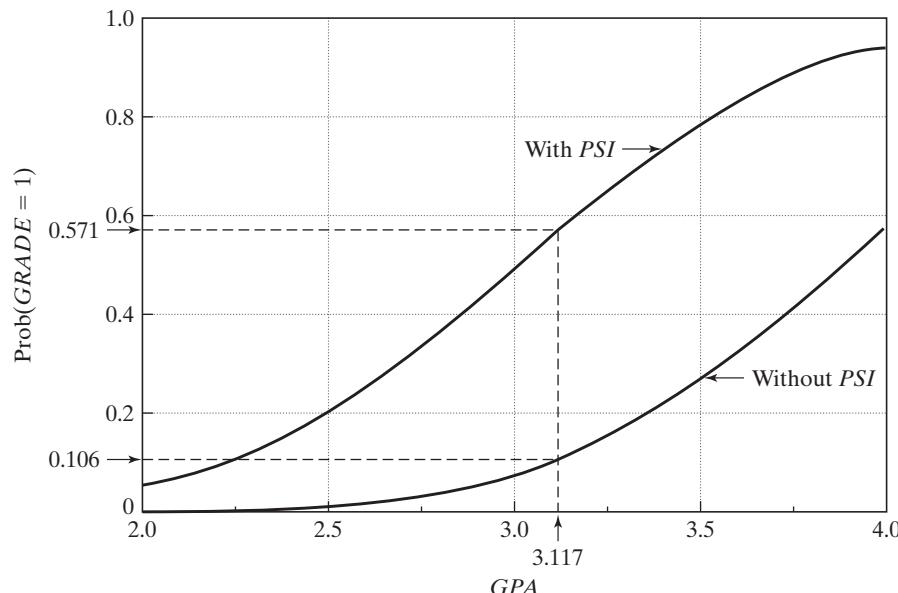
$$PSI = 0: \text{Prob}(GRADE = 1) = \Phi[-7.452 + 1.626GPA + 0.052(21.938)],$$

$$PSI = 1: \text{Prob}(GRADE = 1) = \Phi[-7.452 + 1.626GPA + 0.052(21.938) + 1.426].$$

Figure 17.2 shows these two functions plotted over the range of GPA observed in the sample, 2.0 to 4.0. The marginal effect of PSI is the difference between the two functions, which ranges from only about 0.06 at GPA = 2 to about 0.50 at GPA of 3.5. This effect shows that the probability that a student's grade will increase after exposure to PSI is far greater for students with high GPAs than for those with low GPAs. At the sample mean of GPA 3.117, the effect of PSI on the probability is 0.465. The simple derivative calculation of (17-9) is given in Table 17.1; the estimate is 0.468. But, of course, this calculation does not show the wide range of differences displayed in Figure 17.2.

Table 17.2 presents the estimated coefficients and marginal effects for the probit and logit models in Table 17.2. In both cases, the asymptotic covariance matrix is computed from the negative inverse of the actual Hessian of the log-likelihood. The standard errors for the estimated marginal effect of PSI are computed using (17-24) and (17-25) since PSI is a binary variable. In comparison, the simple derivatives produce estimates and standard errors of (0.449, 0.181) for the logit model and (0.464, 0.188) for the probit model. These differ only slightly from the results given in the table.

<sup>12</sup>One might be tempted in this case to suggest an asymmetric distribution for the model, such as the Gumbel distribution. However, the asymmetry in the model, to the extent that it is present at all, refers to the values of  $\varepsilon$ , not to the observed sample of values of the dependent variable.

**696 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**

**FIGURE 17.2** Effect of *PSI* on Predicted Probabilities.

**17.3.2.a Average Partial Effects**

The preceding has emphasized computing the partial effects for the average individual in the sample. Current practice has many applications based, instead, on “average partial effects.” [See, e.g., Wooldridge (2002a).] The underlying logic is that the quantity of interest is

$$APE = E_x \left[ \frac{\partial E[y | \mathbf{x}]}{\partial \mathbf{x}} \right].$$

In practical terms, this suggests the computation

$$\widehat{APE} = \bar{y} = \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}'_i \hat{\beta}) \hat{\beta}.$$

**TABLE 17.2** Estimated Coefficients and Standard Errors (standard errors in parentheses)

<i>Variable</i>	<i>Logistic</i>				<i>Probit</i>			
	<i>Coefficient</i>	<i>t Ratio</i>	<i>Slope</i>	<i>t Ratio</i>	<i>Coefficient</i>	<i>t Ratio</i>	<i>Slope</i>	<i>t Ratio</i>
Constant	-13.021 (4.931)	-2.641	—	—	-7.452 (2.542)	-2.931	—	—
GPA	2.826 (1.263)	2.238	0.534 (0.237)	2.252	1.626 (0.694)	2.343	0.533 (0.232)	2.294
TUCE	0.095 (0.142)	0.672	0.018 (0.026)	0.685	0.052 (0.084)	0.617 (0.027)	0.017 (0.027)	0.626
PSI	2.379 (1.065)	2.234	0.456 (0.181)	2.521	1.426 (0.595)	2.397	0.464 (0.170)	2.727
log-likelihood			-12.890				-12.819	

## CHAPTER 17 ♦ Discrete Choice 697

This does raise two questions. Because the computation is (marginally) more burdensome than the simple marginal effects at the means, one might wonder whether this produces a noticeably different answer. That will depend on the data. Save for small sample variation, the difference in these two results is likely to be small. Let

$$\bar{\gamma}_k = APE_k = \frac{1}{n} \sum_{i=1}^n \frac{\partial \Pr(y_i = 1 | \mathbf{x}_i)}{\partial x_{ik}} = \frac{1}{n} \sum_{i=1}^n F'(\mathbf{x}'_i \boldsymbol{\beta}) \beta_k = \frac{1}{n} \sum_{i=1}^n \gamma_k(\mathbf{x}_i)$$

denote the computation of the average partial effect. We compute this at the MLE,  $\hat{\boldsymbol{\beta}}$ . Now, expand this function in a second-order Taylor series around the point of sample means,  $\bar{\mathbf{x}}$ , to obtain

$$\begin{aligned} \bar{\gamma}_k &= \frac{1}{n} \sum_{i=1}^n \left[ \gamma_k(\bar{\mathbf{x}}) + \sum_{m=1}^K \frac{\partial \gamma_k(\bar{\mathbf{x}})}{\partial \bar{x}_m} (x_{im} - \bar{x}_m) \right. \\ &\quad \left. + \frac{1}{2} \sum_{l=1}^K \sum_{m=1}^K \frac{\partial^2 \gamma_k(\bar{\mathbf{x}})}{\partial \bar{x}_l \partial \bar{x}_m} (x_{il} - \bar{x}_l)(x_{im} - \bar{x}_m) \right] + \Delta, \end{aligned}$$

where  $\Delta$  is the remaining higher-order terms. The first of the three terms is the marginal effect computed at the sample means. The second term is zero by construction. That leaves the remainder plus an average of a term that is a function of the variances and covariances of the data and the curvature of the probability function at the means. Little can be said to characterize these two terms in any particular sample, but one might guess they are likely to be small. We will examine an application in Example 17.4.

Based on the sample of observations on the partial effects, a natural estimator of the variance of the partial effects would seem to be

$$\hat{\sigma}_{\gamma,k}^2 = \frac{1}{n-1} \sum_{i=1}^n (\hat{\gamma}_k(\mathbf{x}_i) - \bar{\gamma}_k)^2 = \frac{1}{n-1} \sum_{i=1}^n (\widehat{PE}_{i,k} - \widehat{APE}_k)^2.$$

See, for example, Contoyannis et al. (2004, p. 498), who report that they computed the “sample standard deviation of the partial effects.” Since  $\widehat{APE}_k = \bar{\gamma}_k$  is the mean of a sample, notwithstanding the following consideration, the preceding estimator should be further divided by the sample size since we are computing the *standard error of the mean of a sample*. This seems not to be the norm in the literature. This estimator should not be viewed as an alternative to the delta method applied to the partial effects evaluated at the means of the data,  $\hat{\boldsymbol{\gamma}}(\bar{\mathbf{x}})$ . The delta method produces an estimator of the asymptotic variance of an estimator of the population parameter,  $\boldsymbol{\gamma}(\boldsymbol{\mu}_{\mathbf{x}})$ , that is, of a function of  $\hat{\boldsymbol{\beta}}$ . The asymptotic covariance matrix computed using the delta method for  $\hat{\boldsymbol{\gamma}}(\bar{\mathbf{x}})$  would be  $\hat{\mathbf{G}}(\bar{\mathbf{x}})\hat{\mathbf{V}}\hat{\mathbf{G}}'(\bar{\mathbf{x}})$  where  $\hat{\mathbf{G}}(\bar{\mathbf{x}})$  is the matrix of partial derivatives and  $\hat{\mathbf{V}}$  is the estimator of the asymptotic variance of  $\hat{\boldsymbol{\beta}}$ . This variance estimator converges to zero because  $\hat{\boldsymbol{\beta}}$  converges to  $\boldsymbol{\beta}$  and  $\bar{\mathbf{x}}$  converges to a vector of constants. The ~~naive~~ estimator above does not converge to zero; it converges to the variance of the random variable  $PE_{i,k}$ .

The “asymptotic variance” of the partial effects estimator is intended to reflect the variation of the parameter estimator,  $\hat{\boldsymbol{\beta}}$ , whereas the ~~naive~~ estimator generates the variation from the heterogeneity of the sample data while holding the parameter fixed

**698 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

at  $\hat{\beta}$ . For example, for a logit model,

$$\hat{y}_k(\mathbf{x}_i) = \hat{\beta}_k \Lambda(\mathbf{x}'_i \hat{\beta}) [1 - \Lambda(\mathbf{x}'_i \hat{\beta})] = \hat{\beta}_k \hat{\delta}_i,$$

and  $\hat{\delta}_i$  is the same for all  $k$ . It follows that

$$\hat{\sigma}_{\gamma,k}^2 = \hat{\beta}_k^2 \left[ \frac{1}{n-1} \sum_{i=1}^n (\hat{\delta}_i - \bar{\delta})^2 \right] = \hat{\beta}_k^2 s_{\bar{\delta}}^2.$$

A surprising consequence is that if one computes  $t$  ratios for the average partial effects using  $\hat{\sigma}_{\gamma,k}^2$ , the values will all equal the same  $1/s_{\bar{\delta}}$ . This might signal that something is amiss. (This is somewhat apparent in the Contoyannis et al. results on page 498; however, not enough digits were reported to see the effect clearly.)

A search for applications that use the delta method to estimate standard errors for average partial effects in nonlinear models yields hundreds of occurrences. However, we could not locate any that document in detail the precise formulas used. (One author, noting the complexity of computation, recommended bootstrapping instead.) A complicated flaw in the sample variance estimator (notwithstanding all the preceding) is that the ~~naive~~ estimator (whether scaled by  $1/n$  or not) neglects the fact that all  $n$  observations used to compute the estimated  $APE$  are correlated; they all use the same estimator of  $\beta$ . The preceding estimator treats the estimates of  $PE_i$  as if they were a random sample. They would be if they were based on the true  $\beta$ . But the estimators based on the same  $\hat{\beta}$  are not uncorrelated. The delta method will account for the asymptotic (co)variation of the terms in the sum of functions of  $\hat{\beta}$ . To use the delta method to estimate the asymptotic standard errors for the average partial effects,  $\widehat{APE}_k$ , we should use

$$\begin{aligned} \text{Est. Asy. Var}[\bar{\gamma}] &= \frac{1}{n^2} \text{Est. Asy. Var} \left[ \sum_{i=1}^n \hat{y}_i \right] \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \text{Est. Asy. Cov} [\hat{y}_i, \hat{y}_j] \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{G}_i(\hat{\beta}) \hat{\mathbf{V}} \mathbf{G}'_j(\hat{\beta}) \\ &= \left[ \frac{1}{n} \sum_{i=1}^n \mathbf{G}_i(\hat{\beta}) \right] \hat{\mathbf{V}} \left[ \frac{1}{n} \sum_{j=1}^n \mathbf{G}'_j(\hat{\beta}) \right], \end{aligned}$$

where

$$\mathbf{G}_i(\hat{\beta}) = \frac{\partial f(\mathbf{x}'_i \hat{\beta}) \hat{\beta}}{\partial \hat{\beta}'} = f(\mathbf{x}'_i \hat{\beta}) \mathbf{I} + f'(\mathbf{x}'_i \hat{\beta}) \hat{\beta} \mathbf{x}'_i.$$

This treats the  $APE$  as a point estimator of a population parameter—one that converges in probability to what we assume is its population counterpart. But, it is conditioned on the sample data; convergence is with respect to  $\hat{\beta}$ . This looks like a formidable amount of computation—Example 17.4 uses a sample of 27,326 observations, so it appears we need a double sum of roughly 750 million terms. However, the computation is actually

**TABLE 17.3** Estimated Parameters and Partial Effects

<b>Variable</b>	<b>Parameter Estimates</b>		<b>Marginal Effects</b>		<b>Average Partial Effects</b>		
	<b>Estimate</b>	<b>Std.Error</b>	<b>Estimate</b>	<b>Std.Error</b>	<b>Estimate</b>	<b>Std.Error</b>	<b>Naive S.E.</b>
<b>Constant</b>	0.25112	0.09114					
<b>Age</b>	0.02071	0.00129	0.00497	0.00031	0.00471	0.00029	0.00043
<b>Income</b>	-0.18592	0.07506	-0.04466	0.01803	-0.04229	0.01707	0.00386
<b>Kids</b>	-0.22947	0.02954	-0.05512	0.00710	-0.05220	0.00669	0.00476
<b>Education</b>	-0.04559	0.00565	-0.01095	0.00136	-0.01037	0.00128	0.00095
<b>Married</b>	0.08529	0.03329	0.02049	0.00800	0.01940	0.00757	0.00177

linear in  $n$ , not quadratic, because the same matrix is used in the center of each product. The estimator of the asymptotic covariance matrix for the APE is simply

$$\text{Est. Asy. Var } [\hat{\gamma}] = \mathbf{G}(\hat{\beta}) \hat{\mathbf{V}} \mathbf{G}'(\hat{\beta}).$$

The appropriate covariance matrix is computed by making the same adjustment as in the partial effects—the derivative matrices are averaged over the observations rather than being computed at the means of the data.

#### Example 17.4 Average Partial Effects

We estimated a binary logit model for  $y = 1(DocVis > 0)$  using the German health care utilization data examined in Example 7.6 (and several later examples). The model is

$$\text{Prob}(DocVis_{it} > 0) = \Lambda(\beta_1 + \beta_2 Age_{it} + \beta_3 Income_{it} + \beta_4 Kids_{it} + \beta_5 Education_{it} + \beta_6 Married_{it}).$$

No account of the panel nature of the data set was taken for this exercise. The sample contains 27,326 observations, which should be large enough to reveal the large sample behavior of the computations. Table 17.3 presents the parameter estimates for the logit probability model and both the marginal effects and the average partial effects, each with standard errors computed using the results given earlier. (The partial effects for the two dummy variables, *Kids* and *Married*, are computed using the approximation, rather than using the ~~difference~~ differences.) The results do suggest the similarity of the computations. The values in parentheses in the last column are based on the naive estimator that ignores the covariances and is not divided by the  $1/n$  for the variance of the mean.

#### 17.3.2.b Interaction Effects

Models with **interaction effects**, such as

$$\begin{aligned} \text{Prob}(DocVis_{it} > 0) = & \Lambda(\beta_1 + \beta_2 Age_{it} + \beta_3 Income_{it} + \beta_4 Kids_{it} \\ & + \beta_5 Education_{it} + \beta_6 Married_{it} + \beta_7 Age_{it} \times Education_{it}), \end{aligned}$$

have attracted considerable attention in recent applications of binary choice models.<sup>13</sup> A practical issue concerns the computation of partial effects by standard computer packages. Write the model as

$$\text{Prob}(DocVis_{it} > 0) = \Lambda(\beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + \beta_7 x_{7it}).$$

Estimation of the model parameters is routine. Rote computation of partial effects using (17-11) will produce

$$PE_7 = \partial \text{Prob}(DocVis > 0) / \partial x_7 = \beta_7 \Lambda(\mathbf{x}' \boldsymbol{\beta}) [1 - \Lambda(\mathbf{x}' \boldsymbol{\beta})],$$

<sup>13</sup>See, for example, Ai and Norton (2004) and Greene (2010).

## 700 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

which is what common computer packages will dutifully report. The problem is that  $x_7 = x_2x_5$ , and  $PE_7$  in the previous equation is not the partial effect for  $x_7$ . Moreover, the partial effects for  $x_2$  and  $x_5$  will also be misreported by the rote computation. To revert back to our original specification,

$$\partial \text{Prob}(DocVis > 0 | \mathbf{x}) / \partial Age = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})](\beta_2 + \beta_7 Education),$$

$$\partial \text{Prob}(DocVis > 0 | \mathbf{x}) / \partial Education = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})](\beta_5 + \beta_7 Age),$$

and what is computed as “ $\partial \text{Prob}(DocVis > 0 | \mathbf{x}) / \partial Age \times Education$ ” is meaningless. The practical problem motivating Ai and Norton (2004) was that the computer package does not know that  $x_7$  is  $x_2x_5$ , so it computes a partial effect for  $x_7$  as if it could vary “partially” from the other variables. The (now) obvious solution is for the analyst to force the correct computations of the relevant partial effects by whatever software they are using, perhaps by programming the computations themselves.

The practical complication raises a theoretical question that is less clear cut. What is the “interaction effect” in the model? In a linear model based on the preceding, we would have

$$\partial^2 E[y | \mathbf{x}] / \partial x_2 \partial x_5 = \beta_7$$

which is unambiguous. However, in this *nonlinear* binary choice model, the correct result is

$$\begin{aligned} \partial^2 E[y | \mathbf{x}] / \partial x_2 \partial x_5 &= \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})]\beta_7 \\ &\quad + \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})][1 - 2\Lambda(\mathbf{x}'\boldsymbol{\beta})](\beta_2 + \beta_7 Education)(\beta_5 + \beta_7 Age). \end{aligned}$$

Not only is  $\beta_7$  not the interesting effect, but there is also a complicated additional term. Loosely, we can associate the first term as a “direct” effect—note that it is the naive term  $PE_7$  from earlier. The second part can be attributed to the fact that we are differentiating a nonlinear model—essentially, the second part of the partial effect results from the nonlinearity of the function. The existence of an “interaction effect” in this model is inescapable—notice that the second part is nonzero (generally) even if  $\beta_7$  does equal zero. Whether this is intended to represent an “interaction” in some economic sense is unclear. In the absence of the product term in the model, probably not. We can see an implication of this in Figure 17.1. At the point where  $\mathbf{x}'\boldsymbol{\beta} = 0$ , where the probability equals one half, the probability function is linear. At that point,  $(1 - 2\Lambda)$  will equal zero and the functional form effect will be zero as well. When  $\mathbf{x}'\boldsymbol{\beta}$  departs from zero, the probability becomes nonlinear. (These same effects can be shown for the probit model—at  $\mathbf{x}'\boldsymbol{\beta} = 0$ , the second derivative of the probit probability is  $-\mathbf{x}'\boldsymbol{\beta}\phi'(\mathbf{x}'\boldsymbol{\beta}) = 0$ .)

We developed an extensive application of interaction effects in a nonlinear model in Example 7.6. In that application, using the same data for the numerical exercise, we analyzed a nonlinear regression  $E[y | \mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$ . The results obtained in that study were general, and will apply to the application here, where the nonlinear regression is  $E[y | \mathbf{x}] = \Lambda(\mathbf{x}'\boldsymbol{\beta})$  or  $\Phi(\mathbf{x}'\boldsymbol{\beta})$ .

### Example 17.5 Interaction Effect

We added the interaction term,  $Age \times Education$ , to the model in Example 17.4. The model is now

$$\begin{aligned} \text{Prob}(DocVis_{it} > 0) &= \Lambda(\beta_1 + \beta_2 Age_{it} + \beta_3 Income_{it} + \beta_4 Kids_{it} \\ &\quad + \beta_5 Education_{it} + \beta_6 Married_{it} + \beta_7 Age_{it} \times Education_{it}). \end{aligned}$$

Estimation of the model produces an estimate of  $\beta_7$  of  $-0.00112$ . The naive average partial effect for  $x_7$  is  $-0.000254$ . This is the first part in the earlier decomposition. The second, functional form term (averaged over the sample observations) is  $0.0000634$ , so the estimated interaction effect, the sum of the two terms is  $-0.000191$ . The naive calculation errs by about  $(-0.000254 / -0.000191 - 1) \times 100$  percent = 33 percent.

### 17.3.3 MEASURING GOODNESS OF FIT

There have been many fit measures suggested for QR models.<sup>14</sup> At a minimum, one should report the maximized value of the log-likelihood function,  $\ln L$ . Because the hypothesis that all the slopes in the model are zero is often interesting, the log-likelihood computed with only a constant term,  $\ln L_0$  [see (17-29)], should also be reported. An analog to the  $R^2$  in a conventional regression is McFadden's (1974) likelihood ratio index,

$$\text{LRI} = 1 - \frac{\ln L}{\ln L_0}.$$

This measure has an intuitive appeal in that it is bounded by zero and one. (See Section 14.6.5.) If all the slope coefficients are zero, then it equals zero. There is no way to make LRI equal 1, although one can come close. If  $F_i$  is always one when  $y$  equals one and zero when  $y$  equals zero, then  $\ln L$  equals zero (the log of one) and LRI equals one. It has been suggested that this finding is indicative of a "perfect fit" and that LRI increases as the fit of the model improves. To a degree, this point is true. Unfortunately, the values between zero and one have no natural interpretation. If  $F(\mathbf{x}'_i \boldsymbol{\beta})$  is a proper  $\pi_i$ , then even with many regressors the model cannot fit perfectly unless  $\mathbf{x}'_i \boldsymbol{\beta}$  goes to  $+\infty$  or  $-\infty$ . As a practical matter, it does happen. But when it does, it indicates a flaw in the model, not a good fit. If the range of one of the independent variables contains a value, say,  $x^*$ , such that the sign of  $(x - x^*)$  predicts  $y$  perfectly and vice versa, then the model will become a perfect predictor. This result also holds in general if the sign of  $\mathbf{x}' \boldsymbol{\beta}$  gives a perfect predictor for some vector  $\boldsymbol{\beta}$ .<sup>15</sup> For example, one might mistakenly include as a regressor a dummy variable that is identical, or nearly so, to the dependent variable. In this case, the maximization procedure will break down precisely because  $\mathbf{x}' \boldsymbol{\beta}$  is diverging during the iterations. [See McKenzie (1998) for an application and discussion.] Of course, this situation is not at all what we had in mind for a good fit.

Other fit measures have been suggested. Ben-Akiva and Lerman (1985) and Kay and Little (1986) suggested a fit measure that is keyed to the prediction rule,

$$R_{BL}^2 = \frac{1}{n} \sum_{i=1}^n [y_i \hat{F}_i + (1 - y_i)(1 - \hat{F}_i)],$$

which is the average probability of correct prediction by the prediction rule. The difficulty in this computation is that in unbalanced samples, the less frequent outcome will usually be predicted very badly by the standard procedure, and this measure does not pick up that point. Cramer (1999) has suggested an alternative measure that directly

<sup>14</sup>See, for example, Cragg and Uhler (1970), Amemiya (1981), Maddala (1983), McFadden (1974), Ben-Akiva and Lerman (1985), Kay and Little (1986), Veall and Zimmermann (1992), Zavoina and McKelvey (1975), Efron (1978), and Cramer (1999). A survey of techniques appears in Windmeijer (1995).

<sup>15</sup>See McFadden (1984) and Amemiya (1985). If this condition holds, then gradient methods *will* find that  $\boldsymbol{\beta}$ .

**702 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**


measures this failure,

$$\begin{aligned}\lambda &= (\text{average } \hat{F} | y_i = 1) - (\text{average } \hat{F} | y_i = 0) \\ &= (\text{average}(1 - \hat{F}) | y_i = 0) - (\text{average}(1 - \hat{F}) | y_i = 1).\end{aligned}$$

Cramer's measure heavily penalizes the incorrect predictions, and because each proportion is taken within the subsample, it is not unduly influenced by the large proportionate size of the group of more frequent outcomes.

A useful summary of the predictive ability of the model is a  $2 \times 2$  table of the hits and misses of a prediction rule such as

$$\hat{y} = 1 \quad \text{if } \hat{F} > F^* \text{ and } 0 \text{ otherwise.} \quad (17-26)$$

The usual threshold value is 0.5, on the basis that we should predict a one if the model says a one is more likely than a zero. It is important not to place too much emphasis on this measure of goodness of fit, however. Consider, for example, the naive predictor

$$\hat{y} = 1 \quad \text{if } P > 0.5 \text{ and } 0 \text{ otherwise,} \quad (17-27)$$

where  $P$  is the simple proportion of ones in the sample. This rule will always predict correctly  $100P$  percent of the observations, which means that the naive model does not have zero fit. In fact, if the proportion of ones in the sample is very high, it is possible to construct examples in which the second model will generate more correct predictions than the first! Once again, this flaw is not in the model; it is a flaw in the fit measure.<sup>16</sup> The important element to bear in mind is that the coefficients of the estimated model are not chosen so as to maximize this (or any other) fit measure, as they are in the linear regression model where  $\mathbf{b}$  maximizes  $R^2$ .

Another consideration is that 0.5, although the usual choice, may not be a very good value to use for the threshold. If the sample is **unbalanced**—that is, has many more ones than zeros, or vice versa—then by this prediction rule it might never predict a one (or zero). To consider an example, suppose that in a sample of 10,000 observations, only 1,000 have  $Y = 1$ . We know that the average predicted probability in the sample will be 0.10. As such, it may require an extreme configuration of regressors even to produce an  $F$  of 0.2, to say nothing of 0.5. In such a setting, the prediction rule may fail every time to predict when  $Y = 1$ . The obvious adjustment is to reduce  $F^*$ . Of course, this adjustment comes at a cost. If we reduce the threshold  $F^*$  so as to predict  $y = 1$  more often, then we will increase the number of correct classifications of observations that do have  $y = 1$ , but we will also increase the number of times that we *incorrectly* classify as ones observations that have  $y = 0$ .<sup>17</sup> In general, any prediction rule of the form in (17-26) will make two types of errors: It will incorrectly classify zeros as ones and ones as zeros. In practice, these errors need not be symmetric in the costs that result. For example, in a credit scoring model [see Boyes, Hoffman, and Low (1989)], incorrectly classifying an applicant as a bad risk is not the same as incorrectly classifying a bad risk as a good one. Changing  $F^*$  will always reduce the probability of one type of error

<sup>16</sup>See Amemiya (1981).

<sup>17</sup>The technique of **discriminant analysis** is used to build a procedure around this consideration. In this setting, we consider not only the number of correct and incorrect classifications, but also the cost of each type of misclassification.

while increasing the probability of the other. There is no correct answer as to the best value to choose. It depends on the setting and on the criterion function upon which the prediction rule depends.

The likelihood ratio index and various modifications of it are obviously related to the likelihood ratio statistic for testing the hypothesis that the coefficient vector is zero. Cramer's measure is oriented more toward the relationship between the fitted probabilities and the actual values. This is usefully tied to the standard prediction rule  $\hat{y} = \mathbf{1}[\hat{F} > 0.5]$ . Whether these have a close relationship to any type of fit in the familiar sense is a question that needs to be studied. In some cases, it appears so. But the maximum likelihood estimator, on which all the fit measures are based, is not chosen so as to maximize a fitting criterion based on prediction of  $y$  as it is in the classical regression (which maximizes  $R^2$ ). It is chosen to maximize the joint density of the observed dependent variables. It remains an interesting question for research whether fitting  $y$  well or obtaining good parameter estimates is a preferable estimation criterion. Evidently, they need not be the same thing.

#### **Example 17.6 Prediction with a Probit Model**

Tunali (1986) estimated a probit model in a study of migration, subsequent remigration, and earnings for a large sample of observations of male members of households in Turkey. Among his results, he reports the summary shown here for a probit model: The estimated model is highly significant, with a likelihood ratio test of the hypothesis that the coefficients (16 of them) are zero based on a chi-squared value of 69 with 16 degrees of freedom.<sup>18</sup> The model predicts 491 of 690, or 71.2 percent, of the observations correctly, although the likelihood ratio index is only 0.083. A naive model, which always predicts that  $y = 0$  because  $P < 0.5$ , predicts 487 of 690, or 70.6 percent, of the observations correctly. This result is hardly suggestive of no fit. The maximum likelihood estimator produces several significant influences on the probability but makes only four more correct predictions than the naive predictor.<sup>19</sup>

		<i>Predicted</i>		
		<i>D</i> = 0	<i>D</i> = 1	Total
<i>Actual</i>	<i>D</i> = 0	471	16	487
	<i>D</i> = 1	183	20	203
	Total	654	36	690

#### **17.3.4 HYPOTHESIS TESTS**

For testing hypotheses about the coefficients, the full menu of procedures is available. The simplest method for a single restriction would be based on the usual *t* tests, using the standard errors from the information matrix. Using the normal distribution of the estimator, we would use the standard normal table rather than the *t* table for critical points. For more involved restrictions, it is possible to use the Wald test. For a set of restrictions  $\mathbf{R}\beta = \mathbf{q}$ , the statistic is

$$W = (\mathbf{R}\hat{\beta} - \mathbf{q})' \{ \mathbf{R}(\text{Est. Asy. Var}[\hat{\beta}]) \mathbf{R}' \}^{-1} (\mathbf{R}\hat{\beta} - \mathbf{q}).$$

<sup>18</sup>This view actually understates slightly the significance of his model, because the preceding predictions are based on a bivariate model. The likelihood ratio test fails to reject the hypothesis that a univariate model applies, however.

<sup>19</sup>It is also noteworthy that nearly all the correct predictions of the maximum likelihood estimator are the zeros. It hits only 10 percent of the ones in the sample.

## 704 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

For example, for testing the hypothesis that a subset of the coefficients, say, the last  $M$ , are zero, the Wald statistic uses  $\mathbf{R} = [\mathbf{0} \mid \mathbf{I}_M]$  and  $\mathbf{q} = \mathbf{0}$ . Collecting terms, we find that the test statistic for this hypothesis is

$$W = \hat{\beta}'_M \mathbf{V}_M^{-1} \hat{\beta}_M, \quad (17-28)$$

where the subscript  $M$  indicates the subvector or submatrix corresponding to the  $M$  variables and  $\mathbf{V}$  is the estimated asymptotic covariance matrix of  $\hat{\beta}$ .

Likelihood ratio and Lagrange multiplier statistics can also be computed. The likelihood ratio statistic is

$$\text{LR} = -2[\ln \hat{L}_R - \ln \hat{L}_U],$$

where  $\hat{L}_R$  and  $\hat{L}_U$  are the log-likelihood functions evaluated at the restricted and unrestricted estimates, respectively. A common test, which is similar to the  $F$  test that all the slopes in a regression are zero, is the **likelihood ratio test** that all the slope coefficients in the probit or logit model are zero. For this test, the constant term remains unrestricted. In this case, the restricted log-likelihood is the same for both probit and logit models,

$$\ln L_0 = n[P \ln P + (1 - P) \ln(1 - P)], \quad (17-29)$$

where  $P$  is the proportion of the observations that have dependent variable equal to 1.

It might be tempting to use the likelihood ratio test to choose between the probit and logit models. But there is no restriction involved, and the test is not valid for this purpose. To underscore the point, there is nothing in its construction to prevent the chi-squared statistic for this “test” from being negative.

The **Lagrange multiplier test** statistic is  $\text{LM} = \mathbf{g}'\mathbf{V}\mathbf{g}$ , where  $\mathbf{g}$  is the first derivatives of the *unrestricted* model evaluated at the *restricted* parameter vector and  $\mathbf{V}$  is any of the three estimators of the asymptotic covariance matrix of the maximum likelihood estimator, once again computed using the restricted estimates. Davidson and MacKinnon (1984) find evidence that  $E[\mathbf{H}]$  is the best of the three estimators to use, which gives

$$\text{LM} = \left( \sum_{i=1}^n g_i \mathbf{x}_i \right)' \left[ \sum_{i=1}^n E[-h_i] \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \left( \sum_{i=1}^n g_i \mathbf{x}_i \right), \quad (17-30)$$

where  $E[-h_i]$  is defined in (17-21) for the logit model and in (17-23) for the probit model.

For the logit model, when the hypothesis is that all the slopes are zero,

$$\text{LM} = n R^2,$$

where  $R^2$  is the uncentered coefficient of determination in the regression of  $(y_i - \bar{y})$  on  $\mathbf{x}_i$  and  $\bar{y}$  is the proportion of 1s in the sample. An alternative formulation based on the BHHH estimator, which we developed in Section 14.6.3 is also convenient. For any of the models (probit, logit, Gumbel, etc.), the first derivative vector can be written as

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n g_i \mathbf{x}_i = \mathbf{X}' \mathbf{G} \mathbf{i},$$

## CHAPTER 17 ♦ Discrete Choice 705

where  $\mathbf{G}(n \times n) = \text{diag}[g_1, g_2, \dots, g_n]$  and  $\mathbf{i}$  is an  $n \times 1$  column of 1s. The BHHH estimator of the Hessian is  $(\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{X})$ , so the LM statistic based on this estimator is

$$\text{LM} = n \left[ \frac{1}{n} \mathbf{i}'(\mathbf{G}\mathbf{X})(\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{G}')\mathbf{i} \right] = n R_i^2, \quad (17-31)$$

where  $R_i^2$  is the uncentered coefficient of determination in a regression of a column of ones on the first derivatives of the logs of the individual probabilities.

All the statistics listed here are asymptotically equivalent and under the null hypothesis of the restricted model have limiting chi-squared distributions with degrees of freedom equal to the number of restrictions being tested. We consider some examples in the next section.

**Example 17.7 Testing for Structural Break in a Logit Model**

The model in Example 17.4, based on Riphahn, Wambach, and Million (2003), is

$$\begin{aligned} \text{Prob}(DocVis_{it} > 0) = & \Lambda(\beta_1 + \beta_2 \text{Age}_{it} + \beta_3 \text{Income}_{it} + \beta_4 \text{Kids}_{it} \\ & + \beta_5 \text{Education}_{it} + \beta_6 \text{Married}_{it}). \end{aligned}$$

In the original study, the authors split the sample on the basis of gender, and fit separate models for male and female headed households. We will use the preceding results to test for the appropriateness of the sample splitting. This test of the pooling hypothesis is a counterpart to the **Chow test** of structural change in the linear model developed in Section 6.4.1. Since we are not using least squares (in a linear model), we use the likelihood based procedures rather than an F test as we did earlier. Estimates of the three models are shown in Table 17.4. The chi-squared statistic for the likelihood ratio test is

$$\text{LR} = -2[-17673.09788 - (-9541.77802 - 7855.96999)] = 550.69744.$$

The 95 percent critical value for six degrees of freedom is 12.592. To carry out the Wald test for this hypothesis there are two numerically identical ways to proceed. First, using the estimates for Male and Female samples separately, we can compute a chi-squared statistic to test the hypothesis that the difference of the two coefficients is zero. This would be

$$\begin{aligned} W &= [\hat{\beta}_{Male} - \hat{\beta}_{Female}]'[\text{Est. Asy. Var}(\hat{\beta}_{Male}) + \text{Est. Asy. Var}(\hat{\beta}_{Female})]^{-1}[\hat{\beta}_{Male} - \hat{\beta}_{Female}] \\ &= 538.13629. \end{aligned}$$

Another way to obtain the same result is to add to the pooled model the original 6 variables now multiplied by the *Female* dummy variable. We use the augmented  $\mathbf{X}$  matrix

**TABLE 17.4** Estimated Models for Pooling Hypothesis

<i>Variable</i>	<i>Pooled Sample</i>		<i>Male</i>		<i>Female</i>	
	<i>Estimate</i>	<i>Std.Error</i>	<i>Estimate</i>	<i>Std.Error</i>	<i>Estimate</i>	<i>Std.Error</i>
Constant	0.25112	0.09114	-0.20881	0.11475	0.44767	0.16016
Age	0.02071	0.00129	0.02375	0.00178	0.01331	0.00202
Income	-0.18592	0.07506	-0.23059	0.10415	-0.17182	0.11225
Kids	-0.22947	0.02954	-0.26149	0.04054	-0.27153	0.04539
Education	-0.04559	0.00565	-0.04251	0.00737	-0.00170	0.00970
Married	0.08529	0.03329	0.17451	0.04833	0.03621	0.04864
ln L	-17673.09788		-9541.77802		-7855.96999	

## 706 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

$\mathbf{X}^* = [\mathbf{X}, \text{female} \times \mathbf{X}]$ . The model with 12 variables is now estimated, and a test of the pooling hypothesis is done by testing the joint hypothesis that the coefficients on these 6 additional variables are zero. The Lagrange multiplier test is carried out by using this augmented model as well. To apply (17-31), the necessary derivatives are in (17-18). For the logit model, the derivative matrix is simply  $\mathbf{G}^* = \text{diag}[y_i - \Lambda(\mathbf{x}_i^* \boldsymbol{\beta})]$ . For the LM test, the vector  $\boldsymbol{\beta}$  that is used is the one for the restricted model. Thus,  $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\beta}}'_{\text{Pooled}}, 0, 0, 0, 0, 0)'.$  The estimated probabilities that appear in  $\mathbf{G}^*$  are simply those obtained from the pooled model. Then,

$$\text{LM} = \mathbf{i}' \mathbf{G}^* \mathbf{X}^* \times [(\mathbf{X}^{*'} \mathbf{G}^{*'}) (\mathbf{G}^* \mathbf{X}^*)]^{-1} \mathbf{X}^{*'} \mathbf{G}^* \mathbf{i} = 548.17052.$$

The pooling hypothesis is rejected by all three procedures.

### 17.3.5 ENDOGENOUS RIGHT-HAND-SIDE VARIABLES IN BINARY CHOICE MODELS

The analysis in Example 17.8 (Labor Supply Model) suggests that the presence of endogenous right-hand-side variables in a binary choice model presents familiar problems for estimation. The problem is made worse in nonlinear models because even if one has an instrumental variable readily at hand, it may not be immediately clear what is to be done with it. The instrumental variable estimator described in Chapter 8 is based on moments of the data, variances, and covariances. In this binary choice setting, we are not using any form of least squares to estimate the parameters, so the IV method would appear not to apply. Generalized method of moments is a possibility.

$$\begin{aligned} y_i^* &= \mathbf{x}_i' \boldsymbol{\beta} + \gamma w_i + \varepsilon_i, \\ y_i &= 1(y_i^* > 0), \\ E[\varepsilon_i | w_i] &= g(w_i) \neq 0. \end{aligned}$$

Thus,  $w_i$  is endogenous in this model. The maximum likelihood estimators considered earlier will not consistently estimate  $(\boldsymbol{\beta}, \gamma)$ . [Without an additional specification that allows us to formalize  $\text{Prob}(y_i = 1 | \mathbf{x}_i, w_i)$ , we cannot state what the MLE will, in fact, estimate.] Suppose that we have a “relevant” (see Section 8.2) instrumental variable,  $z_i$  such that

$$\begin{aligned} E[\varepsilon_i | z_i, \mathbf{x}_i] &= 0, \\ E[w_i z_i] &\neq 0. \end{aligned}$$

A natural instrumental variable estimator would be based on the “moment” condition

$$E \left[ (y_i^* - \mathbf{x}_i' \boldsymbol{\beta} - \gamma w_i) \begin{pmatrix} \mathbf{x}_i \\ z_i \end{pmatrix} \right] = \mathbf{0}.$$

However,  $y_i^*$  is not observed,  $y_i$  is. But the “residual,”  $y_i - \mathbf{x}_i' \boldsymbol{\beta} - \gamma w_i$ , would have no meaning even if the true parameters were known.<sup>20</sup> One approach that was used in Avery et al. (1983), Butler and Chatterjee (1997), and Bertschek and Lechner (1998) is to assume that the instrumental variable is orthogonal to the residual  $[y - \Phi(\mathbf{x}_i' \boldsymbol{\beta} + \gamma w_i)]$ ;

<sup>20</sup>One would proceed in precisely this fashion if the central specification were a linear probability model (LPM) to begin with. See, for example, Eisenberg and Rowe (2006) or Angrist (2001) for an application and some analysis of this case.

that is,

$$E\left[\left[y_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta} + \gamma w_i)\right] \begin{pmatrix} \mathbf{x}_i \\ z_i \end{pmatrix}\right] = \mathbf{0}.$$

This form of the moment equation, based on observables, can form the basis of a straight-forward two-step GMM estimator. (See Chapter 13 for details.)

The GMM estimator is not less parametric than the full information maximum likelihood estimator described later because the probit model based on the normal distribution is still invoked to specify the moment equation.<sup>21</sup> Nothing is gained in simplicity or robustness of this approach to full information maximum likelihood estimation, which we now consider. (As Bertschek and Lechner argue, however, the gains might come in terms of practical implementation and computation time. The same considerations motivate Avery et al.)

This maximum likelihood estimator requires a full specification of the model, including the assumption that underlies the endogeneity of  $w_i$ . This becomes essentially a simultaneous equations model. The model equations are

$$\begin{aligned} y_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \gamma w_i + \varepsilon_i, y_i = 1[y_i^* > 0], \\ w_i &= \mathbf{z}_i'\boldsymbol{\alpha} + u_i, \\ (\varepsilon_i, u_i) &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_u \\ \rho\sigma_u & \sigma_u^2 \end{pmatrix}\right]. \end{aligned}$$

(We are assuming that there is a vector of instrumental variables,  $\mathbf{z}_i$ .) Probit estimation based on  $y_i$  and  $(\mathbf{x}_i, w_i)$  will not consistently estimate  $(\boldsymbol{\beta}, \gamma)$  because of the correlation between  $w_i$  and  $\varepsilon_i$  induced by the correlation between  $u_i$  and  $\varepsilon_i$ . Several methods have been proposed for estimation of this model. One possibility is to use the partial reduced form obtained by inserting the second equation in the first. This becomes a probit model with probability  $\text{Prob}(y_i = 1 | \mathbf{x}_i, \mathbf{z}_i) = \Phi(\mathbf{x}_i'\boldsymbol{\beta}^* + \mathbf{z}_i'\boldsymbol{\alpha}^*)$ . This will produce consistent estimates of  $\boldsymbol{\beta}^* = \boldsymbol{\beta}/(1 + \gamma^2\sigma_u^2 + 2\gamma\sigma_u\rho)^{1/2}$  and  $\boldsymbol{\alpha}^* = \gamma\boldsymbol{\alpha}/(1 + \gamma^2\sigma_u^2 + 2\gamma\sigma_u\rho)^{1/2}$  as the coefficients on  $\mathbf{x}_i$  and  $\mathbf{z}_i$ , respectively. (The procedure will estimate a mixture of  $\boldsymbol{\beta}^*$  and  $\boldsymbol{\alpha}^*$  for any variable that appears in both  $\mathbf{x}_i$  and  $\mathbf{z}_i$ .) In addition, linear regression of  $w_i$  on  $\mathbf{z}_i$  produces estimates of  $\boldsymbol{\alpha}$  and  $\sigma_u^2$ . There is no method of moments estimator of  $\rho$  or  $\gamma$  produced by this procedure, so this estimator is incomplete. Newey (1987) suggested a “minimum chi-squared” estimator that does estimate all parameters. A more direct, and actually simpler approach is full information maximum likelihood.

The log-likelihood is built up from the joint density of  $y_i$  and  $w_i$ , which we write as the product of the conditional and the marginal densities,

$$f(y_i, w_i) = f(y_i | w_i) f(w_i).$$

To derive the conditional distribution, we use results for the bivariate normal, and write

$$\varepsilon_i | u_i = [(\rho\sigma_u)/\sigma_u^2] u_i + v_i,$$

<sup>21</sup>This is precisely the platform that underlies the GLIM/GEE treatment of binary choice models in, for example, the widely used programs *SAS* and *Stata*.

**708 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**

where  $v_i$  is normally distributed with  $\text{Var}[v_i] = (1 - \rho^2)$ . Inserting this in the first equation, we have

$$y_i^* | w_i = \mathbf{x}'_i \boldsymbol{\beta} + \gamma w_i + (\rho/\sigma_u) u_i + v_i.$$

Therefore,

$$\text{Prob}[y_i = 1 | \mathbf{x}_i, w_i] = \frac{\Phi\left(\frac{\mathbf{x}'_i \boldsymbol{\beta} + \gamma w_i + (\rho/\sigma_u) u_i}{\sqrt{1 - \rho^2}}\right)}{\sqrt{1 - \rho^2}}. \quad (17-32)$$

Inserting the expression for  $u_i = (w_i - \mathbf{z}'_i \boldsymbol{\alpha})$ , and using the normal density for the marginal distribution of  $w_i$  in the second equation, we obtain the log-likelihood function for the sample,

$$\ln L = \sum_{i=1}^n \ln \left[ \frac{1}{\sigma_u} \Phi\left(\frac{y_i - \mathbf{z}'_i \boldsymbol{\alpha}}{\sigma_u}\right) \right] + \ln \left[ \frac{1}{\sqrt{1 - \rho^2}} \left( \frac{\mathbf{x}'_i \boldsymbol{\beta} + \gamma w_i + (\rho/\sigma_u)(w_i - \mathbf{z}'_i \boldsymbol{\alpha})}{\sqrt{1 - \rho^2}} \right) \right].$$

**Example 17.8 Labor Supply Model**

In Examples 5.2 and 17.1, we examined a labor supply model for married women using Mroz's (1987) data on labor supply. The wife's labor force participation equation suggested in Example 17.1 is

$$\text{Prob}(LFP_i = 1) = \Phi(\beta_1 + \beta_2 \text{Age}_i + \beta_3 \text{Age}_i^2 + \beta_4 \text{Education}_i + \beta_5 \text{Kids}_i).$$

A natural extension of this model would be to include the husband's hours in the equation,

$$\text{Prob}(LFP_i = 1) = \Phi(\beta_1 + \beta_2 \text{Age}_i + \beta_3 \text{Age}_i^2 + \beta_4 \text{Education}_i + \beta_5 \text{Kids}_i + \gamma \text{HHrs}_i).$$

It would also be natural to assume that the husband's hours would be correlated with the determinants (observed and unobserved) of the wife's labor force participation. The auxiliary equation might be

$$\text{HHrs}_i = \alpha_1 + \alpha_2 \text{HAge}_i + \alpha_3 \text{HEducation}_i + \alpha_4 \text{Family Income}_i + u_i.$$

As before, we use the Mroz (1987) labor supply data described in Example 5.2. Table 17.5 reports the single-equation and maximum likelihood estimates of the parameters of the two equations. Comparing the two sets of probit estimates, it appears that the (assumed) endogeneity of the husband's hours is not substantially affecting the estimates. There are two

**TABLE 17.5 Estimated Labor Supply Model**

	<b>Probit</b>	<b>Regression</b>	<b>Maximum Likelihood</b>
Constant	-3.86704 (1.41153)		-5.08405 (1.43134)
Age	0.18681 (0.065901)		0.17108 (0.063321)
Age <sup>2</sup>	-0.00243 (0.000774)		-0.00219 (0.0007629)
Education	0.11098 (0.021663)		0.09037 (0.029041)
Kids	-0.42652 (0.13074)		-0.40202 (0.12967)
Husband hours	-0.000173 (0.0000797)		0.00055 (0.000482)
Constant		2325.38 (167.515)	2424.90 (158.152)
Husband age		-6.71056 (2.73573)	-7.3343 (2.57979)
Husband education		9.29051 (7.87278)	2.1465 (7.28048)
Family income		55.72534 (19.14917)	63.4669 (18.61712)
$\sigma_u$		588.2355	586.994
$\rho$		0.0000	-0.4221 (0.26931)
$\ln L$	-489.0766	-5868.432	-6357.093

## CHAPTER 17 ♦ Discrete Choice 709

simple ways to test the hypothesis that  $\rho$  equals zero. The FIML estimator produces an estimated asymptotic standard error with the estimate of  $\rho$ , so a Wald test can be carried out. For the preceding results, the Wald statistic would be  $(-0.4221/0.26921)^2 = 2.458$ . The critical value from the chi-squared table for one degree of freedom would be 3.84, so we would not reject the hypothesis. The second approach would use the likelihood ratio test. Under the null hypothesis of exogeneity, the probit model and the regression equation can be estimated independently. The log-likelihood for the full model would be the sum of the two log-likelihoods, which would be -6357.508 based on the following results. Without the restriction  $\rho = 0$ , the combined log likelihood is -6357.093. Twice the difference is 0.831, which is also well under the 3.84 critical value, so on this basis as well, we would not reject the null hypothesis that  $\rho = 0$ .

Blundell and Powell (2004) label the foregoing the **control function** approach to accommodating the endogeneity. As noted, the estimator is fully parametric. They propose an alternative semiparametric approach that retains much of the functional form specification, but works around the specific distributional assumptions. Adapting their model to our earlier notation, their departure point is a general specification that produces, once again, a control function,

$$E[y_i | \mathbf{x}_i, w_i, u_i] = F(\mathbf{x}'_i \boldsymbol{\beta} + \gamma w_i, u_i).$$

Note that (17-32) satisfies the assumption; however, they reach this point without assuming either joint or marginal normality. The authors propose a three-step, semiparametric approach to estimating the structural parameters. In an application somewhat similar to Example 17.8, they apply the technique to a labor force participation model for British men in which a variable of interest is a dummy variable for education greater than 16 years, the endogenous variable in the participation equation, also of interest, is earned income of the spouse, and an instrumental variable is a welfare benefit entitlement. Their findings are rather more substantial than ours; they find that when the endogeneity of other family income is accommodated in the equation, the education coefficient increases by 40 percent and remains significant, but the coefficient on other income increases by more than tenfold.

In the control function model noted earlier, where  $E[y_i | \mathbf{x}_i, w_i, u_i] = F(\mathbf{x}'_i \boldsymbol{\beta} + \gamma w_i, u_i)$  and  $w_i = \mathbf{z}'_i \boldsymbol{\alpha} + u_i$ , since the covariance of  $w_i$  and  $u_i$  is the issue, it might seem natural to solve the problem by replacing  $w_i$  with  $\mathbf{z}'_i \mathbf{a}$  where  $\mathbf{a}$  is an estimator of  $\boldsymbol{\alpha}$ , or some other prediction of  $w_i$  based only on exogenous variables. The earlier development shows that the appropriate approach is to add the estimated residual to the equation, instead. The issue is explored in detail by Terza, Basu, and Rathouz (2008), who reach the same conclusion in a general model.

The residual inclusion method also suggests a two-step approach. Rewrite the log-likelihood function as

$$\ln L = \sum_{i=1}^n \ln \Phi [(2y_i - 1)(\mathbf{x}'_i \boldsymbol{\beta}^* + \gamma^* w_i + \tau \tilde{\varepsilon}_i)] + \sum_{i=1}^n \ln \left[ \frac{1}{\sigma_u} \phi(\tilde{\varepsilon}_i) \right],$$

where  $\boldsymbol{\beta}^* = (1/\sqrt{1-\rho^2})\boldsymbol{\beta}$ ,  $\gamma^* = (1/\sqrt{1-\rho^2})\gamma$ ,  $\tau = (\rho/\sqrt{1-\rho^2})$  and  $\tilde{\varepsilon}_i = (w_i - \mathbf{z}'_i \boldsymbol{\alpha})/\sigma_u$ .

The parameters in the regression,  $\boldsymbol{\alpha}$  and  $\sigma_u$ , can be consistently estimated by a linear regression of  $w$  on  $\mathbf{z}$ . The scaled residual  $\tilde{\varepsilon}_i = (w_i - \mathbf{z}'_i \mathbf{a})/s_u$  can now be computed and inserted into the log-likelihood. Note that the second term in the log-likelihood involves parameters that have already been estimated at the first step. The second-step

## 710 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

log-likelihood is, then,

$$\ln L = \sum_{i=1}^n \ln \Phi [(2y_i - 1)(\mathbf{x}'_i \boldsymbol{\beta}^* + \gamma^* w_i + \tau \tilde{e}_i)].$$

This can be maximized using the methods developed in Section 17.3. The estimator of  $\rho$  can be recovered from  $\rho = \tau/(1 + \tau^2)^{1/2}$ . Estimators of  $\boldsymbol{\beta}$  and  $\gamma$  follow, and the delta method can be used to construct standard errors. Since this is a two-step estimator, the resulting estimator of the asymptotic covariance matrix would be further adjusted using the Murphy and Topel (2002) results in Section 14.7. Bootstrapping the entire apparatus (see Section 15.4) would be an alternative way to estimate an asymptotic covariance matrix. The original (one-step) log-likelihood is not very complicated, and full information estimation is fairly straightforward. The preceding demonstrates how the alternative two-step method would proceed and emphasizes once again, the appropriateness of the “residual inclusion” method.

The case in which the endogenous variable in the main equation is, itself, a binary variable occupies a large segment of the recent literature. Consider the model

$$\begin{aligned} T_i^* &= \mathbf{z}'_i \boldsymbol{\alpha} + u_i, \quad T_i = 1[w_i^* > 0], \\ y_i^* &= \mathbf{x}'_i \boldsymbol{\beta} + \gamma T_i + \varepsilon_i, \quad y_i = 1[y_i^* > 0], \\ \begin{pmatrix} \varepsilon_i \\ u_i \end{pmatrix} &\sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right], \end{aligned}$$

where  $T_i$  is a binary variable indicating some kind of program participation (e.g., graduating from high school or college, receiving some kind of job training, purchasing health insurance, etc.). The model in this form (and several similar ones) is a “treatment effects” model. The subject of treatment effects models is surveyed in many studies, including Angrist (2001) and Angrist and Pischke (2009, 2010). The main object of estimation is  $\gamma$  (at least superficially). In these settings, the observed outcome may be  $y_i^*$  (e.g., income or hours) or  $y_i$  (e.g., labor force participation). We have considered the first case in Chapter 8, and will revisit it in Chapter 19. The case just examined is that in which  $y_i$  and  $T_i^*$  are the observed variables. The preceding analysis has suggested that problems of endogeneity will intervene in all cases. We will examine this model in some detail in Section 17.5.5 and in Chapter 19.

### 17.3.6 ENDOGENOUS CHOICE-BASED SAMPLING

In some studies [e.g., Boyes, Hoffman, and Low (1989), Greene (1992)], the mix of ones and zeros in the observed sample of the dependent variable is deliberately skewed in favor of one outcome or the other to achieve a more balanced sample than random sampling would produce. The sampling is said to be **choice based**. In the studies noted, the dependent variable measured the occurrence of loan default, which is a relatively uncommon occurrence. To enrich the sample, observations with  $y = 1$  (default) were oversampled. Intuition should suggest (correctly) that the bias in the sample should be transmitted to the parameter estimates, which will be estimated so as to mimic the sample, not the population, which is known to be different. Manski and Lerman (1977) derived the weighted endogenous sampling maximum likelihood (WESML) estimator for this situation. The estimator requires that the true population proportions,  $\omega_1$

and  $\omega_0$ , be known. Let  $p_1$  and  $p_0$  be the sample proportions of ones and zeros. Then the estimator is obtained by maximizing a weighted log-likelihood,

$$\ln L = \sum_{i=1}^n w_i \ln F(q_i \mathbf{x}'_i \boldsymbol{\beta}),$$

where  $w_i = y_i(\omega_1/p_1) + (1 - y_i)(\omega_0/p_0)$ . Note that  $w_i$  takes only two different values. The derivatives and the Hessian are likewise weighted. A final correction is needed after estimation; the appropriate estimator of the asymptotic covariance matrix is the sandwich estimator discussed in Section 17.3.1,  $\mathbf{H}^{-1} \mathbf{B} \mathbf{H}^{-1}$  (with weighted  $\mathbf{B}$  and  $\mathbf{H}$ ), instead of  $\mathbf{B}$  or  $\mathbf{H}$  alone. (The weights are not squared in computing  $\mathbf{B}$ .)<sup>22</sup>

### Example 17.2 Credit Scoring

In Example 7.2 we examined the spending patterns of a sample of 10,499 cardholders for a major credit card vendor. The sample of cardholders is a subsample of 13,444 applicants for the credit card. Applications for credit cards, then (1992) and now are processed by a major nationwide processor, Fair Isaacs, Inc. The algorithm used by the processors is proprietary. However, conventional wisdom holds that a few variables are important in the process, such as Age, Income, whether the applicant owns their home, whether they are self-employed, and how long they have lived at their current address. The number of major and minor derogatory reports (60-day and 30-day delinquencies) are influential variables in credit scoring. The probit model we will use to ‘model the model’ is

$$\begin{aligned} \text{Prob}(Cardholder} = 1) &= \text{Prob}(C = 1 | \mathbf{x}) \\ &= \Phi(\beta_1 + \beta_2 \text{Age} + \beta_3 \text{Income} + \beta_4 \text{OwnRent} \\ &\quad + \beta_5 \text{Months Living at Current Address} \\ &\quad + \beta_6 \text{Self-Employed} \\ &\quad + \beta_7 \text{Number of major derogatory reports} \\ &\quad + \beta_8 \text{Number of minor derogatory reports}) \end{aligned}$$

In the data set, 78.1 percent of the applicants are cardholders. In the population, at that time, the true proportion was roughly 23.2 percent, so the sample is substantially choice based on this variable. The sample was deliberately skewed in favor of cardholders for purposes of the original study [Greene (1992)]. The weights to be applied for the WESML estimator are  $0.232/0.781 = 0.297$  for the observations with  $C = 1$  and  $0.768/0.219 = 3.507$  for observations with  $C = 0$ . Table 17.6 presents the unweighted and weighted estimates for this application. The change in the estimates produced by the weighting is quite modest, save for the constant term. The results are consistent with the conventional wisdom that *Income* and *OwnRent* are two important variables in a credit application and self-employment receives a substantial negative weight. But, as might be expected, the single most significant influence on cardholder status is major derogatory reports. Since lenders are strongly focused on default probability, past evidence of default behavior will be a major consideration.

#### 17.3.7 SPECIFICATION ANALYSIS

In his survey of qualitative response models, Amemiya (1981) reports the following widely cited approximations for the linear probability (LP) model: Over the range of

<sup>22</sup>WESML and the choice-based sampling estimator are not the free lunch they may appear to be. That which the biased sampling does, the weighting undoes. It is common for the end result to be very large standard errors, which might be viewed as unfortunate, insofar as the purpose of the biased sampling was to balance the data precisely to avoid this problem.

**712 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**
**TABLE 17.6** Estimated Card Application Equation (*t* ratios in parentheses)

<i>Variable</i>	<i>Unweighted</i>		<i>Weighted</i>			
	<i>Estimate</i>	<i>Standard Error</i>	<i>Estimate</i>	<i>Standard Error</i>		
Constant	0.31783	0.05094	(6.24)	-1.13089	0.04725	(-23.94)
Age	0.00184	0.00154	(1.20)	0.00156	0.00145	(1.07)
Income	0.00095	0.00025	(3.86)	0.00094	0.00024	(3.92)
OwnRent	0.18233	0.03061	(5.96)	0.23967	0.02968	(8.08)
CurrentAddress	0.02237	0.00120	(18.67)	0.02106	0.00109	(19.40)
SelfEmployed	-0.43625	0.05585	(-7.81)	-0.47650	0.05851	(-8.14)
Major Derogs	-0.69912	0.01920	(-36.42)	-0.64792	0.02525	(-25.66)
Minor Derogs	-0.04126	0.01865	(-2.21)	-0.04285	0.01778	(-2.41)

probabilities of 30 to 70 percent,

$$\hat{\beta}_{LP} \approx 0.4\beta_{probit} \text{ for the slopes,}$$

$$\hat{\beta}_{LP} \approx 0.25\beta_{logit} \text{ for the slopes.}$$

Aside from confirming our intuition that least squares approximates the nonlinear model and providing a quick comparison for the three models involved, the practical usefulness of the formula is somewhat limited. Still, it is a striking result.<sup>23</sup> A series of studies has focused on reasons why the least squares estimates should be proportional to the probit and logit estimates. A related question concerns the problems associated with assuming that a probit model applies when, in fact, a logit model is appropriate or vice versa.<sup>24</sup> The approximation would seem to suggest that with this type of misspecification, we would once again obtain a scaled version of the correct coefficient vector. (Amemiya also reports the widely observed relationship  $\hat{\beta}_{logit} \approx 1.6\hat{\beta}_{probit}$ , which follows from the results for the linear probability model. This result is apparent in Table 17.1 where the ratios of the three slopes range from 1.6 to 1.9.)

In the linear regression model, we considered two important specification problems: the effect of omitted variables and the effect of heteroscedasticity. In the classical model,  $y = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$ , when least squares estimates  $\mathbf{b}_1$  are computed omitting  $\mathbf{X}_2$ ,

$$E[\mathbf{b}_1] = \beta_1 + [\mathbf{X}'_1\mathbf{X}_1]^{-1}\mathbf{X}'_1\mathbf{X}_2\beta_2.$$

Unless  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are orthogonal or  $\beta_2 = \mathbf{0}$ ,  $\mathbf{b}_1$  is biased. If we ignore heteroscedasticity, then although the least squares estimator is still unbiased and consistent, it is inefficient and the usual estimate of its sampling covariance matrix is inappropriate. Yatchew and Griliches (1984) have examined these same issues in the setting of the probit and logit models. Their general results are far more pessimistic. In the context of a binary choice model, they find the following:

<sup>23</sup>This result does not imply that it is useful to report 2.5 times the linear probability estimates with the probit estimates for comparability. The linear probability estimates are already in the form of marginal effects, whereas the probit coefficients must be scaled *downward*. If the sample proportion happens to be close to 0.5, then the right scale factor will be roughly  $\phi[\Phi^{-1}(0.5)] = 0.3989$ . But the density falls rapidly as  $P$  moves away from 0.5.

<sup>24</sup>See Ruud (1986) and Gourieroux et al. (1987).

1. If  $x_2$  is omitted from a model containing  $x_1$  and  $x_2$ , (i.e.  $\beta_2 \neq 0$ ) then

$$\text{plim } \hat{\beta}_1 = c_1\beta_1 + c_2\beta_2,$$

where  $c_1$  and  $c_2$  are complicated functions of the unknown parameters. The implication is that even if the omitted variable is uncorrelated with the included one, the coefficient on the included variable will be inconsistent.

2. If the disturbances in the underlying regression are heteroscedastic, then the maximum likelihood estimators are inconsistent and the covariance matrix is inappropriate.

The second result is particularly troubling because the probit model is most often used with microeconomic data, which are frequently heteroscedastic.

Any of the three methods of hypothesis testing discussed here can be used to analyze these specification problems. The Lagrange multiplier test has the advantage that it can be carried out using the estimates from the restricted model, which sometimes brings a large saving in computational effort. This situation is especially true for the test for **heteroscedasticity**.<sup>25</sup>

To reiterate, the Lagrange multiplier statistic is computed as follows. Let the null hypothesis,  $H_0$ , be a specification of the model, and let  $H_1$  be the alternative. For example,  $H_0$  might specify that only variables  $\mathbf{x}_1$  appear in the model, whereas  $H_1$  might specify that  $\mathbf{x}_2$  appears in the model as well. The statistic is

$$\text{LM} = \mathbf{g}'_0 \mathbf{V}_0^{-1} \mathbf{g}_0,$$

where  $\mathbf{g}_0$  is the vector of derivatives of the log-likelihood as specified by  $H_1$  but evaluated at the maximum likelihood estimator of the parameters assuming that  $H_0$  is true, and  $\mathbf{V}_0^{-1}$  is any of the three consistent estimators of the asymptotic variance matrix of the maximum likelihood estimator under  $H_1$ , also compu~~ting~~ using the maximum likelihood estimators based on  $H_0$ . The statistic is asymptotically distributed as chi-squared with degrees of freedom equal to the number of restrictions.

#### 17.3.7.a Omitted Variables

The hypothesis to be tested is

$$\begin{aligned} H_0: y^* &= \mathbf{x}'_1 \boldsymbol{\beta}_1 + \varepsilon, \\ H_1: y^* &= \mathbf{x}'_1 \boldsymbol{\beta}_1 + \mathbf{x}'_2 \boldsymbol{\beta}_2 + \varepsilon, \end{aligned} \tag{17-33}$$

so the test is of the null hypothesis that  $\boldsymbol{\beta}_2 = \mathbf{0}$ . The Lagrange multiplier test would be carried out as follows:

1. Estimate the model in  $H_0$  by maximum likelihood. The restricted coefficient vector is  $[\hat{\beta}_1, \mathbf{0}]$ .
2. Let  $\mathbf{x}$  be the compound vector,  $[\mathbf{x}_1, \mathbf{x}_2]$ .

The statistic is then computed according to (17-30) or (17-31). It is noteworthy that in this case as in many others, the Lagrange multiplier is the coefficient of determination in a regression. The likelihood ratio test is equally straightforward. Using the estimates of the two models, the statistic is simply  $2(\ln L_1 - \ln L_0)$ .

<sup>25</sup>The results in this section are based on Davidson and MacKinnon (1984) and Engle (1984). A symposium on the subject of specification tests in discrete choice models is Blundell (1987).

## 714 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

### 17.3.7.b Heteroscedasticity

We use the general formulation analyzed by Harvey (1976) (see Section 14.9.2.a),<sup>26</sup>

$$\text{Var}[\varepsilon] = [\exp(\mathbf{z}'\boldsymbol{\gamma})]^2.$$

This model can be applied equally to the probit and logit models. We will derive the results specifically for the probit model; the logit model is essentially the same. Thus,

$$\begin{aligned} y^* &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \\ \text{Var}[\varepsilon | \mathbf{x}, \mathbf{z}] &= [\exp(\mathbf{z}'\boldsymbol{\gamma})]^2. \end{aligned} \quad (17-34)$$

he presence of heteroscedasticity makes some care necessary in interpreting the coefficients for a variable  $w_k$  that could be in  $\mathbf{x}$  or  $\mathbf{z}$  or both,

$$\frac{\partial \text{Prob}(Y=1 | \mathbf{x}, \mathbf{z})}{\partial w_k} = \phi \left[ \frac{\mathbf{x}'\boldsymbol{\beta}}{\exp(\mathbf{z}'\boldsymbol{\gamma})} \right] \frac{\beta_k - (\mathbf{x}'\boldsymbol{\beta})\gamma_k}{\exp(\mathbf{z}'\boldsymbol{\gamma})}.$$

Only the first (second) term applies if  $w_k$  appears only in  $\mathbf{x}$  ( $\mathbf{z}$ ). This implies that the simple coefficient may differ radically from the effect that is of interest in the estimated model. This effect is clearly visible in the next example.

The log-likelihood is

$$\ln L = \sum_{i=1}^n \left\{ y_i \ln F \left( \frac{\mathbf{x}_i'\boldsymbol{\beta}}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})} \right) + (1-y_i) \ln \left[ 1 - F \left( \frac{\mathbf{x}_i'\boldsymbol{\beta}}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})} \right) \right] \right\}. \quad (17-35)$$

To be able to estimate all the parameters,  $\mathbf{z}$  cannot have a constant term. The derivatives are

$$\begin{aligned} \frac{\partial \ln L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^n \left[ \frac{f_i(y_i - F_i)}{F_i(1 - F_i)} \right] \exp(-\mathbf{z}_i'\boldsymbol{\gamma}) \mathbf{x}_i, \\ \frac{\partial \ln L}{\partial \boldsymbol{\gamma}} &= \sum_{i=1}^n \left[ \frac{f_i(y_i - F_i)}{F_i(1 - F_i)} \right] \exp(-\mathbf{z}_i'\boldsymbol{\gamma}) \mathbf{z}_i (-\mathbf{x}_i'\boldsymbol{\beta}), \end{aligned} \quad (17-36)$$

which implies a difficult log-likelihood to maximize. But if the model is estimated assuming that  $\boldsymbol{\gamma} = \mathbf{0}$ , then we can easily test for homoscedasticity. Let

$$\mathbf{w}_i = \begin{bmatrix} \mathbf{x}_i \\ (-\mathbf{x}_i'\hat{\boldsymbol{\beta}})\mathbf{z}_i \end{bmatrix}, \quad (17-37)$$

computed at the maximum likelihood estimator, assuming that  $\boldsymbol{\gamma} = \mathbf{0}$ . Then (17-30) or (17-31) can be used as usual for the Lagrange multiplier statistic.

Davidson and MacKinnon carried out a Monte Carlo study to examine the true sizes and power functions of these tests. As might be expected, the test for omitted variables is relatively powerful. The test for heteroscedasticity may well pick up some other form of misspecification, however, including perhaps the simple omission of  $\mathbf{z}$  from the index function, so its power may be problematic. It is perhaps not surprising that the same problem arose earlier in our test for heteroscedasticity in the linear regression model.

<sup>26</sup>See Knapp and Seaks (1992) for an application. Other formulations are suggested by Fisher and Nagin (1981), Hausman and Wise (1978), and Horowitz (1993).

**Example 17.10 Specification Tests in a Labor Force Participation Model**

Using the data described in Example 17.1, we fit a probit model for labor force participation based on the specification

$$\text{Prob}[LFP = 1] = F(\text{constant}, \text{age}, \text{age}^2, \text{family income}, \text{education}, \text{kids}).$$

For these data,  $P = 428/753 = 0.568393$ . The restricted (all slopes equal zero, free constant term) log-likelihood is  $325 \times \ln(325/753) + 428 \times \ln(428/753) = -514.8732$ . The unrestricted log-likelihood for the probit model is  $-490.8478$ . The chi-squared statistic is, therefore, 48.05072. The critical value from the chi-squared distribution with five degrees of freedom is 11.07, so the joint hypothesis that the coefficients on *age*, *age*<sup>2</sup>, *family income*, and *kids* are all zero is rejected.

Consider the alternative hypothesis, that the constant term and the coefficients on *age*, *age*<sup>2</sup>, *family income*, and *education* are the same whether *kids* equals one or zero, against the alternative that an altogether different equation applies for the two groups of women, those with *kids* = 1 and those with *kids* = 0. To test this hypothesis, we would use a counterpart to the **Chow test** of Section 6.1 and Example 6.9. The restricted model in this instance would be based on the pooled data set of all 753 observations. The log-likelihood for the pooled model—which has a constant term, *age*, *age*<sup>2</sup>, *family income*, and *education* is  $-496.8663$ . The log-likelihoods for this model based on the 524 observations with *kids* = 1 and the 229 observations with *kids* = 0 are  $-347.87441$  and  $-141.60501$ , respectively. The log-likelihood for the unrestricted model with separate coefficient vectors is thus the sum,  $-489.47942$ . The chi-squared statistic for testing the five restrictions of the pooled model is twice the difference,  $\text{LR} = 2[-489.47942 - (-496.8663)] = 14.7738$ . The 95 percent critical value from the chi-squared distribution with 5 degrees of freedom is 11.07, so at this significance level, the hypothesis that the constant terms and the coefficients on *age*, *age*<sup>2</sup>, *family income*, and *education* are the same is rejected. (The 99 percent critical value is 15.09.)

Table 17.7 presents estimates of the probit model with a correction for heteroscedasticity of the form

$$\text{Var}[\varepsilon_i] = \exp(\gamma_1 \text{kids} + \gamma_2 \text{family income}).$$

The three tests for homoscedasticity give

$$\text{LR} = 2[-487.6356 - (-490.8478)] = 6.424,$$

$$\text{LM} = 2.236 \text{ based on the BHHH estimator,}$$

$$\text{Wald} = 6.533 \text{ (2 restrictions).}$$

The 95 percent critical value for two restrictions is 5.99, so the LM statistic conflicts with the other two.

**TABLE 17.7** Estimated Coefficients

		Estimate (Std. Er.)	Marg. Effect*	Estimate (St. Er.)	Marg. Effect*
Constant	$\beta_1$	-4.157(1.402)	-0.00823(.00649)	-6.030(2.498)	-0.00823(.00649)
Age	$\beta_2$	0.185(0.0660)	-0.0079(0.0027)	0.264(0.118)	-0.0088(0.00251)
Age <sup>2</sup>	$\beta_3$	-0.0024(0.00077)	—	-0.0036(0.0014)	—
Income	$\beta_4$	0.0458(0.0421)	0.0180(0.0165)	0.424(0.222)	0.0552(0.0240)
Education	$\beta_5$	0.0982(0.0230)	0.0385(0.0090)	0.140(0.0519)	0.0289(0.00869)
Kids	$\beta_6$	-0.449(0.131)	-0.171(0.0480)	-0.879(0.303)	-0.167(0.0779)
Kids	$\gamma_1$	0.000	—	-0.141(0.324)	—
Income	$\gamma_2$	0.000	—	0.313(0.123)	—
ln L			-490.8478		-487.6356
Correct Preds.			0s: 106, 1s: 357		0s: 115, 1s: 358

\*Marginal effect and estimated standard error include both mean ( $\beta$ ) and variance ( $\gamma$ ) effects.

## 716 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

### 17.4 BINARY CHOICE MODELS FOR PANEL DATA

Qualitative response models have been a growth industry in econometrics. The recent literature, particularly in the area of panel data analysis, has produced a number of new techniques. The availability of high-quality panel data sets on microeconomic behavior has maintained an interest in extending the models of Chapter 11 to binary (and other discrete choice) models. In this section, we will survey a few results from this rapidly growing literature.

The structural model for a possibly unbalanced panel of data would be written

$$\begin{aligned} y_{it}^* &= \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \dots, n, t = 1, \dots, T_i, \\ y_{it} &= 1 \quad \text{if } y_{it}^* > 0, \text{ and } 0 \text{ otherwise,} \end{aligned} \tag{17-38}$$

The second line of this definition is often written

$$y_{it} = \mathbf{1}(\mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it} > 0)$$

to indicate a variable that equals one when the condition in parentheses is true and zero when it is not. Ideally, we would like to specify that  $\varepsilon_{it}$  and  $\varepsilon_{is}$  are freely correlated within a group, but uncorrelated across groups. But doing so will involve computing joint probabilities from a  $T_i$  variate distribution, which is generally problematic.<sup>27</sup> (We will return to this issue later.) A more promising approach is an effects model,

$$\begin{aligned} y_{it}^* &= \mathbf{x}'_{it}\boldsymbol{\beta} + v_{it} + u_i, \quad i = 1, \dots, n, t = 1, \dots, T_i, \\ y_{it} &= 1 \quad \text{if } y_{it}^* > 0, \text{ and } 0 \text{ otherwise,} \end{aligned} \tag{17-39}$$

where, as before (see Sections 11.4 and 11.5),  $u_i$  is the unobserved, individual specific heterogeneity. Once again, we distinguish between “random” and “fixed” effects models by the relationship between  $u_i$  and  $\mathbf{x}_{it}$ . The assumption that  $u_i$  is unrelated to  $\mathbf{x}_{it}$ , so that the conditional distribution  $f(u_i | \mathbf{x}_{it})$  is not dependent on  $\mathbf{x}_{it}$ , produces the **random effects model**. Note that this places a restriction on the distribution of the heterogeneity.

If that distribution is unrestricted, so that  $u_i$  and  $\mathbf{x}_{it}$  may be correlated, then we have what is called the **fixed effects model**. The distinction does not relate to any intrinsic characteristic of the effect itself.

As we shall see shortly, this is a modeling framework that is fraught with difficulties and unconventional estimation problems. Among them are the following: Estimation of the random effects model requires very strong assumptions about the heterogeneity;

---

<sup>27</sup>A “limited information” approach based on the GMM estimation method has been suggested by Avery, Hansen, and Hotz (1983). With recent advances in simulation-based computation of multinormal integrals (see Section 15.6.2.b), some work on such a panel data estimator has appeared in the literature. See, for example, Geweke, Keane, and Runkle (1994, 1997). The GEE estimator of Diggle, Liang, and Zeger (1994) [see also, Liang and Zeger (1986) and Stata (2006)] seems to be another possibility. However, in all these cases, it must be remembered that the procedure specifies estimation of a correlation matrix for a  $T_i$  vector of unobserved variables based on a dependent variable that takes only two values. We should not be too optimistic about this if  $T_i$  is even moderately large.

the fixed effects model encounters an **incidental parameters problem** that renders the maximum likelihood estimator inconsistent.

#### 17.4.1 THE POOLED ESTIMATOR

To begin, it is useful to consider the pooled estimator that results if we simply ignore the heterogeneity,  $u_i$  in (17-39) and fit the model as if the cross-section specification of Section 17.2.2 applies. In this instance, the adage that “ignoring the heterogeneity does not make it go away,” applies even more forcefully than in the linear regression case.

If the fixed effects model is appropriate, then all the preceding results for omitted variables, including the Yatchew and Griliches result (1984) apply. The pooled MLE that ignores fixed effects will be inconsistent—possibly wildly so. (Note that since the estimator is ML, not least squares, converting the data to deviations from group means is not a solution—converting the binary dependent variable to deviations will produce a continuous variable with unknown properties.)

The random effects case is more benign. From (17-39), the marginal probability implied by the model is

$$\begin{aligned}\text{Prob}(y_{it} = 1 | \mathbf{x}_{it}) &= \text{Prob}(v_{it} + u_i > -\mathbf{x}'_{it}\boldsymbol{\beta}) \\ &= F[\mathbf{x}'_{it}\boldsymbol{\beta}/(1 + \sigma_u^2)^{1/2}] \\ &= F(\mathbf{x}'_{it}\boldsymbol{\delta}).\end{aligned}$$

The implication is that based on the marginal distributions, we can consistently estimate  $\boldsymbol{\delta}$  (but not  $\boldsymbol{\beta}$  or  $\sigma_u$  separately) by pooled MLE. [This result is explored at length in Wooldridge (2002).] This would be a “pseudo MLE” since the log-likelihood function is not the true log-likelihood for the full set of observed data, but it is the correct product of the marginal distributions for  $y_{it} | \mathbf{x}_{it}$ . (This would be the binary choice case counterpart to consistent estimation of  $\boldsymbol{\beta}$  in a linear random effects model by pooled ordinary least squares.) The implication, which is absent in the linear case is that ignoring the random effects in a pooled model produces an attenuated (inconsistent—downward biased) estimate of  $\boldsymbol{\beta}$ ; the scale factor that produces  $\boldsymbol{\delta}$  is  $1/(1 + \sigma_u^2)^{1/2}$  which is between zero and one. The implication for the partial effects is less clear. In the model specification, the partial effect is

$$PE(\mathbf{x}_{it}, u_i) = \partial E[y_{it} | \mathbf{x}_{it}, u_i]/\partial \mathbf{x}_{it} = \boldsymbol{\beta} \times f(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i),$$

which is not computable. The useful result would be

$$E_u[PE(\mathbf{x}_{it}, u_i)] = \boldsymbol{\beta} E_u[f(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i)].$$

Wooldridge (2002a)  shows that the end result, assuming normality of both  $v_{it}$  and  $u_i$  is  $E_u[PE(\mathbf{x}_{it}, u_i)] = \boldsymbol{\beta} F(\mathbf{x}'_{it}\boldsymbol{\delta})$ . Thus far, surprisingly, it would seem that simply pooling the data and using the simple MLE “works.” The estimated standard errors will be incorrect, so a correction such as the cluster estimator shown in Section 14.8.4 would be appropriate. Three considerations suggest that one might want to proceed to the full MLE in spite of these results: (1) The pooled estimator will be inefficient compared to the full MLE; (2) the pooled estimator does not produce an estimator of  $\sigma_u$  which might be of interest in its own right; (3) the FIML estimator is available in contemporary

## 718 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

software and is no more difficult to estimate than the pooled estimator. Note that the pooled estimator is not justified (over the FIML approach) on robustness considerations because the same normality and random effects assumptions that are needed to obtain the FIML estimator will be needed to obtain the preceding results for the pooled estimator.

### 17.4.2 RANDOM EFFECTS MODELS

A specification that has the same structure as the random effects model of Section 11.5 has been implemented by Butler and Moffitt (1982). We will sketch the derivation to suggest how random effects can be handled in discrete and limited dependent variable models such as this one. Full details on estimation and inference may be found in Butler and Moffitt (1982) and Greene (1995a). We will then examine some extensions of the Butler and Moffitt model.

The random effects model specifies

$$\varepsilon_{it} = v_{it} + u_i,$$

where  $v_{it}$  and  $u_i$  are independent random variables with

$$E[v_{it} | \mathbf{X}] = 0; \text{Cov}[v_{it}, v_{js} | \mathbf{X}] = \text{Var}[v_{it} | \mathbf{X}] = 1, \quad \text{if } i = j \text{ and } t = s; 0 \text{ otherwise,}$$

$$E[u_i | \mathbf{X}] = 0; \text{Cov}[u_i, u_j | \mathbf{X}] = \text{Var}[u_i | \mathbf{X}] = \sigma_u^2, \quad \text{if } i = j; 0 \text{ otherwise,}$$

$$\text{Cov}[v_{it}, u_j | \mathbf{X}] = 0 \text{ for all } i, t, j,$$

and  $\mathbf{X}$  indicates all the exogenous data in the sample,  $\mathbf{x}_{it}$  for all  $i$  and  $t$ .<sup>28</sup> Then,

$$E[\varepsilon_{it} | \mathbf{X}] = 0,$$

$$\text{Var}[\varepsilon_{it} | \mathbf{X}] = \sigma_v^2 + \sigma_u^2 = 1 + \sigma_u^2,$$

and

$$\text{Corr}[\varepsilon_{it}, \varepsilon_{is} | \mathbf{X}] = \rho = \frac{\sigma_u^2}{1 + \sigma_u^2}.$$

The new free parameter is  $\sigma_u^2 = \rho/(1 - \rho)$ .

Recall that in the cross-section case, the marginal probability associated with an observation is

$$P(y_i | \mathbf{x}_i) = \int_{L_i}^{U_i} f(\varepsilon_i) d\varepsilon_i, (L_i, U_i) = (-\infty, -\mathbf{x}'_i \boldsymbol{\beta}) \quad \text{if } y_i = 0 \text{ and } (-\mathbf{x}'_i \boldsymbol{\beta}, +\infty) \quad \text{if } y_i = 1.$$

This simplifies to  $\Phi[(2y_i - 1)\mathbf{x}'_i \boldsymbol{\beta}]$  for the normal distribution and  $\Lambda[(2y_i - 1)\mathbf{x}'_i \boldsymbol{\beta}]$  for the logit model. In the fully general case with an unrestricted covariance matrix, the contribution of group  $i$  to the likelihood would be the joint probability for all  $T_i$  observations;

$$L_i = P(y_{i1}, \dots, y_{iT_i} | \mathbf{X}) = \int_{L_{iT_i}}^{U_{iT_i}} \dots \int_{L_{i1}}^{U_{i1}} f(\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iT_i}) d\varepsilon_{i1} d\varepsilon_{i2} \dots d\varepsilon_{iT_i}. \quad (17-40)$$

<sup>28</sup>See Wooldridge (1999) for discussion of this assumption.

The integration of the joint density, as it stands, is impractical in most cases. The special nature of the random effects model allows a simplification, however. We can obtain the joint density of the  $v_{it}$ 's by integrating  $u_i$  out of the joint density of  $(\varepsilon_{i1}, \dots, \varepsilon_{iT_i}, u_i)$  which is

$$f(\varepsilon_{i1}, \dots, \varepsilon_{iT_i}, u_i) = f(\varepsilon_{i1}, \dots, \varepsilon_{iT_i} | u_i) f(u_i).$$

So,

$$f(\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iT_i}) = \int_{-\infty}^{+\infty} f(\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iT_i} | u_i) f(u_i) du_i.$$

The advantage of this form is that conditioned on  $u_i$ , the  $\varepsilon_{it}$ 's are independent, so

$$f(\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iT_i}) = \int_{-\infty}^{+\infty} \prod_{t=1}^{T_i} f(\varepsilon_{it} | u_i) f(u_i) du_i.$$

Inserting this result in (17-40) produces

$$L_i = P[y_{i1}, \dots, y_{iT_i} | \mathbf{X}] = \int_{L_{iT_i}}^{U_{iT_i}} \dots \int_{L_{i1}}^{U_{i1}} \int_{-\infty}^{+\infty} \prod_{t=1}^{T_i} f(\varepsilon_{it} | u_i) f(u_i) du_i d\varepsilon_{i1} d\varepsilon_{i2} \dots d\varepsilon_{iT_i}.$$

This may not look like much simplification, but in fact, it is. Because the ranges of integration are independent, we may change the order of integration;

$$L_i = P[y_{i1}, \dots, y_{iT_i} | \mathbf{X}] = \int_{-\infty}^{+\infty} \left[ \int_{L_{iT_i}}^{U_{iT_i}} \dots \int_{L_{i1}}^{U_{i1}} \prod_{t=1}^{T_i} f(\varepsilon_{it} | u_i) d\varepsilon_{i1} d\varepsilon_{i2} \dots d\varepsilon_{iT_i} \right] f(u_i) du_i.$$

Conditioned on the common  $u_i$ , the  $\varepsilon$ 's are independent, so the term in square brackets is just the product of the individual probabilities. We can write this as

$$L_i = P[y_{i1}, \dots, y_{iT_i} | \mathbf{X}] = \int_{-\infty}^{+\infty} \left[ \prod_{t=1}^{T_i} \left( \int_{L_{it}}^{U_{it}} f(\varepsilon_{it} | u_i) d\varepsilon_{it} \right) \right] f(u_i) du_i. \quad (17-41)$$

Now, consider the individual densities in the product. Conditioned on  $u_i$ , these are the now-familiar probabilities for the individual observations, computed now at  $\mathbf{x}'_{it}\beta + u_i$ . This produces a general model for random effects for the binary choice model. Collecting all the terms, we have reduced it to

$$L_i = P[y_{i1}, \dots, y_{iT_i} | \mathbf{X}] = \int_{-\infty}^{+\infty} \left[ \prod_{t=1}^{T_i} \text{Prob}(Y_{it} = y_{it} | \mathbf{x}'_{it}\beta + u_i) \right] f(u_i) du_i. \quad (17-42)$$

It remains to specify the distributions, but the important result thus far is that the entire computation requires only one-dimensional integration. The inner probabilities may be any of the models we have considered so far, such as probit, logit, Gumbel, and so on. The intricate part that remains is to determine how to do the outer integration. **Butler and Moffitt's method** assuming that  $u_i$  is normally distributed is detailed in Section 14.9.6.c.

A number of authors have found the Butler and Moffitt formulation to be a satisfactory compromise between a fully unrestricted model and the cross-sectional variant that ignores the correlation altogether. An application that includes both group and time effects is Tauchen, Witte, and Griesinger's (1994) study of arrests and criminal

## 720 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

behavior. The Butler and Moffitt approach has been criticized for the restriction of equal correlation across periods. But it does have a compelling virtue that the model can be efficiently estimated even with fairly large  $T_i$  using conventional computational methods. [See Greene (2007b).]

A remaining problem with the Butler and Moffitt specification is its assumption of normality. In general, other distributions are problematic because of the difficulty of finding either a closed form for the integral or a satisfactory method of approximating the integral. An alternative approach that allows some flexibility is the method of **maximum simulated likelihood** (MSL), which was discussed in Section 15.6. The transformed likelihood we derived in (17-42) is an expectation:

$$\begin{aligned} L_i &= \int_{-\infty}^{+\infty} \left[ \prod_{t=1}^{T_i} \text{Prob}(Y_{it} = y_{it} | \mathbf{x}'_{it}\beta + u_i) \right] f(u_i) du_i \\ &= E_{u_i} \left[ \prod_{t=1}^{T_i} \text{Prob}(Y_{it} = y_{it} | \mathbf{x}'_{it}\beta + u_i) \right]. \end{aligned}$$

This expectation can be approximated by simulation rather than **quadrature**. First, let  $\theta$  now denote the scale parameter in the distribution of  $u_i$ . This would be  $\sigma_u$  for a normal distribution, for example, or some other scaling for the logistic or uniform distribution. Then, write the term in the likelihood function as

$$L_i = E_{u_i} \left[ \prod_{t=1}^{T_i} F(y_{it}, \mathbf{x}'_{it}\beta + \theta u_i) \right] = E_{u_i}[h(u_i)].$$

The function is smooth, continuous, and continuously differentiable. If this expectation is finite, then the conditions of the law of large numbers should apply, which would mean that for a sample of observations  $u_{i1}, \dots, u_{iR}$ ,

$$\text{plim } \frac{1}{R} \sum_{r=1}^R h(u_{ir}) = E_u[h(u_i)].$$

This suggests, based on the results in Chapter 15, an alternative method of maximizing the log-likelihood for the random effects model. A sample of person-specific draws from the population  $u_i$  can be generated with a random number generator. For the Butler and Moffitt model with normally distributed  $u_i$ , the simulated log-likelihood function is

$$\ln L_{\text{Simulated}} = \sum_{i=1}^n \ln \left\{ \frac{1}{R} \sum_{r=1}^R \left[ \prod_{t=1}^{T_i} F[\underline{2}y_{it} - \underline{1}(\mathbf{x}'_{it}\beta + \sigma_u u_{ir})] \right] \right\}. \quad (17-43)$$

This function is maximized with respect to  $\beta$  and  $\sigma_u$ . Note that in the preceding, as in the quadrature approximated log-likelihood, the model can be based on a probit, logit, or any other functional form desired.

We have examined two approaches to estimation of a probit model with random effects. GMM estimation is another possibility. Avery, Hansen, and Hotz (1983), Bertschek and Lechner (1998), and Inkmann (2000) examine this approach; the latter two offer some comparison with the quadrature and simulation-based estimators considered here. (Our application in Example 17.23 will use the Bertschek and Lechner data.)

### 17.4.3 FIXED EFFECTS MODELS

The fixed effects model is

$$\begin{aligned} y_{it}^* &= \alpha_i d_{it} + \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T_i, \\ y_{it} &= 1 \quad \text{if } y_{it}^* > 0, \quad \text{and 0 otherwise,} \end{aligned} \tag{17-44}$$

where  $d_{it}$  is a dummy variable that takes the value one for individual  $i$  and zero otherwise. For convenience, we have redefined  $\mathbf{x}_{it}$  to be the nonconstant variables in the model. The parameters to be estimated are the  $K$  elements of  $\boldsymbol{\beta}$  and the  $n$  individual constant terms. Before we consider the several virtues and shortcomings of this model, we consider the practical aspects of estimation of what are possibly a huge number of parameters,  $(n + K) - n$  is not limited here, and could be in the thousands in a typical application. The log-likelihood function for the fixed effects model is

$$\ln L = \sum_{i=1}^n \sum_{t=1}^{T_i} \ln P(y_{it} | \alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta}), \tag{17-45}$$

where  $P(\cdot)$  is the probability of the observed outcome, for example,  $\Phi[q_{it}(\alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta})]$  for the probit model or  $\Lambda[q_{it}(\alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta})]$  for the logit model, where  $q_{it} = 2y_{it} - 1$ . What follows can be extended to any index function model, but for the present, we'll confine our attention to symmetric distributions such as the normal and logistic, so that the probability can be conveniently written as  $\text{Prob}(Y_{it} = y_{it} | \mathbf{x}_{it}) = P[q_{it}(\alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta})]$ . It will be convenient to let  $z_{it} = \alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta}$  so  $\text{Prob}(Y_{it} = y_{it} | \mathbf{x}_{it}) = P(q_{it} z_{it})$ .

In our previous application of this model, in the linear regression case, we found that estimation of the parameters was made possible by a transformation of the data to deviations from group means which eliminated the person specific constants from the estimator. (See Section 11.4.1.) Save for the special case discussed later, that will not be possible here, so that if one desires to estimate the parameters of this model, it will be necessary actually to compute the possibly huge number of constant terms at the same time. This has been widely viewed as a practical obstacle to estimation of this model because of the need to invert a potentially large second derivatives matrix, but this is a misconception. [See, e.g., Maddala (1987), p. 317.] The method for estimation of nonlinear fixed effects models such as the probit and logit models is detailed in Section 14.9.6.d.

The problems with the fixed effects estimator are statistical, not practical. The estimator relies on  $T_i$  increasing for the constant terms to be consistent—in essence, each  $\alpha_i$  is estimated with  $T_i$  observations. But, in this setting, not only is  $T_i$  fixed, it is likely to be quite small. As such, the estimators of the constant terms are not consistent (not because they converge to something other than what they are trying to estimate, but because they do not converge at all). The estimator of  $\boldsymbol{\beta}$  is a function of the estimators of  $\alpha$ , which means that the MLE of  $\boldsymbol{\beta}$  is not consistent either. This is the incidental parameters problem. [See Neyman and Scott (1948) and Lancaster (2000).] There is, as well, a small sample (small  $T_i$ ) bias in the estimators. How serious this bias is remains a question in the literature. Two pieces of received wisdom are Hsiao's (1986) results for a binary logit model [with additional results in Abrevaya (1997)] and Heckman and MacCurdy's (1980) results for the probit model. Hsiao found that for  $T_i = 2$ , the bias in the MLE of  $\boldsymbol{\beta}$  is 100 percent, which is extremely pessimistic. Heckman and MacCurdy

## 722 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

found in a Monte Carlo study that in samples of  $n = 100$  and  $T = 8$ , the bias appeared to be on the order of 10 percent, which is substantive, but certainly less severe than Hsiao's results suggest. No other theoretical results have been shown for other models, although in very few cases, it can be shown that there is no incidental parameters problem. (The Poisson model mentioned in Chapter 14 is one of these special cases.) The fixed effects approach does have some appeal in that it does not require an assumption of orthogonality of the independent variables and the heterogeneity. An ongoing pursuit in the literature is concerned with the severity of the tradeoff of this virtue against the incidental parameters problem. Some commentary on this issue appears in Arellano (2001). Results of our own investigation appear in Section 15.5.2 and Greene (2004).

### 17.4.4 A CONDITIONAL FIXED EFFECTS ESTIMATOR

Why does the incidental parameters problem arise here and not in the linear regression model?<sup>29</sup> Recall that estimation in the regression model was based on the deviations from group means, not the original data as it is here. The result we exploited there was that although  $f(y_{it} | \mathbf{X}_i)$  is a function of  $\alpha_i$ ,  $f(y_{it} | \mathbf{X}_i, \bar{y}_i)$  is not a function of  $\alpha_i$ , and we used the latter in estimation of  $\beta$ . In that setting,  $\bar{y}_i$  is a **minimal sufficient statistic** for  $\alpha_i$ . Sufficient statistics are available for a few distributions that we will examine, but not for the probit model. They are available for the logit model, as we now examine.

A fixed effects binary logit model is

$$\text{Prob}(y_{it} = 1 | \mathbf{x}_{it}) = \frac{e^{\alpha_i + \mathbf{x}'_{it}\beta}}{1 + e^{\alpha_i + \mathbf{x}'_{it}\beta}}.$$

The unconditional likelihood for the  $nT$  independent observations is

$$L = \prod_i \prod_t (F_{it})^{y_{it}} (1 - F_{it})^{1-y_{it}}.$$

Chamberlain (1980) [following Rasch (1960) and Andersen (1970)] observed that the **conditional likelihood function**,

$$L^c = \prod_{i=1}^n \text{Prob} \left( Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \dots, Y_{iT_i} = y_{iT_i} \middle| \sum_{t=1}^{T_i} y_{it} \right),$$

is free of the incidental parameters,  $\alpha_i$ . The joint likelihood for each set of  $T_i$  observations conditioned on the number of ones in the set is

$$\begin{aligned} & \text{Prob} \left( Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \dots, Y_{iT_i} = y_{iT_i} \middle| \sum_{t=1}^{T_i} y_{it}, \text{data} \right) \\ &= \frac{\exp \left( \sum_{t=1}^{T_i} y_{it} \mathbf{x}'_{it} \beta \right)}{\sum_{\sum_d y_{id} = S_i} \exp \left( \sum_{t=1}^{T_i} d_{it} \mathbf{x}'_{it} \beta \right)}. \end{aligned} \tag{17-46}$$

<sup>29</sup>The incidental parameters problem does show up in ML estimation of the FE linear model, where Neyman and Scott (1948) discovered it, in estimation of  $\sigma_\varepsilon^2$ . The MLE of  $\sigma_\varepsilon^2$  is  $\mathbf{e}'\mathbf{e}/nT$ , which converges to  $[(T-1)/T]\sigma_\varepsilon^2 < \sigma_\varepsilon^2$ .

The function in the denominator is summed over the set of all  $\binom{T_i}{S_i}$  different sequences of  $T_i$  zeros and ones that have the same sum as  $S_i = \sum_{t=1}^{T_i} y_{it}$ .<sup>30</sup>

Consider the example of  $T_i = 2$ . The unconditional likelihood is

$$L = \prod_i \text{Prob}(Y_{i1} = y_{i1}) \text{Prob}(Y_{i2} = y_{i2}).$$

For each pair of observations, we have these possibilities:

1.  $y_{i1} = 0$  and  $y_{i2} = 0$ .  $\text{Prob}(0, 0 | \text{sum} = 0) = 1$ .
2.  $y_{i1} = 1$  and  $y_{i2} = 1$ .  $\text{Prob}(1, 1 | \text{sum} = 2) = 1$ .

The  $i$ th term in  $L^c$  for either of these is just one, so they contribute nothing to the conditional likelihood function.<sup>31</sup> When we take logs, these terms (and these observations) will drop out. But suppose that  $y_{i1} = 0$  and  $y_{i2} = 1$ . Then

$$3. \quad \text{Prob}(0, 1 | \text{sum} = 1) = \frac{\text{Prob}(0, 1 \text{ and sum} = 1)}{\text{Prob}(\text{sum} = 1)} = \frac{\text{Prob}(0, 1)}{\text{Prob}(0, 1) + \text{Prob}(1, 0)}.$$

Therefore, for this pair of observations, the conditional probability is

$$\frac{\frac{1}{1 + e^{\alpha_i + \mathbf{x}'_{i1}\beta}} \frac{e^{\alpha_i + \mathbf{x}'_{i2}\beta}}{1 + e^{\alpha_i + \mathbf{x}'_{i2}\beta}}}{\frac{1}{1 + e^{\alpha_i + \mathbf{x}'_{i1}\beta}} \frac{e^{\alpha_i + \mathbf{x}'_{i1}\beta}}{1 + e^{\alpha_i + \mathbf{x}'_{i1}\beta}} + \frac{1}{1 + e^{\alpha_i + \mathbf{x}'_{i1}\beta}} \frac{1}{1 + e^{\alpha_i + \mathbf{x}'_{i2}\beta}}} = \frac{e^{\mathbf{x}'_{i2}\beta}}{e^{\mathbf{x}'_{i1}\beta} + e^{\mathbf{x}'_{i2}\beta}}.$$

By conditioning on the sum of the two observations, we have removed the heterogeneity. Therefore, we can construct the conditional likelihood function as the product of these terms for the pairs of observations for which the two observations are (0, 1). Pairs of observations with one and zero are included analogously. The product of the terms such as the preceding, for those observation sets for which the sum is not zero or  $T_i$ , constitutes the conditional likelihood. Maximization of the resulting function is straightforward and may be done by conventional methods.

As in the linear regression model, it is of some interest to test whether there is indeed heterogeneity. With homogeneity ( $\alpha_i = \alpha$ ), there is no unusual problem, and the model can be estimated, as usual, as a logit model. It is not possible to test the hypothesis using the likelihood ratio test, however, because the two likelihoods are not comparable. (The conditional likelihood is based on a restricted data set.) None of the usual tests of restrictions can be used because the individual effects are never actually estimated.<sup>32</sup> Hausman's (1978) specification test is a natural one to use here,

<sup>30</sup>The enumeration of all these computations stands to be quite a burden—see Arellano (2000, p. 47) or Baltagi (2005, p. 235). In fact, using a recursion suggested by Kralio and Pike (1984), the computation even with 100 is routine.

<sup>31</sup>Recall that in the probit model when we encounter this situation, the individual constant term could not be estimated and the group was removed from the sample. The same effect is at work here.

<sup>32</sup>This produces a difficulty for this estimator that is shared by the semiparametric estimators discussed in the next section. Because the fixed effects are not estimated, it is not possible to compute probabilities or marginal effects with these estimated coefficients, and it is a bit ambiguous what one can do with the results of the computations. The brute force estimator that actually computes the individual effects might be preferable.

## 724 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

however. Under the null hypothesis of homogeneity, both Chamberlain's conditional maximum likelihood estimator (CMLE) and the usual maximum likelihood estimator are consistent, but Chamberlain's is inefficient. (It fails to use the information that  $\alpha_i = \alpha$ , and it may not use all the data.) Under the alternative hypothesis, the unconditional maximum likelihood estimator is inconsistent,<sup>33</sup> whereas Chamberlain's estimator is consistent and efficient. The Hausman test can be based on the chi-squared statistic

$$\chi^2 = (\hat{\beta}_{\text{CML}} - \hat{\beta}_{\text{ML}})'(\text{Var}[\text{CML}] - \text{Var}[\text{ML}])^{-1}(\hat{\beta}_{\text{CML}} - \hat{\beta}_{\text{ML}}). \quad (17-47)$$

The estimated covariance matrices are those computed for the two maximum likelihood estimators. For the unconditional maximum likelihood estimator, the row and column corresponding to the constant term are dropped. A large value will cast doubt on the hypothesis of homogeneity. (There are  $K$  degrees of freedom for the test.) It is possible that the covariance matrix for the maximum likelihood estimator will be larger than that for the conditional maximum likelihood estimator. If so, then the difference matrix in brackets is assumed to be a zero matrix, and the chi-squared statistic is therefore zero.

### **Example 17.1** Binary Choice Models for Panel Data

In Example 17.1 we fit a pooled binary logit model  $y = 1(DocVis > 0)$  using the German health care utilization data examined in appendix Table F7.1. The model is

$$\begin{aligned} \text{Prob}(DocVis_{it} > 0) = & \Lambda(\beta_1 + \beta_2 \text{Age}_{it} + \beta_3 \text{Income}_{it} + \beta_4 \text{Kids}_{it} \\ & + \beta_5 \text{Education}_{it} + \beta_6 \text{Married}_{it}). \end{aligned}$$

No account of the panel nature of the data set was taken in that exercise. The sample contains a total of 27,326 observations on 7,293 families with  $T_i$  dispersed from one to seven. Table 17.8 lists estimates of parameter estimates and estimated standard errors for probit and logit random and fixed effects models. There is a surprising amount of variation across the estimators. The coefficients are in bold to facilitate reading the table. It is generally difficult to compare across the estimators. The three estimators would be expected to produce very different estimates in any of the three specifications—recall, for example, the pooled estimator is inconsistent in either the fixed or random effects cases. The logit results include two fixed effects estimators. The line market “U” is the unconditional (inconsistent) estimator. The one marked “C” is Chamberlain's consistent estimator. Note for all three fixed effects estimators, it is necessary to drop from the sample any groups that have  $DocVis_{it}$  equal to zero or one for every period. There were 3,046 such groups, which is about 42 percent of the sample. We also computed the probit random effects model in two ways, first by using the Butler and Moffitt method, then by using maximum simulated likelihood estimation. In this case, the estimators are very similar, as might be expected. The estimated correlation coefficient,  $\rho$ , is computed as  $\sigma_u^2 / (\sigma_\varepsilon^2 + \sigma_u^2)$ . For the probit model,  $\sigma_\varepsilon^2 = 1$ . The MSL estimator computes  $s_u = 0.9088376$ , from which we obtained  $\rho$ . The estimated partial effects for the models are shown in Table 17.9. The average of the fixed effects constant terms is used to obtain a constant term for the fixed effects case. Once again there is a considerable amount of variation across the different estimators. On average, the fixed effects models tend to produce much larger values than the pooled or random effects models.

<sup>33</sup>Hsiao (2003) derives the result explicitly for some particular cases.

**TABLE 17.8** Estimated Parameters for Panel Data Binary Choice Models

<i>Model</i>	<i>Estimate</i>	<i>Variable</i>					
		<i>ln L</i>	<i>Constant</i>	<i>Age</i>	<i>Income</i>	<i>Kids</i>	<i>Education</i>
Logit Pooled	$\beta$ St.Err. Rob.SE <sup>c</sup>	-17673.10 0.091135 0.12827	<b>0.25112</b> 0.0012852 0.0017429	<b>0.020709</b> 0.075064 0.091546	<b>-0.18592</b> 0.029537 0.038313	<b>-0.22947</b> 0.005646 0.008075	<b>-0.045587</b> 0.033286 0.045314
Logit.R.E. $\rho = 0.41607$	$\beta$ St.Err.	-15261.90 0.17764	<b>-0.13460</b> 0.0024659	<b>0.03267</b> 0.11866	<b>0.021914</b> 0.047738	<b>-0.21598</b> 0.011322	<b>-0.063578</b> 0.056282
Logit F.E.(U) <sup>a</sup>	$\beta$ St.Err.	-9458.64 0.0072548	<b>0.10475</b> 0.17829	<b>-0.066973</b> 0.074399	<b>-0.088407</b> 0.066749	<b>-0.11671</b> 0.011322	<b>-0.057318</b> 0.10609
Logit E.E.(C) <sup>b</sup>	$\beta$ St.Err.	-6312.57 .08384	<b>-0.06521</b> (.006382)	<b>-0.07802</b> (.15793)	<b>-0.12179</b> (.066186)	<b>-0.04897</b> (.05466)	
Probit Pooled	$\beta$ St.Err. Rob.SE <sup>e</sup>	-17670.94 0.056516 0.079591	<b>0.15500</b> 0.0007903 0.0010739	<b>0.012835</b> 0.046329 0.056543	<b>-0.11643</b> 0.018218 0.023614	<b>-0.14118</b> 0.003503 0.005014	<b>-0.028115</b> 0.020462 0.027904
Probit.RE <sup>c</sup> $\rho = 0.44789$	$\beta$ St.Err.	-16273.96 0.096354	<b>0.034113</b> 0.0013189	<b>0.020143</b> 0.066672	<b>-0.003176</b> 0.027043	<b>-0.15379</b> 0.006289	<b>-0.033694</b> 0.031347
Probit.RE <sup>d</sup> $\rho = 0.44799$	$\beta$ St.Err.	-16279.97 0.063229	<b>0.033290</b> 0.0009013	<b>0.020078</b> 0.052012	<b>-0.002973</b> 0.020286	<b>-0.153579</b> 0.003931	<b>-0.033489</b> 0.022771
Probit F.E.(U)	$\beta$ St.Err.	-9453.71 0.0043219	<b>0.062528</b> 0.10745	<b>-0.034328</b> 0.044559	<b>-0.048270</b> 0.040731	<b>-0.072189</b> 0.063627	<b>-0.032774</b> 0.063627

<sup>a</sup>Unconditional fixed effects estimator<sup>b</sup>Conditional fixed effects estimator<sup>c</sup>Butler and Moffitt estimator<sup>d</sup>Maximum simulated likelihood estimator<sup>e</sup>Robust, "cluster" corrected standard error

**726 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**
**TABLE 17.9** Estimated Partial Effects for Panel Data Binary Choice Models

<i>Model</i>	<i>Age</i>	<i>Income</i>	<i>Kids</i>	<i>Education</i>	<i>Married</i>
Logit, P <sup>a</sup>	0.0048133	-0.043213	-0.053598	-0.010596	0.019936
Logit: RE,Q <sup>b</sup>	0.0064213	0.0035835	-0.035448	-0.010397	0.0041049
Logit: F,U <sup>c</sup>	0.024871	-0.014477	-0.020991	-0.027711	-0.013609
Logit: FC <sup>d</sup>	0.0072991	-0.0043387	-0.0066967	-0.0078206	-0.0044842
Probit, P <sup>a</sup>	0.0048374	-0.043883	-0.053414	-0.010597	0.019783
Probit RE,Q <sup>b</sup>	0.0056049	-0.0008836	-0.042792	-0.0093756	0.0045426
Probit:RE,S <sup>e</sup>	0.0071455	-0.0010582	-0.054655	-0.011917	0.0059878
Probit: F,U <sup>c</sup>	0.023958	-0.013152	-0.018495	-0.027659	-0.012557

<sup>a</sup>Pooled estimator<sup>b</sup>Butler and Moffitt estimator<sup>c</sup>Unconditional fixed effects estimator<sup>d</sup>Conditional fixed effects estimator<sup>e</sup>Maximum simulated likelihood estimator
**Example 17.12 Fixed Effects Logit Models: Magazine Prices Revisited**

The fixed effects model does have some appeal, but the incidental parameters problem is a significant shortcoming of the unconditional probit and logit estimators. The conditional MLE for the fixed effects logit model is a fairly common approach. A widely cited application of the model is Cecchetti's (1986) analysis of changes in newsstand prices of magazines. Cecchetti's model was

$$\text{Prob}(\text{Price change in year } i \text{ of magazine } t) = \Lambda(\alpha_j + \mathbf{x}'_{it}\beta),$$

where the variables in  $\mathbf{x}_{it}$  are (1) time since last price change, (2) inflation since last change, (3) previous fixed price change, (4) current inflation, (5) industry sales growth, and (6) sales volatility. The fixed effect in the model is indexed "j" rather than "i" as it is defined as a three-year interval for magazine  $i$ . Thus, a magazine that had been on the newstands for nine years would have three constants, not just one. In addition to estimating several specifications of the price change model, Cecchetti used the Hausman test in (17-47) to test for the existence of the common effects. Some of Cecchetti's results appear in Table 17.10.

Willis (2006) argued that Cecchetti's estimates were inconsistent and the Hausman test is invalid because right-hand-side variables (1), (2), and (6) are all functions of lagged dependent variables. This state dependence invalidates the use of the sum of the observations for the group as a sufficient statistic in the Chamberlain estimator and the Hausman tests. He proposes, instead, a method suggested by Heckman and Singer (1984b) to incorporate the unobserved heterogeneity in the *unconditional* likelihood function. The Heckman and Singer model can be formulated as a latent class model (see Sections 14.10 and 17.4.7) in which the classes are defined by different constant terms—the remaining parameters in the model

**TABLE 17.10** Models for Magazine Price Changes (standard errors in parentheses)

	<i>Pooled</i>	<i>Unconditional FE</i>	<i>Conditional FE Cecchetti</i>	<i>Conditional FE Willis</i>	<i>Heckman and Singer</i>
$\beta_1$	-1.10 (0.03)	-0.07 (0.03)	1.12 (3.66)	1.02 (0.28)	-0.09 (0.04)
$\beta_2$	6.93 (1.12)	8.83 (1.25)	11.57 (1.68)	19.20 (7.51)	8.23 (1.53)
$\beta_5$	-0.36 (0.98)	-1.14 (1.06)	5.85 (1.76)	7.60 (3.46)	-0.13 (1.14)
Constant 1	-1.90 (0.14)				-1.94 (0.20)
Constant 2					-29.15 (1.1e11)
In $L$	-500.45	-473.18	-82.91	-83.72	-499.65
Sample size	1026	1026		543	1026

are constrained to be equal across classes. Willis fit the Heckman and Singer model with two classes to a restricted version of Cecchetti's model using variables (1), (2), and (5). The results in Table 17.10 show some of the results from Willis's Table I. (Willis reports that he could not reproduce Cecchetti's results—the ones in Cecchetti's second column would be the counterparts—because of some missing values. In fact, Willis's estimates are quite far from Cecchetti's results, so it will be difficult to compare them. Both are reported here.)

The two "mass points" reported by Willis are shown in Table 17.10. He reports that these two values ( $-1.94$  and  $-29.15$ ) correspond to class probabilities of  $0.88$  and  $0.12$ , though it is difficult to make the translation based on the reported values. He does note that the change in the log-likelihood in going from one mass point (pooled logit model) to two is marginal, only from  $-500.45$  to  $-499.65$ . There is another anomaly in the results that is consistent with this finding. The reported standard error for the second "mass point" is  $1.1 \times 10^{11}$ , or essentially  $+\infty$ . The finding is consistent with overfitting the latent class model. The results suggest that the better model is a one-class (pooled) model.

#### 17.4.5 MUNDLAK'S APPROACH, VARIABLE ADDITION AND BIAS REDUCTION

Thus far, both the fixed effects (FE) and the random effects (RE) specifications present problems for modeling binary choice with panel data. The MLE of the FE model is inconsistent even when the model is properly specified—this is the incidental parameters problem. (And, like the linear model, the FE probit and logit models do not allow time-invariant regressors.) The random effects specification requires a strong, often unreasonable, assumption that the effects and the regressors are uncorrelated. Of the two, the FE model is the more appealing, though with modern longitudinal data sets with many demographics, the problem of time-invariant variables would seem to be compelling. This would seem to recommend the conditional estimator in Section 17.4.4, save for yet another complication. With no estimates of the constant terms, neither probabilities nor partial effects can be computed with the results. We are left making inferences about ratios of coefficient. Two approaches have been suggested for finding a middle ground: Mundlak's (1978) approach that involves projecting the effects on the group means of the time-varying variables and recent developments such as Fernandez-Val's approach that involves correcting the bias in the FE MLE.

The Mundlak (1978) [and Chamberlain (1984) and Wooldridge, e.g., (2002a)] approach augments (17-44) as follows:

$$\begin{aligned} y_{it}^* &= \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it} \\ \text{Prob}(y_{it} = 1 | \mathbf{x}_{it}) &= F(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta}) \\ \alpha_i &= \alpha + \bar{\mathbf{x}}'_i\boldsymbol{\delta} + u_i, \end{aligned}$$

where we have used  $\bar{\mathbf{x}}_i$  generically for the group means of the time varying variables in  $\mathbf{x}_{it}$ . The reduced form of the model is

$$\text{Prob}(y_{it} = 1 | \mathbf{x}_{it}) = F(\alpha + \bar{\mathbf{x}}'_i\boldsymbol{\delta} + \mathbf{x}'_{it}\boldsymbol{\beta} + u_i).$$

(Wooldridge and Chamberlain also suggest using all years of  $\mathbf{x}_{it}$  rather than the group means. This raises a problem in unbalanced panels, however. We will ignore this possibility.) The projection of  $\alpha_i$  on  $\bar{\mathbf{x}}_i$  produces a random effects formulation. As in the linear model (see Section 11.5.6), it also suggests a means of testing for fixed vs. random effects. Since  $\boldsymbol{\delta} = \mathbf{0}$  produces the pure random effects model, a joint Wald test of the null hypothesis that  $\boldsymbol{\delta}$  equals zero can be used.

**728 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**
**TABLE 17.11** Estimated Random Effects Models

	<i>Constant</i>	<i>Age</i>	<i>Income</i>	<i>Kids</i>	<i>Education</i>	<i>Married</i>
Random Effects	0.03411 (0.09635)	0.02014 (0.00132)	-0.00318 (0.06667)	-0.15379 (0.02704)	-0.03369 (0.00629)	0.01633 (0.03135)
Augmented Model	0.37485 (0.10501)	0.05035 (0.00357)	-0.03057 (0.09318)	-0.04202 (0.03751)	-0.05449 (0.03307)	-0.02645 (0.05180)
Means		-0.03659 (0.00384)	-0.35065 (0.13984)	-0.22509 (0.05499)	0.02387 (0.03374)	0.14668 (0.06607)

**Example 17.13 Panel Data Random Effects Estimators**

Example 17.11 presents several estimators of panel data estimators for the probit and logit models. Pooled, random effects and fixed effects estimates are given for the probit model

$$\begin{aligned} \text{Prob}(DocVis_{it} > 0) = \Phi(\beta_1 + \beta_2 \text{Age}_{it} + \beta_3 \text{Income}_{it} + \beta_4 \text{Kids}_{it} \\ + \beta_5 \text{Education}_{it} + \beta_6 \text{Married}_{it}). \end{aligned}$$

We continue that analysis here by considering Mundlak's approach to the common effects model. Table 17.11 presents the random effects model from earlier, and the augmented estimator that contains the group means of the variables, all of which are time varying. The addition of the group means to the regression brings large changes to the estimates of the parameters, which might suggest the appropriateness of the fixed effects model. A formal test is carried by computing a Wald statistic for the null hypothesis that the last five coefficients in the augmented model equal zero. The chi-squared statistic equals 113.282 with five degrees of freedom. The critical value from the chi-squared table for 95 percent significance is 11.07, so the hypothesis that  $\delta$  equals zero, that is, the hypothesis of the random effects model (restrictions), is rejected. The two log likelihoods are -16273.96 for the REM and -16222.06 for the augmented REM. The LR statistic would be twice the difference, or 103.8. This produces the same conclusion. The FEM appears to be the preferred model.

A series of recent studies has sought to maintain the fixed effects specification while correcting the bias due to the incidental parameters problem. There are two broad approaches. Hahn and Kuersteiner (2004), Hahn and Newey (2005), and Fernandez-Val (2009) have developed an approximate, "large  $T$ " result for  $\text{plim}(\hat{\beta}_{FEMALE} - \beta)$  that produces a direct correction to the estimator, itself. Fernandez-Val (2009) develops corrections for the estimated constant terms as well. Arellano and Hahn (2006, 2007) propose a modification of the log-likelihood function with, in turn, different first-order estimation equations, that produces an approximately unbiased estimator of  $\beta$ . In a similar fashion to the second of these approaches, Carro (2007) modifies the first-order conditions (estimating equations) from the original log-likelihood function, once again to produce an approximately unbiased estimator of  $\beta$ . (In general, given the overall approach of using a  $T$  approximation, the payoff to these estimators is to reduce the bias of the FEMALE from  $O(1/T)$  to  $O(1/T^2)$ , which is a considerable reduction.) These estimators are not yet in widespread use. The received evidence suggests that in the simple case we are considering here, the incidental parameters problem is a secondary concern when  $T$  reaches say 10 or so. For some modern public use data sets, such as the BHPS or GSOEP which are beyond their 15th wave, the incidental parameters problem may not be too severe. However, most of the studies mentioned above are concerned with dynamic models (see Section 17.4.6), where the problem is possibly more severe than in the static case. Research in this area is ongoing.

#### 17.4.6 DYNAMIC BINARY CHOICE MODELS

A random or fixed effects model that explicitly allows for lagged effects would be

$$y_{it} = \mathbf{1}'(\mathbf{x}'_{it}\beta + \alpha_i + \gamma y_{i,t-1} + \varepsilon_{it}) > 0.$$

Lagged effects, or **persistence**, in a binary choice setting can arise from three sources, serial correlation in  $\varepsilon_{it}$ , the **heterogeneity**,  $\alpha_i$ , or true **state dependence** through the term  $\gamma y_{i,t-1}$ . Chiappori (1998) [and see Arellano (2001)] suggests an application to the French automobile insurance market in which the incentives built into the pricing system are such that having an accident in one period should lower the probability of having one in the next (state dependence), but, some drivers remain more likely to have accidents than others in every period, which would reflect the heterogeneity instead. State dependence is likely to be particularly important in the typical panel which has only a few observations for each individual. Heckman (1981a) examined this issue at length. Among his findings were that the somewhat muted small sample bias in fixed effects models with  $T = 8$  was made much worse when there was state dependence. A related problem is that with a relatively short panel, the **initial conditions**,  $y_{i0}$ , have a crucial impact on the entire path of outcomes. Modeling dynamic effects and initial conditions in binary choice models is more complex than in the linear model, and by comparison there are relatively fewer firm results in the applied literature.<sup>34</sup>

The correlation between  $\alpha_i$  and  $y_{i,t-1}$  in the dynamic binary choice model makes  $y_{i,t-1}$  endogenous. Thus, the estimators we have examined thus far will not be consistent. Two familiar alternative approaches that have appeared in recent applications are due to Heckman (1981) and Wooldridge (2005), both of which build on the random effects specification. Heckman's approach provides a separate equation for the initial condition,

$$\text{Prob}(y_{i1} = 1 | \mathbf{x}_{i1}, \mathbf{z}_i, \alpha_i) = \Phi(\mathbf{x}'_{i1}\delta + \mathbf{z}'_i\tau + \theta\alpha_i)$$

$$\text{Prob}(y_{it} = 1 | \mathbf{x}_{it}, y_{i,t-1}, \alpha_i) = \Phi(\mathbf{x}'_{it}\beta + \gamma y_{i,t-1} + \alpha_i), t = 2, \dots, T_i,$$

where  $\mathbf{z}_i$  is a set of “instruments” observed at the first period that are not contained in  $\mathbf{x}_{it}$ . The conditional log-likelihood is

$$\begin{aligned} \ln L | \alpha &= \sum_{i=1}^n \ln \left\{ \Phi[(2y_{i1} - 1)(\mathbf{x}'_{i1}\delta + \mathbf{z}'_i\tau + \theta\alpha_i)] \prod_{t=2}^{T_i} \Phi[(2y_{it} - 1)(\mathbf{x}'_{it}\beta + \gamma y_{i,t-1} + \alpha_i)] \right\} \\ &= \sum_{i=1}^n \ln L_i | \alpha_i. \end{aligned}$$

We now adopt the random effects approach and further assume that  $\alpha_i$  is normally distributed with mean zero and variance  $\sigma_\alpha^2$ . The random effects log-likelihood function can be maximized with respect to  $(\delta, \tau, \theta, \beta, \gamma, \sigma_\alpha)$  using either the Butler and Moffitt

<sup>34</sup>A survey of some of these results is given by Hsiao (2003). Most of Hsiao (2003) is devoted to the linear regression model. A number of studies specifically focused on discrete choice models and panel data have appeared recently, including Beck, Epstein, Jackman and O'Halloran (2001), Arellano (2001) and Greene (2001). Vella and Verbeek (1998) provide an application to the joint determination of wages and union membership. Other important references are Aguirregabiria and Mira (2010), Carro (2007), and Feenstra and Val (2009). Stewart (2006) and Arulampalam and Stewart (2007) provide several results for practitioners.

## 730 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

quadrature method or the maximum simulated likelihood method described in Section 17.4.2. Stewart and Arulampalam (2007) suggest a useful shortcut for formulating the Heckman model. Let  $D_{it} = 1$  in period 1 and 0 in every other period and let  $C_{it} = 1 - D_{it}$ . Then, the two parts may be combined in

$$\ln L | \alpha = \sum_{i=1}^n \ln \prod_{t=1}^{T_i} \{ \Phi [ (2y_{it} - 1) \langle C_{it}(\mathbf{x}'_{it}\beta + \gamma y_{i,t-1}) + D_{it}(\mathbf{x}'_{it}\delta + \mathbf{z}'_i\tau) + (1 + \lambda D_{it})\alpha_i \rangle ] \}.$$

In this form, the model can be viewed as a random parameters (random constant term) model in which there is heteroscedasticity in the random part of the constant term.

Wooldridge's approach builds on the Mundlak device of the previous section. Starting from the same point, he suggests a model for the random effect conditioned on the initial value. Thus,

$$\alpha_i | y_{i1}, \mathbf{z}_i \sim N[\alpha_0 + \eta y_{i1} + \mathbf{z}'_i\tau, \sigma_\alpha^2].$$

Assembling the parts, Wooldridge's model is a bit simpler than Heckman's;

$$\begin{aligned} \text{Prob}(y_{it} = 1 | \mathbf{x}_{it}, y_{i1}, u_i) \\ = \Phi[(2y_{it} - 1)(\alpha_0 + \mathbf{x}'_{it}\beta + \gamma y_{i,t-1} + \eta y_{i1} + \mathbf{z}'_i\tau + u_i)], t = 2, \dots, T_i. \end{aligned}$$

Much of the contemporary literature has focused on methods of avoiding the strong parametric assumptions of the probit and logit models. Manski (1987) and Honore and Kyriazidou (2000) show that Manski's (1986) maximum score estimator can be applied to the differences of unequal pairs of observations in a two-period panel with fixed effects. However, the limitations of the maximum score estimator have motivated research on other approaches. An extension of lagged effects to a parametric model is Chamberlain (1985), Jones and Landwehr (1988), and Magnac (1997), who added state dependence to Chamberlain's fixed effects logit estimator. Unfortunately, once the identification issues are settled, the model is only operational if there are no other exogenous variables in it, which limits its usefulness for practical application. Lewbel (2000) has extended his fixed effects estimator to dynamic models as well.

Dong and Lewbel (2010) have extended Lewbel's "special regressor" method to dynamic binary choice models and have devised an estimator based on an IV linear regression. Honore and Kyriazidou (2000) have combined the logic of the **conditional logit model** and Manski's maximum score estimator. They specify

$$\begin{aligned} \text{Prob}(y_{i0} = 1 | \mathbf{x}_i, \alpha_i) &= p_0(\mathbf{x}_i, \alpha_i) \quad \text{where } \mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT}), \\ \text{Prob}(y_{it} = 1 | \mathbf{x}_i, \alpha_i, y_{i0}, y_{i1}, \dots, y_{i,t-1}) &= F(\mathbf{x}'_{it}\beta + \alpha_i + \gamma y_{i,t-1}) \quad t = 1, \dots, T. \end{aligned}$$

The analysis assumes a single regressor and focuses on the case of  $T = 3$ . The resulting estimator resembles Chamberlain's but relies on observations for which  $\mathbf{x}_{it} = \mathbf{x}_{i,t-1}$ , which rules out direct time effects as well as, for practical purposes, any continuous variable. The restriction to a single regressor limits the generality of the technique as well. The need for observations with equal values of  $\mathbf{x}_{it}$  is a considerable restriction, and the authors propose a kernel density estimator for the difference,  $\mathbf{x}_{it} - \mathbf{x}_{i,t-1}$ , instead which does relax that restriction a bit. The end result is an estimator that converges (they conjecture) but to a nonnormal distribution and at a rate slower than  $n^{-1/3}$ .

Semiparametric estimators for dynamic models at this point in the development are still primarily of theoretical interest. Models that extend the parametric formulations to

include state dependence have a much longer history, including Heckman (1978, 1981a, 1981b), Heckman and MacCurdy (1980), Jakubson (1988), Keane (1993), and Beck et al. (2001) to name a few.<sup>35</sup> In general, even without heterogeneity, dynamic models ultimately involve modeling the joint outcome  $(y_{i0}, \dots, y_{iT})$ , which necessitates some treatment involving multivariate integration. Example 17.14 describes an application. Stewart (2006) provides another.

**Example 17.14 An Intertemporal Labor Force Participation Equation**

Hyslop (1999) presents a model of the labor force participation of married women. The focus of the study is the high degree of persistence in the participation decision. Data used in the study were the years 1979–1985 of the Panel Study of Income Dynamics. A sample of 1,812 continuously married couples were studied. Exogenous variables that appeared in the model were measures of permanent and transitory income and fertility captured in yearly counts of the number of children from 0–2, 3–5, and 6–17 years old. Hyslop's formulation, in general terms, is

$$\begin{aligned} & \text{(initial condition)} \quad y_{i0} = 1(\mathbf{x}'_{i0}\boldsymbol{\beta}_0 + v_{i0} > 0), \\ & \text{(dynamic model)} \quad y_{it} = 1(\mathbf{x}'_{it}\boldsymbol{\beta} + \gamma y_{i,t-1} + \alpha_i + v_{it} > 0) \\ & \text{(heterogeneity correlated with participation)} \quad \alpha_i = \mathbf{z}'_i\boldsymbol{\delta} + \eta_i, \\ & \text{(stochastic specification)} \\ & \quad \eta_i | \mathbf{X}_i \sim N[0, \sigma_\eta^2], \\ & \quad v_{i0} | \mathbf{X}_i \sim N[0, \sigma_0^2], \\ & \quad w_{it} | \mathbf{X}_i \sim N[0, \sigma_w^2], \\ & \quad v_{it} = \rho v_{i,t-1} + w_{it}, \quad \sigma_\eta^2 + \sigma_w^2 = 1. \\ & \quad \text{Corr}[v_{i0}, v_{it}] = \rho^t, \quad t = 1, \dots, T - 1. \end{aligned}$$

The presence of the autocorrelation and state dependence in the model invalidate the simple maximum likelihood procedures we examined earlier. The appropriate likelihood function is constructed by formulating the probabilities as

$$\text{Prob}(y_{i0}, y_{i1}, \dots) = \text{Prob}(y_{i0}) \times \text{Prob}(y_{i1} | y_{i0}) \times \dots \times \text{Prob}(y_{iT} | y_{i,T-1}).$$

This still involves a  $T = 7$  order normal integration, which is approximated in the study using a simulator similar to the GHK simulator discussed in 15.6.2.b. Among Hyslop's results are a comparison of the model fit by the simulator for the multivariate normal probabilities with the same model fit using the maximum simulated likelihood technique described in Section 15.6.

#### 17.4.7 A SEMIPARAMETRIC MODEL FOR INDIVIDUAL HETEROGENEITY

The panel data analysis considered thus far has focused on modeling heterogeneity with the fixed and random effects specifications. Both assume that the heterogeneity is continuously distributed among individuals. The random effects model is fully parametric, requiring a full specification of the likelihood for estimation. The fixed effects model

<sup>35</sup>Beck et al. (2001) is a bit different from the others mentioned in that in their study of “state failure,” they observe a large sample of countries (147) observed over a fairly large number of years, 40. As such, they are able to formulate their models in a way that makes the asymptotics with respect to  $T$  appropriate. They can analyze the data essentially in a time-series framework. Sepanski (2000) is another application that combines state dependence and the random coefficient specification of Akin, Guilkey, and Sickles (1979).

**732 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**

is essentially semiparametric. It requires no specific distributional assumption, however, it does require that the realizations of the latent heterogeneity be treated as parameters, either estimated in the unconditional fixed effects estimator or conditioned out of the likelihood function when possible. As noted in the preceding example, Heckman and Singer's (1984b) model provides a less stringent model specification based on a discrete distribution of the latent heterogeneity. A straightforward method of implementing their model is to cast it as a latent class model in which the classes are distinguished by different constant terms and the associated probabilities. The class probabilities are treated as parameters to be estimated with the model parameters.

**Example 17.15 Semiparametric Models of Heterogeneity**

We have extended the random effects and fixed effects logit models in Example 17.11 by fitting the Heckman and Singer (1984b) model. Table 17.12 shows the specification search and the results under different specifications. The first column of results shows the estimated fixed effects model from Example 17.11. The conditional estimates are shown in parentheses. Of the 7,293 groups in the sample, 3,056 are not used in estimation of the fixed effects models because the sum of  $Doctor_{it}$  is either 0 or  $T_i$  for the group. The mean and standard deviation of the estimated underlying heterogeneity distribution are computed using the estimates of  $\alpha_i$  for the remaining 4,237 groups. The remaining five columns in the table show the results for different numbers of latent classes in the Heckman and Singer model. The listed constant terms are the "mass points" of the underlying distributions. The associated class probabilities are shown in parentheses under them. The mean and standard deviation are derived from the

**TABLE 17.12** Estimated Heterogeneity Models

	<i>Fixed Effect</i>	<i>Number of Classes</i>				
		<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>
$\beta_1$	0.10475 (0.084760)	0.020708	0.030325	0.033684	0.034083	0.034159
$\beta_2$	-0.060973 (-0.050383)	-0.18592	0.025550	-0.0058013	-0.0063516	-0.013627
$\beta_3$	-0.088407 (-0.077764)	-0.22947	-0.24708	-0.26388	-0.26590	-0.26626
$\beta_4$	-0.11671 (-0.090816)	-0.045588	-0.050924	-0.058022	-0.059751	-0.059176
$\beta_5$	-0.057318 (-0.52072)	0.085293	0.042974	0.037944	0.029227	0.030699
$\alpha_1$	-2.62334	0.25111 (1.00000)	0.91764 (0.62681)	1.71669 (0.34838)	1.94536 (0.29309)	2.76670 (0.11633)
$\alpha_2$			-1.47800 (0.37319)	-2.23491 (0.18412)	-1.76371 (0.21714)	1.18323 (0.26468)
$\alpha_3$				-0.28133 (0.46749)	-0.036739 (0.46341)	-1.96750 (0.19573)
$\alpha_4$					-4.03970 (0.026360)	-0.25588 (0.40930)
$\alpha_5$						-6.48191 (0.013960)
<i>Mean</i>	-2.62334	0.00000	0.023613	0.055059	0.063685	0.054705
<i>Std. Dev.</i>	3.13415	0.00000	1.158655	1.40723	1.48707	1.62143
<i>ln L</i>	-9458.638 (-6299.02)	-17673.10	-16353.14	-16278.56	-16276.07	-16275.85
<i>AIC</i>	1.00349	1.29394	1.19748	1.19217	1.19213	1.19226

2- to 5-point discrete distributions shown. It is noteworthy that the mean of the distribution is relatively stable, but the standard deviation rises monotonically. The search for the best model would be based on the AIC. As noted in Section 14.10, using a likelihood ratio test in this context is dubious, as the number of degrees of freedom is ambiguous. Based on the AIC, the four-class model is the preferred specification.

#### 17.4.8 MODELING PARAMETER HETEROGENEITY

In Section 11.11, we examined specifications that extend the underlying heterogeneity to all the parameters of the model. We have considered two approaches. The random parameters, or mixed models discussed in Chapter 15 allow parameters to be distributed continuously across individuals. The latent class model in Section 16.8 specifies a discrete distribution instead. (The Heckman and Singer model in the previous section applies this method to the constant term.) Most of the focus to this point, save for Example 16.8, has been on linear models.

The random effects model can be cast as a model with a random constant term;

$$y_{it}^* = \alpha_i + \mathbf{x}'_{it}\beta + \varepsilon_{it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T_i,$$

$$y_{it} = 1 \quad \text{if } y_{it}^* > 0, \text{ and } 0 \text{ otherwise,}$$

where  $\alpha_i = \alpha + \sigma_u u_i$ . This is simply a reinterpretation of the model we just analyzed. We might, however, now extend this formulation to the full parameter vector. The resulting structure is

$$y_{it}^* = \mathbf{x}'_{it}\beta_i + \varepsilon_{it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T_i,$$

$$y_{it} = 1 \quad \text{if } y_{it}^* > 0, \text{ and } 0 \text{ otherwise,}$$

where  $\beta_i = \beta + \Gamma \mathbf{u}_i$  where  $\Gamma$  is a nonnegative definite diagonal matrix—some of its diagonal elements could be zero for nonrandom parameters. The method of estimation is **maximum simulated likelihood**. The simulated log-likelihood is now

$$\ln L_{\text{Simulated}} = \sum_{i=1}^n \ln \left\{ \frac{1}{R} \sum_{r=1}^R \left[ \prod_{t=1}^{T_i} F[q_{it}(\mathbf{x}'_{it}(\beta + \Gamma \mathbf{u}_{ir}))] \right] \right\}.$$

The simulation now involves  $R$  draws from the multivariate distribution of  $\mathbf{u}$ . Because the draws are uncorrelated— $\Gamma$  is diagonal—this is essentially the same estimation problem as the random effects model considered previously. This model is estimated in Example 17.16. Example 17.16 also presents a similar model that assumes that the distribution of  $\beta_i$  is discrete rather than continuous.

##### **Example 17.16 Parameter Heterogeneity in a Binary Choice Model**

We have extended the logit model for doctor visits from Example 17.15 to allow the parameters to vary randomly across individuals. The random parameters logit model is

$$\text{Prob}(Doctor_{it} = 1) = \Lambda(\beta_{1i} + \beta_{2i} Age_{it} + \beta_{3i} Income_{it} + \beta_{4i} Kids_{it} + \beta_{5i} Educ_{it} + \beta_{6i} Married_{it}),$$

where the two models for the parameter variation we have employed are:

Continuous:  $\beta_{ki} = \beta_k + \sigma_k u_{ki}$ ,  $u_{ki} \sim N[0, 1]$ ,  $k = 1, \dots, 6$ ,  $\text{Cov}[u_{ki}, u_{mj}] = 0$ ,

Discrete:  $\beta_{ki} = \beta_k^1$  with probability  $\pi_1$   
 $\beta_k^2$  with probability  $\pi_2$   
 $\beta_k^3$  with probability  $\pi_3$ .

**734 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**
**TABLE 17.13** Estimated Heterogeneous Parameter Models

<i>Variable</i>	<i>Pooled</i>	<i>Random Parameters</i>		<i>Latent Class</i>		
	<i>Estimate: <math>\beta</math></i>	<i>Estimate: <math>\beta</math></i>	<i>Estimate: <math>\sigma</math></i>	<i>Estimate: <math>\beta</math></i>	<i>Estimate: <math>\beta</math></i>	<i>Estimate: <math>\beta</math></i>
Constant	0.25111 (0.091135)	-0.034964 (0.075533)	0.81651 (0.016542)	0.96605 (0.43757)	-0.18579 (0.23907)	-1.52595 (0.43498)
Age	0.020709 (0.0012852)	0.026306 (0.0011038)	0.025330 (0.0004226)	0.049058 (0.0069455)	0.032248 (0.0031462)	0.019981 (0.0062550)
Income	-0.18592 (0.075064)	-0.0043649 (0.062445)	0.10737 (0.038276)	-0.27917 (0.37149)	-0.068633 (0.16748)	0.45487 (0.31153)
Kids	-0.22947 (0.029537)	-0.17461 (0.024522)	0.55520 (0.023866)	-0.28385 (0.14279)	-0.28336 (0.066404)	-0.11708 (0.12363)
Education	-0.045588 (0.0056465)	-0.040510 (0.0047520)	0.037915 (0.0013416)	-0.025301 (0.027768)	-0.057335 (0.012465)	-0.09385 (0.027965)
Married	0.085293 (0.033286)	0.014618 (0.027417)	0.070696 (0.017362)	-0.10875 (0.17228)	0.025331 (0.075929)	0.23571 (0.14369)
Class	1.00000		1.00000	0.34833	0.46181	0.18986
Prob.	(0.00000)		(0.00000)	(0.038495)	(0.028062)	(0.022335)
ln L	-17673.10		-16271.72			-16265.59

We have chosen a three-class latent class model for the illustration. In an application, one might undertake a systematic search, such as in Example 17.15, to find a preferred specification. Table 17.13 presents the fixed parameter (pooled) logit model and the two random parameters versions. (There are infinite variations on these specifications that one might explore—See Chapter 15 for discussion—we have shown only the simplest to illustrate the models.<sup>36</sup>

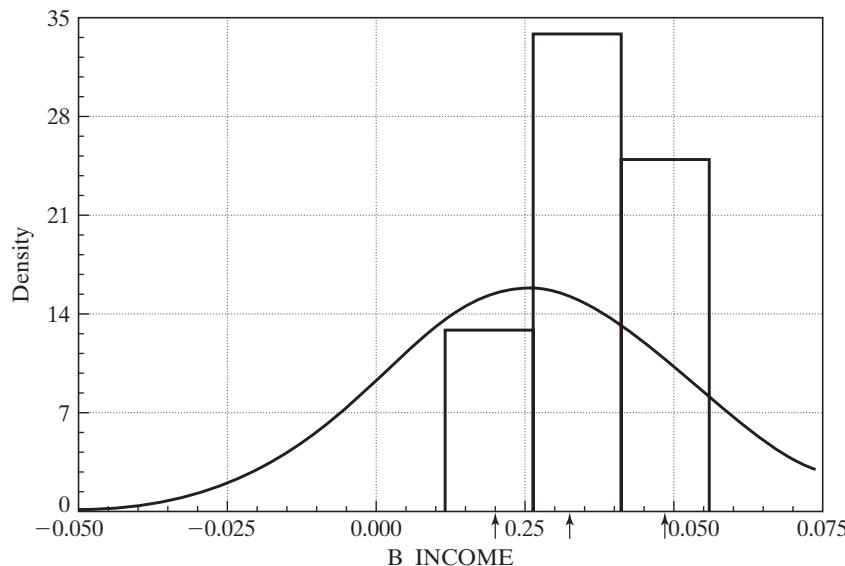
Figure 17.3 shows the implied distribution for the coefficient on age. For the continuous distribution, we have simply plotted the normal density. For the discrete distribution, we first obtained the mean (0.0358) and standard deviation (0.0107). Notice that the distribution is tighter than the estimated continuous normal (mean, 0.026, standard deviation, 0.0253). To suggest the variation of the parameter (purely for purpose of the display, because the distribution is discrete), we placed the mass of the center interval, 0.462, between the midpoints of the intervals between the center mass point and the two extremes. With a width of 0.0145 the density is  $0.461 / 0.0145 = 31.8$ . We used the same interval widths for the outer segments. This range of variation covers about five standard deviations of the distribution.

#### 17.4.9 NONRESPONSE, ATTRITION AND INVERSE PROBABILITY WEIGHTING

Missing observations is a common problem in the analysis of panel data. Nicoletti and Peracchi (2005) suggest several reasons that, for example, panels become unbalanced:

- Demographic events such as death 
- Movement out of the scope of the survey, such as institutionalization or emigration 

<sup>36</sup>We have arrived (once again) at a point where the question of replicability arises. Nonreplicability is an ongoing challenge in empirical work in economics. (See, e.g., Example 17.12.) The problem is particularly acute in analyses that involve simulation such as Monte Carlo studies and random parameter models. In the interest of replicability, we note that the random parameter estimates in Table 17.14 were computed with NLOGIT [Econometric Software (2007)] and are based on 50 Halton draws. We used the first six sequences (prime numbers 2, 3, 5, 7, 11, 13) and discarded the first 10 draws in each sequence.



**FIGURE 17.3** Distributions of Income Coefficient.

- Refusal to respond at subsequent waves
- Absence of the person at the address
- Other types of noncontact

The GSOEP that we (from Riphahn, Wambach, and Million (2003)) have used in many examples in this text is one such data set. Jones, Koolman, and Rice (2006) (JKR) list several other applications, including the British Household Panel Survey (BHPS), the European Community Household Panel (ECHP), and the Panel Study of Income Dynamics (PSID).

If observations are missing completely at random (MCAR), then the problem of nonresponse can be ignored, though for estimation of dynamic models, either the analysis will have to be restricted to observations with uninterrupted sequences of observations, or some very strong assumptions and interpolation methods will have to be employed to fill the gaps. (See Section 4.7.4 for discussion of the terminology and issues in handling missing data.) The problem for estimation arises when observations are missing for reasons that are related to the outcome variable of interest. **Nonresponse bias** and a related problem, **attrition bias** (individuals leave permanently during the study) result when conventional estimators, such as least squares or the probit maximum likelihood estimator being used here, are applied to samples in which observations are present or absent from the sample for reasons related to the outcome variable. It is a form of **sample selection bias**, that we will examine further in Chapter 19.

Verbeek and Nijman (1992) have suggested a test for endogeneity of the sample response pattern. (We will adopt JKR's notation and terminology for this.) Let  $h$  denote the outcome of interest and  $\mathbf{x}$  denote the relevant set of covariates. Let  $R$  denote the pattern of response. If nonresponse is (completely) random, then  $E[h | \mathbf{x}, R] = E[h | \mathbf{x}]$ . This suggests a variable addition test (neglecting other panel data effects); a pooled

## 736 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

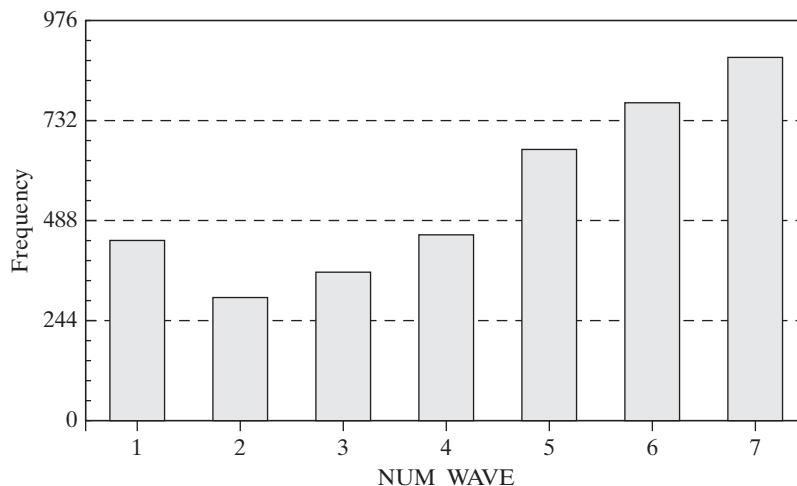
model that contains  $R$  in addition to  $\mathbf{x}$  can provide the means for a simple test of endogeneity. JKR (and Verbeek and Nijman) suggest using the number of waves at which the individual is present as the measure of  $R$ . Thus, adding  $R$  to the pooled model, we can use a simple  $t$  test for the hypothesis.

Devising an estimator given that (non)response is nonignorable requires a more detailed understanding of the process generating the response pattern. The crucial issue is whether the sample selection is based “on unobservables” or “on observables.” **Selection on unobservables** results when, after conditioning on the relevant variables,  $\mathbf{x}$  and other information,  $\mathbf{z}$ , the sampling mechanism is still nonrandom with respect to the disturbances in the models. Selection on unobservables is at the heart of the sample selectivity methodology pioneered by Heckman (1979) that we will study in Chapter 19. (Some applications of the role of unobservables in biased estimation are discussed in Chapter 8, where we examine sources of endogeneity in regression models.) If selection is on observables and then conditioned on an appropriate specification involving the observable information,  $(\mathbf{x}, \mathbf{z})$ , a consistent estimator of the model parameters will be available by “purging” the estimator of the endogeneity of the sampling mechanism.

JKR adopt an **inverse probability weighted (IPW)** estimator devised by Robins, Rotnitsky and Zhao (1995), Fitzgerald, Gottschalk, and Moffitt (1998), Moffitt, Fitzgerald and Gottschalk (1999), and Wooldridge (2002). The estimator is based on the general MCAR assumption that  $P(R = 1 | h, \mathbf{x}, \mathbf{z}) = P(R = 1 | \mathbf{x}, \mathbf{z})$ . That is, the observable covariates convey all the information that determines the response pattern—the probability of nonresponse does not vary systematically with the outcome variable once the exogenous information is accounted for. Implementing this idea in an estimator would require that  $\mathbf{x}$  and  $\mathbf{z}$  be observable when  $R = 0$ , that is, the exogenous data be available for the nonresponders. This will typically not be the case; in an unbalanced panel, the entire observation is missing. Wooldridge (2002) proposed a somewhat stronger assumption that makes estimation feasible:  $P(R = 1 | h, \mathbf{x}, \mathbf{z}) = P(R = 1 | \mathbf{z})$  where  $\mathbf{z}$  is a set of covariates available at wave 1 (entry to the study). To compute Wooldridge’s IPW estimator, we will begin with the sample of all individuals who are present at wave 1 of the study. (In our Example 17.17, based on the GSOEP data, not all individuals are present at the first wave.) At wave 1,  $(\mathbf{x}_{i1}, \mathbf{z}_{i1})$  are observed for all individuals to be studied;  $\mathbf{z}_{i1}$  contains information on observables that are not included in the outcome equation and that predict the response pattern at subsequent waves, including the response variable at the first wave. At wave 1, then,  $P(R_{i1} = 1 | \mathbf{x}_{i1}, \mathbf{z}_{i1}) = 1$ . Wooldridge suggests using a probit model for  $P(R_{it} = 1 | \mathbf{x}_{it}, \mathbf{z}_{it})$ ,  $t = 2, \dots, T$  for the remaining waves to obtain predicted probabilities of response,  $\hat{p}_{it}$ . The IPW estimator then maximizes the weighted log likelihood

$$\ln L_{IPW} = \sum_{i=1}^n \sum_{t=1}^T \frac{R_{it}}{\hat{p}_{it}} \ln L_{it}.$$

Inference based on the weighted log-likelihood function can proceed as in Section 17.3. A remaining detail concerns whether the use of the predicted probabilities in the weighted log-likelihood function makes it necessary to correct the standard errors for two-step estimation. The case here is not an application of the two-step estimators we considered in Section 14.7, since the first step is not used to produce an estimated parameter vector in the second. Wooldridge (2002) shows that the standard errors computed



**FIGURE 17.4** Number of Waves Responded for Those Present at Wave

without the adjustment are “conservative” in that they are larger than they would be with the adjustment.

**Example 17.17 Nonresponse in the GSOEP Sample**

Of the 7,293 individuals in the GSOEP data that we have used in several earlier examples, 3,874 were present at wave 1 (1984) of the sample. The pattern of the number of waves present by these 3,874 is shown in Figure 17.4. The waves are 1984–1988, 1991, and 1994. A dynamic model would be based on the 1,600 of those present at wave 1 who were also present for the next four waves. There is a substantial amount of nonresponse in these data. Not all individuals exit the sample with the first nonresponse, however, so the resulting panel remains unbalanced. The impression suggested by Figure 17.4 could be a bit misleading—the nonresponse pattern is quite different from simple attrition. For example, of the 3,874 individuals who responded at wave 1, 364 did not respond at wave 2 but returned to the sample at wave 3.

To employ the Verbeek and Nijman test, we used the entire sample of 27,326 household years of data. The pooled probit model for DocVis > 0 produced the results at the left in Table 17.14. A *t* (Wald) test of the hypothesis that the coefficient on number of waves present is zero is strongly rejected, so we proceed to the inverse probability weighted estimator. For computing the inverse probability weights, we used the following specification:

$$x_{i1} = \text{constant, age, income, educ, kids, married}$$

$$z_{i1} = \text{female, handicapped dummy, percentage handicapped, university, working, blue collar, white collar, public servant, } y_{i1}$$

$$y_{i1} = \text{Doctor Visits} > 0 \text{ in period 1.}$$

This first-year data vector is used as the observed explanatory variables in probit models for waves 2–7 for the 3,874 individuals who were present at wave 1. There are 3,874 observations for each of these probit models, since all were observed at wave 1. Fitted probabilities for  $R_{it}$  are computed for waves 2–7, while  $R_{i1} = 1$ . The sample means of these probabilities which equals the proportion of the 3,874 who responded at each wave are 1.000, 0.730, 0.672, 0.626, 0.682, 0.568, and 0.386, respectively. Table 17.14 presents the estimated models for several specifications In each case, it appears that the weighting brings some moderate changes in the parameters and, uniformly, reductions in the standard errors.

**738 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**
**TABLE 17.14** Inverse Probability Weighted Estimators

<i>Variable</i>	<i>Endog. Test</i>	<i>Pooled Model</i>		<i>Random Effects—Mundlak</i>		<i>Fixed Effects</i>	
		<i>Unwtd.</i>	<i>IPW</i>	<i>Unwtd.</i>	<i>IPW</i>	<i>Unwtd.</i>	<i>IPW</i>
Constant	0.26411 (0.05893)	0.03369 (0.07684)	-0.02373 (0.06385)	0.09838 (0.16081)	0.13237 (0.17019)		
Age	0.01369 (0.00080)	0.01667 (0.00107)	0.01831 (0.00088)	0.05141 (0.00422)	0.05656 (0.00388)	0.06210 (0.00506)	0.06841 (0.00465)
Income	-0.12446 (0.04636)	-0.17097 (0.05981)	-0.22263 (0.04801)	0.05794 (0.11256)	0.01699 (0.10580)	0.07880 (0.12891)	0.03603 (0.12193)
Education	-0.02925 (0.00351)	-0.03614 (0.00449)	-0.03513 (0.00365)	-0.06456 (0.06104)	-0.07058 (0.05792)	-0.07752 (0.06582)	-0.08574 (0.06149)
Kids	-0.13130 (0.01828)	-0.13077 (0.02303)	-0.13277 (0.01950)	-0.04961 (0.04500)	-0.03427 (0.04356)	-0.05776 (0.05296)	-0.03546 (0.05166)
Married	0.06759 (0.02060)	0.06237 (0.02616)	0.07015 (0.02097)	-0.06582 (0.06596)	-0.09235 (0.06330)	-0.07939 (0.08146)	-0.11283 (0.07838)
Mean Age				-0.03056 (0.00479)	-0.03401 (0.00455)		
Mean Income				-0.66388 (0.18646)	-0.78077 (0.18866)		
Mean Education				0.02656 (0.06160)	0.02899 (0.05848)		
Mean Kids				-0.17524 (0.07266)	-0.20615 (0.07464)		
Mean Married				0.22346 (0.08719)	0.25763 (0.08433)		
Number of Waves	-0.02977 (0.00450)					0.46538	0.48616
$\rho$							

## 17.5 BIVARIATE AND MULTIVARIATE PROBIT MODELS

In Chapter 10, we analyzed a number of different multiple-equation extensions of the classical and generalized regression model. A natural extension of the probit model would be to allow more than one equation, with correlated disturbances, in the same spirit as the seemingly unrelated regressions model. The general specification for a two-equation model would be

$$\begin{aligned} y_1^* &= \mathbf{x}_1' \boldsymbol{\beta}_1 + \varepsilon_1, \quad y_1 = 1 \text{ if } y_1^* > 0, 0 \text{ otherwise,} \\ y_2^* &= \mathbf{x}_2' \boldsymbol{\beta}_2 + \varepsilon_2, \quad y_2 = 1 \text{ if } y_2^* > 0, 0 \text{ otherwise,} \\ \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} | \mathbf{x}_1, \mathbf{x}_2 &\sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]. \end{aligned} \tag{17-48}$$

This bivariate probit model is interesting in its own right for modeling the joint determination of two variables, such as doctor and hospital visits in the next example. It also provides the framework for modeling in two common applications. In many cases, a treatment effect, or endogenous influence, takes place in a binary choice context. The bivariate probit model provides a specification for analyzing a case in which a probit

model contains an endogenous binary variable in one of the equations. In Example 17.21, we will extend (17-48) to

$$\begin{aligned} W^* &= \mathbf{x}'_1 \boldsymbol{\beta}_1 + \varepsilon_1, \quad W = 1 \text{ if } W^* > 0, 0 \text{ otherwise,} \\ y^* &= \mathbf{x}'_2 \boldsymbol{\beta}_2 + \gamma W + \varepsilon_2, \quad y = 1 \text{ if } y^* > 0, 0 \text{ otherwise,} \\ \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} &\sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]. \end{aligned} \quad (17-49)$$

This model extends the case in Section 17.3.5, where  $W^*$ , rather than  $W$ , appears on the right-hand side of the second equation. In the example,  $W$  denotes whether a liberal arts college supports a women's studies program on the campus while  $y$  is a binary indicator of whether the economics department provides a gender economics course. A second common application, in which the first equation is an endogenous sampling rule, is another variant of the bivariate probit model:

$$\begin{aligned} S^* &= \mathbf{x}'_1 \boldsymbol{\beta}_1 + \varepsilon_1, \quad S = 1 \text{ if } S^* > 0, 0 \text{ otherwise,} \\ y^* &= \mathbf{x}'_2 \boldsymbol{\beta}_2 + \varepsilon_2, \quad y = 1 \text{ if } y^* > 0, 0 \text{ otherwise,} \\ \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} &\sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right], \\ (y, \mathbf{x}_2) &\text{ observed only when } S = 1. \end{aligned} \quad (17-50)$$

In Example 17.22, we will study an application in which  $S$  is the result of a credit card application (or any sort of loan application) while  $y_2$  is a binary indicator for whether the individual defaults on the credit account (loan). This is a form of endogenous sampling (in this instance, sampling on unobservables) that has some commonality with the attrition problem that we encountered in Section 17.4.9.

At the end of this section, we will extend (17-48) to more than two equations. This will allow direct treatment of multiple binary outcomes. It will also allow a more general panel data model for  $T$  periods than is provided by the random effects specification.

### 17.5.1 MAXIMUM LIKELIHOOD ESTIMATION

The bivariate normal cdf is

$$\text{Prob}(X_1 < x_1, X_2 < x_2) = \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} \phi_2(z_1, z_2, \rho) dz_1 dz_2,$$

which we denote  $\Phi_2(x_1, x_2, \rho)$ . The density is<sup>37</sup>

$$\phi_2(x_1, x_2, \rho) = \frac{e^{-(1/2)(x_1^2 + x_2^2 - 2\rho x_1 x_2)/(1-\rho^2)}}{2\pi(1-\rho^2)^{1/2}}.$$

To construct the log-likelihood, let  $q_{i1} = 2y_{i1} - 1$  and  $q_{i2} = 2y_{i2} - 1$ . Thus,  $q_{ij} = 1$  if  $y_{ij} = 1$  and  $-1$  if  $y_{ij} = 0$  for  $j = 1$  and 2. Now let

$$z_{ij} = \mathbf{x}'_{ij} \boldsymbol{\beta}_j \quad \text{and} \quad w_{ij} = q_{ij} z_{ij}, \quad j = 1, 2,$$

<sup>37</sup>See Section B.9.

**740 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**

and

$$\rho_{i^*} = q_{i1}q_{i2}\rho.$$

Note the notational convention. The subscript 2 is used to indicate the bivariate normal distribution in the density  $\phi_2$  and cdf  $\Phi_2$ . In all other cases, the subscript 2 indicates the variables in the second equation. As before,  $\phi(\cdot)$  and  $\Phi(\cdot)$  without subscripts denote the univariate standard normal density and cdf.

The probabilities that enter the likelihood function are

$$\text{Prob}(Y_1 = y_{i1}, Y_2 = y_{i2} | \mathbf{x}_1, \mathbf{x}_2) = \Phi_2(w_{i1}, w_{i2}, \rho_{i^*}),$$

which accounts for all the necessary sign changes needed to compute probabilities for  $y$ 's equal to zero and one. Thus,<sup>38</sup>

$$\ln L = \sum_{i=1}^n \ln \Phi_2(w_{i1}, w_{i2}, \rho_{i^*}).$$

The derivatives of the log-likelihood then reduce to

$$\begin{aligned} \frac{\partial \ln L}{\partial \beta_j} &= \sum_{i=1}^n \left( \frac{q_{ij}g_{ij}}{\Phi_2} \right) \mathbf{x}_{ij}, \quad j = 1, 2, \\ \frac{\partial \ln L}{\partial \rho} &= \sum_{i=1}^n \frac{q_{i1}q_{i2}\phi_2}{\Phi_2}, \end{aligned} \tag{17-51}$$

where

$$g_{i1} = \phi(w_{i1})\Phi\left[\frac{w_{i2} - \rho_{i^*}w_{i1}}{\sqrt{1 - \rho_{i^*}^2}}\right] \tag{17-52}$$

and the subscripts 1 and 2 in  $g_{i1}$  are reversed to obtain  $g_{i2}$ . Before considering the Hessian, it is useful to note what becomes of the preceding if  $\rho = 0$ . For  $\partial \ln L / \partial \beta_1$ , if  $\rho = \rho_{i^*} = 0$ , then  $g_{i1}$  reduces to  $\phi(w_{i1})\Phi(w_{i2})$ ,  $\phi_2$  is  $\phi(w_{i1})\phi(w_{i2})$ ,  $\Phi_2$  is  $\Phi(w_{i1})\Phi(w_{i2})$ . Inserting these results in (17-51) with  $q_{i1}$  and  $q_{i2}$  produces (17-21). Because both functions in  $\partial \ln L / \partial \rho$  factor into the product of the univariate functions,  $\partial \ln L / \partial \rho$  reduces to  $\sum_{i=1}^n \lambda_{i1}\lambda_{i2}$ , where  $\lambda_{ij}$ ,  $j = 1, 2$ , is defined in (17-20). (This result will reappear in the LM statistic shown later.)

The maximum likelihood estimates are obtained by simultaneously setting the three derivatives to zero. The second derivatives are relatively straightforward but tedious. Some simplifications are useful. Let

$$\begin{aligned} \delta_i &= \frac{1}{\sqrt{1 - \rho_{i^*}^2}}, \\ v_{i1} &= \delta_i(w_{i2} - \rho_{i^*}w_{i1}), \quad \text{so } g_{i1} = \phi(w_{i1})\Phi(v_{i1}), \\ v_{i2} &= \delta_i(w_{i1} - \rho_{i^*}w_{i2}), \quad \text{so } g_{i2} = \phi(w_{i2})\Phi(v_{i2}). \end{aligned}$$

By multiplying it out, you can show that

$$\delta_i\phi(w_{i1})\phi(v_{i1}) = \delta_i\phi(w_{i2})\phi(v_{i2}) = \phi_2.$$

<sup>38</sup>To avoid further ambiguity, and for convenience, the observation subscript will be omitted from  $\Phi_2 = \Phi_2(w_{i1}, w_{i2}, \rho_{i^*})$  and from  $\phi_2 = \phi_2(w_{i1}, w_{i2}, \rho_{i^*})$ .

Then

$$\begin{aligned}
 \frac{\partial^2 \log L}{\partial \beta_1 \partial \beta'_1} &= \sum_{i=1}^n \mathbf{x}_{i1} \mathbf{x}'_{i1} \left[ \frac{-w_{i1}g_{i1}}{\Phi_2} - \frac{\rho_{i*}\phi_2}{\Phi_2} - \frac{g_{i1}^2}{\Phi_2^2} \right], \\
 \frac{\partial^2 \log L}{\partial \beta_1 \partial \beta'_2} &= \sum_{i=1}^n q_{i1}q_{i2} \mathbf{x}_{i1} \mathbf{x}'_{i2} \left[ \frac{\phi_2}{\Phi_2} - \frac{g_{i1}g_{i2}}{\Phi_2^2} \right], \\
 \frac{\partial^2 \log L}{\partial \beta_1 \partial \rho} &= \sum_{i=1}^n q_{i2} \mathbf{x}_{i1} \frac{\phi_2}{\Phi_2} \left[ \rho_{i*}\delta_i v_{i1} - w_{i1} - \frac{g_{i1}}{\Phi_2} \right], \\
 \frac{\partial^2 \log L}{\partial \rho^2} &= \sum_{i=1}^n \frac{\phi_2}{\Phi_2} \left[ \delta_i^2 \rho_{i*} (1 - \mathbf{w}'_i \mathbf{R}_i^{-1} \mathbf{w}_i) + \delta_i^2 w_{i1} w_{i2} - \frac{\phi_2}{\Phi_2} \right],
 \end{aligned} \tag{17-53}$$

where  $\mathbf{w}'_i \mathbf{R}_i^{-1} \mathbf{w}_i = \delta_i^2 (w_{i1}^2 + w_{i2}^2 - 2\rho_{i*}w_{i1}w_{i2})$ . (For  $\beta_2$ , change the subscripts in  $\partial^2 \ln L / \partial \beta_1 \partial \beta'_1$  and  $\partial^2 \ln L / \partial \beta_1 \partial \rho$  accordingly.) The complexity of the second derivatives for this model makes it an excellent candidate for the Berndt et al. estimator of the variance matrix of the maximum likelihood estimator.

#### Example 17.18 Tetrachoric Correlation

Returning once again to the health care application of Examples 17.4 and several others, we now consider a second binary variable,

$$Hospital_{it} = 1 \text{ if } HospVis_{it} > 0 \text{ and } 0 \text{ otherwise.}$$

Our previous analyses have focused on

$$Doctor_{it} = 1 \text{ if } DocVis_{it} > 0 \text{ and } 0 \text{ otherwise.}$$

A simple bivariate frequency count for these two variables is

		<b>Hospital</b>		
		<b>0</b>	<b>1</b>	<b>Total</b>
<b>Doctor</b>				
0		9,715	420	10,135
1		15,216	1,975	17,191
Total		24,931	2,395	27,326

Looking at the very large value in the lower-left cell, one might surmise that these two binary variables (and the underlying phenomena that they represent) are negatively correlated. The usual Pearson, product moment correlation would be inappropriate as a measure of this correlation since it is used for continuous variables. Consider, instead, a bivariate probit "model,"

$$\begin{aligned}
 H_{it}^* &= \mu_1 + \varepsilon_{1,it}, \quad Hospital_{it} = 1(H_{it}^* > 0), \\
 D_{it}^* &= \mu_2 + \varepsilon_{2,it}, \quad Doctor_{it} = 1(D_{it}^* > 0),
 \end{aligned}$$

where  $(\varepsilon_1, \varepsilon_2)$  have a bivariate normal distribution with means  $(0, 0)$ , variances  $(1, 1)$  and correlation  $\rho$ . This is the model in (17-48) without independent variables. In this representation, the **tetrachoric correlation**, which is a correlation measure for a pair of binary variables, is precisely the  $\rho$  in this model—it is the correlation that would be measured between the underlying continuous variables if they could be observed. This suggests an interpretation of the correlation coefficient in a bivariate probit model—as the conditional tetrachoric correlation. It also suggests a method of easily estimating the tetrachoric correlation coefficient using a program that is built into nearly all commercial software packages.

Applied to the hospital/doctor data defined earlier, we obtained an estimate of  $\rho$  of 0.31106, with an estimated asymptotic standard error of 0.01357. Apparently, our earlier intuition was incorrect.

## 742 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

### 17.5.2 TESTING FOR ZERO CORRELATION

The Lagrange multiplier statistic is a convenient device for testing for the absence of correlation in this model. Under the null hypothesis that  $\rho$  equals zero, the model consists of independent probit equations, which can be estimated separately. Moreover, in the multivariate model, all the bivariate (or multivariate) densities and probabilities factor into the products of the marginals if the correlations are zero, which makes construction of the test statistic a simple matter of manipulating the results of the independent probits. The Lagrange multiplier statistic for testing  $H_0: \rho = 0$  in a bivariate probit model is<sup>39</sup>

$$\text{LM} = \frac{\left[ \sum_{i=1}^n q_{i1} q_{i2} \frac{\phi(w_{i1})\phi(w_{i2})}{\Phi(w_{i1})\Phi(w_{i2})} \right]^2}{\sum_{i=1}^n \frac{[\phi(w_{i1})\phi(w_{i2})]^2}{\Phi(w_{i1})\Phi(-w_{i1})\Phi(w_{i2})\Phi(-w_{i2})}}.$$

As usual, the advantage of the LM statistic is that it obviates computing the bivariate probit model. But the full unrestricted model is now fairly common in commercial software, so that advantage is minor. The likelihood ratio or Wald test can often be used with equal ease. To carry out the likelihood ratio test, we note first that if  $\rho$  equals zero, then the bivariate probit model becomes two independent univariate probits models. The log-likelihood in that case would simply be the sum of the two separate log-likelihoods. The test statistic would be

$$\lambda_{LR} = 2[\ln L_{\text{BIVARIATE}} - (\ln L_1 + \ln L_2)].$$

This would converge to a chi-squared variable with one degree of freedom. The Wald test is carried out by referring

$$\lambda_{WALD} = \left[ \hat{\rho}_{MLE} / \sqrt{\text{Est. Asy. Var}[\hat{\rho}_{MLE}]} \right]^2$$

to the chi-squared distribution with one degree of freedom. For 95 percent significance, the critical value is 3.84 (or one can refer the positive square root to the standard normal critical value of 1.96). Example 17.19 demonstrates.

### 17.5.3 PARTIAL EFFECTS

There are several “marginal effects” one might want to evaluate in a bivariate probit model.<sup>40</sup> A natural first step would be the derivatives of  $\text{Prob}[y_1 = 1, y_2 = 1 | \mathbf{x}_1, \mathbf{x}_2]$ . These can be deduced from (17.5.1) by multiplying by  $\Phi_2$ , removing the sign carrier,  $q_{ij}$  and differentiating with respect to  $\mathbf{x}_j$  rather than  $\beta_j$ . The result is

$$\frac{\partial \Phi_2(\mathbf{x}'_1 \boldsymbol{\beta}_1, \mathbf{x}'_2 \boldsymbol{\beta}_2, \rho)}{\partial \mathbf{x}_1} = \phi(\mathbf{x}'_1 \boldsymbol{\beta}_1) \Phi \left( \frac{\mathbf{x}'_2 \boldsymbol{\beta}_2 - \rho \mathbf{x}'_1 \boldsymbol{\beta}_1}{\sqrt{1 - \rho^2}} \right) \boldsymbol{\beta}_1.$$

Note, however, the bivariate probability, albeit possibly of interest in its own right, is not a conditional mean function. As such, the preceding does not correspond to a regression coefficient or a slope of a conditional expectation.

<sup>39</sup>This is derived in Kiefer (1982).

<sup>40</sup>See Greene (1996b) and Christofides et al. (1997, 2000).

## CHAPTER 17 ♦ Discrete Choice 743

For convenience in evaluating the conditional mean and its partial effects, we will define a vector  $\mathbf{x} = \mathbf{x}_1 \cup \mathbf{x}_2$  and let  $\mathbf{x}'\beta_1 = \mathbf{x}'\gamma_1$ . Thus,  $\gamma_1$  contains all the nonzero elements of  $\beta_1$  and possibly some zeros in the positions of variables in  $\mathbf{x}$  that appear only in the other equation;  $\gamma_2$  is defined likewise. The bivariate probability is

$$\text{Prob}[y_1 = 1, y_2 = 1 | \mathbf{x}] = \Phi_2[\mathbf{x}'\gamma_1, \mathbf{x}'\gamma_2, \rho].$$

Sig<sup>n</sup>s are changed appropriately if the probability of the zero outcome is desired in either case. (See 17-48.) The marginal effects of changes in  $\mathbf{x}$  on this probability are given by

$$\frac{\partial \Phi_2}{\partial \mathbf{x}} = g_1\gamma_1 + g_2\gamma_2,$$

where  $g_1$  and  $g_2$  are defined in (17-19). The familiar univariate cases will arise if  $\rho = 0$ , and effects specific to one equation or the other will be produced by zeros in the corresponding position in one or the other parameter vector. There are also some conditional mean functions to consider. The unconditional mean functions are given by the univariate probabilities:

$$E[y_j | \mathbf{x}] = \Phi(\mathbf{x}'\gamma_j), \quad j = 1, 2,$$

so the analysis of (17-9) and (17-14) applies. One pair of conditional mean functions that might be of interest are

$$\begin{aligned} E[y_1 | y_2 = 1, \mathbf{x}] &= \text{Prob}[y_1 = 1 | y_2 = 1, \mathbf{x}] = \frac{\text{Prob}[y_1 = 1, y_2 = 1 | \mathbf{x}]}{\text{Prob}[y_2 = 1 | \mathbf{x}]} \\ &= \frac{\Phi_2(\mathbf{x}'\gamma_1, \mathbf{x}'\gamma_2, \rho)}{\Phi(\mathbf{x}'\gamma_2)} \end{aligned}$$

and similarly for  $E[y_2 | y_1 = 1, \mathbf{x}]$ . The marginal effects for this function are given by

$$\frac{\partial E[y_1 | y_2 = 1, \mathbf{x}]}{\partial \mathbf{x}} = \left( \frac{1}{\Phi(\mathbf{x}'\gamma_2)} \right) \left[ g_1\gamma_1 + \left( g_2 - \frac{\phi(\mathbf{x}'\gamma_2)}{\Phi(\mathbf{x}'\gamma_2)} \right) \gamma_2 \right].$$

Finally, one might construct the nonlinear conditional mean function

$$E[y_1 | y_2, \mathbf{x}] = \frac{\Phi_2[\mathbf{x}'\gamma_1, (2y_2 - 1)\mathbf{x}'\gamma_2, (2y_2 - 1)\rho]}{\Phi[(2y_2 - 1)\mathbf{x}'\gamma_2]}.$$

The derivatives of this function are the same as those presented earlier, with sign changes in several places if  $y_2 = 0$  is the argument.

#### **Example 17.19 Bivariate Probit Model for Health Care Utilization**

We have extended the bivariate probit model of the previous example by specifying a set of independent variables,

$$\mathbf{x}_i = \text{Constant}, \text{Female}_i, \text{Age}_{it}, \text{Income}_{it}, \text{Kids}_{it}, \text{Education}_{it}, \text{Married}_{it}.$$

We have specified that the same exogenous variables appear in both equations. (There is no requirement that different variables appear in the equations, nor that a variable be excluded from each equation.) The correct analogy here is to the seemingly unrelated regressions model, not to the linear simultaneous equations model. Unlike the SUR model of Chapter 10, it is not the case here that having the same variables in the two equations implies that the model can be fit equation by equation, one equation at a time. That result only applies to the estimation of sets of linear regression equations.

**744 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**
**TABLE 17.15** Estimated Bivariate Probit Model<sup>a</sup>

Variable	Doctor			Hospital			
	Model Estimates		Partial Effects			Model Estimates	
	Univariate	Bivariate	Direct	Indirect	Total	Univariate	Bivariate
Constant	-0.1243 (0.05815)	-0.1243 (0.05814)				-1.3328 (0.08320)	-1.3385 (0.07957)
Female	0.3559 (0.01602)	0.3551 (0.01604)	0.09650 (0.004957)	-0.00724 (0.001515)	0.08926 (0.005127)	0.1023 (0.02195)	0.1050 (0.02174)
Age	0.01189 (0.0007957)	0.01188 (0.000802)	0.003227 (0.000231)	-0.00032 (0.000073)	0.002909 (0.000238)	0.004605 (0.001082)	0.00461 (0.001058)
Income	-0.1324 (0.04655)	-0.1337 (0.04628)	-0.03632 (0.01260)	-0.003064 (0.004105)	-0.03939 (0.01254)	0.03739 (0.06329)	0.04441 (0.05946)
Kids	-0.1521 (0.01833)	-0.1523 (0.01825)	-0.04140 (0.005053)	0.001047 (0.001773)	-0.04036 (0.005168)	-0.01714 (0.02562)	-0.01517 (0.02570)
Education	-0.01497 (0.003575)	-0.01484 (0.003575)	-0.004033 (0.000977)	0.001512 (0.00035)	-0.002521 (0.0010)	-0.02196 (0.005215)	-0.02191 (0.005110)
Married	0.07352 (0.02064)	0.07351 (0.02063)	0.01998 (0.005626)	0.003303 (0.001917)	0.02328 (0.005735)	-0.04824 (0.02788)	-0.04789 (0.02777)

<sup>a</sup> Estimated correlation coefficient = 0.2981 (0.0139).

Table 17.15 contains the estimates of the parameters of the univariate and bivariate probit models. The tests of the null hypothesis of zero correlation strongly reject the hypothesis that  $\rho$  equals zero. The  $t$  statistic for  $\rho$  based on the full model is  $0.2981 / 0.0139 = 21.446$ , which is much larger than the critical value of 1.96. For the likelihood ratio test, we compute

$$\lambda_{LR} = 2[-25285.07 - [-17422.72 - 8073.604]] = 422.508.$$

Once again, the hypothesis is rejected. (The Wald statistic is  $21.446^2 = 459.957$ .) The LM statistic is 383.953. The coefficient estimates agree with expectations. The income coefficient is statistically significant in the doctor equation, but not in the hospital equation, suggesting, perhaps, that physician visits are at least to some extent discretionary while hospital visits occur on an emergency basis that would be much less tied to income. The table also contains the decomposition of the partial effects for  $E[y_1 | y_2 = 1]$ . The direct effect is  $[g_1 / \Phi(\mathbf{x}'\beta_1)]y_1$  in the definition given earlier. The mean estimate of  $E[y_1 | y_2 = 1]$  is 0.821285. In the table in Example 17.8, the  $\frac{1}{2}$  should correspond to the raw proportion  $P(D = 1, H = 1) / P(H = 1) = (1975 / 27326) / (2055 / 27326) = 0.8246$ .

#### 17.5.4 A PANEL DATA MODEL FOR BIVARIATE BINARY RESPONSE

Extending multiple equation models to accommodate unobserved common effects in panel data settings is straightforward in theory, but complicated in practice. For the bivariate probit case, for example, the natural extension of (17-48) would be

$$\begin{aligned} y_{1,it}^* &= \mathbf{x}'_{1,it}\beta_1 + \varepsilon_{1,it} + \alpha_{1,i}, \quad y_{1,it} = 1 \text{ if } y_{1,it}^* > 0, 0 \text{ otherwise,} \\ y_{2,it}^* &= \mathbf{x}'_{2,it}\beta_2 + \varepsilon_{2,it} + \alpha_{2,i}, \quad y_{2,it} = 1 \text{ if } y_{2,it}^* > 0, 0 \text{ otherwise,} \\ \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} | \mathbf{x}_1, \mathbf{x}_2 &\sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]. \end{aligned}$$

The complication will be in how to treat  $(\alpha_1, \alpha_2)$ . A fixed effects treatment will require estimation of two full sets of dummy coefficients, will likely encounter the incidental parameters problem in double measure, and will be complicated in practical terms.

As in all earlier cases, the fixed effects case also preempts any specification involving time-invariant variables. It is also unclear in a fixed effects model, how any correlation between  $\alpha_1$  and  $\alpha_2$  would be handled. It should be noted that strictly from a consistency standpoint, these considerations are moot. The two equations can be estimated separately, only with some loss of efficiency. The analogous situation would be the seemingly unrelated regressions model in Chapter 10. A random effects treatment (perhaps accommodated with Mundlak's approach of adding the group means to the equations as in Section 17.4.5) offers greater promise. If  $(\alpha_1, \alpha_2) = (u_1, u_2)$  are normally distributed random effects, with

$$\begin{pmatrix} u_{1,i} \\ u_{2,i} \end{pmatrix} | \mathbf{X}_{1,i}, \mathbf{X}_{2,i} \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right],$$

then the unconditional log likelihood for the bivariate probit model,

$$\ln L = \sum_{i=1}^n \ln \int_{u_1, u_2} \prod_{t=1}^{T_i} \Phi_2(w_{1,it} | u_{1,i}, w_{2,it} | u_{2,i}, \rho_{it}^*) f(u_{1,i}, u_{2,i}) du_{1,i} du_{2,i},$$

can be maximized using simulation or quadrature as we have done in previous applications. A possible variation on this specification would specify that the same common effect enter both equations. In that instance, the integration would only be over a single dimension. In this case, there would only be a single new parameter to estimate,  $\sigma^2$ , the variance of the common random effect while  $\rho$  would equal one. A refinement on this form of the model would allow the scaling to be different in the two equations by placing  $u_i$  in the first equation and  $\theta u_i$  in the second. This would introduce the additional scaling parameter, but  $\rho$  would still equal one. This is the formulation of a common random effect used in Heckman's formulation of the dynamic panel probit model in the Section 17.4.6.

#### **Example 17.20 Bivariate Random Effects Model for Doctor and Hospital Visits**

We will extend the pooled bivariate probit model presented in Example 17.19 by allowing a general random effects formulation, with free correlation between the time-varying components ( $\varepsilon_1, \varepsilon_2$ ) and between the time-invariant effects, ( $u_1, u_2$ ). We used simulation to fit the model. Table 17.16 presents the pooled and random effects estimates. The log-likelihood functions for the pooled and random effects models are -25285.07 and -23769.67, respectively. Two times the difference is 3030.76. This would be a chi squared with three degrees of freedom (for the three free elements in the covariance matrix of  $u_1$  and  $u_2$ ). The 95 percent critical value is 7.81, so the pooling hypothesis would be rejected. The change in the correlation coefficient from .2981 to .1501 suggests that we have decomposed the disturbance in the model into a time-varying part and a time-invariant part. The latter seems to be the smaller of the two. Although the time-invariant elements are more highly correlated, their variances are only  $0.2233^2 = 0.0499$  and  $0.6338^2 = 0.4017$  compared to 1.0 for both  $\varepsilon_1$  and  $\varepsilon_2$ .

#### **17.5.5 ENDOGENOUS BINARY VARIABLE A RECURSIVE BIVARIATE PROBIT MODEL**

Section 17.3.5 examines a case in which there is an endogenous variable in a binary choice (probit) model. The model is

$$\begin{aligned} W^* &= \mathbf{x}'_1 \beta_1 + \varepsilon_1, \\ y^* &= \mathbf{x}'_2 \beta_2 + \gamma W^* + \varepsilon_2, \quad y = 1 \text{ if } y^* > 0, 0 \text{ otherwise,} \\ \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} | \mathbf{x}_1, \mathbf{x}_2 &\sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]. \end{aligned}$$

**746 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**
**TABLE 17.16** Estimated Random Effects Bivariate Probit Model

	<i>Doctor</i>	<i>Hospital</i>		
	<i>Pooled</i>	<i>Random Effects</i>	<i>Pooled</i>	<i>Random Effects</i>
Constant	-0.1243 (0.05814)	-0.2976 (0.09650)	-1.3385 (0.07957)	-1.5855 (0.10853)
Female	0.3551 (0.01604)	0.4548 (0.02857)	0.1050 (0.02174)	0.1280 (0.02954)
Age	0.01188 (0.000802)	0.01983 (0.00130)	0.00461 (0.001058)	0.00496 (0.00139)
Income	-0.1337 (0.04628)	-0.01059 (0.06488)	0.04441 (0.05946)	0.13358 (0.07728)
Kids	-0.1523 (0.01825)	-0.1544 (0.02692)	-0.01517 (0.02570)	0.02155 (0.03211)
Education	-0.01484 (0.003575)	-0.02573 (0.00612)	-0.02191 (0.005110)	-0.02444 (0.00675)
Married	0.07351 (0.02063)	0.02876 (0.03167)	-0.04789 (0.02777)	-0.10504 (0.03547)
Corr( $\varepsilon_1, \varepsilon_2$ )	0.2981	0.1501	0.2981	0.1501
Corr( $u_1, u_2$ )	0.0000	0.5382	0.0000	0.5382
Std. Dev. $u$	0.0000	0.2233	0.0000	0.6338
Std. Dev. $\varepsilon$	1.0000	1.0000	1.0000	1.0000

The application examined there involved a labor force participation model that was conditioned on an endogenous variable, the spouse's hours of work. In many cases, the endogenous variable in the equation is also binary. In the application we will examine next, the presence of a gender economics course in the economics curriculum at liberal arts colleges is conditioned on whether or not there is a women's studies program on the campus. The model in this case becomes

$$W^* = \mathbf{x}'_1 \boldsymbol{\beta}_1 + \varepsilon_1, \quad W = 1 \text{ if } W^* > 0, 0 \text{ otherwise},$$

$$y^* = \mathbf{x}'_2 \boldsymbol{\beta}_2 + \gamma W + \varepsilon_2, \quad y = 1 \text{ if } y^* > 0, 0 \text{ otherwise},$$

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} | \mathbf{x}_1, \mathbf{x}_2 \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right].$$

This model illustrates a number of interesting aspects of the bivariate probit model. Note that this model is qualitatively different from the bivariate probit model in (17-48); the first dependent variable,  $W$ , appears on the right-hand side of the second equation.<sup>41</sup> This model is a **recursive**, simultaneous-equations model. Surprisingly, the endogenous nature of one of the variables on the right-hand side of the second equation can be ignored in formulating the log-likelihood. [The model appears in Maddala (1983, p. 123).] We can establish this fact with the following (admittedly trivial) argument: The term that enters the log-likelihood is  $P(y = 1, W = 1) = P(y = 1 | W = 1)P(W = 1)$ . Given the model as stated, the marginal probability for  $W$  is just  $\Phi(\mathbf{x}'_1 \boldsymbol{\beta}_1)$ , whereas the conditional probability is  $\Phi_2(\cdots)/\Phi(\mathbf{x}'_1 \boldsymbol{\beta}_1)$ . The product returns the bivariate normal probability

<sup>41</sup> Eisenberg and Rowe (2006) is another application of this model. In their study, they analyzed the joint (recursive) effect of  $W$  = veteran status on  $y$ , smoking behavior. The estimator they used was two-stage least squares and GMM.

## CHAPTER 17 ♦ Discrete Choice 747

we had earlier. The other three terms in the log-likelihood are derived similarly, which produces (Maddala's results with some sign changes):

$$P(y = 1, W = 1) = \Phi(\mathbf{x}_2' \boldsymbol{\beta}_2 + \gamma, \mathbf{x}_1' \boldsymbol{\beta}_1, \rho),$$

$$P(y = 1, W = 0) = \Phi(-\mathbf{x}_2' \boldsymbol{\beta}_2 - \gamma, -\mathbf{x}_1' \boldsymbol{\beta}_1, -\rho),$$

$$P(y = 0, W = 1) = \Phi[-(\mathbf{x}_2' \boldsymbol{\beta}_2 + \gamma), \mathbf{x}_1' \boldsymbol{\beta}_1, -\rho],$$

$$P(y = 0, W = 0) = \Phi(-\mathbf{x}_2' \boldsymbol{\beta}_2, -\mathbf{x}_1' \boldsymbol{\beta}_1, \rho).$$

These terms are exactly those of (17-48) that we obtain just by carrying  $W$  in the second equation with no special attention to its endogenous nature. We can ignore the simultaneity in this model and we cannot in the linear regression model because, in this instance, we are maximizing the log-likelihood, whereas in the linear regression case, we are manipulating certain sample moments that do not converge to the necessary population parameters in the presence of simultaneity.

**Example 17.21 Gender Economics Courses at Liberal Arts Colleges**

Burnett (1997) proposed the following bivariate probit model for the presence of a gender economics course in the curriculum of a liberal arts college:

$$\text{Prob}[G = 1, W = 1 | \mathbf{x}_G, \mathbf{x}_W] = \Phi[\mathbf{x}_G' \boldsymbol{\beta}_G + \gamma W, \mathbf{x}_W' \boldsymbol{\beta}_W, \rho].$$

The dependent variables in the model are

$G$  = presence of a gender economics course

$W$  = presence of a women's studies program on the campus.

The independent variables in the model are

$Z_1$  = constant term

$Z_2$  = academic reputation of the college, coded 1 (best), 2, ... to 141

$Z_3$  = size of the full-time economics faculty, a count

$Z_4$  = percentage of the economics faculty that are women, proportion (0 to 1)

$Z_5$  = religious affiliation of the college, 0 = no, 1 = yes

$Z_6$  = percentage of the college faculty that are women, proportion (0 to 1)

$Z_7-Z_{10}$  = regional dummy variables, South, Midwest, Northeast, West

The regressor vectors are

$$\mathbf{x}_G = Z_1, Z_2, Z_3, Z_4, Z_5 \quad (\text{gender economics course equation}),$$

$$\mathbf{x}_W = Z_2, Z_5, Z_6, Z_7 - Z_{10} \quad (\text{women's studies program equation}).$$

Maximum likelihood estimates of the parameters of Burnett's model were computed by Greene (1998) using her sample of 132 liberal arts colleges; 31 of the schools offer gender economics, 58 have women's studies, and 29 have both. (See Appendix Table F17.1.) The estimated parameters are given in Table 17.17. Both bivariate probit and the single-equation estimates are given. The estimate of  $\rho$  is only 0.1359, with a standard error of 1.2359. The Wald statistic for the test of the hypothesis that  $\rho$  equals zero is  $(0.1359/1.2359)^2 = 0.011753$ . For a single restriction, the critical value from the chi-squared table is 3.84, so the hypothesis cannot be rejected. The likelihood ratio statistic for the same hypothesis is  $2[-85.6317 - (-85.6458)] = 0.0282$ , which leads to the same conclusion. The Lagrange multiplier statistic is 0.003807, which is consistent. This result might seem counterintuitive, given the setting. Surely "gender economics" and "women's studies" are highly correlated, but this finding does not contradict that proposition. The correlation coefficient measures the correlation between the disturbances in the equations, the omitted factors. That is,  $\rho$  measures (roughly) the correlation between the outcomes after the influence of the included factors is accounted

**748 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**
**TABLE 17.17** Estimates of a Recursive Simultaneous Bivariate Probit Model  
 (estimated standard errors in parentheses)

<i>Variable</i>	<i>Single Equation</i>		<i>Bivariate Probit</i>	
	<i>Coefficient</i>	<i>Standard Error</i>	<i>Coefficient</i>	<i>Standard Error</i>
<b><i>Gender Economics Equation</i></b>				
Constant	-1.4176	(0.8768)	-1.1911	(2.2155)
AcRep	-0.01143	(0.003610)	-0.01233	(0.007937)
WomStud	1.1095	(0.4699)	0.8835	(2.2603)
EconFac	0.06730	(0.05687)	0.06769	(0.06952)
PctWecon	2.5391	(0.8997)	2.5636	(1.0144)
Relig	-0.3482	(0.4212)	-0.3741	(0.5264)
<b><i>Women's Studies Equation</i></b>				
AcRep	-0.01957	(0.004117)	-0.01939	(0.005704)
PctWfac	1.9429	(0.9001)	1.8914	(0.8714)
Relig	-0.4494	(0.3072)	-0.4584	(0.3403)
South	1.3597	(0.5948)	1.3471	(0.6897)
West	2.3386	(0.6449)	2.3376	(0.8611)
North	1.8867	(0.5927)	1.9009	(0.8495)
Midwest	1.8248	(0.6595)	1.8070	(0.8952)
$\rho$	0.0000	(0.0000)	0.1359	(1.2539)
$\ln L$	-85.6458		-85.6317	

for. Thus, the value 0.1359 measures the effect after the influence of women's studies is already accounted for. As discussed in the next paragraph, the proposition turns out to be right. The single most important determinant (at least within this model) of whether a gender economics course will be offered is indeed whether the college offers a women's studies program.

The marginal effects in this model are fairly involved, and as before, we can consider several different types. Consider, for example,  $z_2$ , academic reputation. There is a direct effect produced by its presence in the gender economics course equation. But there is also an indirect effect. Academic reputation enters the women's studies equation and, therefore, influences the probability that  $W$  equals one. Because  $W$  appears in the gender economics course equation, this effect is transmitted back to  $y$ . The total effect of academic reputation and, likewise, religious affiliation is the sum of these two parts. Consider first the gender economics variable,  $y$ . The conditional mean is

$$\begin{aligned} E[G | \mathbf{x}_G, \mathbf{x}_W] &= \text{Prob}[W = 1]E[G | W = 1, \mathbf{x}_G, \mathbf{x}_W] \\ &\quad + \text{Prob}[W = 0]E[G | W = 0, \mathbf{x}_G, \mathbf{x}_W] \\ &= \Phi_2(\mathbf{x}'_G \boldsymbol{\beta}_G + \gamma, \mathbf{x}'_W \boldsymbol{\beta}_W, \rho) + \Phi_2(\mathbf{x}'_G \boldsymbol{\beta}_G, -\mathbf{x}'_W \boldsymbol{\beta}_W, -\rho). \end{aligned}$$

Derivatives can be computed using our earlier results. We are also interested in the effect of religious affiliation. Because this variable is binary, simply differentiating the conditional mean function may not produce an accurate result. Instead, we would compute the conditional mean function with this variable set to one and then zero, and take the difference. Finally, what is the effect of the presence of a women's studies program on the probability that the college will offer a gender economics course? To compute this effect, we would compute

$$\text{Prob}[G = 1 | W = 1, \mathbf{x}_G, \mathbf{x}_W] - \text{Prob}[G = 1 | W = 0, \mathbf{x}_G, \mathbf{x}_W].$$

**TABLE 17.18** Marginal Effects in Gender Economics Model

	<i>Direct</i>	<i>Indirect</i>	<i>Total</i>	<i>(Std. Error)</i>	<i>(Type of Variable, Mean)</i>
<b>Gender Economics Equation</b>					
AcRep	-0.002022	-0.001453	-0.003476	(0.001126)	(Continuous, 119.242)
PctWecon	+0.4491		+0.4491	(0.1568)	(Continuous, 0.24787)
EconFac	+0.01190		+0.1190	(0.01292)	(Continuous, 6.74242)
Relig	-0.06327	-0.02306	-0.08632	(0.08220)	(Binary, 0.57576)
WomStud	+0.1863		+0.1863	(0.0868)	(Endogenous, 0.43939)
PctWfac		+0.14434	+0.14434	(0.09051)	(Continuous, 0.35772)
<b>Women's Studies Equation</b>					
AcRep	-0.00780		-0.00780	(0.001654)	(Continuous, 119.242)
PctWfac	+0.77489		+0.77489	(0.3591)	(Continuous, 0.35772)
Relig	-0.17777		-0.17777	(0.11946)	(Binary, 0.57576)

In all cases, standard errors for the estimated marginal effects can be computed using the delta method or the method of Krinsky and Robb.

Table 17.18 presents the estimates of the marginal effects and some descriptive statistics for the data. The calculations were simplified slightly by using the restricted model with  $\rho = 0$ . Computations of the marginal effects still require the preceding decomposition, but they are simplified by the result that if  $\rho$  equals zero, then the bivariate probabilities factor into the products of the marginals. Numerically, the strongest effect appears to be exerted by the representation of women on the faculty; its coefficient of +0.4491 is by far the largest. This variable, however, cannot change by a full unit because it is a proportion. An increase of 1 percent in the presence of women on the faculty raises the probability by only +0.004, which is comparable in scale to the effect of academic reputation. The effect of women on the faculty is likewise fairly small, only 0.0780 per 1 percent change. As might have been expected, the single most important influence is the presence of a women's studies program, which increases the likelihood of a gender economics course by a full 0.1863. Of course, the raw data would have anticipated this result; of the 31 schools that offer a gender economics course, 29 also have a women's studies program and only two do not. Note finally that the effect of religious affiliation (whatever it is) is mostly direct.

### 17.5.6 ENDOGENOUS SAMPLING IN A BINARY CHOICE MODEL

We have encountered several instances of nonrandom sampling in the binary choice setting. In Section 17.3.6, we examined an application in credit scoring in which the balance in the sample of responses of the outcome variable,  $C = 1$  for acceptance of an application and  $C = 0$  for rejection, is different from the known proportions in the population. The sample was specifically skewed in favor of observations with  $C = 1$  to enrich the data set. A second type of nonrandom sampling arose in the analysis of nonresponse/attrition in the GSOEP in Example 17.17. The data suggest that the observed sample is not random with respect to individuals' presence in the sample at different waves of the panel. The first of these represents selection specifically on an observable outcome—the observed dependent variable. We constructed a model for the second of these that relied on an assumption of selection on a set of certain observables—the variables that entered the probability weights. We will now examine a third form of nonrandom sample selection, based crucially on the unobservables in the two equations of a bivariate probit model.

## 750 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

We return to the banking application of Example 17.9. In that application, we examined a binary choice model,

$$\begin{aligned}\text{Prob}(\text{Cardholder} = 1) &= \text{Prob}(C = 1 | \mathbf{x}) \\ &= \Phi(\beta_1 + \beta_2 \text{Age} + \beta_3 \text{Income} + \beta_4 \text{OwnRent} \\ &\quad + \beta_5 \text{Months at Current Address} \\ &\quad + \beta_6 \text{Self-Employed} \\ &\quad + \beta_7 \text{Number of Major Derogatory Reports} \\ &\quad + \beta_8 \text{Number of Minor Derogatory Reports}).\end{aligned}$$

From the point of view of the lender, cardholder status is not the interesting outcome in the credit history, default is. The more interesting equation describes  $\text{Prob}(\text{Default} = 1 | \mathbf{z}, C = 1)$ . The natural approach, then, would be to construct a binary choice model for the interesting default variable using the historical data for a sample of cardholders. The problem with the approach is that the sample is not randomly drawn—applicants are screened with an eye specifically toward whether or not they seem likely to default. In this application, and in general, there are three economic agents, the credit scorer (e.g., Fair Isaacs), the lender, and the borrower. Each of them has latent characteristics in the equations that determine their behavior. It is these latent characteristics that drive, in part, the application/scoring process and, ultimately, the consumer behavior.

A model that can accommodate these features is (17-50),

$$\begin{aligned}S^* &= \mathbf{x}_1' \boldsymbol{\beta}_1 + \varepsilon_1, \quad S = 1 \text{ if } S^* > 0, 0 \text{ otherwise,} \\ y^* &= \mathbf{x}_2' \boldsymbol{\beta}_2 + \varepsilon_2, \quad y = 1 \text{ if } y^* > 0, 0 \text{ otherwise,} \\ \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} &\sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right], \\ (y, x_2) &\text{ observed only when } S = 1,\end{aligned}$$

which contains an observation rule,  $S = 1$ , and a behavioral outcome,  $y = 0$  or 1. The endogeneity of the sampling rule implies that

$$\text{Prob}(y = 1 | S = 1, \mathbf{x}_2) \neq \Phi(\mathbf{x}_2' \boldsymbol{\beta}).$$

From properties of the bivariate normal distribution, the appropriate probability is

$$\text{Prob}(y = 1 | S = 1, \mathbf{x}_1, \mathbf{x}_2) = \Phi \left[ \frac{\mathbf{x}_2' \boldsymbol{\beta}_2 + \rho \mathbf{x}_1' \boldsymbol{\beta}_1}{\sqrt{1 - \rho^2}} \right].$$

If  $\rho$  is not zero, then in using the simple univariate probit model, we are omitting from our model any variables that are in  $\mathbf{x}_1$  but not in  $\mathbf{x}_2$ , and in any case, the estimator is inconsistent by a factor  $(1 - \rho^2)^{-1/2}$ . To underscore the source of the bias, if  $\rho$  equals zero, the conditional probability returns to the model that would be estimated with the selected sample. Thus, the bias arises because of the correlation of (i.e., the selection on) the unobservables,  $\varepsilon_1$  and  $\varepsilon_2$ . This model was employed by Wynand and van Praag (1981) in the first application of Heckman's (1979) sample selection model in a nonlinear

setting, to insurance purchases, by Boyes, Hoffman, and Lowe (1989) in a study of bank lending, by Greene (1992) to the credit card application begun in Example 17.9 and continued in Example 17.22, and hundreds of applications since. [Some discussion appears in Maddala (1983) as well.]

Given that the forms of the probabilities are known, the appropriate log-likelihood function for estimation of  $\beta_1$ ,  $\beta_2$  and  $\rho$  is easily obtained. The log-likelihood must be constructed for the joint or the marginal probabilities, not the conditional ones. For the “selected observations,” that is,  $(y = 0, S = 1)$  or  $(y = 1, S = 1)$ , the relevant probability is simply

$$\text{Prob}(y = 0 \text{ or } 1 | S = 1) \times \text{Prob}(S = 1) = \Phi_2[(2y - 1)\mathbf{x}_2'\boldsymbol{\beta}_2, \mathbf{x}_1'\boldsymbol{\beta}_1, (2y - 1)\rho]$$

For the observations with  $S = 0$ , the probability that enters the likelihood function is simply  $\text{Prob}(S = 0 | \mathbf{x}_1) = \Phi(-\mathbf{x}_1'\boldsymbol{\beta}_1)$ . Estimation is then based on a simpler form of the bivariate probit log-likelihood that we examined in Section 17.5.1. Partial effects and postestimation analysis would follow the analysis for the bivariate probit model. The desired partial effects would differ by the application, whether one desires the partial effects from the conditional, joint, or marginal probability would vary. The necessary results are in Section 17.5.3.

#### **Example 17.22 Cardholder Status and Default Behavior**

In Example 17.9, we estimated a logit model for cardholder status,

$$\begin{aligned} \text{Prob}(\text{Cardholder} = 1) &= \text{Prob}(C = 1 | \mathbf{x}) \\ &= \Phi(\beta_1 + \beta_2 \text{Age} + \beta_3 \text{Income} + \beta_4 \text{OwnRent} \\ &\quad + \beta_5 \text{Current Address} + \beta_6 \text{SelfEmployed} \\ &\quad + \beta_7 \text{Major Derogatory Reports} \\ &\quad + \beta_8 \text{Minor Derogatory Reports}), \end{aligned}$$

using a sample of 13,444 applications for a credit card. The complication in that example was that the sample was choice based. In the data set, 78.1 percent of the applicants are cardholders. In the population, at that time, the true proportion was roughly 23.2 percent, so the sample is substantially choice based on this variable. The sample was deliberately skewed in favor of cardholders for purposes of the original study [Greene (1992)]. The weights to be applied for the WESML estimator are  $0.232/0.781 = 0.297$  for the observations with  $C = 1$  and  $0.768/0.219 = 3.507$  for observations with  $C = 0$ . Of the 13,444 applicants in the sample, 10,499 were accepted (given the credit cards). The “default rate” in the sample is  $996/10,499$  or 9.48 percent. This is slightly less than the population rate at the time, 10.3 percent. For purposes of a less complicated numerical example, we will ignore the choice-based sampling nature of the data set for the present. An orthodox treatment of both the selection issue and the choice-based sampling treatment is left for the exercises [and pursued in Greene (1992).]

We have formulated the cardholder equation so that it probably resembles the policy of credit scorers, both then and now. A major derogatory report results when a credit account that is being monitored by the credit reporting agency is more than 60 days late in payment. A minor derogatory report is generated when an account is 30 days delinquent. Derogatory reports are a major contributor to credit decisions. Contemporary credit processors such as Fair Isaacs place extremely heavy weight on the “credit score,” a single variable that summarizes the credit history and credit-carrying capacity of an individual. We did not have access to credit scores at the time of this study. The selection equation

**752 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**
**TABLE 17.19** Estimated Joint Cardholder and Default Probability Models

<i>Variable/Equation</i>	<i>Endogenous Sample Model</i>		<i>Uncorrelated Equations</i>	
	<i>Estimate</i>	<i>Standard Error</i>	<i>Estimate</i>	<i>Standard Error</i>
<b>Cardholder Equation</b>				
Constant	0.30516	0.04781	(6.38)	0.31783
Age	0.00226	0.00145	(1.56)	0.00184
Current Address	0.00091	0.00024	(3.80)	0.00095
OwnRent	0.18758	0.03030	(6.19)	0.18233
Income	0.02231	0.00093	(23.87)	0.02237
SelfEmployed	-0.43015	0.05357	(-8.03)	-0.43625
Major Derogatory	-0.69598	0.01871	(-37.20)	-0.69912
Minor Derogatory	-0.04717	0.01825	(-2.58)	-0.04126
<b>Default Equation</b>				
Constant	-0.96043	0.04728	(-20.32)	-0.81528
Dependents	0.04995	0.01415	(3.53)	0.04993
Income	-0.01642	0.00122	(-13.41)	-0.01837
Expend/Income	-0.16918	0.14474	(-1.17)	-0.14172
Correlation	0.41947	0.11762	(3.57)	0.000
Log Likelihood	-8660.90650		-8670.78831	

was given earlier. The default equation is a behavioral model. There is no obvious standard for this part of the model. We have used three variables, *Dependents*, the number of dependents in the household, *Income*, and *Exp\_Income* which equals the ratio of the average credit card expenditure in the 12 months after the credit card was issued to average monthly income. Default status is measured for the first 12 months after the credit card was issued.

Estimation results are presented in Table 17.19. These are broadly consistent with the earlier results—the model with no correlation from Example 17.9 are repeated in Table 17.19. There are two tests we can employ for endogeneity of the selection. The estimate of  $\rho$  is 0.41947 with a standard error of 0.11762. The *t* ratio for the test that  $\rho$  equals zero is 3.57, by which we can reject the hypothesis. Alternatively, the likelihood ratio statistic based on the values in Table 17.19 is  $2(8670.78831 - 8660.90650) = 19.76362$ . This is larger than the critical value of 3.84, so the hypothesis of zero correlation is rejected. The results are as might be expected, with one counterintuitive result, that a larger credit burden, expenditure to income ratio, appears to be associated with lower default probabilities, though not significantly so.

### 17.5.7 A MULTIVARIATE PROBIT MODEL

In principle, a multivariate probit model would simply extend (17-48) to more than two outcome variables just by adding equations. The resulting equation system, again analogous to the seemingly unrelated regressions model, would be

$$\begin{aligned}
 y_m^* &= \mathbf{x}'_m \boldsymbol{\beta}_m + \varepsilon_m, \quad y_m = 1 \text{ if } y_m^* > 0, 0 \text{ otherwise}, \quad m = 1, \dots, M, \\
 E[\varepsilon_m | \mathbf{x}_1, \dots, \mathbf{x}_M] &= 0, \\
 \text{Var}[\varepsilon_m | \mathbf{x}_1, \dots, \mathbf{x}_M] &= 1, \\
 \text{Cov}[\varepsilon_j, \varepsilon_m | \mathbf{x}_1, \dots, \mathbf{x}_M] &= \rho_{jm}, \\
 (\varepsilon_1, \dots, \varepsilon_M) &\sim N_M[\mathbf{0}, \mathbf{R}].
 \end{aligned}$$

The joint probabilities of the observed events,  $[y_{i1}, y_{i2} \dots, y_{iM} | \mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iM}]$ ,  $i = 1, \dots, n$  that form the basis for the log-likelihood function are the  $M$ -variate normal probabilities,

$$L_i = \Phi_M(q_{i1}\mathbf{x}'_{i1}\beta_1, \dots, q_{iM}\mathbf{x}'_{iM}\beta_M, \mathbf{R}^*),$$

where

$$q_{im} = 2y_{im} - 1,$$

$$\mathbf{R}^*_{jm} = q_{ij}q_{im}\rho_{jm}.$$

The practical obstacle to this extension is the evaluation of the  $M$ -variate normal integrals and their derivatives. Some progress has been made on using quadrature for trivariate integration (see Section 14.9.6.c), but existing results are not sufficient to allow accurate and efficient evaluation for more than two variables in a sample of even moderate size. However, given the speed of modern computers, simulation-based integration using the GHK simulator or simulated likelihood methods (see Chapter 15) do allow for estimation of relatively large models. We consider an application in Example 17.23.<sup>42</sup>

The **multivariate probit model** in another form presents a useful extension of the random effects probit model for panel data (Section 17.4.2). If the parameter vectors in all equations are constrained to be equal, we obtain what Bertschek and Lechner (1998) call the “panel probit model,”

$$y_{it}^* = \mathbf{x}'_{it}\beta + \varepsilon_{it}, \quad y_{it} = 1 \text{ if } y_{it}^* > 0, 0 \text{ otherwise}, \quad i = 1, \dots, n, t = 1, \dots, T, \\ (\varepsilon_{i1}, \dots, \varepsilon_{iT}) \sim N[\mathbf{0}, \mathbf{R}].$$

The Butler and Moffitt (1982) approach for this model (see Section 17.4.2) has proved useful in many applications. But, their underlying assumption that  $\text{Cov}[\varepsilon_{it}, \varepsilon_{is}] = \rho$  is a substantive restriction. By treating this structure as a multivariate probit model with the restriction that the coefficient vector be the same in every period, one can obtain a model with free correlations across periods.<sup>43</sup> Hyslop (1999), Bertschek and Lechner (1998), Greene (2004 and Example 17.16), and Cappellari and Jenkins (2006) are applications.

#### **Example 17.23 A Multivariate Probit Model for Product Innovations**

Bertschek and Lechner applied the panel probit model to an analysis of the product innovation activity of 1,270 German firms observed in five years, 1984–1988, in response to imports and foreign direct investment. [See Bertschek (1995).] The probit model to be estimated is based

---

<sup>42</sup>Studies that propose improved methods of simulating probabilities include Pakes and Pollard (1989) and especially Börsch-Supan and Hajivassiliou (1993), Geweke (1989), and Keane (1994). A symposium in the November 1994 issue of *Review of Economics and Statistics* presents discussion of numerous issues in specification and estimation of models based on simulation of probabilities. Applications that employ simulation techniques for evaluation of multivariate normal integrals are now fairly numerous. See, for example, Hyslop (1999) (Example 17.17), which applies the technique to a panel data application with  $T = 7$ . Example 17.23 develops a five-variate application.

<sup>43</sup>By assuming the coefficient vectors are the same in all periods, we actually obviate the normalization that the diagonal elements of  $\mathbf{R}$  are all equal to one as well. The restriction identifies  $T - 1$  relative variances  $\rho_{tt} = \sigma_T^2/\sigma_T^2$ . This aspect is examined in Greene (2004).

**754 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**
**TABLE 17.20** Estimated Pooled Probit Model

<i>Variable</i>	<i>Estimate</i> <sup>a</sup>	<i>Estimated Standard Errors</i>				<i>Marginal Effects</i>		
		<i>SE(1)</i> <sup>b</sup>	<i>SE(2)</i> <sup>c</sup>	<i>SE(3)</i> <sup>d</sup>	<i>SE(4)</i> <sup>e</sup>	<i>Partial</i>	<i>Std. Err.</i>	<i>t ratio</i>
Constant	-1.960	0.239	0.377	0.230	0.373	—	—	—
Sales	0.177	0.0250	0.0375	0.0222	0.0358	0.0683 <sup>f</sup>	0.0138	4.96
Rel Size	1.072	0.206	0.306	0.142	0.269	0.413 <sup>f</sup>	0.103	4.01
Imports	1.134	0.153	0.246	0.151	0.243	0.437 <sup>f</sup>	0.0938	4.66
FDI	2.853	0.467	0.679	0.402	0.642	1.099 <sup>f</sup>	0.247	4.44
Prod.	-2.341	1.114	1.300	0.715	1.115	-0.902 <sup>f</sup>	0.429	-2.10
Raw Mtl	-0.279	0.0966	0.133	0.0807	0.126	-0.110 <sup>g</sup>	0.0503	-2.18
Inv Good	0.188	0.0404	0.0630	0.0392	0.0628	0.0723 <sup>g</sup>	0.0241	3.00

<sup>a</sup>Recomputed. Only two digits were reported in the earlier paper.

<sup>b</sup>Obtained from results in Bertschek and Lechner, Table 9.

<sup>c</sup>Based on the Avery et al. (1983) GMM estimator.

<sup>d</sup>Square roots of the diagonals of the negative inverse of the Hessian

<sup>e</sup>Based on the cluster estimator.

<sup>f</sup>Coefficient scaled by the density evaluated at the sample means

<sup>g</sup>Computed as the difference in the fitted probability with the dummy variable equal to one, then zero.

on the latent regression

$$y_{it}^* = \beta_1 + \sum_{k=2}^8 x_{k,it} \beta_k + \varepsilon_{it}, \quad y_{it} = 1(y_{it}^* > 0), \quad i = 1, \dots, 1,270, \quad t = 1984, \dots, 1988,$$

where

$y_{it}$  = 1 if a product innovation was realized by firm  $i$  in year  $t$ , 0 otherwise

$x_{2,it}$  = Log of industry sales in DM

$x_{3,it}$  = Import share = ratio of industry imports to (industry sales plus imports)

$x_{4,it}$  = Relative firm size = ratio of employment in business unit to employment in the industry (times 30)

$x_{5,it}$  = FDI share = Ratio of industry foreign direct investment to (industry sales plus imports)

$x_{6,it}$  = Productivity = Ratio of industry value added to industry employment

$x_{7,it}$  = Raw materials sector = 1 if the firm is in this sector

$x_{8,it}$  = Investment goods sector = 1 if the firm is in this sector

The coefficients on import share ( $\beta_3$ ) and FDI share ( $\beta_5$ ) were of particular interest. The objectives of the study were the empirical investigation of innovation and the methodological development of an estimator that could obviate computing the five-variate normal probabilities necessary for a full maximum likelihood estimation of the model.

Table 17.20 presents the single-equation, pooled probit model estimates.<sup>44</sup> Given the structure of the model, the parameter vector could be estimated consistently with any single period's data. Hence, pooling the observations, which produces a mixture of the estimators, will also be consistent. Given the panel data nature of the data set, however, the conventional standard errors from the pooled estimator are dubious. Because the marginal distribution

<sup>44</sup>We are grateful to the authors of this study who have generously loaned us their data for our continued analysis. The data are proprietary and cannot be made publicly available, unlike the other data sets used in our examples.

**TABLE 17.21** Estimated Constrained Multivariate Probit Model (estimated standard errors in parentheses)

Coefficients	Full Maximum Likelihood Using GHK Simulator	Random Effects $\rho = 0.578$ (0.0189)
Constant	-1.797** (0.341)	-2.839 (0.534)
Sales	0.154** (0.0334)	0.245 (0.0523)
Relative size	0.953** (0.160)	1.522 (0.259)
Imports	1.155** (0.228)	1.779 (0.360)
FDI	2.426** (0.573)	3.652 (0.870)
Productivity	-1.578 (1.216)	-2.307 (1.911)
Raw material	-0.292** (0.130)	-0.477 (0.202)
Investment goods	0.224** (0.0605)	0.331 (0.0952)
log-likelihood	-3522.85	-3535.55
<i>Estimated Correlations</i>		
1984, 1985	0.460** (0.0301)	
1984, 1986	0.599** (0.0323)	
1985, 1986	0.643** (0.0308)	
1984, 1987	0.540** (0.0308)	
1985, 1987	0.546** (0.0348)	
1986, 1987	0.610** (0.0322)	
1984, 1988	0.483** (0.0364)	
1985, 1988	0.446** (0.0380)	
1986, 1988	0.524** (0.0355)	
1987, 1988	0.605** (0.0325)	

\* Indicates significant at 95 percent level.

\*\* indicates significant at 99 percent level based on a two-tailed test.

will produce a consistent estimator of the parameter vector, this is a case in which the cluster estimator (see Section 14.8.4) provides an appropriate asymptotic covariance matrix. Note that the standard errors in column SE(4) of the table are considerably higher than the uncorrected ones in columns 1–3.

The pooled estimator is consistent, so the further development of the estimator is a matter of (1) obtaining a more efficient estimator of  $\beta$  and (2) computing estimates of the cross-period correlation coefficients. The FIML estimates of the model can be computed using the GHK simulator.<sup>45</sup> The FIML estimates and the random effects model using the Butler and Moffit (1982) quadrature method are reported in Table 17.21. The correlations reported are based on the FIML estimates. Also noteworthy in Table 17.21 is the divergence of the random effects estimates from the FIML estimates. The log-likelihood function is -3535.55 for the random effects model and -3522.85 for the unrestricted model. The chi-squared statistic for the nine restrictions of the equicorrelation model is 25.4. The critical value from the chi-squared table for nine degrees of freedom is 16.9 for 95 percent and 21.7 for 99 percent significance, so the hypothesis of the random effects model would be rejected.

## 17.6 SUMMARY AND CONCLUSIONS

This chapter has surveyed a large range of techniques for modeling a binary choice variable. The model for choice between two outcomes provides the framework for a

<sup>45</sup>The full computation required about one hour of computing time. Computation of the single-equation (pooled) estimators required only about 1/100 of the time reported by the authors for the same models, which suggests that the evolution of computing technology may play a significant role in advancing the FIML estimators.

**756 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

large proportion of the analysis of microeconomic data. Thus, we have given a very large amount of space to this model in its own right. In addition, many issues in model specification and estimation that appear in more elaborate settings, such as those we will examine in the next chapter, can be formulated as extensions of the binary choice model of this chapter. Binary choice modeling provides a convenient point to study endogeneity in a nonlinear model, issues of nonresponse in panel data sets, and general problems of estimation and inference with longitudinal data. The binary probit model in particular has provided the laboratory case for theoretical econometricians such as those who have developed methods of bias reduction for the fixed effects estimator in dynamic nonlinear models.

We began the analysis with the fundamental parametric probit and logit models for binary choice. Estimation and inference issues such as the computation of appropriate covariance matrices for estimators and partial effects are considered here. We then examined familiar issues in modeling, including goodness of fit and specification issues such as the distributional assumption, heteroscedasticity and missing variables. As in other modeling settings, endogeneity of some right-hand variables presents a substantial complication in the estimation and use of nonlinear models such as the probit model. We examined the problem of endogenous right-hand-side variables, and in two applications, problems of endogenous sampling. The analysis of binary choice with panel data provides a setting to examine a large range of issues that reappear in other applications. We reconsidered the familiar pooled, fixed and random effects estimator estimators, and found that much of the wisdom obtained in the linear case does not carry over to the nonlinear case. The incidental parameters problem, in particular, motivates a considerable amount of effort to reconstruct the estimators of binary choice models. Finally, we considered some multivariate extensions of the probit model. As before, the models are useful in their own right. Once again, they also provide a convenient setting in which to examine broader issues, such as more detailed models of endogeneity nonrandom sampling, and computation requiring simulation.

Chapter 18 will continue the analysis of discrete choice models with three frameworks: unordered multinomial choice, ordered choice, and models for count data. Most of the estimation and specification issues we have examined in this chapter will reappear in these settings.

**Key Terms and Concepts**

- Attributes
- Attrition bias
- Average partial effect
- Binary choice model
- Bivariate probit
- Butler and Moffitt method
- Characteristics
- Choice-based sampling
- Chow test
- Complementary log log model
- Conditional likelihood function
- Control function
- Event count
- Fixed effects model
- Generalized residual
- Goodness of fit measure
- Gumbel model
- Heterogeneity
- Heteroscedasticity
- Incidental parameters problem
- Index function model
- Initial conditions
- Interaction effect
- Inverse probability weighted (IPW)
- Lagrange multiplier test
- Latent regression
- Likelihood equations
- Likelihood ratio test

## CHAPTER 17 ♦ Discrete Choice 757

- Linear probability model
- Logit
- Marginal effects
- Maximum likelihood
- Maximum simulated likelihood (MSL)
- Method of scoring
- Microeometrics
- Minimal sufficient statistic
- Multinomial choice
- Multivariate probit model
- Nonresponse bias
- Ordered choice model
- Persistence
- Probit
- Quadrature
- Qualitative response (QR)
- Quasi-maximum likelihood estimator (QMLE)
- Random effects model
- Random parameters logit model
- Random utility model
- Recursive model
- Robust covariance estimation
- Sample selection bias
- Selection on unobservables
- State dependence
- Tetrachoric correlation
- Unbalanced sample

**Exercises**

1. A binomial probability model is to be based on the following index function model:

$$\begin{aligned}y^* &= \alpha + \beta d + \varepsilon, \\y &= 1, \text{ if } y^* > 0, \\y &= 0 \text{ otherwise.}\end{aligned}$$

The only regressor,  $d$ , is a dummy variable. The data consist of 100 observations that have the following:

		$y$	
		0	1
		0	24 28
$d$		1	32 16

Obtain the maximum likelihood estimators of  $\alpha$  and  $\beta$ , and estimate the asymptotic standard errors of your estimates. Test the hypothesis that  $\beta$  equals zero by using a Wald test (asymptotic  $t$  test) and a likelihood ratio test. Use the probit model and then repeat, using the logit model. Do your results change? (Hint: Formulate the log-likelihood in terms of  $\alpha$  and  $\delta = \alpha + \beta$ .)

2. Suppose that a linear probability model is to be fit to a set of observations on a dependent variable  $y$  that takes values zero and one, and a single regressor  $x$  that varies continuously across observations. Obtain the exact expressions for the least squares slope in the regression in terms of the mean(s) and variance of  $x$ , and interpret the result.
3. Given the data set

$y$	1	0	0	1	1	0	0	1	1	1
$x$	9	2	5	4	6	7	3	5	2	6

estimate a probit model and test the hypothesis that  $x$  is not influential in determining the probability that  $y$  equals one.

4. Construct the Lagrange multiplier statistic for testing the hypothesis that all the slopes (but not the constant term) equal zero in the binomial logit model. Prove that the Lagrange multiplier statistic is  $nR^2$  in the regression of  $(y_i = p)$  on the  $x$ 's, where  $p$  is the sample proportion of 1's.

## 758 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics

5. The following hypothetical data give the participation rates in a particular type of recycling program and the number of trucks purchased for collection by 10 towns in a small mid-Atlantic state:

Town	1	2	3	4	5	6	7	8	9	10
Trucks	160	250	170	365	210	206	203	305	270	340
Participation%	11	74	8	87	62	83	48	84	71	79

The town of Eleven is contemplating initiating a recycling program but wishes to achieve a 95 percent rate of participation. Using a probit model for your analysis,

- a. How many trucks would the town expect to have to purchase to achieve its goal? (*Hint:* You can form the log-likelihood by replacing  $y_i$  with the participation rate (e.g., 0.11 for observation 1) and  $(1 - y_i)$  with 1—the rate in (17-22)).
- b. If trucks cost \$20,000 each, then is a goal of 90 percent reachable within a budget of \$6.5 million? (That is, should they *expect* to reach the goal?)
- c. According to your model, what is the marginal value of the 301st truck in terms of the increase in the percentage participation?
6. A data set consists of  $n = n_1 + n_2 + n_3$  observations on  $y$  and  $x$ . For the first  $n_1$  observations,  $y = 1$  and  $x = 1$ . For the next  $n_2$  observations,  $y = 0$  and  $x = 1$ . For the last  $n_3$  observations,  $y = 0$  and  $x = 0$ . Prove that neither (17-18) nor (17-20) has a solution.
7. Prove (17-20).
8. In the panel data models estimated in Section 17.4, neither the logit nor the probit model provides a framework for applying a Hausman test to determine whether fixed or random effects is preferred. Explain. (*Hint:* Unlike our application in the linear model, the incidental parameters problem persists here.)

### Applications

1. Appendix Table F1 provides Fair's (1978) *Redbook* survey on extramarital affairs. The data are described in Application 1 at the end of Chapter 18 and in Appendix F. The variables in the data set are as follows:

*id* = an identification number

*C* = constant, value = 1

*yrb* = a constructed measure of time spent in extramarital affairs

*v1* = a rating of the marriage, coded 1 to 4

*v2* = age, in years, aggregated

*v3* = number of years married

*v4* = number of children, top coded at 5

*v5* = religiosity, 1 to 4, 1 = not, 4 = very

*v6* = education, coded 9, 12, 14, 16, 17, 20,

*v7* = occupation

*v8* = husband's occupation

and three other variables that are not used. The sample contains a survey of 6,366 married women, conducted by *Redbook* magazine. For this exercise, we will analyze,

first, the binary variable

$$A = 1 \text{ if } yrb > 0, 0 \text{ otherwise.}$$

The regressors of interest are  $v_1$  to  $v_8$ ; however, not necessarily all of them belong in your model. Use these data to build a binary choice model for  $A$ . Report all computed results for the model. Compute the marginal effects for the variables you choose. Compare the results you obtain for a probit model to those for a logit model. Are there any substantial differences in the results for the two models?

## 18

# DISCRETE CHOICES AND EVENT COUNTS



## 18.1 INTRODUCTION

Chapter 17 presented most of the econometric issues that arise in analyzing discrete dependent variables, including specification, estimation, inference, and a variety of variations on the basic model. All of these were developed in the context of a model of binary choice, the choice between two alternatives. This chapter will use those results in extending the choice model to three specific settings:

**Multinomial Choice:** The individual chooses among more than two choices, once again, making the choice that provides the greatest utility. Applications include the choice among political candidates, how to commute to work, where to live, or what brand of car, appliance, or food product to buy.

**Ordered Choice:** The individual reveals the strength of their preferences with respect to a single outcome. Familiar cases involve survey questions about strength of feelings about a particular commodity such as a movie, a book, or a consumer product, or self-assessments of social outcomes such as health in general or self-assessed well-being. Although preferences will probably vary continuously in the space of individual utility, the expression of those preferences for purposes of analyses is given in a discrete outcome on a scale with a limited number of choices, such as the typical five-point scale used in marketing surveys.

**Event Counts:** The observed outcome is a count of the number of occurrences. In many cases, this is similar to the preceding settings in that the “dependent variable” measures an individual choice, such as the number of visits to the physician or the hospital, the number of derogatory reports in one’s credit history, or the number of visits to a particular recreation site. In other cases, the event count might be the outcome of some less focused natural process, such as incidence of a disease in a population or the number of defects per unit of time in a production process, the number of traffic accidents that occur at a particular location per month, or the number of messages that arrive at a switchboard per unit of time over the course of a day. In this setting, we will be doing a more familiar sort of regression modeling.

Most of the methodological underpinnings needed to analyze these cases were presented in Chapter 17. In this chapter, we will be able to develop variations on these basic model types that accommodate different choice situations. As in Chapter 17, we are focused on models with discrete outcomes, so the analysis is framed in terms of models of the probabilities attached to those outcomes.

## 18.2 MODELS FOR UNORDERED MULTIPLE CHOICES

Some studies of multiple-choice settings include the following:

1. Hensher (1986, 1991), McFadden (1974), and many others have analyzed the travel mode of urban commuters. In Greene (2007b), Hensher and Greene analyze commuting between Sydney and Melbourne by a sample of individuals who choose among air, train, bus, and car as the mode of travel.
2. Schmidt and Strauss (1975a, b) and Boskin (1974) have analyzed occupational choice among multiple alternatives.
3. Rossi and Allenby (1999, 2003) studied consumer brand choices in a repeated choice (panel data) model.
4. Train (2003) studied the choice of electricity supplier by a sample of California electricity customers.
5. Hensher, Rose, and Greene (2006) analyzed choices of automobile models by a sample of consumers offered a hypothetical menu of features.

In each of these cases, there is a single decision among two or more alternatives. In this and the next section, we will encounter two broad types of multinomial choice sets, **unordered choice models** and **ordered choices**. All of the choice sets listed are unordered. In contrast, a bond rating or a preference scale is, by design, a ranking; that is, its purpose. Quite different techniques are used for the two types of models. We will examine models for ordered choices in Section 18.3. This section will examine models for unordered choice sets. General references on the topics discussed here include Hensher, Louviere, and Swait (2000), Train (2009), and Hensher, Rose, and Greene (2006).

### 18.2.1 RANDOM UTILITY BASIS OF THE MULTINOMIAL LOGIT MODEL

Unordered choice models can be motivated by a random utility model. For the  $i$ th consumer faced with  $J$  choices, suppose that the utility of choice  $j$  is

$$U_{ij} = \mathbf{z}'_i \boldsymbol{\theta} + \varepsilon_{ij}.$$

If the consumer makes choice  $j$  in particular, then we assume that  $U_{ij}$  is the maximum among the  $J$  utilities. Hence, the statistical model is driven by the probability that choice  $j$  is made, which is

$$\text{Prob}(U_{ij} > U_{ik}) \quad \text{for all other } k \neq j.$$

The model is made operational by a particular choice of distribution for the disturbances. As in the binary choice case, two models are usually considered, logit and probit. Because of the need to evaluate multiple integrals of the normal distribution, the probit model has found rather limited use in this setting. The logit model, in contrast, has been widely used in many fields, including economics, market research, politics, finance, and transportation engineering. Let  $Y_i$  be a random variable that indicates the choice made. McFadden (1974a) has shown that if (and only if) the  $J$  disturbances are independent

**762 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**


and identically distributed with Gumbel (type 1 extreme value) distribution,

$$F(\varepsilon_{ij}) = \exp(-\exp(-\varepsilon_{ij})), \quad (18-1)$$

then

$$\text{Prob}(Y_i = j) = \frac{\exp(\mathbf{z}'_{ij}\boldsymbol{\theta})}{\sum_{j=1}^J \exp(\mathbf{z}'_{ij}\boldsymbol{\theta})}, \quad (18-2)$$

which leads to what is called the **conditional logit model**. (It is often labeled the **multinomial logit model**, but this wording conflicts with the usual name for the model discussed in the next section, which differs slightly. Although the distinction turns out to be purely artificial, we will maintain it for the present.)

Utility depends on  $\mathbf{z}_{ij}$ , which includes aspects specific to the individual as well as to the choices. It is useful to distinguish them. Let  $\mathbf{z}_{ij} = [\mathbf{x}_{ij}, \mathbf{w}_i]$  and partition  $\boldsymbol{\theta}$  conformably into  $[\boldsymbol{\beta}', \boldsymbol{\alpha}']'$ . Then  $\mathbf{x}_{ij}$  varies across the choices and possibly across the individual as well. The components of  $\mathbf{x}_{ij}$  are typically called the **attributes** of the choices. But  $\mathbf{w}_i$  contains the **characteristics** of the individual and is, therefore, the same for all choices. If we incorporate this fact in the model, then (18-2) becomes

$$\text{Prob}(Y_i = j) = \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{w}'_i\boldsymbol{\alpha})}{\sum_{j=1}^J \exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{w}'_i\boldsymbol{\alpha})} = \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta}) \exp(\mathbf{w}'_i\boldsymbol{\alpha})}{\left[ \sum_{j=1}^J \exp(\mathbf{x}'_{ij}\boldsymbol{\beta}) \right] \exp(\mathbf{w}'_i\boldsymbol{\alpha})}. \quad (18-3)$$

Terms that do not vary across alternatives—that is, those specific to the individual—fall out of the probability. This is as expected in a model that compares the utilities of the alternatives.

For example, in a model of a shopping center choice by individuals in various cities that depends on the number of stores at the mall,  $S_{ij}$ , the distance from the central business district,  $D_{ij}$  and the shoppers' incomes,  $I_i$ , the utilities for three choices would be

$$U_{i1} = D_{i1}\beta_1 + S_{i1}\beta_2 + \alpha + \gamma I_i + \varepsilon_{i1};$$

$$U_{i2} = D_{i2}\beta_1 + S_{i2}\beta_2 + \alpha + \gamma I_i + \varepsilon_{i2};$$

$$U_{i3} = D_{i3}\beta_1 + S_{i3}\beta_2 + \alpha + \gamma I_i + \varepsilon_{i3}.$$

The choice of alternative 1, for example, reveals that

$$U_{i1} - U_{i2} = (D_{i1} - D_{i2})\beta_1 + (S_{i1} - S_{i2})\beta_2 + (\varepsilon_{i1} - \varepsilon_{i2}) > 0 \text{ and}$$

$$U_{i1} - U_{i3} = (D_{i1} - D_{i3})\beta_1 + (S_{i1} - S_{i3})\beta_2 + (\varepsilon_{i1} - \varepsilon_{i3}) > 0.$$

The constant term and *Income* have fallen out of the comparison. The result follows from the fact that random utility model is ultimately based on comparisons of pairs of alternatives, not the alternatives themselves. Evidently, if the model is to allow individual specific effects, then it must be modified. One method is to create a set of dummy variables (alternative specific constants),  $A_j$ , for the choices and multiply each of them by the common  $\mathbf{w}$ . We then allow the coefficients on these choice invariant characteristics to vary across the choices instead of the characteristics. Analogously to the linear model, a complete set of interaction terms creates a singularity, so one of them must be

CHAPTER 18 ♦ Discrete Choices and Event Counts **763**

dropped. For this example, the matrix of attributes and characteristics would be

$$\mathbf{Z}_i = \begin{bmatrix} S_{i1} & D_{i1} & 1 & 0 & I_i & 0 \\ S_{i2} & D_{i2} & 0 & 1 & 0 & I_i \\ S_{i3} & D_{i3} & 0 & 0 & 0 & 0 \end{bmatrix}$$

The probabilities for this model would be

$$\text{Prob}(Y_i = j | \mathbf{Z}_i) = \frac{\exp\left(\frac{Stores_{ij}\beta_1 + Distance_{ij}\beta_2}{A_1\alpha_1 + A_2\alpha_2 + A_3\alpha_3}\right)}{\sum_{j=1}^3 \exp\left(\frac{Stores_{ij}\beta_1 + Distance_{ij}\beta_2}{A_1\alpha_1 + A_2\alpha_2 + A_3\alpha_3}\right)}, \quad \alpha_3 = 0.$$

### 18.2.2 THE MULTINOMIAL LOGIT MODEL

To set up the model that applies when data are individual specific, it will help to consider an example. Schmidt and Strauss (1975a, b) estimated a model of occupational choice based on a sample of 1,000 observations drawn from the Public Use Sample for three years: 1960, 1967, and 1970. For each sample, the data for each individual in the sample consist of the following:

1. *Occupation*: 0 = menial, 1 = blue collar, 2 = craft, 3 = white collar, 4 = professional. (Note the slightly different numbering convention, starting at zero, which is standard.)
2. *Characteristics*: constant, education, experience, race, sex.

The model for occupational choice is

$$\text{Prob}(Y_i = j | \mathbf{w}_i) = \frac{\exp(\mathbf{w}'_i \boldsymbol{\alpha}_j)}{\sum_{j=0}^4 \exp(\mathbf{w}'_i \boldsymbol{\alpha}_j)}, \quad j = 0, 1, \dots, 4. \quad (18-4)$$

(The binomial logit model in Section 17.3 is conveniently produced as the special case of  $J = 1$ .)

The model in (18-4) is a multinomial logit model.<sup>1</sup> The estimated equations provide a set of probabilities for the  $J + 1$  choices for a decision maker with characteristics  $\mathbf{w}_i$ . Before proceeding, we must remove an indeterminacy in the model. If we define  $\boldsymbol{\alpha}_j^* = \boldsymbol{\alpha}_j + \mathbf{q}$  for any vector  $\mathbf{q}$ , then recomputing the probabilities defined later using  $\boldsymbol{\alpha}_j^*$  instead of  $\boldsymbol{\alpha}_j$  produces the identical set of probabilities because all the terms involving  $\mathbf{q}$  drop out. A convenient normalization that solves the problem is  $\boldsymbol{\alpha}_0 = \mathbf{0}$ . (This arises because the probabilities sum to one, so only  $J$  parameter vectors are needed to determine the  $J + 1$  probabilities.) Therefore, the probabilities are

$$\text{Prob}(Y_i = j | \mathbf{w}_i) = P_{ij} = \frac{\exp(\mathbf{w}'_i \boldsymbol{\alpha}_j)}{1 + \sum_{k=1}^J \exp(\mathbf{w}'_i \boldsymbol{\alpha}_k)}, \quad j = 0, 1, \dots, J, \quad \boldsymbol{\alpha}_0 = \mathbf{0}. \quad (18-5)$$

<sup>1</sup>Nerlove and Press (1973).

## 764 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics

The form of the binomial model examined in Section 17.3 results if  $J = 1$ . The model implies that we can compute  $J$  **log-odds**

$$\ln \left[ \frac{P_{ij}}{P_{ik}} \right] = \mathbf{w}'_i(\boldsymbol{\alpha}_j - \boldsymbol{\alpha}_k) = \mathbf{w}'_i \boldsymbol{\alpha}_j \quad \text{if } k = 0.$$

From the point of view of estimation, it is useful that the odds ratio,  $P_{ij}/P_{ik}$ , does not depend on the other choices, which follows from the independence of the disturbances in the original model. From a behavioral viewpoint, this fact is not very attractive. We shall return to this problem in Section 18.2.4.

The log-likelihood can be derived by defining, for each individual,  $d_{ij} = 1$  if alternative  $j$  is chosen by individual  $i$ , and 0 if not, for the  $J + 1$  possible outcomes. Then, for each  $i$ , one and only one of the  $d_{ij}$ 's is 1. The log-likelihood is a generalization of that for the binomial probit or logit model:

$$\ln L = \sum_{i=1}^n \sum_{j=0}^J d_{ij} \ln \text{Prob}(Y_i = j | \mathbf{w}_i).$$

The derivatives have the characteristically simple form

$$\frac{\partial \ln L}{\partial \boldsymbol{\alpha}_j} = \sum_{i=1}^n (d_{ij} - P_{ij}) \mathbf{w}_i \quad \text{for } j = 1, \dots, J.$$

The exact second derivatives matrix has  $J^2 K \times K$  blocks,<sup>2</sup>

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\alpha}_j \partial \boldsymbol{\alpha}_l} = - \sum_{i=1}^n P_{ij} [\mathbf{1}(j = l) - P_{il}] \mathbf{w}_i \mathbf{w}'_i,$$

where  $\mathbf{1}(j = l)$  equals 1 if  $j$  equals  $l$  and 0 if not. Because the Hessian does not involve  $d_{ij}$ , these are the expected values, and Newton's method is equivalent to the method of scoring. It is worth noting that the number of parameters in this model proliferates with the number of choices, which is inconvenient because the typical cross section sometimes involves a fairly large number of regressors.

The coefficients in this model are difficult to interpret. It is tempting to associate  $\boldsymbol{\alpha}_j$  with the  $j$ th outcome, but that would be misleading. By differentiating (18-5), we find that the partial effects of the characteristics on the probabilities are

$$\delta_{ij} = \frac{\partial P_{ij}}{\partial \mathbf{w}_i} = P_{ij} \left[ \boldsymbol{\alpha}_j - \sum_{k=0}^J P_{ik} \boldsymbol{\alpha}_k \right] = P_{ij} [\boldsymbol{\alpha}_j - \bar{\boldsymbol{\alpha}}]. \quad (18-6)$$

Therefore, every subvector of  $\boldsymbol{\alpha}$  enters every partial effect, both through the probabilities and through the weighted average that appears in  $\delta_{ij}$ . These values can be computed from the parameter estimates. Although the usual focus is on the coefficient estimates, equation (18-6) suggests that there is at least some potential for confusion. Note, for example, that for any particular  $w_{ik}$ ,  $\partial P_{ij}/\partial w_{ik}$  need not have the same sign as  $\alpha_{jk}$ . Standard errors can be estimated using the delta method. (See Section 4.4.4.) For purposes of the computation, let  $\boldsymbol{\alpha} = [\mathbf{0}, \boldsymbol{\alpha}'_1, \boldsymbol{\alpha}'_2, \dots, \boldsymbol{\alpha}'_J]'$ . We include the fixed  $\mathbf{0}$  vector for outcome 0 because although  $\boldsymbol{\alpha}_0 = \mathbf{0}$ ,  $\delta_{i0} = -P_{i0} \bar{\boldsymbol{\alpha}}$ , which is not  $\mathbf{0}$ . Note as well that

<sup>2</sup>If the data were in the form of proportions, such as market shares, then the appropriate log-likelihood and derivatives are  $\sum_i \sum_j n_i p_{ij}$  and  $\sum_i \sum_j n_i (p_{ij} - P_{ij}) \mathbf{w}_i$ , respectively. The terms in the Hessian are multiplied by  $n_i$ .

CHAPTER 18 ♦ Discrete Choices and Event Counts **765**

Asy. Cov[ $\hat{\alpha}_0, \hat{\alpha}_j$ ] =  $\mathbf{0}$  for  $j = 0, \dots, J$ . Then

$$\text{Asy. Var}[\hat{\delta}_{ij}] = \sum_{l=0}^J \sum_{m=0}^J \left( \frac{\partial \delta_{ij}}{\partial \alpha'_l} \right) \text{Asy. Cov}[\hat{\alpha}'_l, \hat{\alpha}'_m] \left( \frac{\partial \delta'_{ij}}{\partial \alpha_m} \right),$$

$$\frac{\partial \delta_{ij}}{\partial \alpha'_l} = [\mathbf{1}(j=l) - P_{il}][P_{ij}\mathbf{I} + \delta_{ij}\mathbf{w}'_i] - P_{ij}[\delta_{il}\mathbf{w}'_i].$$

Finding adequate fit measures in this setting presents the same difficulties as in the binomial models. As before, it is useful to report the log-likelihood. If the model contains no covariates and no constant term, then the log-likelihood will be

$$\ln L_c = \sum_{j=0}^J n_j \ln \left( \frac{1}{J+1} \right)$$


where  $n_j$  is the number of individuals who choose outcome  $j$ . If the characteristic vector includes only a constant term, then the restricted log-likelihood is

$$\ln L_0 = \sum_{j=0}^J n_j \ln \left( \frac{n_j}{n} \right) = \sum_{j=0}^J n_j \ln p_j,$$

where  $p_j$  is the sample proportion of observations that make choice  $j$ . A useful table will give a listing of hits and misses of the prediction rule “predict  $Y_i = j$  if  $\hat{P}_{ij}$  is the maximum of the predicted probabilities.”<sup>3</sup>

**Example 18.1 Hollingshead Scale of Occupations**

Fair's (1977) study of extramarital affairs is based on a cross section of 601 responses to a survey by *Psychology Today*. One of the covariates is a category of occupations on a seven-point scale, the Hollingshead (1975) scale. [See, also, Bornstein and Bradley (2003).] The Hollingshead scale is intended to be a measure on a prestige scale, a fact which we'll ignore (or disagree with) for the present. The seven levels on the scale are, broadly,

1. Higher executives
  2. Managers and proprietors of medium-sized businesses
  3. Administrative personnel and owners of small businesses
  4. Clerical and sales workers and technicians
  5. Skilled manual employees
  6. Machine operators and semiskilled employees
  7. Unskilled employees
- 

Among the other variables in the data set are *Age*, *Sex*, and *Education*. The data are given in Appendix Table F18.1. Table 18.1 lists estimates of a multinomial logit model. (We emphasize that the data are a self-selected sample of *Psychology Today* readers in 1976, so it is unclear what contemporary population would be represented. The following serves as an uncluttered numerical example that readers could reproduce. Note, as well, that at least by some viewpoint, the outcome for this experiment is ordered.) The log-likelihood for the model is  $-770.28141$  while that for the model with only the constant terms is  $-982.20533$ . The likelihood ratio statistic for the hypothesis that all 18 coefficients of the model are zero is 423.85, which is far larger than the critical value of 28.87. In the estimated parameters, it appears that only gender is consistently statistically significant. However, it is unclear how

<sup>3</sup>It is common for this rule to predict all observation with the same value in an unbalanced sample or a model with little explanatory power. This is not a contradiction of an estimated model with many “significant” coefficients, because the coefficients are not estimated so as to maximize the number of correct predictions.

**766 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**
**TABLE 18.1** Estimated Multinomial Logit Model for Occupation (*t* ratios in parentheses)

	$\alpha_0$	$\alpha_1$	$\alpha_2$	$\alpha_3$	$\alpha_4$	$\alpha_5$	$\alpha_6$
<i>Parameters</i>							
Constant	0.0 (0.0)	3.1506 (1.14)	2.0156 (1.28)	-1.9849 (-1.38)	-6.6539 (-5.49)	-15.0779 (-9.18)	-12.8919 (-4.61)
Age	0.0 (0.0)	-0.0244 (-0.73)	-0.0361 (-1.64)	-0.0123 (-0.63)	0.0038 (0.25)	0.0225 (1.22)	0.0588 (1.92)
Sex	0.0 (0.0)	6.2361 (5.08)	4.6294 (4.39)	4.9976 (4.82)	4.0586 (3.98)	5.2086 (5.02)	5.8457 (4.57)
Education	0.0 (0.0)	-0.4391 (-2.62)	-0.1661 (-1.75)	0.0684 (0.79)	0.4288 (5.92)	0.8149 (8.56)	0.4506 (2.92)
<i>Partial Effects</i>							
Age	-0.0001 (-0.19)	-0.0002 (-0.92)	-0.0028 (-2.23)	-0.0022 (-1.15)	0.0006 (0.23)	0.0036 (1.89)	0.0011 (1.90)
Sex	-0.2149 (-4.24)	0.0164 (1.98)	0.0233 (1.00)	0.1041 (2.87)	-0.1264 (-2.15)	0.1667 (4.20)	0.0308 (2.35)
Education	-0.0187 (-2.22)	-0.0069 (-2.31)	-0.0387 (-6.29)	-0.0460 (-5.1)	0.0278 (2.12)	0.0810 (8.61)	0.0015 (0.56)

to interpret the fact that *Education* is significant in some of the parameter vectors and not others. The partial effects give a similarly unclear picture, though in this case, the effect can be associated with a particular outcome. However, we note that the implication of a test of significance of a partial effect in this model is itself ambiguous. For example, *Education* is not “significant” in the partial effect for outcome 6, though the coefficient on *Education* in  $\alpha_6$  is. This is an aspect of modeling with multinomial choice models that calls for careful interpretation by the model builder.

### 18.2.3 THE CONDITIONAL LOGIT MODEL

When the data consist of choice-specific attributes instead of individual-specific characteristics, the natural model formulation would be

$$\text{Prob}(Y_i = j | \mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iJ}) = \text{Prob}(Y_i = j | \mathbf{X}_i) = P_{ij} = \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta})}{\sum_{j=1}^J \exp(\mathbf{x}'_{ij}\boldsymbol{\beta})}. \quad (18-7)$$

Here, in accordance with the convention in the literature, we let  $j = 1, 2, \dots, J$  for a total of  $J$  alternatives. The model is otherwise essentially the same as the multinomial logit. Even more care will be required in interpreting the parameters, however. Once again, an example will help to focus ideas.

In this model, the coefficients are not directly tied to the marginal effects. The marginal effects for continuous variables can be obtained by differentiating (18-7) with respect to a particular  $\mathbf{x}_m$  to obtain

$$\frac{\partial P_{ij}}{\partial \mathbf{x}_{im}} = [P_{ij}(\mathbf{1}(j = m) - P_{im})]\boldsymbol{\beta}, \quad m = 1, \dots, J.$$

It is clear that through its presence in  $P_{ij}$  and  $P_{im}$ , every attribute set  $\mathbf{x}_m$  affects all the probabilities. Hensher (1991) suggests that one might prefer to report elasticities of the probabilities. The effect of attribute  $k$  of choice  $m$  on  $P_{ij}$  would be

$$\frac{\partial \ln P_j}{\partial \ln x_{mk}} = x_{mk}[\mathbf{1}(j = m) - P_{im}]\beta_k.$$

## CHAPTER 18 ♦ Discrete Choices and Event Counts 767

Because there is no ambiguity about the scale of the probability itself, whether one should report the derivatives or the elasticities is largely a matter of taste.

Estimation of the conditional logit model is simplest by Newton's method or the method of scoring. The log-likelihood is the same as for the multinomial logit model. Once again, we define  $d_{ij} = 1$  if  $Y_i = j$  and 0 otherwise. Then

$$\ln L = \sum_{i=1}^n \sum_{j=1}^J d_{ij} \ln \text{Prob}(Y_i = j).$$

Market share and frequency data are common in this setting. If the data are in this form, then the only change needed is, once again, to define  $d_{ij}$  as the proportion or frequency.

Because of the simple form of  $L$ , the gradient and Hessian have particularly convenient forms: Let  $\bar{\mathbf{x}}_i = \sum_{j=1}^J P_{ij} \mathbf{x}_{ij}$ . Then,

$$\begin{aligned} \frac{\partial \log L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^n \sum_{j=1}^J d_{ij} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i), \\ \frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} &= - \sum_{i=1}^n \sum_{j=1}^J P_{ij} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i) (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)', \end{aligned} \tag{18-8}$$

The usual problems of fit measures appear here. The log-likelihood ratio and tabulation of actual versus predicted choices will be useful. There are two possible constrained log-likelihoods. The model cannot contain a constant term, so the constraint  $\boldsymbol{\beta} = \mathbf{0}$  renders all probabilities equal to  $1/J$ . The constrained log-likelihood for this constraint is then  $L_c = -n \ln J$ . Of course, it is unlikely that this hypothesis would fail to be rejected. Alternatively, we could fit the model with only the  $J - 1$  choice-specific constants, which makes the constrained log-likelihood the same as in the multinomial logit model,  $\ln L_0^* = \sum_j n_j \ln p_j$  where, as before,  $n_j$  is the number of individuals who choose alternative  $j$ .

### 18.2.4 THE INDEPENDENCE FROM IRRELEVANT ALTERNATIVES ASSUMPTION

We noted earlier that the odds ratios in the multinomial logit or conditional logit models are independent of the other alternatives. This property is convenient as regards estimation, but it is not a particularly appealing restriction to place on consumer behavior. The property of the logit model whereby  $P_{ij}/P_{im}$  is independent of the remaining probabilities is called the **independence from irrelevant alternatives (IIA)**.

The independence assumption follows from the initial assumption that the disturbances are independent and homoscedastic. Later we will discuss several models that have been developed to relax this assumption. Before doing so, we consider a test that has been developed for testing the validity of the assumption. Hanan and McFadden (1984) suggest that if a subset of the choice set truly is irrelevant, hitting it from the model altogether will not change parameter estimates systematically. Exclusion of these choices will be inefficient but will not lead to inconsistency. But if the remaining odds ratios are not truly independent from these alternatives, then the parameter estimates obtained when these choices are excluded will be inconsistent. This observation is the

## 768 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics

usual basis for Hausman's specification test. The statistic is

$$\chi^2 = (\hat{\beta}_s - \hat{\beta}_f)' [\hat{\mathbf{V}}_s - \hat{\mathbf{V}}_f]^{-1} (\hat{\beta}_s - \hat{\beta}_f),$$

where  $s$  indicates the estimators based on the restricted subset,  $f$  indicates the estimator based on the full set of choices, and  $\hat{\mathbf{V}}_s$  and  $\hat{\mathbf{V}}_f$  are the respective estimates of the asymptotic covariance matrices. The statistic has a limiting chi-squared distribution with  $K$  degrees of freedom.<sup>4</sup>

### 18.2.5 NESTED LOGIT MODELS

If the independence from irrelevant alternatives test fails, then an alternative to the multinomial logit model will be needed. A natural alternative is a multivariate probit model:

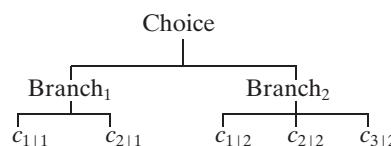
$$U_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \varepsilon_{ij}, \quad j = 1, \dots, J, [\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iJ}] \sim N[\mathbf{0}, \Sigma]. \quad (18-9)$$

We had considered this model earlier but found that as a general model of consumer choice, its failings were the practical difficulty of computing the multinormal integral and estimation of an unrestricted correlation matrix. Hausman and Wise (1978) point out that for a model of consumer choice, the probit model may not be as impractical as it might seem. First, for  $J$  choices, the comparisons implicit in  $U_{ij} > U_{im}$  for  $m \neq j$  involve the  $J - 1$  differences,  $\varepsilon_j - \varepsilon_m$ . Thus, starting with a  $J$ -dimensional problem, we need only consider derivatives of  $(J - 1)$ -order probabilities. Therefore, to come to a concrete example, a model with four choices requires only the evaluation of bivariate normal integrals, which, albeit still complicated to estimate, is well within the received technology. For larger models, however, other specifications have proved more useful.

One way to relax the homoscedasticity assumption in the conditional logit model that also provides an intuitively appealing structure is to group the alternatives into subgroups that allow the variance to differ across the groups while maintaining the IIA assumption within the groups. This specification defines a **nested logit model**. To fix ideas, it is useful to think of this specification as a two- (or more) level choice problem (although, once again, the model arises as a modification of the stochastic specification in the original conditional logit model, not necessarily as a model of behavior). Suppose, then, that the  $J$  alternatives can be divided into  $B$  subgroups (branches) such that the choice set can be written

$$[c_1, \dots, c_J] = [(c_{1|1}, \dots, c_{J_1|1}), (c_{1|2}, \dots, c_{J_2|2}), \dots, (c_{1|B}, \dots, c_{J_B|B})].$$

Logically, we may think of the choice process as that of choosing among the  $B$  choice sets and then making the specific choice within the chosen set. This method produces a tree structure, which for two branches and, say, five choices (twigs) might look as follows:



<sup>4</sup>McFadden (1987) shows how this hypothesis can also be tested using a Lagrange multiplier test.

## CHAPTER 18 ♦ Discrete Choices and Event Counts 769

Suppose as well that the data consist of observations on the attributes of the choices  $\mathbf{x}_{ij|b}$  and attributes of the choice sets  $\mathbf{z}_{ib}$ .

To derive the mathematical form of the model, we begin with the unconditional probability

$$\text{Prob}[twig_j, branch_b] = P_{ijb} = \frac{\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta} + \mathbf{z}'_{ib}\boldsymbol{\gamma})}{\sum_{b=1}^B \sum_{j=1}^{J_b} \exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta} + \mathbf{z}'_{ib}\boldsymbol{\gamma})}.$$

Now write this probability as

$$\begin{aligned} P_{ijb} &= P_{ij|b} P_b \\ &= \left( \frac{\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})}{\sum_{j=1}^{J_b} \exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})} \right) \left( \frac{\exp(\mathbf{z}'_{ib}\boldsymbol{\gamma})}{\sum_{l=1}^L \exp(\mathbf{z}'_{ib}\boldsymbol{\gamma})} \right) \frac{\left( \sum_{j=1}^{J_b} \exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta}) \right) \left( \sum_{l=1}^L \exp(\mathbf{z}'_{ib}\boldsymbol{\gamma}) \right)}{\left( \sum_{l=1}^L \sum_{j=1}^{J_l} \exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta} + \mathbf{z}'_{ib}\boldsymbol{\gamma}) \right)}. \end{aligned}$$

Define the **inclusive value** for the  $l$ th branch as

$$IV_{ib} = \ln \left( \sum_{j=1}^{J_b} \exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta}) \right).$$

Then, after canceling terms and using this result, we find

$$P_{ij|b} = \frac{\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})}{\sum_{j=1}^{J_b} \exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})} \quad \text{and} \quad P_b = \frac{\exp[\tau_b(\mathbf{z}'_{ib}\boldsymbol{\gamma} + IV_{ib})]}{\sum_{b=1}^B \exp[\tau_b(\mathbf{z}'_{ib}\boldsymbol{\gamma} + IV_{ib})]},$$

where the new parameters  $\tau_l$  must equal 1 to produce the original model. Therefore, we use the restriction  $\tau_l = 1$  to recover the conditional logit model, and the preceding equation just writes this model in another form. The nested logit model arises if this restriction is relaxed. The inclusive value coefficients, unrestricted in this fashion, allow the model to incorporate some degree of heteroscedasticity. Within each branch, the IIA restriction continues to hold. The equal variance of the disturbances within the  $j$ th branch are now<sup>5</sup>

$$\sigma_b^2 = \frac{\pi^2}{6\tau_b}. \tag{18-10}$$

With  $\tau_j = 1$ , this reverts to the basic result for the multinomial logit model.

As usual, the coefficients in the model are not directly interpretable. The derivatives that describe covariation of the attributes and probabilities are

$$\begin{aligned} &\frac{\partial \ln \text{Prob}[choice = m, branch = b]}{\partial x(k) \text{ in choice } M \text{ and branch } B} \\ &= \{\mathbf{1}(b = B)[\mathbf{1}(m = M) - P_{M|B}] + \tau_B[\mathbf{1}(b = B) - P_B]P_M | B\}\boldsymbol{\beta}_k. \end{aligned}$$

The nested logit model has been extended to three and higher levels. The complexity of the model increases rapidly with the number of levels. But the model has been found to be extremely flexible and is widely used for modeling consumer choice in the marketing and transportation literatures, to name a few.

---

<sup>5</sup>See Hensher, Louviere, and Swait (2000). See Greene and Hensher (2002) for alternative formulations of the nested logit model.

## 770 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

There are two ways to estimate the parameters of the nested logit model. A **limited information**, two-step maximum likelihood approach can be done as follows:

1. Estimate  $\beta$  by treating the choice within branches as a simple conditional logit model.
2. Compute the inclusive values for all the branches in the model. Estimate  $\gamma$  and the  $\tau$  parameters by treating the choice among branches as a conditional logit model with attributes  $\mathbf{z}_{ib}$  and  $I_{ib}$ .

Because this approach is a two-step estimator, the estimate of the asymptotic covariance matrix of the estimates at the second step must be corrected. [See Section 14.7 and McFadden (1984).] For **full information maximum likelihood** (FIML) estimation of the model, the log-likelihood is

$$\ln L = \sum_{i=1}^n \ln [\text{Prob}(twig | branch)_i \times \text{Prob}(branch)_i].$$

[See Hensher (1986, 1991) and Greene (2007a).] The information matrix is not block diagonal in  $\beta$  and  $(\gamma, \tau)$ , so FIML estimation will be more efficient than two-step estimation. The FIML estimator is now available in several commercial computer packages. The two-step estimator is rarely used in current research.

To specify the nested logit model, it is necessary to partition the choice set into branches. Sometimes there will be a natural partition, such as in the example given by Maddala (1983) when the choice of residence is made first by community, then by dwelling type within the community. In other instances, however, the partitioning of the choice set is ad hoc and leads to the troubling possibility that the results might be dependent on the branches so defined. (Many studies in this literature present several sets of results based on different specifications of the tree structure.) There is no well-defined testing procedure for discriminating among tree structures, which is a problematic aspect of the model.

### 18.2.6 THE MULTINOMIAL PROBIT MODEL

A natural alternative model that relaxes the independence restrictions built into the multinomial logit (MNL) model is the **multinomial probit model (MNP)**. The structural equations of the MNP model are

$$U_{ij} = \mathbf{x}'_{ij}\beta + \varepsilon_{ij}, \quad j = 1, \dots, J, \quad [\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iJ}] \sim N[\mathbf{0}, \Sigma].$$

The term in the log-likelihood that corresponds to the choice of alternative  $q$  is

$$\text{Prob}[\text{choice}_{iq}] = \text{Prob}[U_{iq} > U_{ij}, \quad j = 1, \dots, J, \quad j \neq q].$$

The probability for this occurrence is

$$\text{Prob}[\text{choice}_{iq}] = \text{Prob}[\varepsilon_{i1} - \varepsilon_{iq} < (\mathbf{x}_{iq} - \mathbf{x}_{i1})'\beta, \dots, \varepsilon_{iJ} - \varepsilon_{iq} < (\mathbf{x}_{iq} - \mathbf{x}_{iJ})'\beta]$$

for the  $J - 1$  other choices, which is a cumulative probability from a  $(J - 1)$ -variate normal distribution. Because we are only making comparisons, one of the variances in this  $J - 1$  variate structure—that is, one of the diagonal elements in the reduced  $\Sigma$ —must be normalized to 1.0. Because only comparisons are ever observable in this model, for identification,  $J - 1$  of the covariances must also be normalized, to zero. The

CHAPTER 18 ♦ Discrete Choices and Event Counts **771**

MNP model allows an unrestricted  $(J - 1) \times (J - 1)$  correlation structure and  $J - 2$  free standard deviations for the disturbances in the model. (Thus, a two-choice model returns to the univariate probit model of Section 17.2.) For more than two choices, this specification is far more general than the MNL model, which assumes that  $\Sigma = \mathbf{I}$ . (The scaling is absorbed in the coefficient vector in the MNL model.) It adds the unrestricted correlations to the heteroscedastic model of the previous section.

The main obstacle to implementation of the MNP model has been the difficulty in computing the multivariate normal probabilities for any dimensionality higher than 2. Recent results on accurate simulation of multinormal integrals, however, have made estimation of the MNP model feasible. (See Section 15.6.2.b and a symposium in the November 1994 issue of the *Review of Economics and Statistics*.) Yet some practical problems remain. Computation is exceedingly time consuming. It is also necessary to ensure that  $\Sigma$  remain a positive definite matrix. One way often suggested is to construct the Cholesky decomposition of  $\Sigma$ ,  $\mathbf{L}\mathbf{L}'$ , where  $\mathbf{L}$  is a lower triangular matrix, and estimate the elements of  $\mathbf{L}$ . The normalizations and zero restrictions can be imposed by making the last row of the  $J \times J$  matrix  $\Sigma$  equal  $(0, 0, \dots, 1)$  and using  $\mathbf{L}\mathbf{L}'$  to create the upper  $(J - 1) \times (J - 1)$  matrix. The additional normalization restriction is obtained by imposing  $\mathbf{L}_{11} = 1$ .

Identification appears to be a serious problem with the MNP model. Although the unrestricted MNP model is fully identified in principle, convergence to satisfactory results in applications with more than three choices appears to require many additional restrictions on the standard deviations and correlations, such as zero restrictions or equality restrictions in the case of the standard deviations.

#### 18.2.7 THE MIXED LOGIT MODEL

Another variant of the multinomial logit model is the **random parameters logit model (RPL)** (also called the **mixed logit model**). [See Revelt and Train (1996); Bhat (1996); Berry, Levinsohn, and Pakes (1995); Jain, Vilcassim, and Chintagunta (1994); and Hensher and Greene (2004).] Train's (2003) formulation of the RPL model (which encompasses the others) is a modification of the MNL model. The model is a **random coefficients** formulation. The change to the basic MNL model is the parameter specification in the distribution of the parameters across individuals,  $i$ :

$$\beta_{ik} = \beta_k + \mathbf{z}_i' \boldsymbol{\theta}_k + \sigma_k u_{ik}, \quad (18-11)$$

where  $u_{ik}$ ,  $k = 1, \dots, K$ , is multivariate normally distributed with correlation matrix  $\mathbf{R}$ ,  $\sigma_k$  is the standard deviation of the  $k$ th distribution,  $\beta_k + \mathbf{z}_i' \boldsymbol{\theta}_k$  is the mean of the distribution, and  $\mathbf{z}_i$  is a vector of person specific characteristics (such as age and income) that do not vary across choices. This formulation contains all the earlier models. For example, if  $\boldsymbol{\theta}_k = \mathbf{0}$  for all the coefficients and  $\sigma_k = 0$  for all the coefficients except for choice-specific constants, then the original MNL model with a normal-logistic mixture for the random part of the MNL model arises (hence the name).

The model is estimated by simulating the log-likelihood function rather than direct integration to compute the probabilities, which would be infeasible because the mixture distribution composed of the original  $\varepsilon_{ij}$  and the random part of the coefficient is unknown. For any individual,

$$\text{Prob}[\text{choice } q | \mathbf{u}_i] = \text{MNL probability} | \beta_i(\mathbf{u}_i),$$

**772 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**

with all restrictions imposed on the coefficients. The appropriate probability is

$$E_u[\text{Prob}(\text{choice } q | \mathbf{u})] = \int_{u_1, \dots, u_k} \text{Prob}[\text{choice } q | \mathbf{u}] f(\mathbf{u}) d\mathbf{u},$$

which can be estimated by simulation, using

$$\text{Est. } E_u[\text{Prob}(\text{choice } q | \mathbf{u})] = \frac{1}{R} \sum_{r=1}^R \text{Prob}[\text{choice } q | \boldsymbol{\beta}_i(\mathbf{u}_{ir})],$$

where  $\mathbf{u}_{ir}$  is the  $r$ th of  $R$  draws for observation  $i$ . (There are  $nkR$  draws in total. The draws for observation  $i$  must be the same from one computation to the next, which can be accomplished by assigning to each individual their own seed for the random number generator and restarting it each time the probability is to be computed.) By this method, the log-likelihood and its derivatives with respect to  $(\beta_k, \theta_k, \sigma_k)$ ,  $k = 1, \dots, K$  and  $\mathbf{R}$  are simulated to find the values that maximize the simulated log-likelihood.

The mixed model enjoys two considerable advantages not available in any of the other forms suggested. In a panel data or repeated-choices setting (see Section 18.2.11), one can formulate a random effects model simply by making the variation in the coefficients time invariant. Thus, the model is changed to

$$U_{ijt} = \mathbf{x}'_{ijt} \boldsymbol{\beta}_{it} + \varepsilon_{ijt}, \quad i = 1, \dots, n, \quad j = 1, \dots, J, \quad t = 1, \dots, T,$$

$$\boldsymbol{\beta}_{it,k} = \mathbf{z}'_{it} \boldsymbol{\theta}_k + \sigma_k u_{ik}.$$

The time variation in the coefficients is provided by the choice-invariant variables, which may change through time. Habit persistence is carried by the time-invariant random effect,  $u_{ik}$ . If only the constant terms vary and they are assumed to be uncorrelated, then this is logically equivalent to the familiar random effects model. But, much greater generality can be achieved by allowing the other coefficients to vary randomly across individuals and by allowing correlation of these effects.<sup>6</sup> A second degree of flexibility is in (18-11). The random components,  $u_i$  are not restricted to normality. Other distributions that can be simulated will be appropriate when the range of parameter variation consistent with consumer behavior must be restricted, for example to narrow ranges or to positive values.

### 18.2.8 A GENERALIZED MIXED LOGIT MODEL

The development of functional forms for multinomial choice models begins with the conditional (now usually called the multinomial) logit model that we considered in Section 18.2.3. Subsequent proposals including the multinomial probit and nested logit models (and a wide range of variations on these themes) were motivated by a desire to extend the model beyond the IIA assumptions. These were achieved by allowing correlation across the utility functions or heteroscedasticity such as that in the heteroscedastic extreme value model in (18-12). That issue has been settled in the current generation of multinomial choice models, culminating with the mixed logit model that appears to provide all the flexibility needed to depart from the IIA assumptions. [See McFadden and Train (2000) for a strong endorsement of this idea.]

---

<sup>6</sup>See Hensher (2001) for an application to transportation mode choice in which each individual is observed in several choice situations. A stated choice experiment in which consumers make several choices in sequence about automobile features appears in Hensher, Rose, and Greene (2006).

CHAPTER 18 ♦ Discrete Choices and Event Counts **773**

Recent research in choice modeling has focused on enriching the models to accommodate individual heterogeneity in the choice specification. To a degree, including observable characteristics, such as household income in our application to follow, serves this purpose. In this case, the observed heterogeneity enters the deterministic part of the utility functions. The heteroscedastic HEV model shown in (18-13) moves the observable heterogeneity to the scaling of the utility function instead of the mean. The mixed logit model in (18-11) accommodates both observed and unobserved heterogeneity in the preference parameters. A recent thread of research including Keane (2006), Feibig, Keane, Louviere, and Wasi (2009), and Greene and Hensher (2010) has considered functional forms that accommodate individual heterogeneity in both  parameters (marginal utilities) and overall scaling of the preference structure. Keane et al.'s generalized mixed logit model is

$$\begin{aligned}U_{i,j} &= \mathbf{x}'_{ij}\boldsymbol{\beta}_i + \varepsilon_{ij}, \\ \boldsymbol{\beta}_i &= \sigma_i\boldsymbol{\beta} + [\gamma + \sigma_i(1 - \gamma)]\mathbf{v}_i \\ \sigma_i &= \exp[\bar{\sigma} + \tau w_i]\end{aligned}$$

where  $0 \leq \gamma \leq 1$  and  $w_i$  is an additional source of unobserved random variation in preferences. In this formulation, the weighting parameter,  $\gamma$ , distributes the individual heterogeneity in the preference weights,  $\mathbf{v}_i$  and the overall scaling parameter  $\sigma_i$ . Heterogeneity across individuals in the overall scaling of preference structures is introduced by a nonzero  $\tau$  while  $\bar{\sigma}$  is chosen so that  $E_w[\sigma_i] = 1$ . Greene and Hensher (2010) proposed including the observable heterogeneity already in the mixed logit model, and adding it to the scaling parameter as well. Also allowing the random parameters to be correlated (via the nonzero elements in  $\Gamma$ ), produces a multilayered form of the generalized mixed logit model,

$$\begin{aligned}\boldsymbol{\beta}_i &= \sigma_i[\boldsymbol{\beta} + \Delta\mathbf{z}_i] + [\gamma + \sigma_i(1 - \gamma)]\boldsymbol{\Gamma}\mathbf{v}_i \\ \sigma_i &= \exp[\bar{\sigma} + \delta'\mathbf{h}_i + \tau w_i].\end{aligned}$$

Ongoing research has continued to produce refinements that will accommodate realistic forms of individual heterogeneity in the basic multinomial logit framework.

#### 18.2.9 APPLICATION: CONDITIONAL LOGIT MODEL FOR TRAVEL MODE CHOICE

Hensher and Greene [Greene (2007a)] report estimates of a model of travel mode choice for travel between Sydney and Melbourne, Australia. The data set contains 210 observations on choice among four travel modes, *air*, *train*, *bus*, and *car*. (See Appendix Table F18.2.) The attributes used for their example were: choice-specific constants; two choice-specific continuous measures; *GC*, a measure of the generalized cost of the travel that is equal to the sum of in-vehicle cost, *INVC*, and a wagelike measure times *INVT*, the amount of time spent traveling; and *TTME*, the terminal time (zero for car); and for the choice between air and the other modes, *HINC*, the household income. A summary of the sample data is given in Table 18.2. The sample is **choice based** so as to balance it among the four choices—the true population allocation, as shown in the last column of Table 18.2, is dominated by drivers.

**774 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**
**TABLE 18.2** Summary Statistics for Travel Mode Choice Data

	<b>GC</b>	<b>TTME</b>	<b>INVC</b>	<b>INVT</b>	<b>HINC</b>	<b>Number Choosing</b>	<b>p</b>	<b>True Prop.</b>
Air	102.648	61.010	85.522	133.710	34.548	58	0.28	0.14
	113.522	46.534	97.569	124.828	41.274			
Train	130.200	35.690	51.338	608.286	34.548	63	0.30	0.13
	106.619	28.524	37.460	532.667	23.063			
Bus	115.257	41.657	33.457	629.462	34.548	30	0.14	0.09
	108.133	25.200	33.733	618.833	29.700			
Car	94.414	0	20.995	573.205	34.548	59	0.28	0.64
	89.095	0	15.694	527.373	42.220			

*Note:* The upper figure is the average for all 210 observations. The lower figure is the mean for the observations that made that choice.

The model specified is

$$U_{ij} = \alpha_{air}d_{i,air} + \alpha_{train}d_{i,train} + \alpha_{bus}d_{i,bus} + \beta_G GC_{ij} + \beta_T TTME_{ij} + \gamma_H d_{i,air} HINC_i + \varepsilon_{ij},$$

where for each  $j$ ,  $\varepsilon_{ij}$  has the same independent, type 1 extreme value distribution,

$$F_\varepsilon(\varepsilon_{ij}) = \exp(-\exp(-\varepsilon_{ij})),$$

which has standard deviation  $\pi^2/6$ . The mean is absorbed in the constants. Estimates of the conditional logit model are shown in Table 18.3. The model was fit with and without the corrections for choice-based sampling. Because the sample shares do not differ radically from the population proportions, the effect on the estimated parameters is fairly modest. Nonetheless, it is apparent that the choice-based sampling is not completely innocent. A cross tabulation of the predicted versus actual outcomes is given in Table 18.4. The predictions are generated by tabulating the integer parts of  $m_{jk} = \sum_{i=1}^{210} \hat{p}_{ij} d_{ik}$ ,  $j, k = air, train, bus, car$ , where  $\hat{p}_{ij}$  is the predicted probability of outcome  $j$  for observation  $i$  and  $d_{ik}$  is the binary variable which indicates if individual  $i$  made choice  $k$ .

Are the odds ratios *train/bus* and *car/bus* really independent from the presence of the *air* alternative? To use the Hausman test, we would eliminate choice *air*, from the choice set and estimate a three-choice model. Because 58 respondents chose this mode,

**TABLE 18.3** Parameter Estimates

	<b>Unweighted Sample</b>		<b>Choice-Based Weighting</b>	
	<b>Estimate</b>	<b>t Ratio</b>	<b>Estimate</b>	<b>t Ratio</b>
$\beta_G$	-0.015501	-3.517	-0.01333	-2.711
$\beta_T$	-0.09612	-9.207	-0.13405	-5.216
$\gamma_H$	0.01329	1.295	-0.00108	-0.097
$\alpha_{air}$	5.2074	6.684	6.5940	4.075
$\alpha_{train}$	3.8690	8.731	3.6190	4.317
$\alpha_{bus}$	3.1632	7.025	3.3218	3.822
Log-likelihood at $\beta = 0$		-291.1218		-291.1218
Log-likelihood (sample shares)		-283.7588		-218.9929
Log-likelihood at convergence		-199.1284		-147.5896

## CHAPTER 18 ♦ Discrete Choices and Event Counts 775

**TABLE 18.4** Predicted Choices Based on Model Probabilities (predictions based on choice-based sampling in parentheses)

	Air	Train	Bus	Car	Total (Actual)
Air	32 (30)	8 (3)	5 (3)	13 (23)	58
Train	7 (3)	37 (30)	5 (3)	14 (27)	63
Bus	3 (1)	5 (2)	15 (14)	6 (12)	30
Car	16 (5)	13 (5)	6 (3)	25 (45)	59
Total (Predicted)	58 (39)	63 (40)	30 (23)	59 (108)	210

**TABLE 18.5** Results for IIA Test

	Full-Choice Set				Restricted-Choice Set			
	$\beta_G$	$\beta_T$	$\alpha_{train}$	$\alpha_{bus}$	$\beta_G$	$\beta_T$	$\alpha_{train}$	$\alpha_{bus}$
Estimate	-0.0155	-0.0961	3.869	3.163	-0.0639	-0.0699	4.464	3.105
<i>Estimated Asymptotic Covariance Matrix</i>								
$\beta_G$	0.194e-4				0.000101			
$\beta_T$	-0.46e-6	0.000109			-0.000013	0.000221		
$\alpha_{train}$	-0.00060	-0.0038	0.196		-0.00244	-0.00759	0.410	
$\alpha_{bus}$	-0.00026	-0.0038	0.161	0.203	-0.00113	-0.00753	0.336	0.371

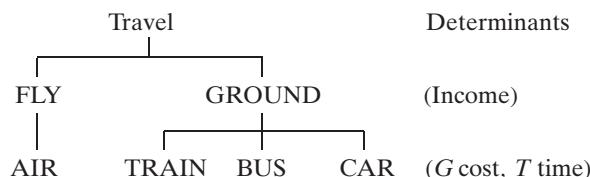
Note: 0.nnnne- $p$  indicates times 10 to the negative  $p$  power.

$H = 33.3367$ . Critical chi-squared[4] = 9.488.

we would lose 58 observations. In addition, for every data vector left in the sample, the air-specific constant and the interaction,  $d_{i,air} \times HINC_i$  would be zero for every remaining individual. Thus, these parameters could not be estimated in the restricted model. We would drop these variables. The test would be based on the two estimators of the remaining four coefficients in the model,  $[\beta_G, \beta_T, \alpha_{train}, \alpha_{bus}]$ . The results for the test are shown in Table 18.5

The hypothesis that the odds ratios for the other three choices are independent from *air* would be rejected based on these results, as the chi-squared statistic exceeds the critical value.

Because IIA was rejected, they estimated a nested logit model of the following type:



Note that one of the branches has only a single choice, so the conditional probability,  $P_{j|fly} = P_{air|fly} = 1$ . The estimates marked “unconditional” in Table 18.6 are the simple conditional (multinomial) logit (MNL) model for choice among the four alternatives that was reported earlier. Both inclusive value parameters are constrained (by construction) to equal 1.0000. The FIML estimates are obtained by maximizing the

**776 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**
**TABLE 18.6** Estimates of a Mode Choice Model (standard errors in parentheses)

Parameter	FIML Estimate	Unconditional
$\alpha_{air}$	6.042	(1.199)
$\alpha_{bus}$	4.096	(0.615)
$\alpha_{train}$	5.065	(0.662)
$\beta_{GC}$	-0.03159	(0.00816)
$\beta_{TTME}$	-0.1126	(0.0141)
$\gamma_H$	0.01533	(0.00938)
$\tau_{fly}$	0.5860	(0.141)
$\tau_{ground}$	0.3890	(0.124)
$\sigma_{fly}$	2.1886	(0.525)
$\sigma_{ground}$	3.2974	(1.048)
$\ln L$	-193.6561	-199.1284

full log-likelihood for the nested logit model. In this model,

$$\text{Prob}(choice | branch) = P(\alpha_{air}d_{air} + \alpha_{train}d_{train} + \alpha_{bus}d_{bus} + \beta_G GC + \beta_T TTME),$$

$$\text{Prob}(branch) = P(\gamma d_{air} HINC + \tau_{fly} IV_{fly} + \tau_{ground} IV_{ground}),$$

$$\text{Prob}(choice, branch) = \text{Prob}(choice | branch) \times \text{Prob}(branch).$$

The likelihood ratio statistic for the nesting (heteroscedasticity) against the null hypothesis of homoscedasticity is  $-2[-199.1284 - (-193.6561)] = 10.945$ . The 95 percent critical value from the chi-squared distribution with two degrees of freedom is 5.99, so the hypothesis is rejected. We can also carry out a Wald test. The asymptotic covariance matrix for the two inclusive value parameters is  $[0.01977 / 0.009621, 0.01529]$ . The Wald statistic for the joint test of the hypothesis that  $\tau_{fly} = \tau_{ground} = 1$ , is

$$W = (0.586 - 1.0 \quad 0.389 - 1.0) \begin{bmatrix} 0.01977 & 0.009621 \\ 0.009621 & 0.01529 \end{bmatrix}^{-1} \begin{pmatrix} 0.586 - 1.0 \\ 0.389 - 1.0 \end{pmatrix} = 24.475.$$

The hypothesis is rejected, once again.

The choice model was reestimated under the assumptions of a heteroscedastic extreme value (HEV) specification. In its simplest form, this model allows a separate variance,

$$\sigma_j^2 = \pi^2 / (6\theta_j^2) \tag{18-12}$$

for each  $\varepsilon_{ij}$  in (18-1). (One of the  $\theta$ 's must be normalized to 1.0 because we can only compare ratios of variances.) The results for this model are shown in Table 18.7. This model is less restrictive than the nested logit model. To make them comparable, we note that we found that  $\sigma_{air} = \pi / (\tau_{fly}\sqrt{6}) = 2.1886$  and  $\sigma_{train} = \sigma_{bus} = \sigma_{car} = \pi / (\tau_{ground}\sqrt{6}) = 3.2974$ . The HEV model thus relaxes an additional restriction because it has three free variances whereas the nested logit model has two. On the other hand, the important degree of freedom is that the HEV model does not impose the IIA assumption anywhere in the choices, whereas the nested logit does, within each branch. Table 18.7 contains two additional results for HEV specifications. In the one denoted “Heteroscedastic HEV Model,” we have allowed heteroscedasticity across individuals as well as across

## CHAPTER 18 ♦ Discrete Choices and Event Counts 777

**TABLE 18.7** Estimates of a Heteroscedastic Extreme Value Model  
(standard errors in parentheses)

Parameter	HEV Model	Heteroscedastic HEV Model	Restricted HEV Model	Nested Logit Model				
$\alpha_{air}$	7.8326	(10.951)	5.1815	(6.042)	2.973	(0.995)	6.062	(1.199)
$\alpha_{bus}$	7.1718	(9.135)	5.1302	(5.132)	4.050	(0.494)	4.096	(0.615)
$\alpha_{train}$	6.8655	(8.829)	4.8654	(5.071)	3.042	(0.429)	5.065	(0.662)
$\beta_{GC}$	-0.05156	(0.0694)	-0.03326	(0.0378)	-0.0289	(0.00580)	-0.03159	(0.00816)
$\beta_{TTME}$	-0.1968	(0.288)	-0.1372	(0.164)	-0.0828	(0.00576)	-0.1126	(0.0141)
$\gamma$	0.04024	(0.0607)	0.03557	(0.0451)	0.0238	(0.0186)	0.01533	(0.00938)
$\tau_{fly}$							0.5860	(0.141)
$\tau_{ground}$							0.3890	(0.124)
$\theta_{air}$	0.2485	(0.369)	0.2890	(0.321)	0.4959	(0.124)		
$\theta_{train}$	0.2595	(0.418)	0.3629	(0.482)	1.0000	(0.000)		
$\theta_{bus}$	0.6065	(1.040)	0.6895	(0.945)	1.0000	(0.000)		
$\theta_{car}$	1.0000	(0.000)	1.0000	(0.000)	1.0000	(0.000)		
$\phi$	0.0000	(0.000)	0.00552	(0.00573)	0.0000	(0.000)		
<i>Implied Standard Deviations</i>								
$\sigma_{air}$	5.161	(7.667)						
$\sigma_{train}$	4.942	(7.978)						
$\sigma_{bus}$	2.115	(3.623)						
$\sigma_{car}$	1.283	(0.000)						
$\ln L$		-195.6605		-194.5107		-200.3791		-193.6561

choices by specifying

$$\theta_{ij} = \theta_j \times \exp(\phi HINC_i). \quad (18-13)$$

[See Salisbury and Feinberg (2010) and Louviere and Swait (2010) for an application of this type of HEV model.]

In the “Restricted HEV Model,” the variance of  $\varepsilon_{i,Air}$  is allowed to differ from the others. Finally, the nested logit model has different variance for *Air* and (*Train, Bus, Car*).

A primary virtue of the HEV model, the nested logit model, and other alternative models is that they relax the IIA assumption. This assumption has implications for the cross elasticities between attributes in the different probabilities. Table 18.8 lists the estimated elasticities of the estimated probabilities with respect to changes in the generalized cost variable. Elasticities are computed by averaging the individual sample values rather than computing them once at the sample means. The implication of the IIA assumption can be seen in the table entries. Thus, in the estimates for the multinomial logit (MNL) model, the cross elasticities for each attribute are all equal. In the nested logit model, the IIA property only holds within the branch. Thus, in the first column, the effect of GC of air affects all ground modes equally, whereas the effect of GC for train is the same for bus and car, but different from these two for air. All these elasticities vary freely in the HEV model.

Table 18.9 lists the estimates of the parameters of the multinomial probit and random parameters logit models. For the multinomial probit model, we fit three specifications: (1) free correlations among the choices, which implies an unrestricted  $3 \times 3$

**778 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**
**TABLE 18.8** Estimated Elasticities with Respect to Generalized Cost

<i>Effect on</i>	<i>Cost Is That of Alternative</i>			
	<i>Air</i>	<i>Train</i>	<i>Bus</i>	<i>Car</i>
<b>Multinomial Logit</b>				
Air	-1.136	0.498	0.238	0.418
Train	0.456	-1.520	0.238	0.418
Bus	0.456	0.498	-1.549	0.418
Car	0.456	0.498	0.238	-1.061
<b>Nested Logit</b>				
Air	-0.858	0.332	0.179	0.308
Train	0.314	-4.075	0.887	1.657
Bus	0.314	1.595	-4.132	1.657
Car	0.314	1.595	0.887	-2.498
<b>Heteroscedastic Extreme Value</b>				
Air	-1.040	0.367	0.221	0.441
Train	0.272	-1.495	0.250	0.553
Bus	0.688	0.858	-6.562	3.384
Car	0.690	0.930	1.254	-2.717

**TABLE 18.9** Parameter Estimates for Normal-Based Multinomial Choice Models

<i>Parameter</i>	<i>Multinomial Probit</i>			<i>Random Parameters Logit</i>		
	<i>Unrestricted</i>	<i>Homoscedastic</i>	<i>Uncorrelated</i>	<i>Unrestricted</i>	<i>Constants</i>	<i>Uncorrelated</i>
$\alpha_{air}$	1.358	3.005	3.171	5.519	4.807	12.603
$\sigma_{air}$	4.940	1.000 <sup>a</sup>	3.629	4.009 <sup>d</sup>	3.225 <sup>b</sup>	2.803 <sup>c</sup>
$\alpha_{train}$	4.298	2.409	4.277	5.776	5.035	13.504
$\sigma_{train}$	1.899	1.000 <sup>a</sup>	1.581	1.904	1.290 <sup>b</sup>	1.373
$\alpha_{bus}$	3.609	1.834	3.533	4.813	4.062	11.962
$\sigma_{bus}$	1.000 <sup>a</sup>	1.000 <sup>a</sup>	1.000 <sup>a</sup>	1.424	3.147 <sup>b</sup>	1.287
$\alpha_{car}$	0.000 <sup>a</sup>	0.000 <sup>a</sup>	0.000 <sup>a</sup>	0.000 <sup>a</sup>	0.000 <sup>a</sup>	0.000
$\sigma_{car}$	1.000 <sup>a</sup>	1.000	1.000 <sup>a</sup>	1.283 <sup>a</sup>	1.283 <sup>a</sup>	1.283 <sup>a</sup>
$\beta_G$	-0.0351	-0.0113	-0.0325	-0.0326	-0.0317	-0.0544
$\sigma_{\beta G}$	—	—	—	0.000 <sup>a</sup>	0.000 <sup>a</sup>	0.00561
$\beta_T$	-0.0769	-0.0563	-0.0918	-0.126	-0.112	-0.2822
$\sigma_{\beta T}$	—	—	—	0.000 <sup>a</sup>	0.000 <sup>a</sup>	0.182
$\gamma_H$	0.0593	0.0126	0.0370	0.0334	0.0319	0.0846
$\sigma_\gamma$	—	—	—	0.000 <sup>a</sup>	0.000 <sup>a</sup>	0.0768
$\rho_{AT}$	0.581	0.000 <sup>a</sup>	0.000 <sup>a</sup>	0.543	0.000 <sup>a</sup>	0.000 <sup>a</sup>
$\rho_{AB}$	0.576	0.000 <sup>a</sup>	0.000 <sup>a</sup>	0.532	0.000 <sup>a</sup>	0.000 <sup>a</sup>
$\rho_{BT}$	0.718	0.000 <sup>a</sup>	0.000 <sup>a</sup>	0.993	0.000 <sup>a</sup>	0.000 <sup>a</sup>
$\log L$	-196.9244	-208.9181	-199.7623	-193.7160	-199.0073	-175.5333

<sup>a</sup>Restricted to this fixed value.

<sup>b</sup>Computed as the square root of  $(\pi^2/6 + \theta_j^2)$ ,  $\theta_{air} = 2.959$ ,  $\theta_{train} = 0.136$ ,  $\theta_{bus} = 0.183$ ,  $\theta_{car} = 0.000$ .

<sup>c</sup> $\theta_{air} = 2.492$ ,  $\theta_{train} = 0.489$ ,  $\theta_{bus} = 0.108$ ,  $\theta_{car} = 0.000$ .

<sup>d</sup>Derived standard deviations for the random constants are  $\theta_{air} = 3.798$ ,  $\theta_{train} = 1.182$ ,  $\theta_{bus} = 0.0712$ ,  $\theta_{car} = 0.000$ .

correlation matrix and two free standard deviations; (2) uncorrelated disturbances, but free standard deviations, a model that parallels the heteroscedastic extreme value model; and (3) uncorrelated disturbances and equal standard deviations, a model that is the same as the original conditional logit model save for the normal distribution of

**CHAPTER 18 ♦ Discrete Choices and Event Counts 779**

the disturbances instead of the extreme value assumed in the logit model. In this case, the scaling of the utility functions is different by a factor of  $(\pi^2/6)^{1/2} = 1.283$ , as the probit model assumes  $\varepsilon_j$  has a standard deviation of 1.0.

We also fit three variants of the random parameters logit. In these cases, the choice-specific variance for each utility function is  $\sigma_j^2 + \theta_j^2$  where  $\sigma_j^2$  is the contribution of the logit model, which is  $\pi^2/6 = 1.645$ , and  $\theta_j^2$  is the estimated constant specific variance estimated in the random parameters model. The combined estimated standard deviations are given in the table. The estimates of the specific parameters,  $\theta_j$ , are given in the footnotes. The estimated models are (1) unrestricted variation and correlation among the three intercept parameters—this parallels the general specification of the multinomial probit model; (2) only the constant terms randomly distributed but uncorrelated, a model that is parallel to the multinomial probit model with no cross-equation correlation and to the heteroscedastic extreme value model shown in Table 18.7 and (3) random but uncorrelated parameters. This model is more general than the others but is somewhat restricted as the parameters are assumed to be uncorrelated. Identification of the correlation matrix is weak in this model—after all, we are attempting to estimate a  $6 \times 6$  correlation matrix for all unobserved variables. Only the estimated parameters are shown in Table 18.9 Estimated standard errors are similar to (although generally somewhat larger than) those for the basic multinomial logit model.

The standard deviations and correlations shown for the multinomial probit model are parameters of the distribution of  $\varepsilon_{ij}$ , the overall randomness in the model. The counterparts in the random parameters model apply to the distributions of the parameters. Thus, the full disturbance in the model in which only the constants are random is  $\varepsilon_{iair} + u_{air}$  for air, and likewise for train and bus. Likewise, the correlations shown for the first two models are directly comparable, although it should be noted that in the random parameters model, the disturbances have a distribution that is that of a sum of an extreme value and a normal variable, while in the probit model, the disturbances are normally distributed. With these considerations, the “unrestricted” models in each case are comparable and are, in fact, fairly similar.

None of this discussion suggests a preference for one model or the other. The likelihood values are not comparable, so a direct test is precluded. Both relax the IIA assumption, which is a crucial consideration. The random parameters model enjoys a significant practical advantage, as discussed earlier, and also allows a much richer specification of the utility function itself. But, the question still warrants additional study. Both models are making their way into the applied literature.

#### **18.2.10 ESTIMATING WILLINGNESS TO PAY**

One of the standard applications of choice models is to estimate how much consumers value the attributes of the choices. Recall that we are not able to observe the scale of the utilities in the choice model. However, we can use the marginal utility of income, also scaled in the same unobservable way, to effect the valuation. In principle, we could estimate

$$\begin{aligned} \text{WTP} &= (\text{Marginal Utility of Attribute}/\sigma)/(\text{Marginal Utility of Income}/\sigma) \\ &= (\beta_{\text{attribute}}/\sigma)/(\gamma_{\text{Income}}/\sigma), \end{aligned}$$

**780 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**

where  $\sigma$  is the unknown scaling of the utility functions. Note that  $\sigma$  cancels out of the ratio. In our application, for example, we might assess how much consumers would be willing to pay to have shorter waits at the terminal for the public modes of transportation by using

$$\text{WTP}_{\text{time}} = -\beta_{\text{TIME}}/\gamma_{\text{Income}}.$$

(We use the negative because additional time spent waiting at the terminal provides disutility, as evidenced by its coefficient's negative sign.) In settings in which income is not observed, researchers often use the negative of the coefficient on a cost variable as a proxy for the marginal utility of income. Standard errors for estimates of WTP can be computed using the delta method or the method of Krinsky and Robb. (See Sections 4.4.4 and 15.3.)

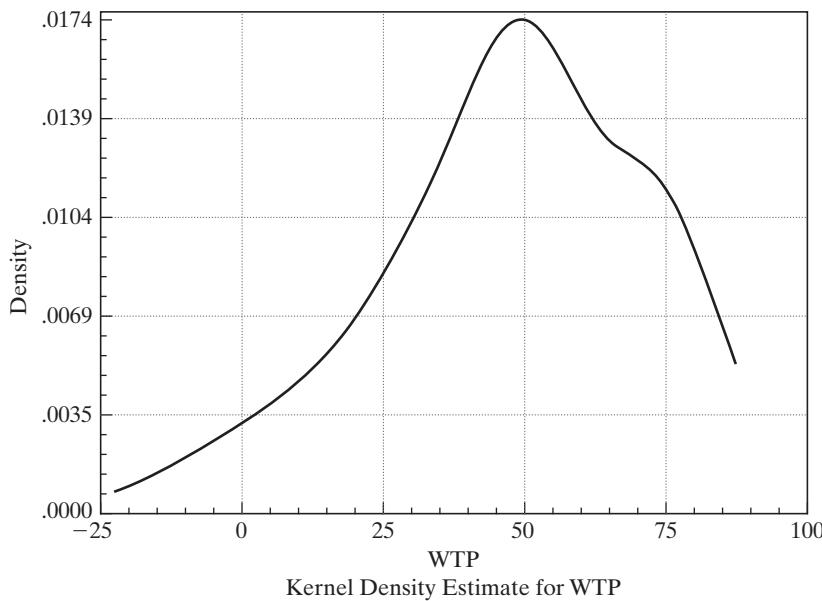
In the basic multinomial logit model, the estimator of WTP is a simple ratio of parameters. In our estimated model in Table 18.3, for example, using the household income coefficient as the numeraire, the estimate of WTP for a shorter wait at the terminal is  $-0.09612/0.01329 = 7.239$ . The units of measurement must be resolved in this computation, since terminal time is measured in minutes while the cost is in \$1,000/year. Multiplying this result by \$60 minutes/hour and dividing by the equivalent hourly income of income times  $8,760/1,000$  gives \$49.54 per hour of waiting time. To compute the estimated asymptotic standard error, for convenience, we first rescaled the terminal time to hours by dividing it by 60 and the income variable to \$/hour by multiplying it by  $1,000/8,760$ . The resulting estimated asymptotic distribution for the estimators is

$$\begin{pmatrix} \hat{\beta}_{\text{TIME}} \\ \hat{\gamma}_{\text{HINC}} \end{pmatrix} \sim N \left[ \begin{pmatrix} -5.76749 \\ 0.11639 \end{pmatrix}, \begin{pmatrix} 0.392365 & 0.00193095 \\ 0.00193095 & 0.00808177 \end{pmatrix} \right].$$

The derivatives of  $\text{WTP}_{\text{TIME}} = -\widehat{\beta}_{\text{TIME}}/\gamma_H$  are  $-1/\gamma_H$  for  $\beta_{\text{TIME}}$  and  $-\text{WTP}/\gamma_H$  for  $\gamma_H$ . This provides an estimator of 38.8304 for the standard error. The confidence interval for this parameter would be  $-26.56$  to  $+125.63$ . This seems extremely wide. We will return to this issue later.

In the mixed logit model, if either of the coefficients in the computation is random, then the preceding simple computation above will not reveal the heterogeneity in the result. In many studies of WTP using mixed logit models, it is common to allow the utility parameter on the attribute (numerator) to be random and treat the numeraire (income or cost coefficient  as nonrandom. Using our mode choice application, we refit the model with  $\widehat{\beta}_{\text{TIME},i} = \widehat{\beta}_{\text{TIME}} + \widehat{\sigma}_{\text{TIME}} v_i$  and all other coefficients nonrandom. We then used the method described in Section 15.10 to estimate  $E[\beta_{\text{TIME},i} | \mathbf{X}_i, \text{choice}_i]/\gamma_H$  to estimate the expected WTP for each individual in the sample. Income and terminal time were scaled as before. Figure 18.1 displays a kernel estimator of the estimates of  $\text{WTP}_i$  by this method. Note that the distribution is roughly centered on our earlier estimate of \$49.53. The density estimator reveals the heterogeneity in the population of this parameter.

Willingness to pay measures computed as suggested above are ultimately based on a ratio of two asymptotically normally distributed parameter estimators. In general, ratios of normally distributed random variables do not have a finite variance. This often becomes apparent when using the delta method, as it seems previously. A number of writers, notably, Daly, Hess, and Train (2009), have documented the problem of extreme



**FIGURE 18.1** Estimated Willingness to Pay for Decreased Terminal Time.

results of WTP computations, and why they should be expected. One solution suggested, for example, by Train and Weeks (2005), Sonnier, Ainsle, and Otter (2007), and Scarpa, Thiene, and Train (2008), is to recast the original model in **willingness to pay space**. In the multinomial logit case, this amounts to a trivial reparameterization of the model. Using our application as an example, we would write

$$\begin{aligned} U_{ij} &= \alpha_j + \beta_{GC} [G_i] + \widehat{\beta_{TIME}} / [TIME_i] + \gamma_H A_{AIR} HINC_i + \varepsilon_{ij} \\ &= \alpha_j + \beta_{GC} [G_i] + \lambda_{TIME} [TIME_i] + \gamma_H A_{AIR} HINC_i + \varepsilon_{ij}. \end{aligned}$$

This obviously returns the original model, though in the process, it transforms a linear estimation problem into a nonlinear one. But, in principle, with the model reparameterized in “WTP space,” we have sidestepped the problem noted earlier –  $\lambda_{TIME}$  is the estimator of WTP with no further transformation of the parameters needed. As noted, this will return the numerically identical results for a multinomial logit model. It will not return the identical results for a mixed logit model, in which we write  $\lambda_{TIME,i} = \lambda_{TIME} + \theta_{TIME} v_{TIME,i}$ . Greene and Hensher (2010b) apply this method to the generalized mixed logit model in Section 18.2.8.

#### 18.2.11 PANEL DATA AND STATED CHOICE EXPERIMENTS

Panel data in the unordered discrete choice setting typically come in the form of sequential choices. Train (2009, Chapter 6) reports an analysis of the site choices of 258 anglers who chose among 59 possible fishing sites for a total of 962 visits. Allenby and Rossi (1999) modeled brand choice for a sample of shoppers who made multiple store trips. The mixed logit model is a framework that allows the counterpart to a random

## 782 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

effects model. The random utility model would appear

$$U_{ij,t} = \mathbf{x}'_{ij,t} \boldsymbol{\beta}_i + \varepsilon_{ij,t},$$

where conditioned on  $\boldsymbol{\beta}_i$ , a multinomial logit model applies. The random coefficients carry the common effects across choice situations. For example, if the random coefficients include choice-specific constant terms, then the random utility model becomes essentially a random effects model. A modification of the model that resembles Mundlak's correction for the random effects model is

$$\boldsymbol{\beta}_i = \boldsymbol{\beta}^0 + \Delta \mathbf{z}_i + \Gamma \mathbf{u}_i,$$

where, typically,  $\mathbf{z}_i$  would contain demographic and socioeconomic information.

The **stated choice experiment** is similar to the repeated choice situation, with a crucial difference. In a stated choice survey, the respondent is asked about his or her preferences over a series of hypothetical choices, often including one or more that are actually available and others that might not be available (yet). Hensher, Rose, and Greene (2006) describe a survey of Australian commuters who were asked about hypothetical  modes in a choice set that included the one they currently took and a variety of alternatives. Revelt and Train (2000) analyzed a stated choice experiment in which California electricity consumers were asked to choose among alternative hypothetical energy suppliers. The advantage of the stated choice experiment is that it allows the analyst to study choice situations over a range of variation of the attributes or a range of choices that might not exist within the observed, actual outcomes. Thus, the original work on the MNL by McFadden et al. concerned survey data on whether commuters would ride a (then-hypothetical) underground train system to work in the San Francisco Bay area. The disadvantage of **stated choice data** is that they are hypothetical. Particularly when they are mixed with **revealed preference data**, the researcher must assume that the same preference patterns govern both types of outcomes. This is likely to be a dubious assumption. One method of accommodating the mixture of underlying preferences is to build different scaling parameters into the model for the  and revealed preference components of the model. Greene and Hensher (2007) suggest a nested logit model that groups the hypothetical choices in one branch of a tree and the observed choices in another.

### 18.2.12 AGGREGATE MARKET SHARE DATA—THE BLP RANDOM PARAMETERS MODEL

We note, finally, an important application of the mixed logit model, the structural demand model of Berry, Levinsohn, and Pakes (1995). (Demand models for differentiated products such as automobiles [BLP (1995), Goldberg (1995)], ready-to-eat cereals [Nevo (2001)], and consumer electronics [Das, Olley, and Pakes (1996)], have been constructed using the mixed logit model with market share data.<sup>7</sup> A basic structure is defined for

Markets, denoted  $t = 1, \dots, T$

Consumers in the markets, denoted  $i = 1, \dots, n_t$



Products, denoted  $j = 1, \dots, J$

<sup>7</sup>We draw heavily on Nevo (2000) for this discussion.

CHAPTER 18 ♦ Discrete Choices and Event Counts **783**

The definition of a market varies by application; BLP analyzed the U.S. national automobile market for 20 years; Nevo examined a cross section of cities over 20 quarters so the city-quarter is a market; Das et al. defined a market as the annual sales to consumers in particular income levels.

For market  $t$ , we base the analysis on average prices,  $p_{jt}$ , aggregate quantities  $q_{jt}$ , consumer incomes  $y_i$  observed product attributes,  $\mathbf{x}_{jt}$  and unobserved (by the analyst) product attributes,  $\Delta_{jt}$ . The indirect utility function for consumer  $i$ , for product  $j$  in market  $t$  is

$$u_{ijt} = \alpha_i(y_i - p_{jt}) + \mathbf{x}_{jt}'\boldsymbol{\beta}_i + \Delta_{jt} + \varepsilon_{ijt}, \quad (18-14)$$

where  $\alpha_i$  is the marginal utility of income and  $\boldsymbol{\beta}_i$  are marginal utilities attached to specific observable attributes of the products. The fact that some unobservable product attributes,  $\Delta_{jt}$  will be reflected in the prices implies that prices will be endogenous in a demand model that is based on only the observable attributes. Heterogeneity in preferences is reflected (as we did earlier) in the formulation of the random parameters,

$$\begin{pmatrix} \alpha_i \\ \boldsymbol{\beta}_i \end{pmatrix} = \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\pi}' \\ \boldsymbol{\Pi} \end{pmatrix} \mathbf{d}_i + \begin{pmatrix} \gamma w_i \\ \boldsymbol{\Gamma} \mathbf{v}_i \end{pmatrix} \quad (18-15)$$

where  $\mathbf{d}_i$  is a vector of demographics such as gender and age while  $\alpha, \boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \gamma$ , and  $\boldsymbol{\Gamma}$  are structural parameters to be estimated (assuming they are identified). A utility function is also defined for an “outside good” that is (presumably) chosen if the consumer chooses none of the brands  $1, \dots, J$ :

$$u_{i0t} = \alpha_i y_i + \Delta_{0t} + \boldsymbol{\pi}'_0 \mathbf{d}_i + \varepsilon_{i0t}.$$

Since there is no variation in income across the choices,  $\alpha_i y_i$  will fall out of the logit probabilities, as we saw earlier. A normalization is used instead,  $u_{i0t} = \varepsilon_{i0t}$ , so that comparisons of utilities are against the outside good. The resulting model can be reconstructed by inserting (18-15) into (18-14),

$$\begin{aligned} u_{ijt} &= \alpha_i y_i + \delta_{jt}(\mathbf{x}_{jt}, p_{jt}, \Delta_{jt} : \alpha, \boldsymbol{\beta}) + \tau_{ijt}(\mathbf{x}_{jt}, p_{jt}, \mathbf{v}_i, w_i : \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\gamma}, \boldsymbol{\Gamma}) + \varepsilon_{ijt} \\ \delta_{jt} &= \mathbf{x}_{jt}'\boldsymbol{\beta} - \alpha p_{jt} + \Delta_{jt} \\ \tau_{jt} &= [-p_{jt}, \mathbf{x}_{jt}'] \left[ \begin{pmatrix} \boldsymbol{\pi}' \\ \boldsymbol{\Pi} \end{pmatrix} d_i + \begin{pmatrix} \gamma w_i \\ \boldsymbol{\Gamma} \mathbf{v}_i \end{pmatrix} \right]. \end{aligned}$$

The preceding model defines the random utility model for consumer  $i$  in market  $t$ . Each consumer is assumed to purchase the one good that maximizes utility. The market share of the  $j$ th product in this market is obtained by summing over the choices made by those consumers. With the assumption of homogeneous tastes ( $\boldsymbol{\Gamma} = \mathbf{0}$  and  $\gamma = 0$ ) and i.i.d., type I extreme value distributions for  $\varepsilon_{ijt}$ , it follows that the market share of product  $j$  is

$$s_{jt} = \frac{\exp(\mathbf{x}_{jt}'\boldsymbol{\beta} - \alpha p_{jt} + \Delta_{jt})}{1 + \sum_{k=1}^J \exp(\mathbf{x}_{kt}'\boldsymbol{\beta} - \alpha p_{kt} + \Delta_{kt})}.$$

The IIA assumptions produce the familiar problems of peculiar and unrealistic substitution patterns among the goods. Alternatives considered include a nested logit, a “generalized extreme value” model and, finally, the mixed logit model, now applied to the aggregate data.

## 784 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics

Estimation cannot proceed along the lines of Section 18.2.7 because  $\Delta_{jt}$  is unobserved and  $p_{jt}$  is, therefore, endogenous. BLP propose, instead to use a GMM estimator, based on the moment equations

$$E\{[S_{jt} - s_{jt}(\mathbf{x}_{jt}, p_{jt}|\boldsymbol{\alpha}, \boldsymbol{\beta})]\mathbf{z}_{jt}\} = 0$$

for a suitable set of instruments. Layering in the random parameters specification, we obtain an estimation based on **method of simulated moments**, rather than a maximum simulated log likelihood. The simulated moments would be based on

$$E_{w,v}[S_{jt}(\mathbf{x}_{jt}, p_{jt}|\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)] = \int_{w,v} E_{\mathbf{x}_{jt}}[\mathbf{x}_{jt} p_{jt}|\boldsymbol{\alpha}_i(w), \boldsymbol{\beta}_i(v)] dF(w) dF(v).$$

These would be simulated using the method of Section 18.2.7.

### 18.3 RANDOM UTILITY MODELS FOR ORDERED CHOICES

The analysts at bond rating agencies such as Moody's and Standard and Poor provide an evaluation of the quality of a bond that is, in practice, a discrete listing of the continuously varying underlying features of the security. The rating scales are as follows:

<b>Rating</b>	<b>S&amp;P Rating</b>	<b>Moody's Rating</b>
Highest quality	AAA	Aaa
High quality	AA	Aa
Upper medium	A	A
Medium grade	BBB	Baa
Somewhat speculative	BB	Ba
Low grade, speculative	B	B
Low grade, default possible	CCC	Caa
Low grade, partial recovery possible	CC	Ca
Default, recovery unlikely	C	C

For another example, *Netflix* ([www.netflix.com](http://www.netflix.com)) is an Internet company that rents movies. Subscribers order the film online for download or home delivery of a DVD. The next time the customer logs onto the web site, they are invited to rate the movie on a five-point scale, where five is the highest, most favorable rating. The ratings of the many thousands of subscribers who rented that movie are averaged to provide a recommendation to prospective viewers. As of April 5, 2009, the average rating of the 2007 movie *National Treasure: Book of Secrets* given by approximately 12,900 visitors to the site was 3.8. Many other Internet sellers of products and services, such as Barnes and Noble, Amazon, Hewlett Packard, and Best Buy, employ rating schemes such as this. Many recently developed national survey data sets, such as the British Household Panel Data Set (BHPS) (<http://www.iser.essex.ac.uk/survey/bhps>) and the German Socioeconomic Panel (GSOEP) (<http://www.diw.de/en/soep>), contain questions that elicit self-assessed ratings of health, health satisfaction, or overall well-being. Like the other examples listed, these survey questions are answered on a discrete scale, such as the

CHAPTER 18 ♦ Discrete Choices and Event Counts **785**

zero to 10 scale of the question about health satisfaction in the GSOEP. Ratings such as these provide applications of the models and methods that interest us in this section.<sup>8</sup>

For any individual respondent, we hypothesize that there is a continuously varying strength of preferences that underlies the rating they submit. For convenience and consistency with what follows, we will label that strength of preference “utility,”  $U^*$ . Continuing the Netflix example, we describe utility as ranging over the entire real line:

$$-\infty < U_{im}^* < +\infty$$

where  $i$  indicates the individual and  $m$  indicates the movie. Individuals are invited to “rate” the movie on an integer scale from 1 to 5. Logically, then, the translation from underlying utility to a rating could be viewed as a *censoring* of the underlying utility,

$$\begin{aligned} R_{im} &= 1 \text{ if } -\infty < U_{im}^* \leq \mu_1, \\ R_{im} &= 2 \text{ if } \mu_1 < U_{im}^* \leq \mu_2, \\ R_{im} &= 3 \text{ if } \mu_2 < U_{im}^* \leq \mu_3, \\ R_{im} &= 4 \text{ if } \mu_3 < U_{im}^* \leq \mu_4, \\ R_{im} &= 5 \text{ if } \mu_4 < U_{im}^* < \infty. \end{aligned}$$

The same mapping would characterize the bond ratings, since the qualities of bonds that produce the ratings will vary continuously. The self-assessed health and well-being questions in the panel survey data sets based on an underlying utility or preference structure. The crucial feature of the description thus far is that underlying the discrete response is a continuous range of preferences. Therefore, the observed rating represents a censored version of the true underlying preferences. Providing a rating of five could be an outcome ranging from general enjoyment to wild enthusiasm. Note that the thresholds,  $\mu_j$ , number ( $J - 1$ ) where  $J$  is the number of possible ratings (here, five) –  $J - 1$  values are needed to divide the range of utility into  $J$  cells. The thresholds are an important element of the model; they divide the range of utility into cells that are then identified with the observed outcomes. Importantly, the difference between two levels of a rating scale (one compared to two, two compared to three) is not the same as on a utility scale, hence we have a strictly nonlinear transformation captured by the thresholds, which are estimable parameters in an **ordered choice model**.

The model as suggested thus far provides a crude description of the mechanism underlying an observed rating. Any individual brings their own set of *characteristics* to the utility function, such as age, income, education, gender, where they live, family situation, and so on, which we denote  $x_{i1}, x_{i2}, \dots, x_{iK}$ . They also bring their own aggregate of unmeasured and unmeasurable (by the statistician) idiosyncrasies, denoted  $\varepsilon_{im}$ . How these features enter the utility function is uncertain, but it is conventional to use a linear function, which produces a familiar *random utility function*:

$$U_{im}^* = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_K x_{iK} + \varepsilon_{im}.$$

<sup>8</sup>Greene and Hensher (2010) provide a survey of ordered choice modeling. Other textbook and monograph treatments include DeMaris (2004), Long (1997), Johnson and Abbot (1999), and Long and Freese (2006). Introductions to the model also appear in journal articles such as Winship and Mare (1984), Becker and Kennedy (1992), Daykin and Moffatt (2002), and Boes and Winkelmann (2006).

**786 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**
**Example 18.2 Movie Ratings**

The web site [www.imdb.com](http://www.imdb.com) invites visitors to rate movies that they have seen, in the same fashion as the Netflix site. This site uses a 10 point scale. On December 1, 2008, they reported the results in Figure 18.2 for the movie *National Treasure: Book of Secrets* for 41,771 users of the site. The earlier panel at the left shows the overall ratings. The panel at the right shows how the average rating varies across age, gender, and whether the rater is a U.S. viewer or not.

The rating mechanism we have constructed is

$$R_{im} = 1 \text{ if } -\infty < \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_{im} \leq \mu_1,$$

$$R_{im} = 2 \text{ if } \mu_1 < \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_{im} \leq \mu_2,$$

$$R_{im} = 3 \text{ if } \mu_2 < \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_{im} \leq \mu_3,$$

$$R_{im} = 4 \text{ if } \mu_3 < \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_{im} \leq \mu_4,$$

$$R_{im} = 5 \text{ if } \mu_4 < \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_{im} < \infty.$$

Relying on a central limit to aggregate the innumerable small influences that add up to the individual idiosyncrasies and movie attraction, we assume that the random component,  $\varepsilon_{im}$ , is normally distributed with zero mean and (for now) constant variance. The assumption of normality will allow us to attach probabilities to the ratings. In particular, arguably the most interesting one is

$$\text{Prob}(R_{im} = 5 | \mathbf{x}_i) = \text{Prob}[\varepsilon_{im} > \mu_4 - \mathbf{x}'_i \boldsymbol{\beta}].$$

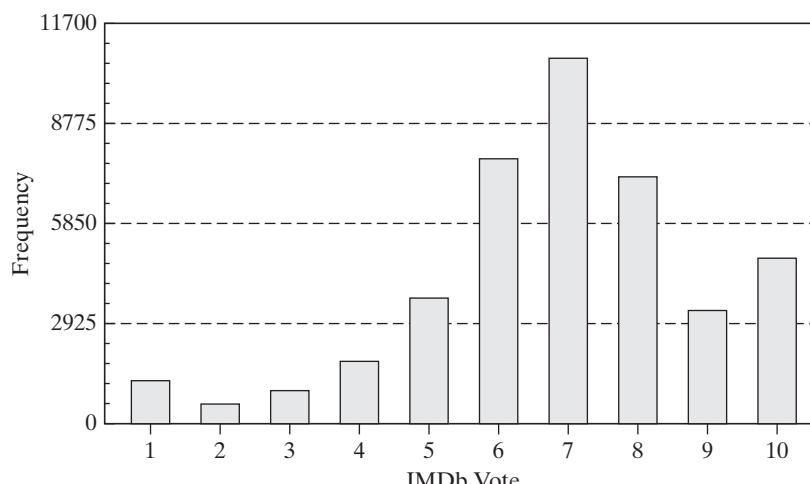
The structure provides the framework for an econometric model of how individuals rate movies (that they rent from Netflix). The resemblance of this model to familiar models of binary choice is more than superficial. For example, one might translate this econometric model directly into a probit model by focusing on the variable

$$E_{im} = 1 \text{ if } R_{im} = 5,$$

$$E_{im} = 0 \text{ if } R_{im} < 5.$$

Thus, the model is an extension of a binary choice model to a setting of more than two choices. But, the crucial feature of the model is the ordered nature of the observed outcomes and the correspondingly ordered nature of the underlying preference scale.

**FIGURE 18.2** IMDb.com Ratings ([www.imdb.com/title/tt0465234/ratings](http://www.imdb.com/title/tt0465234/ratings)).



CHAPTER 18 ♦ Discrete Choices and Event Counts **787**

The model described here is an *ordered choice model*. (The choice of the normal distribution for the random term makes it an *ordered probit model*.) Ordered choice models are appropriate for a wide variety of settings in the social and biological sciences. The essential ingredient is the *mapping from an underlying, naturally ordered preference scale to a discrete ordered observed outcome*, such as the rating scheme described. The model of ordered choice pioneered by Aitcheson and Silvey (1957), Snell (1964), and Walker and Duncan (1967) and articulated in its modern form by Zavoina and McElvey (1975) has become a widely used tool in many fields. The number of applications in the current literature is large and increasing rapidly, including

- Bond ratings [Terza (1985a)]
- Congressional voting on a Medicare bill [McElvey and Zavoina (1975)]
- Credit ratings [Cheung (1996), Metz, and Cantor (2006)]
- Driver injury severity in car accidents [Eluru, Bhat, and Hensher (2008)]
- Drug reactions [Fu, Gordon, Liu, Dale, and Christensen. (2004)]
- Education [Machin and Vignoles (2005), Carneiro, Hansen, and Heckman (2003), Cunha, Heckman, and Navarro (2007)]
- Financial failure of firms [Hensher and Jones (2007)]
- Happiness [Winkelmann (2005), Zigante (2007)]
- Health status [Jones, Koolman, and Rice (2003)]
- Life satisfaction [Clark, Georgellis, and Sanfey (2001), Groot and van den Brink (2003)]
- Monetary policy [Eichengreen, Watson, and Grossman (1985)]
- Nursing labor supply [Brewer, Kovner, Greene, and Cheng (2008)]
- Obesity [Greene, Harris, Hollingsworth, and Maitra (2008)]
- Political efficacy [King, Murray, Salomon, and Tandon (2004)]
- Pollution [Wang and Kockelman (2009)]
- Promotion and rank in nursing [Pudney and Shields (2000)]
- Stock price movements [Tsay (2005)]
- Tobacco use [Harris and Zhao (2007), Kasteridis, Munkin, and Yen (2008)]
- Work disability [Kapteyn et al. (2007)]



### 18.3.1 THE ORDERED PROBIT MODEL

The ordered probit model is built around a latent regression in the same manner as the binomial probit model. We begin with

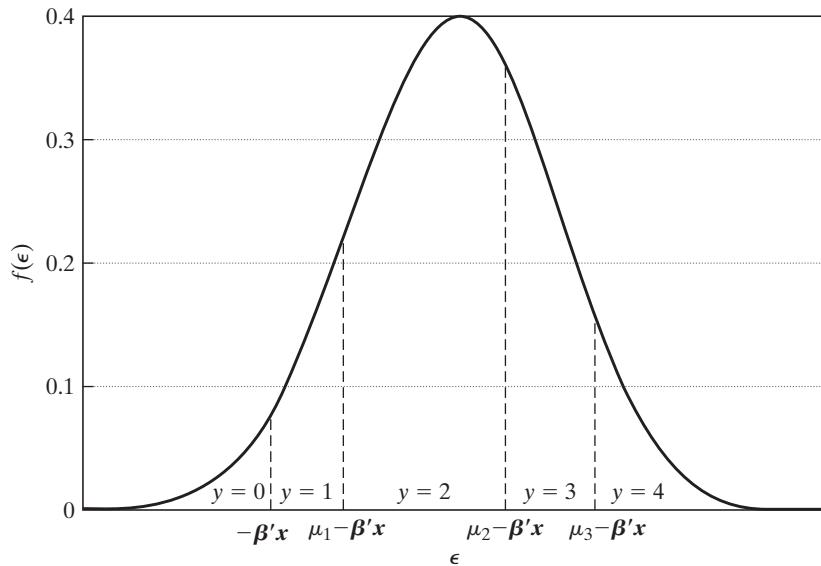
$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon.$$

As usual,  $y^*$  is unobserved. What we do observe is

$$\begin{aligned} y &= 0 && \text{if } y^* \leq 0 \\ &= 1 && \text{if } 0 < y^* \leq \mu_1 \\ &= 2 && \text{if } \mu_1 < y^* \leq \mu_2 \\ &\vdots \\ &= J && \text{if } \mu_{J-1} \leq y^*, \end{aligned}$$

which is a form of censoring. The  $\mu$ 's are unknown parameters to be estimated with  $\boldsymbol{\beta}$ .

**788** PART IV ♦ Cross Sections, Panel Data, and Microeometrics



**FIGURE 18.3** Probabilities in the Ordered Probit Model.

We assume that  $\varepsilon$  is normally distributed across observations.<sup>9</sup> For the same reasons as in the binomial probit model (which is the special case of  $J = 1$ ), we normalize the mean and variance of  $\varepsilon$  to zero and one. We then have the following probabilities:

$$\begin{aligned}\text{Prob}(y = 0 \mid \mathbf{x}) &= \Phi(-\mathbf{x}'\boldsymbol{\beta}), \\ \text{Prob}(y = 1 \mid \mathbf{x}) &= \Phi(\mu_1 - \mathbf{x}'\boldsymbol{\beta}) - \Phi(-\mathbf{x}'\boldsymbol{\beta}), \\ \text{Prob}(y = 2 \mid \mathbf{x}) &= \Phi(\mu_2 - \mathbf{x}'\boldsymbol{\beta}) - \Phi(\mu_1 - \mathbf{x}'\boldsymbol{\beta}), \\ &\vdots \\ \text{Prob}(y = J \mid \mathbf{x}) &= 1 - \Phi(\mu_{J-1} - \mathbf{x}'\boldsymbol{\beta}).\end{aligned}$$

For all the probabilities to be positive, we must have

$$0 < \mu_1 < \mu_2 < \cdots < \mu_{I-1}.$$

Figure 18.3 shows the implications of the structure. This is an extension of the univariate probit model we examined in chapter 17. The log-likelihood function and its derivatives can be obtained reusing the `ll` and `llv` functions, and optimization can be done by the usual means.

As usual, the marginal effects of the regressors  $\mathbf{x}$  on the probabilities are not equal to the coefficients. It is helpful to consider a simple example. Suppose there are three categories. The model thus has only one unknown threshold parameter. The three

<sup>9</sup>Other distributions, particularly the logistic, could be used just as easily. We assume the normal purely for convenience. The logistic and normal distributions generally give similar results in practice.

CHAPTER 18 ♦ Discrete Choices and Event Counts **789**

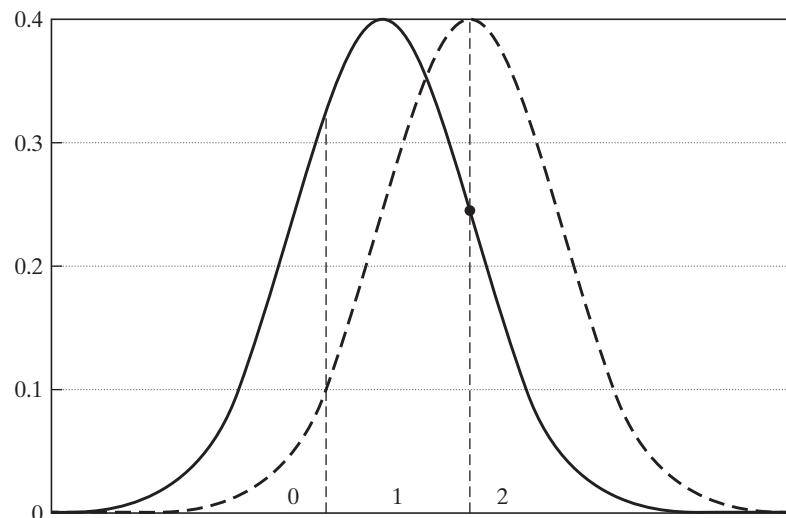
probabilities are

$$\begin{aligned}\text{Prob}(y = 0 | \mathbf{x}) &= 1 - \Phi(\mathbf{x}'\boldsymbol{\beta}), \\ \text{Prob}(y = 1 | \mathbf{x}) &= \Phi(\mu - \mathbf{x}'\boldsymbol{\beta}) - \Phi(-\mathbf{x}'\boldsymbol{\beta}), \\ \text{Prob}(y = 2 | \mathbf{x}) &= 1 - \Phi(\mu - \mathbf{x}'\boldsymbol{\beta}).\end{aligned}$$

For the three probabilities, the marginal effects of changes in the regressors are

$$\begin{aligned}\frac{\partial \text{Prob}(y = 0 | \mathbf{x})}{\partial \mathbf{x}} &= -\phi(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}, \\ \frac{\partial \text{Prob}(y = 1 | \mathbf{x})}{\partial \mathbf{x}} &= [\phi(-\mathbf{x}'\boldsymbol{\beta}) - \phi(\mu - \mathbf{x}'\boldsymbol{\beta})]\boldsymbol{\beta}, \\ \frac{\partial \text{Prob}(y = 2 | \mathbf{x})}{\partial \mathbf{x}} &= \phi(\mu - \mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}.\end{aligned}$$

Figure 18.4 illustrates the effect. The probability distributions of  $y$  and  $y^*$  are shown in the solid curve. Increasing one of the  $x$ 's while holding  $\boldsymbol{\beta}$  and  $\mu$  constant is equivalent to shifting the distribution slightly to the right, which is shown as the dashed curve. The effect of the shift is unambiguously to shift some mass out of the leftmost cell. Assuming that  $\boldsymbol{\beta}$  is positive (for this  $x$ ),  $\text{Prob}(y = 0 | \mathbf{x})$  must decline. Alternatively, from the previous expression, it is obvious that the derivative of  $\text{Prob}(y = 0 | \mathbf{x})$  has the opposite sign from  $\boldsymbol{\beta}$ . By a similar logic, the change in  $\text{Prob}(y = 2 | \mathbf{x})$  [or  $\text{Prob}(y = J | \mathbf{x})$  in the general case] must have the same sign as  $\boldsymbol{\beta}$ . Assuming that the particular  $\boldsymbol{\beta}$  is positive, we are shifting some probability into the rightmost cell. But what happens to the middle cell is ambiguous. It depends on the two densities. In the general case, relative to the signs of the coefficients, only the signs of the changes in  $\text{Prob}(y = 0 | \mathbf{x})$  and  $\text{Prob}(y = J | \mathbf{x})$  are unambiguous! The upshot is that we must be very careful



**FIGURE 18.4** Effects of Change in  $x$  on Predicted Probabilities.

## 790 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

in interpreting the coefficients in this model. Indeed, without a fair amount of extra calculation, it is quite unclear how the coefficients in the ordered probit model should be interpreted.

### **Example 18.3 Rating Assignments**

Marcus and Greene (1985) estimated an ordered probit model for the job assignments of new Navy recruits. The Navy attempts to direct recruits into job classifications in which they will be most productive. The broad classifications the authors analyzed were technical jobs with three clearly ranked skill ratings: "medium skilled," "highly skilled," and "nuclear qualified/highly skilled." Because the assignment is partly based on the Navy's own assessment and needs and partly on factors specific to the individual, an ordered probit model was used with the following determinants: (1) ENSPE = a dummy variable indicating that the individual entered the Navy with an "A school" (technical training) guarantee; (2) EDMA = educational level of the entrant's mother; (3) AFQT = score on the Armed Forces Qualifying Test; (4) EDYRS = years of education completed by the trainee; (5) MARR = a dummy variable indicating that the individual was married at the time of enlistment; and (6) AGEAT = trainee's age at the time of enlistment. (The data used in this study are not available for distribution.) The sample size was 5,641. The results are reported in Table 18.10. The extremely large *t* ratio on the AFQT score is to be expected, as it is a primary sorting device used to assign job classifications.

To obtain the marginal effects of the continuous variables, we require the standard normal density evaluated at  $-\bar{x}'\hat{\beta} = -0.8479$  and  $\hat{\mu} - \bar{x}'\hat{\beta} = 0.9421$ . The predicted probabilities are  $\Phi(-0.8479) = 0.198$ ,  $\Phi(0.9421) - \Phi(-0.8479) = 0.628$ , and  $1 - \Phi(0.9421) = 0.174$ . (The actual frequencies were 0.25, 0.52, and 0.23.) The two densities are  $\phi(-0.8479) = 0.278$  and  $\phi(0.9421) = 0.255$ . Therefore, the derivatives of the three probabilities with respect to AFQT, for example, are

$$\begin{aligned}\frac{\partial P_0}{\partial \text{AFQT}} &= (-0.278)0.039 = -0.01084, \\ \frac{\partial P_1}{\partial \text{AFQT}} &= (0.278 - 0.255)0.039 = 0.0009, \\ \frac{\partial P_2}{\partial \text{AFQT}} &= 0.255(0.039) = 0.00995.\end{aligned}$$

Note that the marginal effects sum to zero, which follows from the requirement that the probabilities add to one. This approach is not appropriate for evaluating the effect of a dummy variable. We can analyze a dummy variable by comparing the probabilities that result when the variable takes its two different values with those that occur with the other variables held at their sample means. For example, for the MARR variable, we have the results given in Table 18.11.

**TABLE 18.10** Estimated Rating Assignment Equation

Variable	Estimate	t Ratio	Mean of Variable
Constant	-4.34	—	—
ENSPA	0.057	1.7	0.66
EDMA	0.007	0.8	12.1
AFQT	0.039	39.9	71.2
EDYRS	0.190	8.7	12.1
MARR	-0.48	-9.0	0.08
AGEAT	0.0015	0.1	18.8
$\mu$	1.79	80.8	—

CHAPTER 18 ♦ Discrete Choices and Event Counts **791****TABLE 18.11** Marginal Effect of a Binary Variable

	$-\hat{\beta}'\mathbf{x}$	$\hat{\mu} - \hat{\beta}'\mathbf{x}$	<i>Prob[y = 0]</i>	<i>Prob[y = 1]</i>	<i>Prob[y = 2]</i>
MARR = 0	-0.8863	0.9037	0.187	0.629	0.184
MARR = 1	-0.4063	1.3837	0.342	0.574	0.084
Change			0.155	-0.055	-0.100

**18.3.2 A SPECIFICATION TEST FOR THE ORDERED CHOICE MODEL**

The basic formulation of the ordered choice model implies that for constructed binary variables,

$$w_{ij} = 1 \text{ if } y_i \leq j, 0 \text{ otherwise, } j = 1, 2, \dots, J-1, \quad (18-16)$$

$$\text{Prob}(w_{ij} = 1 | \mathbf{x}_i) = F(\mathbf{x}_i' \boldsymbol{\beta} - \mu_j).$$

The first of these, when  $j = 1$ , is the binary choice model of Section 17.2. One implication is that we could estimate the slopes, but not the threshold parameters, in the ordered choice model just by using  $w_{i1}$  and  $\mathbf{x}_i$  in a binary probit or logit model. (Note that this result also implies the validity of combining adjacent cells in the ordered choice model.) But, (18-16) also defines a set of  $J-1$  binary choice models with different constants but common slope vector,  $\boldsymbol{\beta}$ . This equality of the parameter vectors in (18-16) has been labeled the **parallel regression assumption**. Although it is merely an implication of the model specification, this has been viewed as an implicit restriction on the model. [See, e.g., Long (1997, p. 141).] Brant (1990) suggests a test of the parallel regressions assumption based on (18-16). One can, in principle, fit  $J-1$  such binary choice models separately. Each will produce its own constant term and a consistent estimator of the common  $\boldsymbol{\beta}$ . Brant's Wald test examines the linear restrictions  $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \dots = \boldsymbol{\beta}_{J-1}$ , or  $H_0: \boldsymbol{\beta}_q - \boldsymbol{\beta}_1 = \mathbf{0}, q = 2, \dots, J-1$ . The Wald statistic will be

$$\chi^2[(J-2)K] = (\mathbf{R}\hat{\boldsymbol{\beta}}^*)'[\mathbf{R} \times \text{Asy.Var}[\hat{\boldsymbol{\beta}}^*] \times \mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}^*),$$

where  $\hat{\boldsymbol{\beta}}^*$  is obtained by stacking the individual binary logit or probit estimates of  $\boldsymbol{\beta}$  (without the constant terms). [See Brant (1990), Long (1997), or Greene and Hensher (2010, page 187) for details on computing the statistic.]

Rejection of the null hypothesis calls the model specification into question. An alternative model in which there is a different  $\boldsymbol{\beta}$  for each value of  $y$  has two problems: it does not force the probabilities to be positive and it is internally inconsistent. On the latter point, consider the suggested latent regression,  $y^* = \mathbf{x}'\boldsymbol{\beta}_j + \varepsilon$ . If the “ $\boldsymbol{\beta}$ ” is different for each  $j$ , then it is not possible to construct a data generating mechanism for  $y^*$  (or, for example, simulate it); the realized value of  $y^*$  cannot be defined without knowing  $y$  (i.e., the realized  $j$ ), since the applicable  $\boldsymbol{\beta}$  depends on  $j$ , but  $y$  is supposed to be determined from  $y^*$  through, for example, (18-16). There is no parametric restriction other than the one we seek to avoid that will preserve the ordering of the probabilities for all values of the data and maintain the coherency of the model. This still leaves the question of what specification failure would logically explain the finding. Some suggestions in Brant (1990) include (1) misspecification of the latent regression,  $\mathbf{x}'\boldsymbol{\beta}$ ; (2) heteroscedasticity

## 792 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

of  $\varepsilon$ ; and (3) misspecification of the distributional form for the latent variable, that is, “nonlogistic link function.”

### **Example 18.4 Brant Test for an Ordered Probit Model of Health Satisfaction**

In Example 17.4, we studied the health care usage of a sample of households in the German Socioeconomic Panel (GSOEP). The data include a self-reported measure of “health satisfaction,” (HSAT) that is coded 0–10. This variable provides a natural application of the ordered choice models in this chapter. The data are an unbalanced panel. For purposes of this exercise, we have used the fifth (1984) wave of the data set, which is a cross section of 4,483 observations. We then collapsed the 10 cells into 5 [(0–2),(3–5), (6–8),(9),(10)] for this example. The utility function is

$$\begin{aligned} HSAT_i^* &= \beta_1 + \beta_2 AGE_i + \beta_3 INCOME_i + \beta_4 KIDS_i \\ &\quad + \beta_5 EDUC_i + \beta_6 MARRIED_i - \beta_7 WORKING_i + \varepsilon_i. \end{aligned}$$

Variables KIDS, MARRIED, and WORKING, are binary indicators of whether there are children in the household, marital status, and whether the individual was working at the time of the survey. (These data are examined further in Example 18.6.) The model contains six variables, and there are four binary choice models fit, so there are  $(J-2)(K) = (3)(6) = 18$  restrictions. The chi-squared for the probit model is 87.836. The critical value for 95 percent is 28.87, so the homogeneity restriction is rejected. The corresponding value for the logit model is 77.84, which leads to the same conclusion.

### 18.3.3 BIVARIATE ORDERED PROBIT MODELS

There are several extensions of the ordered probit model that follow the logic of the bivariate probit model we examined in Section 17.5. A direct analog to the base case two-equation model is used in the study in Example 18.5.

### **Example 18.5 Calculus and Intermediate Economics Courses**

Butler et al. (1994) analyzed the relationship between the level of calculus attained and grades in intermediate economics courses for a sample of Vanderbilt students. The two-step estimation approach involved the following strategy. (We are stylizing the precise formulation a bit to compress the description.) Step 1 involved a direct application of the ordered probit model of Section 18.3.1 to the level of calculus achievement, which is coded 0, 1, . . . , 6:

$$\begin{aligned} m_i^* &= \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i | \mathbf{x}_i \sim N[0, 1], \\ m_i &= 0 \text{ if } -\infty < m_i^* \leq 0 \\ &= 1 \text{ if } 0 < m_i^* \leq \mu_1 \\ &\quad \dots \\ &= 6 \text{ if } \mu_5 < m_i^* < +\infty. \end{aligned}$$

The authors argued that although the various calculus courses can be ordered discretely by the material covered, the differences between the levels cannot be measured directly. Thus, this is an application of the ordered probit model. The independent variables in this first-step model included SAT scores, foreign language proficiency, indicators of intended major, and several other variables related to areas of study.

The second step of the estimator involves regression analysis of the grade in the intermediate microeconomics or macroeconomics course. Grades in these courses were translated to a granular continuous scale (A = 4.0, A– = 3.7, etc.). A linear regression is specified,

$$Grade_i = \mathbf{z}'_i \boldsymbol{\delta} + u_i, \quad \text{where } u_i | \mathbf{z}_i \sim N[0, \sigma_u^2].$$

CHAPTER 18 ♦ Discrete Choices and Event Counts **793**

Independent variables in this regression include, among others, (1) dummy variables for which outcome in the ordered probit model applies to the student (with the zero reference case omitted), (2) grade in the last calculus course, (3) several other variables related to prior courses, (4) class size, (5) freshman GPA, and so on. The unobservables in the *Grade* equation and the math attainment are clearly correlated, a feature captured by the additional assumption that  $(\varepsilon_i, u_i | \mathbf{x}_i, \mathbf{z}_i) \sim N_2[(0, 0), (1, \sigma_u^2), \rho \sigma_u]$ . A nonzero  $\rho$  captures this “selection” effect. With this in place, the dummy variables in (1) have now become endogenous. The solution is a “selection” correction that we will examine in detail in Chapter 19. The modified equation becomes

$$\begin{aligned} Grade_i | m_i &= \mathbf{z}'_i \delta + E[u_i | m_i] + v_i \\ &= \mathbf{z}'_i \delta + (\rho \sigma_u) [\lambda(\mathbf{x}'_i \beta, \mu_1, \dots, \mu_5)] + v_i. \end{aligned}$$

They thus adopt a “control function” approach to accommodate the endogeneity of the math attainment dummy variables. [See Section 17.3.5 and (17-32) for another application of this method.] The term  $\lambda(\mathbf{x}'_i \beta, \mu_1, \dots, \mu_5)$  is a generalized residual that is constructed using the estimates from the first-stage ordered probit model. [A precise statement of the form of this variable is given in Li and Tobias (2006).] Linear regression of the course grade on  $\mathbf{z}_i$  and this constructed regressor is computed at the second step. The standard errors at the second step must be corrected for the use of the estimated regressor using what amounts to a Murphy and Topel (2002) correction. (See Section 14.7.)

Li and Tobias (2006) in a replication of and comment on Butler et al. (1994), after roughly replicating the classical estimation results with a Bayesian estimator, observe that the preceding *Grade* equation above could also be treated as an ordered probit model. The resulting **bivariate ordered probit** model would be

$$\begin{array}{lll} m_i^* = \mathbf{x}'_i \beta + \varepsilon_i, & \text{and} & g_i^* = \mathbf{z}'_i \delta + u_i, \\ m_i = 0 \text{ if } -\infty < m_i^* \leq 0 & & g_i = 0 \text{ if } -\infty < g_i^* \leq 0 \\ = 1 \text{ if } 0 < m_i^* \leq \mu_1 & & = 1 \text{ if } 0 < g_i^* \leq \alpha_1 \\ \dots & & \dots \\ = 6 \text{ if } \mu_5 < m_i^* < +\infty. & & = 11 \text{ if } \mu_9 < g_i^* < +\infty \end{array}$$

where

$$(\varepsilon_i, u_i | \mathbf{x}_i, \mathbf{z}_i) \sim N_2 [(0, 0), (1, \sigma_u^2), \rho \sigma_u].$$

Li and Tobias extended their analysis to this case simply by “transforming” the dependent variable in Butler et al.’s second equation. Computing the log-likelihood using sets of bivariate normal probabilities is fairly straightforward for the bivariate ordered probit model. [See Greene (2007).] However, the classical study of these data using the bivariate ordered approach remains to be done, so a side-by-side comparison to Li and Tobias’s Bayesian alternative estimator is not possible. The endogeneity of the calculus dummy variables in (1) remains a feature of the model, so both the MLE and the Bayesian posterior are less straightforward than they might appear. Whether the results in Section 17.5.5 on the recursive bivariate probit model extend to this case also remains to be determined.

The bivariate ordered probit model has been applied in a number of settings in the recent empirical literature, including husband and wife’s education levels [Magee et al. (2000)], family size [(Calhoun (1991))], and many others. In two early contributions to the field of pet econometrics, Butler and Chatterjee analyze ownership of cats and dogs (1995) and dogs and televisions (1997).

## 794 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

### 18.3.4 PANEL DATA APPLICATIONS

The ordered probit model is used to model discrete scales that represent indicators of a continuous underlying variable such as strength of preference, performance, or level of attainment. Many of the recently assembled national panel data sets contain survey questions that ask about subjective assessments of health, satisfaction, or well-being, all of which are applications of this interpretation. Examples include the following:

- The European Community Household Panel (ECHP) includes questions about job satisfaction [see D'Addio (2004)].
- The British Household Panel Survey (BHPS) includes questions about health status [see Contoyannis et al. (2004)].
- The German Socioeconomic Household Panel (GSOEP) includes questions about subjective well-being [see Winkelmann (2004)] and subjective assessment of health satisfaction [see Riphahn et al. (2003) and Example 18.4.]

Ostensibly, the applications would fit well into the ordered probit frameworks already described. However, given the panel nature of the data, it will be desirable to augment the model with some accommodation of the individual heterogeneity that is likely to be present. The two standard models, fixed and random effects, have both been applied to the analyses of these survey data.

#### 18.3.4.a Ordered Probit Models with Fixed Effects

D'Addio et al. (2003), using methodology developed by Frijters et al. (2004) and Ferrer-i-Carbonel et al. (2004), analyzed survey data on job satisfaction using the Danish component of the European Community Household Panel. Their estimator for an ordered logit model is built around the logic of Chamberlain's estimator for the binary logit model. [See Section 17.4.4.] Because the approach is robust to individual specific threshold parameters and allows time-invariant variables, it differs sharply from the fixed effects models  have considered thus far as well as from the ordered probit model of Section 23.10.1.<sup>10</sup> Unlike Chamberlain's estimator for the binary logit model, however, their conditional estimator is not a function of minimal sufficient statistics. As such, the incidental parameters problem remains an issue.

Das and van Soest (2000) proposed a somewhat simpler approach. [See, as well, Long's (1997) discussion of the “parallel regressions assumption,” which employs this device in a cross-section framework]. Consider the base case ordered logit model with fixed effects,

$$y_{it}^* = \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}, \quad \varepsilon_{it} | \mathbf{X}_i \sim N[0, 1], \\ y_{it} = j \quad \text{if} \quad \mu_{j-1} < y_{it}^* < \mu_j, \quad j = 0, 1, \dots, J \quad \text{and} \quad \mu_{-1} = -\infty, \mu_0 = 0, \mu_J = +\infty.$$

The model assumptions imply that

$$\text{Prob}(y_{it} = j | \mathbf{X}_i) = \Lambda(\mu_j - \alpha_i - \mathbf{x}'_{it}\boldsymbol{\beta}) - \Lambda(\mu_{j-1} - \alpha_i - \mathbf{x}'_{it}\boldsymbol{\beta}),$$

<sup>10</sup>Cross-section versions of the ordered probit model with individual specific thresholds appear in Terza (1985a), Pudney and Shields (2000), and Greene (2007).

CHAPTER 18 ♦ Discrete Choices and Event Counts **795**

where  $\Lambda(t)$  is the cdf of the logistic distribution. Now, define a binary variable

$$w_{it,j} = 1 \text{ if } y_{it} > j, \quad j = 0, \dots, J - 1.$$

It follows that

$$\begin{aligned} \text{Prob}[w_{it,j} = 1 | \mathbf{X}_i] &= \Lambda(\alpha_i - \mu_j + \mathbf{x}'_{it}\boldsymbol{\beta}) \\ &= \Lambda(\theta_i + \mathbf{x}'_{it}\boldsymbol{\beta}). \end{aligned}$$

The “ $j$ ” specific constant, which is the same for all individuals, is absorbed in  $\theta_i$ . Thus, a fixed effects binary logit model applies to each of the  $J - 1$  binary random variables,  $w_{it,j}$ . The method in Section 17.4.4 can now be applied to each of the  $J - 1$  random samples. This provides  $J - 1$  estimators of the parameter vector  $\boldsymbol{\beta}$  (but no estimator of the threshold parameters). The authors propose to reconcile these different estimators by using a minimum distance estimator of the common true  $\boldsymbol{\beta}$ . (See Section 13.3.) The minimum distance estimator at the second step is chosen to minimize

$$q = \sum_{j=0}^{J-1} \sum_{m=0}^{J-1} (\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta})' [\mathbf{V}_{jm}^{-1}] (\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}),$$

where  $[\mathbf{V}_{jm}^{-1}]$  is the  $j, m$  block of the inverse of the  $(J - 1)K \times (J - 1)K$  partitioned matrix  $\mathbf{V}$  that contains Asy. Cov $[\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_m]$ . The appropriate form of this matrix for a set of cross-section estimators is given in Brant (1990). Das and van Soest (2000) used the counterpart for Chamberlain's fixed effects estimator but do not provide the specifics for computing the off-diagonal blocks in  $\mathbf{V}$ .

The full ordered probit model with fixed effects, including the individual specific constants, can be estimated by unconditional maximum likelihood using the results in Section 14.9.6.d. The likelihood function is concave [see Pratt (1981)], so despite its superficial complexity, the estimation is straightforward. (In the following application, with more than 27,000 observations and 7,293 individual effects, estimation of the full model required roughly five seconds of computation.) No theoretical counterpart to the Hsiao (1986, 2003) and Abrevaya (1997) results on the small  $T$  bias (incidental parameters problem) of the MLE in the presence of fixed effects has been derived for the ordered probit model. The Monte Carlo results in Greene (2004) (see, as well, Chapter 15), suggest that biases comparable to those in the binary choice models persist in the ordered probit model as well. As in the binary choice case, the complication of the fixed effects model is the small sample bias, not the computation. The Das and van Soest approach finesse this problem—their estimator is consistent—but at the cost of losing the information needed to compute partial effects or predicted probabilities.

#### 18.3.4.b Ordered Probit Models with Random Effects

The random effects ordered probit model has been much more widely used than the fixed effects model. Applications include Groot and van den Brink (2003), who studied training levels of employees, with firm effects; Winkelmann (2003b), who examined subjective measures of well-being with individual and family effects; Contoyannis et al. (2004), who analyzed self-reported measures of health status; and numerous others. In the simple case, the method of the Butler and Moffitt (1982) quadrature method (Section 14.9.6.b) can be extended to this model.

**796 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

**Example 18.6 Health Satisfaction**

The GSOEP German Health Care data that we have used in Example 17.4, and others includes a self-reported measure of health satisfaction,  $HSAT$ , that takes values  $0, 1, \dots, 10$ .<sup>11</sup> This is a typical application of a scale variable that reflects an underlying continuous variable, "health." The frequencies and sample proportions for the reported values are as follows:

<b>HSAT</b>	<b>Frequency</b>	<b>Proportion</b>
0	447	1.6%
1	255	0.9%
2	642	2.3%
3	1173	4.2%
4	1390	5.0%
5	4233	15.4%
6	2530	9.2%
7	4231	15.4%
8	6172	22.5%
9	3061	11.2%
10	3192	11.6%

We have fit pooled and panel data versions of the ordered probit model to these data. The model used is

$$y_{it}^* = \beta_1 + \beta_2 Age_{it} + \beta_3 Income_{it} + \beta_4 Kids_{it} \beta_6 Education_{it} + \beta_6 Married_{it} + \beta_7 Working_{it} + \varepsilon_{it} + c_i,$$

where  $c_i$  will be the common fixed or random effect. (We are interested in comparing the fixed and random effects estimators, so we have not included any time-invariant variables such as gender in the equation.) Table 18.12 lists five estimated models. (Standard errors for the estimated threshold parameters are omitted.) The first is the pooled ordered probit model. The second and third are fixed effects. Column 2 shows the unconditional fixed effects estimates using the results of Section 14.9.6.d. Column 3 shows the Das and van Soest estimator. For the minimum distance estimator, we used an inefficient weighting matrix, the block-diagonal matrix in which the  $j$ th block is the inverse of the  $j$ th asymptotic covariance matrix for the individual logit estimators. With this weighting matrix, the estimator is

$$\hat{\beta}_{MDE} = \left[ \sum_{j=0}^g \mathbf{V}_j^{-1} \right]^{-1} \sum_{j=0}^g \mathbf{V}_j^{-1} \hat{\beta}_j,$$

and the estimator of the asymptotic covariance matrix is approximately equal to the bracketed inverse matrix. The fourth set of results is the random effects estimator computed using the maximum simulated likelihood method. This model can be estimated using Butler and Moffitt's quadrature method; however, we found that even with a large number of nodes, the quadrature estimator converged to a point where the log-likelihood was far lower than the MSL estimator, and at parameter values that were implausibly different from the other estimates. Using different starting values and different numbers of quadrature points did not change this outcome. The MSL estimator for a random constant term (see Section 15.6) is considerably slower but produces more reasonable results. The fifth set of results is the Mundlak form of the random effects model, which includes the group means in the models as controls to accommodate possible correlation between the latent heterogeneity and the included variables. As noted in Example 17.6, the components of the ordered choice model must be interpreted with some care. By construction, the partial effects of the variables on



<sup>11</sup>In the original data set, 40 (of 27,326) observations on this variable were coded with noninteger values between 6 and 7. For purposes of our example, we have recoded all 40 observations to 7.

CHAPTER 18 ♦ Discrete Choices and Event Counts **797****TABLE 18.12** Estimated Ordered Probit Models for Health Satisfaction

Variable	(1) Pooled	(2)	(3)	(4)	(5) Random Effects Mundlak Controls	
		Fixed Effects Unconditional	Fixed Effects Conditional	Random Effects	Variables	Means
Constant	2.4739 (0.04669)			3.8577 (0.05072)	3.2603 (0.05323)	
Age	-0.01913 (0.00064)	-0.07162 (0.002743)	-0.1011 (0.002878)	-0.03319 (0.00065)	-0.06282 (0.00234)	0.03940 (0.002442)
Income	0.1811 (0.03774)	0.2992 (0.07058)	0.4353 (0.07462)	0.09436 (0.03632)	0.2618 (0.06156)	0.1461 (0.07695)
Kids	0.06081 (0.01459)	-0.06385 (0.02837)	-0.1170 (0.03041)	0.01410 (0.01421)	-0.05458 (0.02566)	0.1854 (0.03129)
Education	0.03421 (0.002828)	0.02590 (0.02677)	0.06013 (0.02819)	0.04728 (0.002863)	0.02296 (0.02793)	0.02257 (0.02807)
Married	0.02574 (0.01623)	0.05157 (0.04030)	0.08505 (0.04181)	0.07327 (0.01575)	0.04605 (0.03506)	-0.04829 (0.03963)
Working	0.1292 (0.01403)	-0.02659 (0.02758)	-0.007969 (0.02830)	0.07108 (0.01338)	-0.02383 (0.02311)	0.2702 (0.02856)
$\mu_1$	0.1949	0.3249		0.2726		0.2752
$\mu_2$	0.5029	0.8449		0.7060		0.7119
$\mu_3$	0.8411	1.3940		1.1778		1.1867
$\mu_4$	1.111	1.8230		1.5512		1.5623
$\mu_5$	1.6700	2.6992		2.3244		2.3379
$\mu_6$	1.9350	3.1272		2.6957		2.7097
$\mu_7$	2.3468	3.7923		3.2757		3.2911
$\mu_8$	3.0023	4.8436		4.1967		4.2168
$\mu_9$	3.4615	5.5727		4.8308		4.8569
$\sigma_u$	0.0000	0.0000		1.0078		0.9936
ln L	-56813.52	-41875.63		-53215.54		-53070.43

**TABLE 18.13** Estimated Marginal Effects: Pooled Model

HSAT	Age	Income	Kids	Education	Married	Working
0	0.0006	-0.0061	-0.0020	-0.0012	-0.0009	-0.0046
1	0.0003	-0.0031	-0.0010	-0.0006	-0.0004	-0.0023
2	0.0008	-0.0072	-0.0024	-0.0014	-0.0010	-0.0053
3	0.0012	-0.0113	-0.0038	-0.0021	-0.0016	-0.0083
4	0.0012	-0.0111	-0.0037	-0.0021	-0.0016	-0.0080
5	0.0024	-0.0231	-0.0078	-0.0044	-0.0033	-0.0163
6	0.0008	-0.0073	-0.0025	-0.0014	-0.0010	-0.0050
7	0.0003	-0.0024	-0.0009	-0.0005	-0.0003	-0.0012
8	-0.0019	0.0184	0.0061	0.0035	0.0026	0.0136
9	-0.0021	0.0198	0.0066	0.0037	0.0028	0.0141
10	-0.0035	0.0336	0.0114	0.0063	0.0047	0.0233

the probabilities of the outcomes must change sign, so the simple coefficients do not show the complete picture implied by the estimated model. Table 18.13 shows the partial effects for the pooled model to illustrate the computations.

Winkelmann (2003b) used the random effects approach to analyze the **subjective well-being** (SWB) question (also coded 0 to 10) in the German Socioeconomic

## 798 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

Panel (GSOEP) data set. The ordered probit model in this study is based on the latent regression

$$y_{imt}^* = \mathbf{x}'_{imt} \boldsymbol{\beta} + \varepsilon_{imt} + u_{im} + v_i.$$

The independent variables include age, gender, employment status, income, family size, and an indicator for good health. An unusual feature of the model is the nested random effects (see Section 14.9.6.b), which include a family effect,  $v_i$ , as well as the individual family member ( $i$  in family  $m$ ) effect,  $u_{im}$ . The GLS/MLE approach we applied to the linear regression model in Section 14.9.6.b is unavailable in this nonlinear setting. Winkelmann instead employed a Hermite quadrature procedure to maximize the log-likelihood function.

Contoyannis, Jones, and Rice (2004) analyzed a self-assessed health scale that ranged from 1 (very poor) to 5 (excellent) in the British Household Panel Survey. Their model accommodated a variety of complications in survey data. The latent regression underlying their ordered probit model is

$$h_{it}^* = \mathbf{x}'_{it} \boldsymbol{\beta} + \mathbf{H}'_{i,t-1} \boldsymbol{\gamma} + \alpha_i + \varepsilon_{it},$$

where  $\mathbf{x}_{it}$  includes marital status, race, education, household size, age, income, and number of children in the household. The lagged value,  $\mathbf{H}_{i,t-1}$ , is a set of binary variables for the observed health status in the previous period. (This is the same device that was used by Butler et al. in Example 18.5.) In this case, the lagged values capture state dependence—the assumption that the health outcome is redrawn randomly in each period is inconsistent with evident runs in the data. The initial formulation of the regression is a fixed effects model. To control for the possible correlation between the effects,  $\alpha_i$ , and the regressors, and the initial conditions problem that helps to explain the state dependence, they use a hybrid of Mundlak's (1978) correction and a suggestion by Wooldridge (2002a) for modeling the initial conditions,

$$\alpha_i = \alpha_0 + \bar{\mathbf{x}}' \boldsymbol{\alpha}_1 + \mathbf{H}'_{i,1} \boldsymbol{\delta} + u_i,$$

where  $u_i$  is exogenous. Inserting the second equation into the first produces a random effects model that can be fit using the quadrature method we considered earlier.

### 18.3.5 EXTENSIONS OF THE ORDERED PROBIT MODEL

The basic specification of the ordered probit model can be extended in the same directions as we considered in constructing models for binary choice in Chapter 17. These include heteroscedasticity in the random utility function [see Section 17.3.7.b, Keele and Park (2005), and Wang and Kockelman (2005), for an application] and heterogeneity in the preferences (i.e., random parameters and latent classes). [An extensive study of heterogeneity in health satisfaction based on 22 waves of the GSOEP is Jones and Schurer (2010).] Two specification issues that are specific to the ordered choice model are accommodating heterogeneity in the threshold parameters and reconciling differences in the meaning of the preference scale across different groups. We will sketch the model extensions in this section. Further details are given in Chapters 6 and 7 of Hensher and Greene (2010).

## CHAPTER 18 ♦ Discrete Choices and Event Counts **799**

### **18.3.5.a Threshold Models—Generalized Ordered Choice Models**

The model analyzed thus far assumes that the thresholds  $\mu_j$  are the same for every individual in the sample. Terza (1985a), Pudney and Shields (2000), King, Murray, Salomon and Tandon (KMST, 2004), Boes and Winkelmann (2006a), Greene, Harris, Hollingsworth and Maitra (2008), and Greene and Hensher (2009) all present applications that include individual variation in the thresholds of the ordered choice model.

In his analysis of bond ratings, Terza (1985) suggested the generalization,

$$\mu_{ij} = \mu_i + \mathbf{x}_i' \boldsymbol{\delta}.$$

With three outcomes, the probabilities are 

$$y_i^* = \alpha + \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i,$$

and

$$\begin{aligned} y_i &= 0 \text{ if } y_i^* \leq 0, \\ &1 \text{ if } 0 < y_i^* \leq \mu + \mathbf{x}_i' \boldsymbol{\delta}, \\ &2 \text{ if } y_i^* > \mu + \mathbf{x}_i' \boldsymbol{\delta}. \end{aligned}$$

For three outcomes, the model has two thresholds,  $\mu_0 = 0$  and  $\mu_1 = \mu + \mathbf{x}_i' \boldsymbol{\delta}$ . The three probabilities can be written

$$\begin{aligned} P_0 &= \text{Prob}(y_i = 0 | \mathbf{x}_i) = \Phi[-(\alpha + \mathbf{x}_i' \boldsymbol{\beta})] \\ P_1 &= \text{Prob}(y_i = 1 | \mathbf{x}_i) = \Phi[(\mu + \mathbf{x}_i' \boldsymbol{\delta}) - (\alpha + \mathbf{x}_i' \boldsymbol{\beta})] - \Phi[-(\alpha + \mathbf{x}_i' \boldsymbol{\beta})] \\ P_2 &= \text{Prob}(y_i = 2 | \mathbf{x}_i) = 1 - \Phi[(\mu + \mathbf{x}_i' \boldsymbol{\delta}) - (\alpha + \mathbf{x}_i' \boldsymbol{\beta})]. \end{aligned}$$

For applications of this approach, see, for example, Kerkhofs and Lindeboom (1995), Groot and van den Brink (2003) and Lindeboom and van Doorslayer (2003). Note that if  $\boldsymbol{\delta}$  is unrestricted, then  $\text{Prob}(y_i = 1 | \mathbf{x}_i)$  can be negative. This is a shortcoming of the model when specified in this form. Subsequent development of the generalized model involves specifications that avoid this internal inconsistency. Note, as well, that if the model is recast in terms of  $\mu$  and  $\boldsymbol{\gamma} = [\alpha, (\boldsymbol{\beta} - \boldsymbol{\delta})]$ , then the model is not distinguished from the original ordered probit model with a constant threshold parameter. This identification issue emerges prominently in Pudney and Shield's (2000) continued development of this model.

Pudney and Shield's (2000) "generalized ordered probit model," was also formulated to accommodate *observable* individual heterogeneity in the threshold parameters. Their application was in the context of job promotion for UK nurses in which the steps on the promotion ladder are individual specific. In their setting, in contrast to Terza's, some of the variables in the threshold equations are explicitly different from those in the regression. The authors constructed a generalized model and a test of "threshold constancy" by defining  $\mathbf{q}_i$  to include a constant term and those variables that are unique to the threshold model. Variables that are common to both the thresholds and the regression are placed in  $\mathbf{x}_i$  and the model is reparameterized as

$$\text{Pr}(y_i = g | \mathbf{x}_i, \mathbf{q}_i) = \Phi[\mathbf{q}_i' \boldsymbol{\delta}_g - \mathbf{x}_i' (\boldsymbol{\beta} - \boldsymbol{\delta}_g)] - \Phi[\mathbf{q}_i' \boldsymbol{\delta}_{g-1} - \mathbf{x}_i' (\boldsymbol{\beta} - \boldsymbol{\delta}_{g-1})].$$

## 800 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

An important point noted by the authors is that the same model results if these common variables are placed in the thresholds instead. This is a minor algebraic result, but it exposes an ambiguity in the interpretation of the model—whether a particular variable affects the regression or the thresholds is one of the issues that was developed in the original model specification.

As will be evident in the application in the next section, the specification of the threshold parameters is a crucial feature of the ordered choice model. KMST (2004), Greene (2007a), Eluru, Bhat, and Hensher (2008), and Greene and Hensher (2009) employ a “hierarchical ordered probit” or HOPIT model,

$$\begin{aligned} y_i^* &= \beta' \mathbf{x}_i + \varepsilon_i, \\ y_i &= j \text{ if } \mu_{i,j-1} \leq y_i^* < \mu_{ij}, \\ \mu_0 &= 0, \\ \mu_{i,j} &= \exp(\lambda_j + \gamma' \mathbf{z}_i) \quad (\text{case 1}), \\ \text{or } \mu_{i,j} &= \exp(\lambda_j + \gamma'_j \mathbf{z}_i) \quad (\text{case 2}). \end{aligned}$$

Case 2 is the Terza (1985) and Pudney and Shields (2000) model with an exponential rather than linear function for the thresholds. This formulation addresses two problems: (1) The thresholds are mathematically distinct from the regression; (2) by this construction, the threshold parameters must be positive. With a slight modification, the ordering of the thresholds can also be imposed. In case 1,

$$\mu_{i,j} = [\exp(\lambda_1) + \exp(\lambda_2) + \cdots + \exp(\lambda_j)] \times \exp(\gamma' \mathbf{z}_i),$$

and in case 2,

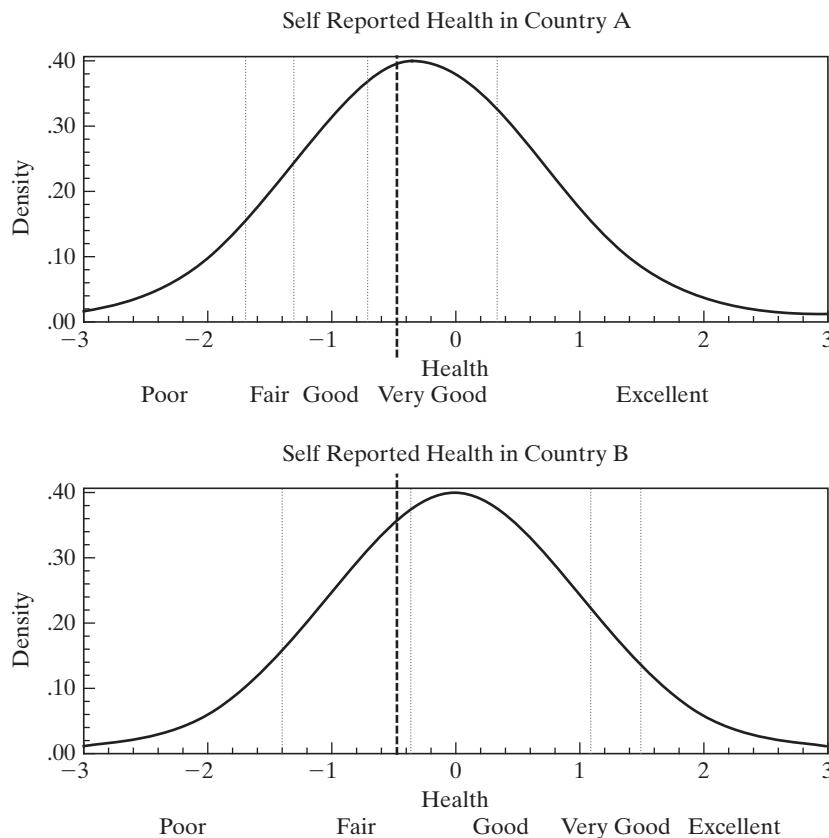
$$\mu_{i,j} = \mu_{i,j-1} + \exp(\lambda_j + \gamma'_j \mathbf{z}_i).$$

In practical terms, the model can now be fit with the constraint that all predicted probabilities are greater than zero. This is a numerical solution to the problem of ordering the thresholds for all data vectors.

This extension of the ordered choice model shows a case of **identification through functional form**. As we saw in the previous two models, the parameters  $(\lambda_j, \gamma_j, \beta)$  would not be separately identified if all the functions were linear. The contemporary literature views models that are unidentified without a change in functional form with some skepticism. However, the underlying theory of this model does not insist on linearity of the thresholds (or the utility function, for that matter), but it *does* insist on the ordering of the thresholds, and one might equally criticize the original model for being unidentified *because the model builder insists on a linear form*. That is, there is no obvious reason that the threshold parameters must be linear functions of the variables, or that linearity enjoys some claim to first precedence in the utility function. This is a methodological issue that cannot be resolved here. The nonlinearity of the preceding specification, or others that resemble it, does provide the benefit of a simple way to achieve other fundamental results, for example, coherency of the model (all positive probabilities).

### 18.3.5.b Thresholds and Heterogeneity—Anchoring Vignettes

The introduction of observed heterogeneity into the threshold parameters attempts to deal with a fundamentally restrictive assumption of the ordered choice model. Survey respondents rarely view the survey questions exactly the same way. This is certainly true

CHAPTER 18 ♦ Discrete Choices and Event Counts **801****FIGURE 18.5** Differential item Functioning in Ordered Choices.

in surveys of health satisfaction or subjective well-being. [See Boes and Winkelmann (2006b) and Ferrer-i-Carbonell and Frijters (2004).] KMST (2004) identify two very basic features of survey data that will make this problematic. First, they often measure concepts that are definable only with reference to examples, such as freedom, health, satisfaction, and so on. Second, individuals do, in fact, often understand survey questions very differently, particularly with respect to answers at the extremes. A widely used term for this interpersonal incomparability is **differential item functioning (DIF)**. Kapteyn, Smith, and Van Soest (KSV, 2007) and Van Soest, Delaney, Harmon, Kapteyn and Smith (2007) suggest the results in Figure 18.5 to describe the implications of DIF. The figure shows the distribution of Health (or drinking behavior in the latter study) in two hypothetical countries. The density for country A (the upper figure) is to the left of that for country B, implying that, on average, people in country A are less healthy than those in country B. But, the people in the two countries culturally offer very different response scales if asked to report their health on a five-point scale, as shown. In the figure, those in country A have a much more positive view of a given, objective health status than those in country B. A person in country A with health status indicated by the dotted line would report that they are in "Very Good" health while a person in

**802 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

country B with the same health status would report only “Fair.” A simple frequency of the distribution of self-assessments of health status in the two countries would suggest that people in country A are much healthier than those in country B when, in fact, the opposite is true. Correcting for the influences of DIF in such a situation would be essential to obtaining a meaningful comparison of the two countries. The impact of DIF is an accepted feature of the model within a population but could be strongly distortionary when comparing very disparate groups, such as across countries, as in KMST (political groups), Murray, Tandon, Mathers, and Sudana (2002) (health outcomes), Tandon et al. (2004), and KSV (work disability), Sirven, Santos-Eggmann, and Spagnoli (2008), and Gupta, Kristensen, and Possoli (2008) (health), Angelini et al. (2008) (life satisfaction), Kristensen and Johansson (2008), and Bago d’Uva et al. (2008), all of whom used the ordered probit model to make cross group comparisons.

KMST proposed the use of *anchoring vignettes* to resolve this difference in perceptions across groups. The essential approach is to use a series of examples that, it is believed, all respondents will agree on to estimate each respondent’s DIF and correct for it. The idea of using vignettes to anchor perceptions in survey questions is not itself new; KMST cite a number of earlier uses. The innovation is their method for incorporating the approach in a formal model for the ordered choices. The bivariate and multivariate probit models that they develop combine the elements described in Sections 18.3.1–18.3.3 and the HOPIT model in Section 18.3.4.a.

## 18.4 MODELS FOR COUNTS OF EVENTS

We have encountered behavioral variables that involve counts of events at several points in this text. In Examples 14.10 and 17.20, we examined the number of times an individual visited the physician using the GSOEP data. The credit default data that we used in Examples 7.10 and 17.22 also include another behavioral variable, the number of derogatory reports in an individual’s credit history. Finally, in Example 17.23, we analyzed data on firm innovation. Innovation is often analyzed [for example, by Hausman, Hall, and Griliches (1984) and many others] in terms of the number of patents that the firm obtains (or applies for). In each of these cases, the variable of interest is a count of events. This obviously differs from the discrete dependent variables we analyzed in the previous two sections. A count is a quantitative measure that is, at least in principle, amenable to analysis using multiple linear regression. However, the typical preponderance of zeros and small values and the discrete nature of the outcome variable suggest that the regression approach can be improved by a method that explicitly accounts for these aspects.

Like the basic multinomial logit model for unordered data in Section 18.2 and the simple probit and logit models for binary and ordered data in Sections 17.2 and 18.3, the Poisson regression model is the fundamental starting point for the analysis of count data. We will develop the elements of modeling for count data in this framework in Sections 18.4.1–18.4.3, and then turn to more elaborate, flexible specifications in subsequent sections. Sections 18.4.4 and 18.4.5 will present the negative binomial and other alternatives to the Poisson functional form. Section 18.4.6 will describe the implications for the model specification of some complicating features of observed data, truncation, and censoring. Truncation arises when certain values, such as zero, are absent from the

## CHAPTER 18 ♦ Discrete Choices and Event Counts **803**

observed data because of the sampling mechanism, not as a function of the data generating process. Data on recreation site visitation that are gathered at the site, for example, will, by construction, not contain any zeros. Censoring arises when certain ranges of outcomes are all coded with the same value. In the example analyzed the response variable is censored at 12, though values larger than 12 are possible “in the field.” As we have done in the several earlier treatments, in Section 18.4.7, we will examine extensions of the count data models that are made possible when the analysis is based on panel data. Finally, Section 18.4.8 discusses some behavioral models that involve more than one equation. For an example, based on the large number of zeros in the observed data, it appears that our count of doctor visits might be generated by a two-part process, a first step in which the individual decides whether or not to visit the physician at all, and a second decision, given the first, how many times to do so. A “hurdle model” that applies here and some related variants are discussed in Section 18.4.8.

### 18.4.1 THE POISSON REGRESSION MODEL

The **Poisson regression model** specifies that each  $y_i$  is drawn from a Poisson distribution with parameter  $\lambda_i$ , which is related to the regressors  $\mathbf{x}_i$ . The primary equation of the model is

$$\text{Prob}(Y = y_i | \mathbf{x}_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \dots \quad (18-17)$$

The most common formulation for  $\lambda_i$  is the **loglinear model**,

$$\ln \lambda_i = \mathbf{x}'_i \boldsymbol{\beta}.$$

It is easily shown that the expected number of events *per period* is given by

$$E[y_i | \mathbf{x}_i] = \text{Var}[y_i | \mathbf{x}_i] = \lambda_i = e^{\mathbf{x}'_i \boldsymbol{\beta}},$$

so

$$\frac{\partial E[y_i | \mathbf{x}_i]}{\partial \mathbf{x}_i} = \lambda_i \boldsymbol{\beta}.$$

With the parameter estimates in hand, this vector can be computed using any data vector desired.

In principle, the Poisson model is simply a nonlinear regression. But it is far easier to estimate the parameters with maximum likelihood techniques. The log-likelihood function is

$$\ln L = \sum_{i=1}^n [-\lambda_i + y_i \mathbf{x}'_i \boldsymbol{\beta} - \ln y_i!].$$

The likelihood equations are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n (y_i - \lambda_i) \mathbf{x}_i = \mathbf{0}.$$

The Hessian is

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = - \sum_{i=1}^n \lambda_i \mathbf{x}_i \mathbf{x}'_i.$$

The Hessian is negative definite for all  $\mathbf{x}$  and  $\boldsymbol{\beta}$ . Newton’s method is a simple algorithm for this model and will usually converge rapidly. At convergence,  $[\sum_{i=1}^n \hat{\lambda}_i \mathbf{x}_i \mathbf{x}'_i]^{-1}$

## 804 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

provides an estimator of the asymptotic covariance matrix for the parameter estimates. Given the estimates, the prediction for observation  $i$  is  $\hat{\lambda}_i = \exp(\mathbf{x}_i'\hat{\beta})$ . A standard error for the prediction interval can be formed by using a linear Taylor series approximation. The estimated variance of the prediction will be  $\hat{\lambda}_i^2 \mathbf{x}_i' \mathbf{V} \mathbf{x}_i$ , where  $\mathbf{V}$  is the estimated asymptotic covariance matrix for  $\hat{\beta}$ .

For testing hypotheses, the three standard tests are very convenient in this model. The Wald statistic is computed as usual. As in any discrete choice model, the likelihood ratio test has the intuitive form

$$\text{LR} = 2 \sum_{i=1}^n \ln \left( \frac{\hat{P}_i}{\hat{P}_{\text{restricted},i}} \right),$$

where the probabilities in the denominator are computed with using the restricted model. Using the BHHH estimator for the asymptotic covariance matrix, the LM statistic is simply

$$\text{LM} = \left[ \sum_{i=1}^n \mathbf{x}_i(y_i - \hat{\lambda}_i) \right]' \left[ \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' (y_i - \hat{\lambda}_i)^2 \right]^{-1} \left[ \sum_{i=1}^n \mathbf{x}_i (y_i - \hat{\lambda}_i) \right] = \mathbf{i}' \mathbf{G} (\mathbf{G}' \mathbf{G})^{-1} \mathbf{G}' \mathbf{i}, \quad (18-18)$$

where each row of  $\mathbf{G}$  is simply the corresponding row of  $\mathbf{X}$  multiplied by  $e_i = (y_i - \hat{\lambda}_i)$ ,  $\hat{\lambda}_i$  is computed using the restricted coefficient vector, and  $\mathbf{i}$  is a column of ones.

### 18.4.2 MEASURING GOODNESS OF FIT

The Poisson model produces no natural counterpart to the  $R^2$  in a linear regression model, as usual, because the conditional mean function is nonlinear and, moreover, because the regression is heteroscedastic. But many alternatives have been suggested.<sup>12</sup> A measure based on the standardized residuals is

$$R_p^2 = 1 - \frac{\sum_{i=1}^n \left[ \frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}} \right]^2}{\sum_{i=1}^n \left[ \frac{y_i - \bar{y}}{\sqrt{\bar{y}}} \right]^2}.$$

This measure has the virtue that it compares the fit of the model with that provided by a model with only a constant term. But it can be negative, and it can rise when a variable is dropped from the model. For an individual observation, the **deviance** is

$$d_i = 2[y_i \ln(y_i/\hat{\lambda}_i) - (y_i - \hat{\lambda}_i)] = 2[y_i \ln(y_i/\hat{\lambda}_i) - e_i],$$

where, by convention,  $0 \ln(0) = 0$ . If the model contains a constant term, then  $\sum_{i=1}^n e_i = 0$ . The sum of the deviances,

$$G^2 = \sum_{i=1}^n d_i = 2 \sum_{i=1}^n y_i \ln(y_i/\hat{\lambda}_i),$$

is reported as an alternative fit measure by some computer programs. This statistic will equal 0.0 for a model that produces a perfect fit. (Note that because  $y_i$  is an integer

<sup>12</sup>See the surveys by Cameron and Windmeijer (1993), Gurmu and Trivedi (1994), and Greene (1995b).

CHAPTER 18 ♦ Discrete Choices and Event Counts **805**

while the prediction is continuous, it could not happen.) Cameron and Windmeijer (1993) suggest that the fit measure based on the deviances,

$$R_d^2 = 1 - \frac{\sum_{i=1}^n \left[ y_i \log \left( \frac{y_i}{\hat{\lambda}_i} \right) - (y_i - \hat{\lambda}_i) \right]}{\sum_{i=1}^n \left[ y_i \log \left( \frac{y_i}{\bar{y}} \right) \right]},$$

has a number of desirable properties. First, denote the log-likelihood function for the model in which  $\psi_i$  is used as the prediction (e.g., the mean) of  $y_i$  as  $\ell(\psi_i, y_i)$ . The Poisson model fit by MLE is, then,  $\ell(\hat{\lambda}_i, y_i)$ , the model with only a constant term is  $\ell(\bar{y}, y_i)$ , and a model that achieves a perfect fit (by predicting  $y_i$  with itself) is  $\ell(y_i, y_i)$ . Then

$$R_d^2 = \frac{\ell(\hat{\lambda}_i, y_i) - \ell(\bar{y}, y_i)}{\ell(y_i, y_i) - \ell(\bar{y}, y_i)}.$$

Both numerator and denominator measure the improvement of the model over one with only a constant term. The denominator measures the maximum improvement, since one cannot improve on a perfect fit. Hence, the measure is bounded by zero and one and increases as regressors are added to the model.<sup>13</sup> We note, finally, the passing resemblance of  $R_d^2$  to the “pseudo- $R^2$ ,” or “likelihood ratio index” reported by some statistical packages (e.g., Stata),

$$R_{\text{LRI}}^2 = 1 - \frac{\ell(\hat{\lambda}_i, y_i)}{\ell(\bar{y}, y_i)}.$$

Many modifications of the Poisson model have been analyzed by economists. In this and the next few sections, we briefly examine a few of them.

#### 18.4.3 TESTING FOR OVERDISPERSION

The Poisson model has been criticized because of its implicit assumption that the variance of  $y_i$  equals its mean. Many extensions of the Poisson model that relax this assumption have been proposed by Hausman, Hall, and Griliches (1984), McCullagh and Nelder (1983), and Cameron and Trivedi (1986), to name but a few.

The first step in this extended analysis is usually a test for overdispersion in the context of the simple model. A number of authors have devised tests for “overdispersion” within the context of the Poisson model. [See Cameron and Trivedi (1990), Gurmu (1991), and Lee (1986).] We will consider three of the common tests, one based on a regression approach, one a conditional moment test, and a third, a **Lagrange multiplier test**, based on an alternative model.

Cameron and Trivedi (1990) offer several different tests for overdispersion. A simple regression-based procedure used for testing the hypothesis

$$\begin{aligned} H_0: \text{Var}[y_i] &= E[y_i], \\ H_1: \text{Var}[y_i] &= E[y_i] + \alpha g(E[y_i]), \end{aligned}$$

---

<sup>13</sup>Note that multiplying both numerator and denominator by 2 produces the ratio of two likelihood ratio statistics, each of which is distributed as chi-squared.

**806 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

is carried out by regressing

$$z_i = \frac{(y_i - \hat{\lambda}_i)^2 - y_i}{\hat{\lambda}_i \sqrt{2}},$$

where  $\hat{\lambda}_i$  is the predicted value from the regression, on either a constant term or  $\hat{\lambda}_i$  without a constant term. A simple  $t$  test of whether the coefficient is significantly different from zero tests  $H_0$  versus  $H_1$ .

The next section presents the **negative binomial model**. This model relaxes the Poisson assumption that the mean equals the variance. The Poisson model is obtained as a parametric restriction on the negative binomial model, so a Lagrange multiplier test can be computed. In general, if an alternative distribution for which the Poisson model is obtained as a parametric restriction, such as the negative binomial model, can be specified, then a Lagrange multiplier statistic can be computed. [See Cameron and Trivedi (1986, p. 41).] The LM statistic is

$$\text{LM} = \left[ \frac{\sum_{i=1}^n \hat{w}_i [(y_i - \hat{\lambda}_i)^2 - y_i]}{\sqrt{2 \sum_{i=1}^n \hat{w}_i \hat{\lambda}_i^2}} \right]^2. \quad (18-19)$$

The weight,  $\hat{w}_i$ , depends on the assumed alternative distribution. For the negative binomial model discussed later,  $\hat{w}_i$  equals 1.0. Thus, under this alternative, the statistic is particularly simple to compute:

$$\text{LM} = \frac{(\mathbf{e}'\mathbf{e} - n\bar{y})^2}{2\hat{\lambda}'\hat{\lambda}}. \quad (18-20)$$

The main advantage of this test statistic is that one need only estimate the Poisson model to compute it. Under the hypothesis of the Poisson model, the limiting distribution of the LM statistic is chi-squared with one degree of freedom.

#### 18.4.4 HETEROGENEITY AND THE NEGATIVE BINOMIAL REGRESSION MODEL

The assumed equality of the conditional mean and variance functions is typically taken to be the major shortcoming of the Poisson regression model. Many alternatives have been suggested [see Hausman, Hall, and Griliches (1984), Cameron and Trivedi (1986, 1998), Gurmu and Trivedi (1994), Johnson and Kotz (1993), and Winkelmann (2003) for discussion]. The most common is the negative binomial model, which arises from a natural formulation of cross-section heterogeneity. [See Hilbe (2007).] We generalize the Poisson model by introducing an individual, unobserved effect into the conditional mean,

$$\ln \mu_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i = \ln \lambda_i + \ln u_i,$$

where the disturbance  $\varepsilon_i$  reflects either **specification error**, as in the classical regression model, or the kind of cross-sectional heterogeneity that normally characterizes microeconomic data. Then, the distribution of  $y_i$  conditioned on  $\mathbf{x}_i$  and  $u_i$  (i.e.,  $\varepsilon_i$ ) remains Poisson with conditional mean and variance  $\mu_i$ :

$$f(y_i | \mathbf{x}_i, u_i) = \frac{e^{-\lambda_i u_i} (\lambda_i u_i)^{y_i}}{y_i!}.$$

**CHAPTER 18 ♦ Discrete Choices and Event Counts 807**

The unconditional distribution  $f(y_i | \mathbf{x}_i)$  is the expected value (over  $u_i$ ) of  $f(y_i | \mathbf{x}_i, u_i)$ ,

$$f(y_i | \mathbf{x}_i) = \int_0^\infty \frac{e^{-\lambda_i u_i} (\lambda_i u_i)^{y_i}}{y_i!} g(u_i) du_i.$$

The choice of a density for  $u_i$  defines the unconditional distribution. For mathematical convenience, a gamma distribution is usually assumed for  $u_i = \exp(\varepsilon_i)$ .<sup>14</sup> As in other models of heterogeneity, the mean of the distribution is unidentified if the model contains a constant term (because the disturbance enters multiplicatively) so  $E[\exp(\varepsilon_i)]$  is assumed to be 1.0. With this normalization,

$$g(u_i) = \frac{\theta^\theta}{\Gamma(\theta)} e^{-\theta u_i} u_i^{\theta-1}.$$

The density for  $y_i$  is then

$$\begin{aligned} f(y_i | \mathbf{x}_i) &= \int_0^\infty \frac{e^{-\lambda_i u_i} (\lambda_i u_i)^{y_i}}{y_i!} \frac{\theta^\theta u_i^{\theta-1} e^{-\theta u_i}}{\Gamma(\theta)} du_i \\ &= \frac{\theta^\theta \lambda_i^{y_i}}{\Gamma(y_i + 1) \Gamma(\theta)} \int_0^\infty e^{-(\lambda_i + \theta) u_i} u_i^{\theta+y_i-1} du_i \\ &= \frac{\theta^\theta \lambda_i^{y_i} \Gamma(\theta + y_i)}{\Gamma(y_i + 1) \Gamma(\theta) (\lambda_i + \theta)^{\theta+y_i}} \\ &= \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1) \Gamma(\theta)} r_i^{y_i} (1 - r_i)^\theta, \quad \text{where } r_i = \frac{\lambda_i}{\lambda_i + \theta}, \end{aligned}$$

which is one form of the negative binomial distribution. The distribution has conditional mean  $\lambda_i$  and conditional variance  $\lambda_i(1 + (1/\theta)\lambda_i)$ . [This model is Negbin 2 in Cameron and Trivedi's (1986) presentation.] The negative binomial model can be estimated by maximum likelihood without much difficulty. A test of the Poisson distribution is often carried out by testing the hypothesis  $\alpha = 1/\theta = 0$  using the Wald or likelihood ratio test.

#### 18.4.5 FUNCTIONAL FORMS FOR COUNT DATA MODELS

The equidispersion assumption of the Poisson regression model,  $E[y_i | \mathbf{x}_i] = \text{Var}[y_i | \mathbf{x}_i]$ , is a major shortcoming. Observed data rarely, if ever, display this feature. The very large amount of research activity on functional forms for count models is often focused on testing for equidispersion and building functional forms that relax this assumption. In practice, the Poisson model is typically only the departure point for an extended specification search.

One easily remedied minor issue concerns the units of measurement of the data. In the Poisson and negative binomial models, the parameter  $\lambda_i$  is the expected number of events *per unit of time*, thus, there is a presumption in the model formulation, for

<sup>14</sup>An alternative approach based on the normal distribution is suggested in Terza (1998), Greene (1995a, 1997a, 2007d), Winkelmann (1997) and Riphahn, Wambach and Million (2003). The normal-Poisson mixture is also easily extended to the random effects model discussed in the next section. There is no closed form for the normal-Poisson mixture model, but it can be easily approximated by using Hermite quadrature or simulation. See Sections 14.9.6.b and 17.4.8.

**808 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

example, the Poisson, that the same amount of time is observed for each  $i$ . In a spatial context, such as measurements of the incidence of a disease per group of  $N_i$  persons, or the number of bomb craters per square mile (London, 1940), the assumption would be that the same physical area or the same size of population applies to each observation. Where this differs by individual, it will introduce a type of heteroscedasticity in the model. The simple remedy is to modify the model to account for the **exposure**,  $T_i$ , of the observation as follows:

$$\text{Prob}(y_i = j | \mathbf{x}_i, T_i) = \frac{\exp(-T_i\phi_i)(T_i\phi_i)^j}{j!}, \quad \phi_i = \exp(\mathbf{x}'_i \boldsymbol{\beta}), \quad j = 0, 1, \dots$$

The original model is returned if we write  $\lambda_i = \exp(\mathbf{x}'_i \boldsymbol{\beta} + \ln T_i)$ . Thus, when the exposure differs by observation, the appropriate accommodation is to include the log of exposure in the regression part of the model with a coefficient of 1.0. (For less than obvious reasons, the term “offset variable” is commonly associated with the exposure variable  $T_i$ .) Note that if  $T_i$  is the same for all  $i$ ,  $\ln T_i$  will simply vanish into the constant term of the model (assuming one is included in  $\mathbf{x}_i$ ).

The recent literature, mostly associating the result with Cameron and Trivedi’s (1986, 1998) work, defines two familiar forms of the negative binomial model. The **Negbin 2 (NB2) form** of the probability is

$$\begin{aligned} \text{Prob}(Y = y_i | \mathbf{x}_i) &= \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)} r_i^{y_i} (1 - r_i)^\theta, \\ \lambda_i &= \exp(\mathbf{x}'_i \boldsymbol{\beta}), \\ r_i &= \lambda_i / (\theta + \lambda_i). \end{aligned} \tag{18-21}$$

This is the default form of the model in the received econometrics packages that provide an estimator for this model. The **Negbin 1 (NB1) form** of the model results if  $\theta$  in the preceding is replaced with  $\theta_i = \theta\lambda_i$ . Then,  $r_i$  reduces to  $r = 1/(1 + \theta)$ , and the density becomes

$$\text{Prob}(Y = y_i | \mathbf{x}_i) = \frac{\Gamma(\theta\lambda_i + y_i)}{\Gamma(y_i + 1)\Gamma(\theta\lambda_i)} r^{y_i} (1 - r)^{\theta\lambda_i}. \tag{18-22}$$

This is not a simple reparameterization of the model. The results in Example 18.7 demonstrate that the log-likelihood functions are not equal at the maxima, and the parameters are not simple transformations in one model versus the other. We are not aware of a theory that justifies using one form or the other for the negative binomial model. Neither is a restricted version of the other, so we cannot carry out a likelihood ratio test of one versus the other. The more general **Negbin P (NBP)** family does nest both of them, so this may provide a more general, encompassing approach to finding the right specification. [See Greene (2005, 2008).] The Negbin  $P$  model is obtained by replacing  $\theta$  in the Negbin 2 form with  $\theta\lambda_i^{2-P}$ . We have examined the cases of  $P = 1$  and  $P = 2$  in (18-5) and (18-6). The full model is

$$\text{Prob}(Y = y_i | \mathbf{x}_i) = \frac{\Gamma(\theta\lambda_i^Q + y_i)}{\Gamma(y_i + 1)\Gamma(\theta\lambda_i^Q)} \left( \frac{\lambda_i}{\theta\lambda_i^Q + \lambda_i} \right)^{y_i} \left( \frac{\theta\lambda_i^Q}{\theta\lambda_i^Q + \lambda_i} \right)^{\theta\lambda_i^Q}, \quad Q = 2 - P.$$

CHAPTER 18 ♦ Discrete Choices and Event Counts **809**

The conditional mean function for the three cases considered is

$$E[y_i | \mathbf{x}_i] = \exp(\mathbf{x}'_i \boldsymbol{\beta}) = \lambda_i.$$

The parameter  $P$  is picking up the scaling. A general result is that for all three variants of the model,

$$\text{Var}[y_i | \mathbf{x}_i] = \lambda_i (1 + \alpha \lambda_i^{P-1}), \quad \text{where } \alpha = 1/\theta.$$

Thus, the NB2 form has a variance function that is quadratic in the mean while the NB1 form's variance is a simple multiple of the mean. There have been many other functional forms proposed for count data models, including the generalized Poisson, gamma, and Polya-Aeppli forms described in Winkelmann (2003) and Greene (2007a, Chapter 24).

The heteroscedasticity in the count models is induced by the relationship between the variance and the mean. The single parameter  $\theta$  picks up an implicit overall scaling, so it does not contribute to this aspect of the model. As in the linear model, microeconomic data are likely to induce heterogeneity in both the mean and variance of the response variable. A specification that allows independent variation of both will be of some virtue. The result

$$\text{Var}[y_i | \mathbf{x}_i] = \lambda_i (1 + (1/\theta) \lambda_i^{P-1})$$

suggests that a natural platform for separately modeling heteroscedasticity will be the dispersion parameter,  $\theta$ , which we now parameterize as

$$\theta_i = \theta \exp(\mathbf{z}'_i \boldsymbol{\delta}).$$

Operationally, this is a relatively minor extension of the model. But, it is likely to introduce quite a substantial increase in the flexibility of the specification. Indeed, a heterogeneous Negbin P model is likely to be sufficiently parameterized to accommodate the behavior of most data sets. (Of course, the specialized models discussed in Section 18.4.8, for example, the zero inflation models, may yet be more appropriate for a given situation.)

#### **Example 18.7 Count Data Models for Doctor Visits**

The study by Riphahn et al. (2003) that provided the data we have used in numerous earlier examples analyzed the two count variables DocVis (visits to the doctor) and HospVis (visits to the hospital). The authors were interested in the joint determination of these two count variables. One of the issues considered in the study was whether the data contained evidence of moral hazard, that is, whether health care utilization as measured by these two outcomes was influenced by the subscription to health insurance. The data contain indicators of two levels of insurance coverage, PUBLIC, which is the main source of insurance, and ADDON, which is a secondary optional insurance. In the sample of 27,326 observations (family/years), 24,203 individuals held the public insurance. (There is quite a lot of within group variation in this. Individuals did not routinely obtain the insurance for all periods.) Of these 24,203, 23,689 had only public insurance and 514 had both types. (One could not have only the ADDON insurance.) To explore the issue, we have analyzed the DocVis variable with the count data models described in this section. The exogenous variables in our model are

$$\mathbf{x}_{it} = (1, \text{Age}, \text{Education}, \text{Income}, \text{Kids}, \text{Public}).$$

(Variables are described in Appendix Table F7.1.)

Table 18.14 presents the estimates of the several count models. In all specifications, the coefficient on PUBLIC is positive, large, and highly statistically significant, which is consistent with the results in the authors' study. The various test statistics strongly reject the hypothesis

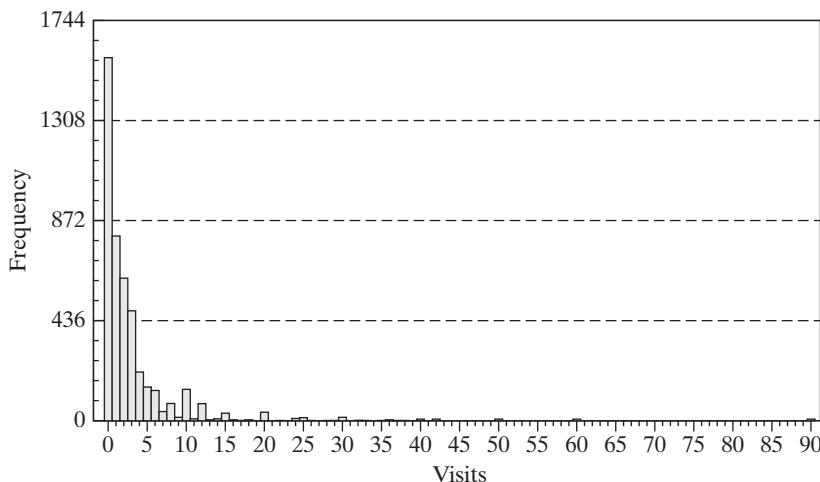
**810 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**
**TABLE 18.14** Estimated Models for DOCVIS (standard errors in parentheses)

<i>Variable</i>	<i>Poisson</i>	<i>Negbin 2</i>	<i>Heterogeneous</i>	<i>Negbin 1</i>	<i>Negbin P</i>
Constant	0.7162 (0.03287)	0.7628 (0.07247)	0.7928 (0.07459)	0.6848 (0.06807)	0.6517 (0.07759)
Age	0.01844 (0.0003316)	0.01803 (0.0007915)	0.01704 (0.0008146)	0.01585 (0.0007042)	0.01907 (0.0008078)
Education	-0.03429 (0.001797)	-0.03839 (0.003965)	-0.03581 (0.004036)	-0.02381 (0.003702)	-0.03388 (0.004308)
Income	-0.4751 (0.02198)	-0.4206 (0.04700)	-0.4108 (0.04752)	-0.1892 (0.04452)	-0.3337 (0.05161)
Kids	-0.1582 (0.007956)	-0.1513 (0.01738)	-0.1568 (0.01773)	-0.1342 (0.01647)	-0.1622 (0.01856)
Public	0.2364 (0.01328)	0.2324 (0.02900)	0.2411 (0.03006)	0.1616 (0.02678)	0.2195 (0.03155)
<i>P</i>	0.0000 (0.0000)	2.0000 (0.0000)	2.0000 (0.0000)	1.0000 (0.0000)	1.5473 (0.03444)
$\theta$	0.0000 (0.0000)	1.9242 (0.02008)	2.6060 (0.05954)	6.1865 (0.06861)	3.2470 (0.1346)
$\delta$ (Female)	0.0000 (0.0000)	0.0000 (0.0000)	-0.3838 (0.02046)	0.0000 (0.0000)	0.0000 (0.0000)
$\delta$ (Married)	0.0000 (0.0000)	0.0000 (0.0000)	-0.1359 (0.02307)	0.0000 (0.0000)	0.0000 (0.0000)
In <i>L</i>	-104440.3	-60265.49	-60121.77	-60260.68	-60197.15

of equidispersion. Cameron and Trivedi's (1990) semiparametric tests from the Poisson model (see Section 18.4.3 have *t* statistics of 22.147 for  $g_i = \mu_i$  and 22.504 for  $g_i = \mu_i^2$ . Both of these are far larger than the critical value of 1.96. The LM statistic is 972,714.48, which is also larger than the (any) critical value. On these bases, we would reject the hypothesis of equidispersion. The Wald and likelihood ratio tests based on the negative binomial models produce the same conclusion. For comparing the different negative binomial models, note that Negbin 2 is the worst of the three by the likelihood function, although NB1 and NB2 are not directly comparable. On the other hand, note that in the NBP model, the estimate of *P* is more than 10 standard errors from 1.0000 or 2.000, so both NB1 and NB2 are rejected in favor of the unrestricted NBP form of the model. The NBP and the heterogeneous NB2 model are not nested either, but comparing the log-likelihoods, it does appear that the heterogeneous model is substantially superior. We computed the Vuong statistic based on the individual contributions to the log-likelihoods, with  $v_i = \ln L_i(\text{NBP}) - \ln L_i(\text{NB2-H})$ . (See Section 14.6.6). The value of the statistic is -3.27. On this basis, we would reject NBP in favor of NB2-H. Finally, with regard to the original question, the coefficient on PUBLIC is larger than 10 times the estimated standard error in every specification. We would conclude that the results are consistent with the proposition that there is evidence of moral hazard.

#### 18.4.6 TRUNCATION AND CENSORING IN MODELS FOR COUNTS

Truncation and censoring are relatively common in applications of models for counts. Truncation arises as a consequence of discarding what appear to be unusable data, such as the zero values in survey data on the number of uses of recreation facilities [Shaw (1988), Bockstael et al. (1990)]. In this setting, a more common case which also gives rise to truncation is on-site sampling. When one is interested in visitation by the entire population, which will naturally include zero visits, but one draws their sample

CHAPTER 18 ♦ Discrete Choices and Event Counts **811**

**FIGURE 18.6** Number of Doctor Visits. 1988 Wave of GSOEP Data.

“on-site,” the distribution of visits is truncated at zero by construction. Every visitor has visited at least once. Shaw (1988), Englin and Shonkwiler (1995), Grogger and Carson (1991), Creel and Loomis (1990), Egan and Herriges (2006) and Martinez-Espinera and Amoako-Tuffour (2008) are among a number of studies that have treated truncation due to on-site sampling in environmental and recreation applications. Truncation will also arise when data are trimmed to remove what appear to be unusual values. Figure 18.6 displays a histogram for the number of doctor visits in the 1988 wave of the GSOEP data that we have used in several examples. There is a suspiciously large spike at zero and an extremely long right tail of what might seem to be atypical observations. For modeling purposes, it might be tempting to remove these “non-Poisson” appearing observations in these tails. (Other models might be a better solution.) The distribution that characterizes what remains in the sample is a truncated distribution. Truncation is not innocent. If the entire population is of interest, then conventional statistical inference (such as estimation) on the truncated sample produces a systematic bias known as (of course) “truncation bias.” This would arise, for example, if an ordinary Poisson model intended to characterize the full population is fit to the sample from a truncated population.

Censoring, in contrast, is generally a feature of the sampling design. In the application in Example 18.9, the dependent variable is the self-reported number of extramarital affairs in a survey taken by the magazine *Psychology Today*. The possible answers are 0, 1, 2, 3, 4–10 (coded as 7) and “monthly, weekly or daily” coded as 12. The two upper categories are censored. Similarly, in the doctor visits data in the previous paragraph, recognizing the possibility of truncation bias due to data trimming, we might, instead, simply censor the distribution of values at 15. The resulting variable would take values  $0, \dots, 14, “15 \text{ or more}.”$  In both cases, applying conventional estimation methods leads to predictable biases. However, it is also possible to reconstruct the estimators specifically to account for the truncation or censoring in the data.

## 812 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

Truncation and censoring produce similar effects on the distribution of the random variable and on the features of the population such as the mean. For the truncation case, suppose that the original random variable has a Poisson distribution (and these results can be [j]ctly extended to the negative binomial or any of the other models considered earlier),

$$P(y_i = j | \mathbf{x}_i) = \exp(-\lambda_i) \lambda_i^j / j! = P_{i,j}.$$

If the distribution is truncated at value  $C$ —that is, only values  $C + 1, \dots$  are observed—then the resulting random variable has probability distribution

$$P(y_i = j | \mathbf{x}_i, y_i > C) = \frac{P(y_i = j | \mathbf{x}_i)}{P(y_i > C | \mathbf{x}_i)} = \frac{P(y_i = j | \mathbf{x}_i)}{1 - P(y_i \leq C | \mathbf{x}_i)}.$$

The original distribution must be scaled up so that it sums to one for the cells that remain in the truncated distribution. The leading case is truncation at zero, that is, “left truncation,” which, for the Poisson model produces

$$P(y_i = j | \mathbf{x}_i, y_i > 0) = \frac{\exp(-\lambda_i) \lambda_i^j}{j! [1 - \exp(-\lambda_i)]} = \frac{P_{i,j}}{1 - P_{i,0}}, \quad j = 1, \dots$$

[See, e.g., Mullahy (1986), Shaw (1988), Grogger and Carson (1991), Greene (1998), and Winkelmann (1987).] The conditional mean function is

$$E(y_i | \mathbf{x}_i, y_i > 0) = \frac{1}{[1 - \exp(-\lambda_i)]} \sum_{j=1}^{\infty} \frac{j \exp(-\lambda_i) \lambda_i^j}{j!} = \frac{\lambda_i}{[1 - \exp(-\lambda_i)]} > \lambda_i.$$

The second equality results because the sum can be started at zero—the first term is zero—and this produces the expected value of the original variable. As might be expected, truncation “from below” has the effect of increasing the expected value. It can be shown that it decreases the conditional variance however. The partial effects are

$$\delta_i = \frac{\partial E[y_i | \mathbf{x}_i, y_i > 0]}{\partial \mathbf{x}_i} = \left[ \frac{1 - P_{i,0} - \lambda_i P_{i,0}}{(1 - P_{i,0})^2} \right] \mathbf{x}_i p. \quad (18-23)$$

The term outside the brackets is the partial effects in the absence of the truncation while the bracketed term rises from slightly greater than 0.5 to 1.0 as  $\lambda_i$  increases from just above zero.

### Example 18.8 Major Derogatory Reports

In Section 17.5.6 and Examples 17.9 and 17.22, we examined a binary choice model for the accept/reject decision for a sample of applicants for a major credit card. Among the variables in that model is “Major [P]rogatory Reports” (MDRs). This is an interesting behavioral variable in its own right that should be appropriately modeled using the count data specifications in this chapter. In the sample of 13,444 individuals, 10,833 had zero MDRs while the values for the remaining 2561 ranged from 1 to 22. This preponderance of zeros exceeds by far what one would anticipate in a Poisson model that was dispersed enough to produce the distribution of remaining individuals. As we will pursue an Example 18.11, a natural approach for these data is to treat the extremely large block of zeros explicitly in an extended model. For present purposes, we will consider the nonzero observations apart from the zeros and examine the effect of accounting for left truncation at zero on the estimated models. Estimation results are shown in Table 18.15. The first column of results compared to the second shows the

CHAPTER 18 ♦ Discrete Choices and Event Counts **813****TABLE 18.15** Estimated Truncated Poisson Regression Model (*t* ratios in parentheses)

	<i>Poisson Full Sample</i>	<i>Poisson</i>	<i>Truncated Poisson</i>			
Constant	0.8756	(17.10)	0.8698	(16.78)	0.7400	(11.99)
Age	0.0036	(2.38)	0.0035	(2.32)	0.0049	(2.75)
Income	-0.0039	(-4.78)	-0.0036	(-3.83)	-0.0051	(-4.51)
OwnRent	-0.1005	(-3.52)	-0.1020	(-3.56)	-0.1415	(-4.18)
Self Employed	-0.0325	(-0.62)	-0.0345	(-0.66)	-0.0515	(-0.82)
Dependents	0.0445	(4.69)	0.0440	(4.62)	0.0606	(5.48)
MthsCurAdr	0.00004	(0.23)	0.00005	(0.25)	0.00007	(0.30)
ln <i>L</i>		-5379.30		-5378.79		-5097.08
			Average Partial Effects			
Age		0.0017		0.0085		0.0084
Income		-0.0018		-0.0087		-0.0089
OwnRent		-0.0465		-0.2477		-0.2460
Self Employed		-0.0150		-0.0837		-0.0895
Dependents		0.0206		0.1068		0.1054
MthsCurAdr		0.00002		0.00012		0.00013
Cond'l. Mean		0.4628		2.4295		2.4295
Scale factor		0.4628		2.4295		1.7381

suspected impact of incorrectly including the zero observations. The coefficients change only slightly, but the partial effects are far smaller when the zeros are included in the estimation. It was not possible to fit the truncated negative binomial with these data.

Censoring is handled similarly. The usual case is “right censoring,” in which realized values greater than or equal to  $C$  are all given the value  $C$ . In this case, we have a two-part distribution [see Terza (1985b)]. The observed random variable,  $y_i$  is constructed from an underlying random variable,  $y_i^*$  by

$$y_i = \text{Min}(y_i^*, C).$$

Probabilities are constructed using the axioms of probability. This produces

$$\text{Prob}(y_i = j | \mathbf{x}_i) = P_{i,j}, \quad j = 0, 1, \dots, C-1,$$

$$\text{Prob}(y_i = C | \mathbf{x}_i) = \sum_{j=C}^{\infty} P_{i,j} = 1 - \sum_{j=0}^{C-1} P_{i,j}.$$

In this case, the conditional mean function is

$$\begin{aligned} E[y_i | \mathbf{x}_i] &= \sum_{j=0}^{C-1} j P_{i,j} + \sum_{j=C}^{\infty} C P_{i,j} \\ &= \sum_{j=0}^{\infty} j P_{i,j} - \sum_{j=C}^{\infty} (j - C) P_{i,j} \\ &= \lambda_i - \sum_{j=C}^{\infty} (j - C) P_{i,j} < \lambda_i. \end{aligned}$$

## 814 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

The infinite sum is computed by using the complement. Thus,

$$\begin{aligned}
 E[y_i | \mathbf{x}_i] &= \lambda_i - \left[ \sum_{j=0}^{\infty} (j - C) P_{i,j} - \sum_{j=0}^{C-1} (j - C) P_{i,j} \right] \\
 &= \lambda_i - (\lambda_i - C) + \sum_{j=0}^{C-1} (j - C) P_{i,j} \\
 &= C - \sum_{j=0}^{C-1} (C - j) P_{i,j}.
 \end{aligned}$$

### Example 18.9 Extramarital Affairs

In 1969, the popular magazine *Psychology Today* published a 101-question survey on sex and asked its readers to mail in their answers. The results of the survey were discussed in the July 1970 issue. From the approximately 2,000 replies that were collected in electronic form (of about 20,000 received), Professor Ray Fair (1978) extracted a sample of 601 observations on men and women then currently married for the first time and analyzed their responses to a question about extramarital affairs. Fair's analysis in this frequently cited study suggests several interesting econometric questions. [In addition, his 1977 companion paper in *Econometrica* on estimation of the tobit model contributed to the development of the EM algorithm, which was published by and is usually associated with Dempster, Laird, and Rubin (1977).]

Fair used the tobit model that we discuss in Chapter 19 as a platform. The nonexperimental nature of the data (which can be downloaded from the Internet at <http://fairmodel.econ.yale.edu/rayfair/work.ss.htm> and are given in Appendix Table F18.1) provides a laboratory case that we can use to examine the relationships among the tobit, truncated regression, and probit models. Although the tobit model seems to be a natural choice for the model for these data, given the cluster of zeros, the fact that the behavioral outcome variable is a count that typically takes a small value suggests that the models for counts that we have examined in this chapter might be yet a better choice. Finally, the preponderance of zeros in the data that initially motivated the tobit model suggests that even the standard Poisson model, although an improvement, might still be inadequate. We will pursue that aspect of the data later. In this example, we will focus on just the censoring issue. Other features of the models and data are reconsidered in the exercises.

The study was based on 601 observations on the following variables (full details on data coding are given in the data file and Appendix Table F18.1):

$y$  = number of affairs in the past year, 0, 1, 2, 3, 4–10 coded as 7

“monthly, weekly, or daily,” coded as 12. Sample mean = 1.46

Frequencies = (451, 34, 17, 19, 42, 38)

$z_1$  = sex = 0 for female, 1 for male. Sample mean = 0.476

$z_2$  = age. Sample mean = 32.5

$z_3$  = number of years married. Sample mean = 8.18

$z_4$  = children, 0 = no, 1 = yes. Sample mean = 0.715

$z_5$  = religiousness, 1 = anti, . . . , 5 = very. Sample mean = 3.12

$z_6$  = education, years, 9 = grade school, 12 = high school, . . . , 20 = Ph.D or other Sample mean = 16.2

$z_7$  = occupation, “Hollingshead scale,” 1–7. Sample mean = 4.19

$z_8$  = self-rating of marriage, 1 = very unhappy, . . . , 5 = very happy. Sample mean = 3.93



CHAPTER 18 ♦ Discrete Choices and Event Counts **815****TABLE 18.16** Censored Poisson and Negative Binomial Distributions

<b>Variable</b>	<b>Poisson Regression</b>			<b>Negative Binomial Regression</b>		
	<b>Estimate</b>	<b>Standard Error</b>	<b>Marginal Effect</b>	<b>Estimate</b>	<b>Standard Error</b>	<b>Marginal Effect</b>
<b>Based on Uncensored Poisson Distribution</b>						
Constant	2.53	0.197	—	2.19	0.664	—
$z_2$	-0.0322	0.00585	-0.0470	-0.0262	0.0192	-0.00393
$z_3$	0.116	0.00991	0.168	0.0848	0.0350	0.127
$z_5$	-0.354	0.0309	-0.515	-0.422	0.111	-0.632
$z_7$	0.0798	0.0194	0.16	0.0604	0.0702	0.0906
$z_8$	-0.409	0.0274	-0.596	-0.31	0.111	-0.646
$\alpha$					0.786	
$\ln L$	-1427.037			-728.2441		
<b>Based on Poisson Distribution Right Censored at <math>y = 4</math></b>						
Constant	1.90	0.283	—	4.79	1.16	—
$z_2$	-0.0328	0.00838	-0.125	-0.0166	0.0250	-0.00428
$z_3$	0.105	0.0140	0.141	0.174	0.0568	0.045
$z_5$	-0.323	0.0437	-0.232	-0.723	0.198	-0.186
$z_7$	0.0798	0.0275	0.0521	0.0900	0.116	0.0232
$z_8$	-0.390	0.0391	-0.279	-0.854	0.216	-0.220
$\alpha$				9.40	1.35	
$\ln L$	-747.7541			-482.0505		

The tobit model was fit to  $y$  using a constant term and all eight variables. A restricted model was fit by excluding  $z_1$ ,  $z_4$ , and  $z_6$ , none of which was individually statistically significant in the model. We are able to match exactly Fair's results for both equations. The tobit model should only be viewed as an approximation for these data. The dependent variable is a count, not a continuous measurement. The Poisson regression model, or perhaps one of the many variants of it, should be a preferable modeling framework. Table 18.16 presents estimates of the Poisson and negative binomial regression models. There is ample evidence of overdispersion in these data; the  $t$  ratio on the estimated overdispersion parameter is  $7.014/0.945 = 7.42$ , which is strongly suggestive. The large absolute value of the coefficient is likewise suggestive.

Responses of 7 and 12 do not represent the actual counts. It is unclear what the effect of the first recoding would be, because it might well be the mean of the observations in this group. But the second is clearly a censored observation. To remove both of these effects, we have recoded both the values 7 and 12 as 4 and treated this observation (appropriately) as a censored observation, with 4 denoting "4 or more." As shown in the third and fourth sets of results in Table 18.16, the effect of this treatment of the data is greatly to reduce the measured effects. Although this step does remove a deficiency in the data, it does not remove the overdispersion; at this point, the negative binomial model is still the preferred specification.

**18.4.7 PANEL DATA MODELS**

The familiar approaches to accommodating heterogeneity in panel data have fairly straightforward extensions in the count data setting. [Hausman, Hall, and Griliches (1984) give full details for these models.] We will examine them for the Poisson model. The authors [and Allison (2000)] also give results for the negative binomial model.

## 816 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

### 18.4.7.a Robust Covariance Matrices for Pooled Estimators

The standard asymptotic covariance matrix estimator for the Poisson model is

$$\text{Est. Asy. Var}[\hat{\beta}] = \left[ -\frac{\partial^2 \ln L}{\partial \hat{\beta} \partial \hat{\beta}'} \right]^{-1} = \left[ \sum_{i=1}^n \hat{\lambda}_i \mathbf{x}_i \mathbf{x}_i' \right]^{-1} = [\mathbf{X}' \hat{\Lambda} \mathbf{X}]^{-1},$$

where  $\hat{\Lambda}$  is a diagonal matrix of predicted values. The BHHH estimator is

$$\begin{aligned} \text{Est. Asy. Var}[\hat{\beta}] &= \left[ \sum_{i=1}^n \left( \frac{\partial \ln P_i}{\partial \hat{\beta}} \right) \left( \frac{\partial \ln P_i}{\partial \hat{\beta}} \right)' \right]^{-1} \\ &= \left[ \sum_{i=1}^n (y_i - \hat{\lambda}_i')^2 \mathbf{x}_i \mathbf{x}_i' \right]^{-1} = [\mathbf{X}' \hat{\mathbf{E}}^2 \mathbf{X}]^{-1}, \end{aligned}$$

where  $\hat{\mathbf{E}}$  is a diagonal matrix of residuals. The Poisson model is one in which the MLE is robust to certain misspecifications of the model, such as the failure to incorporate latent heterogeneity in the mean (i.e., one fits the Poisson model when the negative binomial is appropriate). In this case, a robust covariance matrix is the “sandwich” estimator,

$$\text{Robust Est. Asy. Var}[\hat{\beta}] = [\mathbf{X}' \hat{\Lambda} \mathbf{X}]^{-1} [\mathbf{X}' \hat{\mathbf{E}}^2 \mathbf{X}] [\mathbf{X}' \hat{\Lambda} \mathbf{X}]^{-1},$$

which is appropriate to accommodate this failure of the model. It has become common to employ this estimator with all specifications, including the negative binomial. One might question the virtue of this. Because the negative binomial model already accounts for the latent heterogeneity, it is unclear what *additional* failure of the assumptions of the model this estimator would be robust to. The questions raised in Section 14.8.3 and 14.8.4 about robust covariance matrices would be relevant here.

A related calculation is used when observations occur in groups that may be correlated. This would include a random effects setting in a panel in which observations have a common latent heterogeneity as well as more general, stratified, and clustered data sets. The parameter estimator is unchanged in this case (and an assumption is made that the estimator is still consistent), but an adjustment is made to the estimated asymptotic covariance matrix. The calculation is done as follows: Suppose the  $n$  observations are assembled in  $G$  clusters of observations, in which the number of observations in the  $i$ th cluster is  $n_i$ . Thus,  $\sum_{i=1}^G n_i = n$ . Denote by  $\beta$  the full set of model parameters in whatever variant of the model is being estimated. Let the observation-specific gradients and Hessians be  $\mathbf{g}_{ij} = \partial \ln L_{ij} / \partial \beta = (y_{ij} - \lambda_{ij}) \mathbf{x}_{ij}$  and  $\mathbf{H}_{ij} = \partial^2 \ln L_{ij} / \partial \beta \partial \beta' = -\lambda_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}'$ . The uncorrected estimator of the asymptotic covariance matrix based on the Hessian is

$$\mathbf{V}_H = -\mathbf{H}^{-1} = \left( -\sum_{i=1}^G \sum_{j=1}^{n_i} \mathbf{H}_{ij} \right)^{-1}.$$

The corrected asymptotic covariance matrix is

$$\text{Est. Asy. Var}[\hat{\beta}] = \mathbf{V}_H \left( \frac{G}{G-1} \right) \left[ \sum_{i=1}^G \left( \sum_{j=1}^{n_i} \mathbf{g}_{ij} \right) \left( \sum_{j=1}^{n_i} \mathbf{g}_{ij} \right)' \right] \mathbf{V}_H.$$

**CHAPTER 18 ♦ Discrete Choices and Event Counts 817**

Note that if there is exactly one observation per cluster, then this is  $G/(G - 1)$  times the sandwich (robust) estimator.

**18.4.7.b Fixed Effects**

Consider first a fixed effects approach. The Poisson distribution is assumed to have conditional mean

$$\log \lambda_{it} = \beta' \mathbf{x}_{it} + \alpha_i, \quad (18-24)$$

where now,  $\mathbf{x}_{it}$  has been redefined to exclude the constant term. The approach used in the linear model of transforming  $y_{it}$  to group mean deviations does not remove the heterogeneity, nor does it leave a Poisson distribution for the transformed variable. However, the Poisson model with fixed effects can be fit using the methods described for the probit model in Section 17.4.3. The extension to the Poisson model requires only the minor modifications,  $g_{it} = (y_{it} - \lambda_{it})$  and  $h_{it} = -\lambda_{it}$ . Everything else in that derivation applies with only a simple change in the notation. The first-order conditions for maximizing the log-likelihood function for the Poisson model will include

$$\frac{\partial \ln L}{\partial \alpha_i} = \sum_{t=1}^{T_i} (y_{it} - e^{\alpha_i} \mu_{it}) = 0 \quad \text{where } \mu_{it} = e^{\mathbf{x}'_{it} \beta}.$$

This implies an explicit solution for  $\alpha_i$  in terms of  $\beta$  in this model,

$$\hat{\alpha}_i = \ln \left( \frac{(1/T_i) \sum_{t=1}^{T_i} y_{it}}{(1/T_i) \sum_{t=1}^{T_i} \hat{\mu}_{it}} \right) = \ln \left( \frac{\bar{y}_i}{\hat{\mu}_i} \right). \quad (18-25)$$

Unlike the regression or the probit model, this does not require that there be within-group variation in  $y_{it}$ —all the values can be the same. It does require that at least one observation for individual  $i$  be nonzero, however. The rest of the solution for the fixed effects estimator follows the same lines as that for the probit model. An alternative approach, albeit with little practical gain, would be to concentrate the log-likelihood function by inserting this solution for  $\alpha_i$  back into the original log-likelihood, and then maximizing the resulting function of  $\beta$ . While logically this makes sense, the approach suggested earlier for the probit model is simpler to implement.

An estimator that is not a function of the fixed effects is found by obtaining the joint distribution of  $(y_{i1}, \dots, y_{iT_i})$  conditional on their sum. For the Poisson model, a close cousin to the multinomial logit model discussed earlier is produced:

$$P \left( y_{i1}, y_{i2}, \dots, y_{iT_i} \mid \sum_{t=1}^{T_i} y_{it} \right) = \frac{\left( \sum_{t=1}^{T_i} y_{it} \right)!}{\left( \prod_{t=1}^{T_i} y_{it}! \right)} \prod_{t=1}^{T_i} p_{it}^{y_{it}}, \quad (18-26)$$

where

$$p_{it} = \frac{e^{\mathbf{x}'_{it} \beta + \alpha_i}}{\sum_{t=1}^{T_i} e^{\mathbf{x}'_{it} \beta + \alpha_i}} = \frac{e^{\mathbf{x}'_{it} \beta}}{\sum_{t=1}^{T_i} e^{\mathbf{x}'_{it} \beta}}. \quad (18-27)$$

The contribution of group  $i$  to the conditional log-likelihood is

$$\ln L_i = \sum_{t=1}^{T_i} y_{it} \ln p_{it}.$$

## 818 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

Note, once again, that the contribution to  $\ln L$  of a group in which  $y_{it} = 0$  in every period is zero. Cameron and Trivedi (1998) have shown that these two approaches give identical results.

Hausman, Hall, and Griliches (1984) (HHG) report the following conditional density for the fixed effects negative binomial (FENB) model:

$$p\left(y_{i1}, y_{i2}, \dots, y_{iT_i} \mid \sum_{t=1}^{T_i} y_{it}\right) = \frac{\Gamma\left(1 + \sum_{t=1}^{T_i} y_{it}\right) \Gamma\left(\sum_{t=1}^{T_i} \lambda_{it}\right)}{\Gamma\left(\sum_{t=1}^{T_i} y_{it} + \sum_{t=1}^{T_i} \lambda_{it}\right)} \prod_{t=1}^{T_i} \frac{\Gamma(y_{it} + \lambda_{it})}{\Gamma(1 + y_{it})\Gamma(\lambda_{it})},$$

which is free of the fixed effects. This is the default FENB formulation used in popular software packages such as SAS, Stata, and LIMDEP. Researchers accustomed to the admonishments that fixed effects models cannot contain overall constants or time-invariant covariates are sometimes surprised to find (perhaps accidentally) that this fixed effects model allows both. [This issue is explored at length in Allison (2000) and Allison and Waterman (2002).] The resolution of this apparent contradiction is that the HHG FENB model is not obtained by shifting the conditional mean function by the fixed effect,  $\ln \lambda_{it} = \mathbf{x}'_{it}\beta + \alpha_i$ , as it is in the Poisson model. Rather, the HHG model is obtained by building the fixed effect into the model as an individual specific  $\theta_i$  in the Negbin 1 form in (18-22). The conditional mean functions in the models are as follows (we have changed the notation slightly to conform to our earlier formulation):

$$\text{NB1(HHG): } E[y_{it} \mid \mathbf{x}_{it}] = \theta_i \phi_{it} = \theta_i \exp(\mathbf{x}'_{it}\beta),$$

$$\text{NB2: } E[y_{it} \mid \mathbf{x}_{it}] = \exp(\alpha_i) \phi_{it} = \lambda_{it} = \exp(\mathbf{x}'_{it}\beta + \alpha_i).$$

The conditional variances are

$$\text{NB1(HHG): } \text{Var}[y_{it} \mid \mathbf{x}_{it}] = \theta_i \phi_{it}[1 + \theta_i],$$

$$\text{NB2: } \text{Var}[y_{it} \mid \mathbf{x}_{it}] = \lambda_{it}[1 + \theta_i \lambda_{it}].$$

Letting  $\mu_i = \ln \theta_i$ , it appears that the HHG formulation does provide a fixed effect in the mean, as now,  $E[y_{it} \mid \mathbf{x}_{it}] = \exp(\mathbf{x}'_{it}\beta + \mu_i)$ . Indeed, by this construction, it appears (as the authors suggest) that there are separate effects in both the mean and the variance. They make this explicit by writing  $\theta_i = \exp(\mu_i)\gamma_i$  so that in their model,

$$E[y_{it} \mid \mathbf{x}_{it}] = \gamma_i \exp(\mathbf{x}'_{it}\beta + \mu_i),$$

$$\text{Var}[y_{it} \mid \mathbf{x}_{it}] = \gamma_i \exp(\mathbf{x}'_{it}\beta + \mu_i) / [1 + \gamma_i \exp(\mu_i)].$$

The contradiction arises because the authors assert that  $\mu_i$  and  $\gamma_i$  are separate parameters. In fact, they cannot vary separately only  $\theta_i$  can vary autonomously. The firm-specific effect in the HHG model is still isolated in the scaling parameter, which falls out of the conditional density. The mean is homogeneous, which explains why a separate constant, or a time-invariant regressor (or another set of firm-specific effects) can reside there. [See Greene (2007d) and Allison and Waterman (2002) for further discussion.]

### 18.4.7.c Random Effects

The fixed effects approach has the same flaws and virtues in this setting as in the probit case. It is not necessary to assume that the heterogeneity is uncorrelated with

CHAPTER 18 ♦ Discrete Choices and Event Counts **819**

the included exogenous variables. If the uncorrelatedness of the regressors and the heterogeneity can be maintained, then the random effects model is an attractive alternative model. Once again, the approach used in the linear regression model (Partial deviations from the group means followed by generalized least squares (see Chapter 11)), is not usable here. The approach used is to formulate the joint probability conditioned upon the heterogeneity, then integrate it out of the joint distribution. Thus, we form

$$p(y_{i1}, \dots, y_{iT_i} | u_i) = \prod_{t=1}^{T_i} p(y_{it} | u_i).$$

Then the random effect is swept out by obtaining

$$\begin{aligned} p(y_{i1}, \dots, y_{iT_i}) &= \int_{u_i} p(y_{i1}, \dots, y_{iT_i}, u_i) du_i \\ &= \int_{u_i} p(y_{i1}, \dots, y_{iT_i} | u_i) g(u_i) du_i \\ &= E_{u_i}[p(y_{i1}, \dots, y_{iT_i} | u_i)]. \end{aligned}$$

This is exactly the approach used earlier to condition the heterogeneity out of the Poisson model to produce the negative binomial model. If, as before, we take  $p(y_{it} | u_i)$  to be Poisson with mean  $\lambda_{it} = \exp(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i)$  in which  $\exp(u_i)$  is distributed as gamma with mean 1.0 and variance  $1/\alpha$ , then the preceding steps produce a negative binomial distribution,

$$p(y_{i1}, \dots, y_{iT_i}) = \frac{\left[ \prod_{t=1}^{T_i} \lambda_{it}^{y_{it}} \right] \Gamma\left(\theta + \sum_{t=1}^{T_i} y_{it}\right)}{\left[ \Gamma(\theta) \prod_{t=1}^{T_i} y_{it}! \right] \left[ \left( \sum_{t=1}^{T_i} \lambda_{it} \right)^{\sum_{t=1}^{T_i} y_{it}} \right]} Q_i^\theta (1 - Q_i)^{\sum_{t=1}^{T_i} y_{it}}, \quad (18-28)$$

where

$$Q_i = \frac{\theta}{\theta + \sum_{t=1}^{T_i} \lambda_{it}}.$$

For estimation purposes, we have a negative binomial distribution for  $Y_i = \sum_t y_{it}$  with mean  $\Lambda_i = \sum_t \lambda_{it}$ .

Like the fixed effects model, introducing random effects into the negative binomial model adds some additional complexity. We do note, because the negative binomial model derives from the Poisson model by adding latent heterogeneity to the conditional mean, adding a random effect to the negative binomial model might well amount to introducing the heterogeneity a second time. However, one might prefer to interpret the negative binomial as the density for  $y_{it}$  in its own right and treat the common effects in the familiar fashion. Hausman et al.'s (1984) random effects negative binomial (RENB) model is a hierarchical model that is constructed as follows. The heterogeneity is assumed to enter  $\lambda_{it}$  additively with a gamma distribution with mean 1,  $\Gamma(\theta_i, \theta_i)$ . Then,  $\theta_i/(1+\theta_i)$  is assumed to have a beta distribution with parameters  $a$  and  $b$

**820 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

[see Appendix B.4.6)]. The resulting unconditional density after the heterogeneity is integrated out is

$$p(y_{i1}, y_{i2}, \dots, y_{iT_i}) = \frac{\Gamma(a+b)\Gamma\left(a + \sum_{t=1}^{T_i} \lambda_{it}\right)\Gamma\left(b + \sum_{t=1}^{T_i} y_{it}\right)}{\Gamma(a)\Gamma(b)\Gamma\left(a + \sum_{t=1}^{T_i} \lambda_{it} + b + \sum_{t=1}^{T_i} y_{it}\right)}.$$

As before, the relationship between the heterogeneity and the conditional mean function is unclear, because the random effect impacts the parameter of the scedastic function. An alternative approach that maintains the essential flavor of the Poisson model (and other random effects models) is to augment the NB2 form with the random effect,

$$\begin{aligned} \text{Prob}(Y = y_{it} | \mathbf{x}_{it}, \varepsilon_i) &= \frac{\Gamma(\theta + y_{it})}{\Gamma(y_{it} + 1)\Gamma(\theta)} r_{it}^{y_{it}} (1 - r_{it})^\theta, \\ \lambda_{it} &= \exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_i), \\ r_{it} &= \lambda_{it}/(\theta + \lambda_{it}). \end{aligned}$$

We then estimate the parameters by forming the conditional (on  $\varepsilon_i$ ) log-likelihood and integrating  $\varepsilon_i$  out either by quadrature or simulation. The parameters are simpler to interpret by this construction. Estimates of the two forms of the random effects model are presented in Example 18.10.2 for a comparison.

There is a mild preference in the received literature for the fixed effects estimators over the random effects estimators. The virtue of dispensing with the assumption of uncorrelatedness of the regressors and the group specific effects is substantial. On the other hand, the assumption does come at a cost. To compute the probabilities or the marginal effects, it is necessary to estimate the constants,  $\alpha_i$ . The unscaled coefficients in these models are of limited usefulness because of the nonlinearity of the conditional mean functions.

Other approaches to the random effects model have been proposed. Greene (1994, 1995a), Riphahn et al. (2003), and Terza (1995) specify a normally distributed heterogeneity, on the assumption that this is a more natural distribution for the aggregate of small independent effects. Brannas and Johanssen (1994) have suggested a semiparametric approach based on the GMM estimator by superimposing a very general form of heterogeneity on the Poisson model. They assume that conditioned on a random effect  $\varepsilon_{it}$ ,  $y_{it}$  is distributed as Poisson with mean  $\varepsilon_{it}\lambda_{it}$ . The covariance structure of  $\varepsilon_{it}$  is allowed to be fully general. For  $t, s = 1, \dots, T$ ,  $\text{Var}[\varepsilon_{it}] = \sigma_i^2$ ,  $\text{Cov}[\varepsilon_{it}, \varepsilon_{js}] = \gamma_{ij}(|t-s|)$ . For a long time series, this model is likely to have far too many parameters to be identified without some restrictions, such as first-order homogeneity ( $\boldsymbol{\beta}_i = \boldsymbol{\beta} \forall i$ ), uncorrelatedness across groups, [ $\gamma_{ij} = 0$  for  $i \neq j$ ], groupwise homoscedasticity ( $\sigma_i^2 = \sigma^2 \forall i$ ), and nonautocorrelatedness [ $\gamma(r) = 0 \forall r \neq 0$ ]. With these assumptions, the estimation procedure they propose is similar to the procedures suggested earlier. If the model imposes enough restrictions, then the parameters can be estimated by the method of moments. The authors discuss estimation of the model in its full generality. Finally, the latent class model discussed in Section 14.10 and the random parameters model in Section 15.9.5 extend naturally to the Poisson model. Indeed, most of the received applications of the latent class structure have been in the Poisson regression framework. [See Greene (2001) for a survey.]

CHAPTER 18 ♦ Discrete Choices and Event Counts **821****Example 18.10 Panel Data Models for Doctor Visits**

The German health care panel data set contains 7,293 individuals with group sizes ranging from 1 to 7. Table 18.17 presents the fixed and random effects estimates of the equation for DocVis. The pooled estimates are also shown for comparison. Overall, the panel data treatments bring large changes in the estimates compared to the pooled estimates. There is also a considerable amount of variation across the specifications. With respect to the parameter of interest, *Public*, we find that the size of the coefficient falls substantially with all panel data treatments. Whether using the pooled, fixed, or random effects specifications, the test statistics (Wald, LR) all reject the Poisson model in favor of the negative binomial. Similarly, either common effects specification is preferred to the pooled estimator. There is no simple basis for choosing between the fixed and random effects models, and we have further blurred the distinction by suggesting two formulations of each of them. We do note that the two random effects estimators are producing similar results, which one might hope for. But, the two fixed effects estimators are producing very different estimates. The NB1 estimates include two coefficients, *Income* and *Education*, which are positive, but negative in every other case. Moreover, the coefficient on *Public*, which is large and significant throughout the table, has become small and less significant with the fixed effects estimators.

We also fit a three-class latent class model for these data. (See Section 14.10.) The three class probabilities were modeled as functions of *Married* and *Female*, which appear from the results to be significant determinants of the class sorting. The average prior probabilities for the three classes are 0.09212, 0.49361, and 0.41427. The coefficients on *Public* in the three classes, with associated *t* ratios are 0.3388 (11.541), 0.1907 (3.987), and 0.1084 (4.282). The qualitative result concerning evidence of moral hazard suggested at the outset of Example 18.7 appears to be supported in a variety of specifications (with FE-NB1 the sole exception).

#### **18.4.8 TWO-PART MODELS: ZERO INFLATION AND HURDLE MODELS**

Mullahy (1986), Heilbron (1989), Lambert (1992), Johnson and Kotz (1993), and Greene (1994) have analyzed an extension of the hurdle model in which the zero outcome can arise from one of two regimes.<sup>15</sup> In one regime, the outcome is always zero. In the other, the usual Poisson process is at work, which can produce the zero outcome or some other. In Lambert's application, she analyzes the number of defective items produced by a manufacturing process in a given time interval. If the process is under control, then the outcome is always zero (by definition). If it is not under control, then the number of defective items is distributed as Poisson and may be zero or positive in any period. The model at work is therefore

$$\text{Prob}(y_i = 0 | \mathbf{x}_i) = \text{Prob}(\text{regime 1}) + \text{Prob}(y_i = 0 | \mathbf{x}_i, \text{ regime 2})\text{Prob}(\text{regime 2}),$$

$$\text{Prob}(y_i = j | \mathbf{x}_i) = \text{Prob}(y_i = j | \mathbf{x}_i, \text{ regime 2})\text{Prob}(\text{regime 2}), j = 1, 2, \dots.$$

Let  $z$  denote a binary indicator of regime 1 ( $z = 0$ ) or regime 2 ( $z = 1$ ), and let  $y^*$  denote the outcome of the Poisson process in regime 2. Then the observed  $y$  is  $z \times y^*$ . A natural extension of the splitting model is to allow  $z$  to be determined by a set of covariates. These covariates need not be the same as those that determine the conditional probabilities in the Poisson process. Thus, the model is

$$\text{Prob}(z_i = 0 | \mathbf{w}_i) = F(\mathbf{w}_i, \boldsymbol{\gamma}), (\text{Regime 1 : } y \text{ will equal zero.})$$

$$\text{Prob}(y_i = j | \mathbf{x}_i, z_i = 1) = \frac{\exp(-\lambda_i)\lambda_i^j}{j!}. (\text{Regime 2 : } y \text{ will be a count outcome.})$$

<sup>15</sup>The model is variously labeled the "with zeros," or WZ, model [Mullahy (1986)], the **zero inflated Poisson**, or **ZIP**, model [Lambert (1992)], and "zero-altered poisson," or **ZAP**, model [Greene (1994)]

TABLE 18.17 Estimated Panel Data Models for Doctor Visits (standard errors in parentheses)

Variable	Pooled (Robust S.E.)	Poisson			Negative Binomial		
		Fixed Effects		Random Effects	Pooled NB2	FE-NBI	FE-NB2
		FE	FE		FE	FE	Random Effects
Constant	0.7162 (0.1319)	0.0000	0.4957 (0.05463)	0.7628 (0.07247)	-1.2354 (0.1079)	0.0000	-0.6343 (0.07328)
Age	0.01844 (0.001336)	0.03115 (0.001443)	0.02329 (0.0004458)	0.01803 (0.0007916)	0.02389 (0.001188)	0.04479 (0.002769)	0.01899 (0.0007820)
Educ	-0.03429 (0.007255)	-0.03803 (0.01733)	-0.03427 (0.004352)	-0.03839 (0.003965)	0.01652 (0.006501)	-0.04589 (0.02967)	-0.01779 (0.004056)
Income	-0.4751 (.08212)	-0.3030 (0.04104)	-0.2646 (0.01520)	-0.4206 (0.04700)	0.02373 (0.05530)	-0.1968 (0.07320)	-0.08126 (0.04565)
Kids	-0.1582 (0.03115)	-0.001927 (0.01546)	-0.03854 (0.005272)	-0.1513 (0.01738)	-0.03381 (0.02116)	-0.001274 (0.02920)	-0.1103 (0.01675)
Public	0.2365 (0.04307)	0.1015 (0.02980)	0.1535 (0.01268)	0.2324 (0.02900)	0.05837 (0.03896)	0.09700 (0.05334)	0.1486 (0.02834)
$\theta$	0.0000	0.0000	1.1646 (0.01940)	1.9242 (0.02008)	0.0000 (0.02994)	1.9199 (0.02994)	0.0000 (0.01203)
$a$	0.0000	0.0000	0.0000	0.0000	0.0000	2.1463 (0.05955)	0.0000
$b$	0.0000	0.0000	0.0000	0.0000	0.0000	3.8011 (0.1145)	0.0000
$\sigma$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9737 (0.008235)
$\ln L$	-104440.3	-47703.34	-71763.13	-60265.49	-34016.16	-49476.36	-58182.52
							-58177.66

CHAPTER 18 ♦ Discrete Choices and Event Counts **823**

The zero inflation model can also be viewed as a type of latent class model. The two class probabilities are  $F(\mathbf{w}_i, \boldsymbol{\gamma})$  and  $1 - F(\mathbf{w}_i, \boldsymbol{\gamma})$ , and the two regimes are  $y = 0$  and the Poisson or negative binomial data generating process.<sup>16</sup> The extension of the ZIP formulation to the negative binomial model is widely labeled the ZINB model.<sup>17</sup> [See Zaninotti and Falischetti (2010) for an application.]

The mean of this random variable in the Poisson case is

$$E[y_i|\mathbf{x}_i, \mathbf{w}_i] = F_i \times 0 + (1 - F_i) \times E[y_i^*|\mathbf{x}_i, z_i = 1] = (1 - F_i)\lambda_i.$$

Lambert (1992) and Greene (1994) consider a number of alternative formulations, including logit and probit models discussed in Sections 17.2 and 17.3, for the probability of the two regimes.

It might be of interest to test simply whether there is a regime splitting mechanism at work or not. Unfortunately, the basic model and the zero-inflated model are not nested. Setting the parameters of the splitting model to zero, for example, does not produce  $\text{Prob}[z = 0] = 0$ . In the probit case, this probability becomes 0.5, which maintains the regime split. The preceding tests for over- or underdispersion would be rather indirect. What is desired is a test of non-Poissonness. An alternative distribution may (but need not) produce a systematically different proportion of zeros than the Poisson. Testing for a different distribution, as opposed to a different set of parameters, is a difficult procedure. Because the hypotheses are necessarily nonnested, the power of any test is a function of the alternative hypothesis and may, under some, be small. Vuong (1989) has proposed a test statistic for **nonnested models** that is well suited for this setting when the alternative distribution can be specified. (See Section 14.6.6.) Let  $f_j(y_i|\mathbf{x}_i)$  denote the predicted probability that the random variable  $Y$  equals  $y_i$  under the assumption that the distribution is  $f_j(y_i|\mathbf{x}_i)$ , for  $j = 1, 2$ , and let

$$m_i = \ln \left( \frac{f_1(y_i|\mathbf{x}_i)}{f_2(y_i|\mathbf{x}_i)} \right).$$

Then Vuong's statistic for testing the nonnested hypothesis of model 1 versus model 2 is

$$v = \frac{\sqrt{n} \left[ \frac{1}{n} \sum_{i=1}^n m_i \right]}{\sqrt{\frac{1}{n} \sum_{i=1}^n (m_i - \bar{m})^2}} = \frac{\sqrt{n}\bar{m}}{s_m}.$$

This is the standard statistic for testing the hypothesis that  $E[m_i]$  equals zero. Vuong shows that  $v$  has a limiting standard normal distribution. As he notes, the statistic is bidirectional. If  $|v|$  is less than two, then the test does not favor one model or the other. Otherwise, large values favor model 1 whereas small (negative) values favor model 2. Carrying out the test requires estimation of both models and computation of both sets of predicted probabilities. In Greene (1994), it is shown that the Vuong test has some power to discern the zero inflation phenomenon. The logic of the testing procedure is to allow for overdispersion by specifying a negative binomial count data process and then examine whether, *even allowing for the overdispersion*, there still appear to be excess zeros. In his application, that appears to be the case.

<sup>16</sup>Harris and Zhao (2007) applied this approach to a survey of teenage smokers and nonsmokers in Australia, using an ordered probit model. (See Section 18.3.)

<sup>17</sup>Greene (2005) presents a survey of two-part models, including the zero inflation models.

**824 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**
**TABLE 18.18** Estimated Zero Inflated Count Models

	<i>Poisson</i>			<i>Negative Binomial</i>		
	<i>Zero Inflation</i>		<i>Zero Regime</i>	<i>Zero Inflation</i>		<i>Zero Regime</i>
	<i>Poisson Regression</i>	<i>Regression</i>		<i>Negative Binomial</i>	<i>Regression</i>	
Constant	-1.33276	0.75483	2.06919	-1.54536	-0.39628	4.18910
Age	0.01286	0.00358	-0.01741	0.01807	-0.00280	-0.14339
Income	-0.02577	-0.05127	-0.03023	-0.02482	-0.05502	-0.33903
OwnRent	-0.17801	-0.15593	-0.01738	-0.18985	-0.28591	-0.50026
Self Employment	0.04691	-0.01257		0.07920	0.06817	
Dependents	0.13760	0.06038	-0.09098	0.14054	0.08599	-0.32897
Cur. Add.	0.00195	0.00046		0.00245	0.00257	
$\alpha$				6.41435	4.85653	
In $L$	-15467.71		-11569.74	-10582.88		-10516.46
Vuong			20.6981			4.5943

**Example 18.11 Zero Inflation Models for Major Derogatory Reports**

In Example 18.8, we examined the counts of major derogatory reports for a sample of 13,444 credit card applicants. It was noted that there are over 10,800 zeros in the counts. One might guess that among credit card users, there is a certain (probably large) proportion of individuals who would never generate an MDR, and some other proportion who might or might not, depending on circumstances. We propose to extend the count models in Example 10.8 to accommodate the zeros. The extensions to the ZIP and ZINB models are shown in Table 18.18. Only the coefficients are shown for purpose of the comparisons. Vuong's diagnostic statistic appears to confirm intuition that the Poisson model does not adequately describe the data; the value is 20.6981. Using the model parameters to compute a prediction of the number of zeros, it is clear that the splitting model does perform better than the basic Poisson regression. For the simple Poisson model, the average probability of zero times the sample size gives a prediction of 8609. For the ZIP model, the value is 10914.8, which is a dramatic improvement. By the likelihood ratio test, the negative binomial is clearly preferred; comparing the two zero inflation models, the difference in the log-likelihood functions is over 1,000. As might be expected, the Vuong statistic falls considerably, to 4.5943. However, the simple model with no zero inflation is still rejected by the test.

In some settings, the zero outcome of the data generating process is qualitatively different from the positive ones. The zero or nonzero value of the outcome is the result of a separate decision whether or not to "participate" in the activity. On deciding to participate, the individual decides separately how much, that is, how intensively. Mullahy (1986) argues that this fact constitutes a shortcoming of the Poisson (or negative binomial) model and suggests a **hurdle model** as an alternative.<sup>18</sup> In his formulation, a binary probability model determines whether a zero or a nonzero outcome occurs and then, in the latter case, a (truncated) Poisson distribution describes the positive outcomes. The model is

$$\text{Prob}(y_i = 0 | \mathbf{x}_i) = e^{-\theta}$$

$$\text{Prob}(y_i = j | \mathbf{x}_i) = (1 - e^{-\theta}) \frac{\exp(-\lambda_i) \lambda_i^j}{j! [1 - \exp(-\lambda_i)]}, \quad j = 1, 2, \dots$$

<sup>18</sup>For a similar treatment in continuous data application, see Cragg (1971).

CHAPTER 18 ♦ Discrete Choices and Event Counts **825**

This formulation changes the probability of the zero outcome and scales the remaining probabilities so that they sum to one. Mullahy suggests some formulations and applies them to a sample of observations on daily beverage consumption. Mullahy's formulation adds a new restriction that  $\text{Prob}(y_i = 0|\mathbf{x}_i)$  no longer depends on the covariates, however. The natural next step is to parameterize this probability. This extension of the hurdle model would combine a binary choice model like those in Section 17.2 and 17.3 with a truncated count model as shown in Section 18.4.6. This would produce, for example, for a logit participation equation and a Poisson intensity equation,

$$\begin{aligned}\text{Prob}(y_i = 0|\mathbf{w}_i) &= \Lambda(\mathbf{w}'_i \boldsymbol{\gamma}) \\ \text{Prob}(y_i = j|\mathbf{x}_i, \mathbf{w}_i, y_i > 0) &= \frac{[1 - \Lambda(\mathbf{w}'_i \boldsymbol{\gamma})] \exp(-\lambda_i) \lambda_i^j}{j! [1 - \exp(-\lambda_i)]}.\end{aligned}$$

The conditional mean function in the hurdle model is

$$E[y_i|\mathbf{x}_i, \mathbf{w}_i] = \frac{[1 - F(\mathbf{w}'_i \boldsymbol{\gamma})] \lambda_i}{[1 - \exp(-\lambda_i)]}, \quad \lambda_i = \exp(\mathbf{x}'_i \boldsymbol{\beta}),$$

where  $F(\cdot)$  is the probability model used for the participation equation (probit or logit). The partial effects are obtained by differentiating with respect to the two sets of variables separately,

$$\begin{aligned}\frac{\partial E[y_i|\mathbf{x}_i, \mathbf{w}_i]}{\partial \mathbf{x}_i} &= [1 - F(\mathbf{w}'_i \boldsymbol{\gamma})] \delta_i, \\ \frac{\partial E[y_i|\mathbf{x}_i, \mathbf{w}_i]}{\partial \mathbf{w}_i} &= \left\{ \frac{-f(\mathbf{w}'_i \boldsymbol{\gamma}) \lambda_i}{[1 - \exp(-\lambda_i)]} \right\} \boldsymbol{\gamma},\end{aligned}$$

where  $\delta_i$  is defined in (18-23) and  $f(\cdot)$  is the density corresponding to  $F(\cdot)$ . For variables that appear in both  $\mathbf{x}_i$  and  $\mathbf{w}_i$ , the effects are added. For dummy variables, the preceding would be an approximation; the appropriate result would be obtained by taking the difference of the conditional mean with the variable fixed at one and zero.

It might be of interest to test for hurdle effects. The hurdle model is similar to the zero inflation model in that a model without hurdle effects is not nested within the hurdle model; setting  $\boldsymbol{\gamma} = \mathbf{0}$  produces either  $F = \alpha$ , a constant, or  $F = 1/2$  if the constant term is also set to zero. Neither serves the purpose. Nor does forcing  $\boldsymbol{\gamma} = \boldsymbol{\beta}$  in a model with  $\mathbf{w}_i = \mathbf{x}_i$  and  $F = \Lambda$  with a Poisson intensity equation, which might be intuitively appealing. A complementary log log model with

$$\text{Prob}(y_i = 0|\mathbf{w}_i) = \exp[-\exp(\mathbf{w}'_i \boldsymbol{\gamma})]$$

does produce the desired result if  $\mathbf{w}_i = \mathbf{x}_i$ . In this case, "hurdle effects" are absent if  $\boldsymbol{\gamma} = \boldsymbol{\beta}$ . The strategy in this case, then, would be a test of this restriction. But, this formulation is otherwise restrictive, first in the choice of variables and second in its unconventional functional form. The more general approach to this test would be the Vuong test used earlier to test the zero inflation model against the simpler Poisson or negative binomial model.

The hurdle model bears some similarity to the zero inflation model; however, the behavioral implications are different. The zero inflation model can usefully be viewed as a latent class model. The splitting probability defines a regime determination. In the hurdle model, the splitting equation represents a behavioral outcome on the same

**826 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**
**TABLE 18.19** Estimated Hurdle Model for Doctor Visits

	<i>Participation Equation</i>		<i>Intensity Equation</i>		<i>Total Partial Effect</i> (Poisson Model)
	<i>Parameter</i>	<i>Partial Effect</i>	<i>Parameter</i>	<i>Partial Effect</i>	
Constant	-0.0598		1.1203		
Age	0.0221	0.0244	0.0113	0.0538	0.0782 ( 0.0625)
Income	0.0725	0.0800	-0.5152	-2.4470	-2.3670 (-1.8130)
Kids			-0.0842	-0.4000	-0.4000 (-0.4836)
Public	0.2411	0.2663	0.1966	0.9338	1.2001 ( 0.9744)
Education	-0.0291	-0.0321			-0.0321
Married	-0.0233	-0.0258			-0.0258
Working	-0.3624	-0.4003			-0.4003

level as the intensity (count) equation. Both of these modifications substantially alter the Poisson formulation. First, note that the equality of the mean and variance of the distribution no longer follows; both modifications induce overdispersion. On the other hand, the overdispersion does not arise from heterogeneity; it arises from the nature of the process generating the zeros. As such, an interesting identification problem arises in this model. If the data do appear to be characterized by overdispersion, then it seems less than obvious whether it should be attributed to heterogeneity or to the regime splitting mechanism. Mullahy (1986) argues the point more strongly. He demonstrates that overdispersion will always induce excess zeros. As such, in a splitting model, we may misinterpret the excess zeros as due to the splitting process instead of the heterogeneity.

**Example 18.12 Hurdle Model for Doctor Visits**

The hurdle model is a natural specification for models of utilization of the health care system, and has been used in a number of studies. Table 18.19 shows the parameter estimates for a hurdle model for doctor visits based on the entire pooled sample of 27,326 observations. The decomposition of the partial effects shows that the participation and intensity decisions each contribute substantively to the effects of Age, Income, and Public insurance. The value of the Vuong statistic is 51.16, strongly in favor of the hurdle model compared to the pooled Poisson model with no hurdle effects. The effect of the hurdle model on the partial effects is shown in the last column where the results for the Poisson model are shown in parentheses.

#### 18.4.9 ENDOGENOUS VARIABLES AND ENDOGENOUS PARTICIPATION

As in other situations, one would expect to find endogenous variables in models for counts. For example, in the study on which we have relied for our examples of health care utilization, Riphahn, Wambach, and Million (RWM, 2003), the authors were interested in the role of insurance (specifically the *Add-On* insurance) in the usage variable. One might expect the choice to buy insurance to be at least partly influenced by some of the same factors that motivate usage of the health care system. Insurance purchase might well be endogenous in a model such as the hurdle model in Example 18.12.

The Poisson model presents a complication for modeling endogeneity that arises in some other cases as well. For simplicity, consider a continuous variable, such as *Income*, to continue our ongoing example. A model of income determination and doctor visits might appear

$$\text{Income} = \mathbf{z}'_i \boldsymbol{\gamma} + u_i,$$

$$\text{Prob}(DocVis_i = j | \mathbf{x}_i, \text{Income}_i) = \exp(-\lambda_i), \lambda_i^j / j!, \lambda_i = \exp(\mathbf{x}'_i \boldsymbol{\beta} + \delta \text{Income}_i).$$

CHAPTER 18 ♦ Discrete Choices and Event Counts **827**

Endogeneity as we have analyzed it, for example, in Chapter 8 and Sections 17.3.5 and 17.5.5, arises through correlation between the endogenous variable and the unobserved omitted factors in the main equation. But, the Poisson model does not contain any unobservables. This is a major shortcoming of the specification as a “regression” model; all of the regression variation of the dependent variable arises through variation of the observables. There is no accommodation for unobserved heterogeneity or omitted factors. This is the compelling motivation for the negative binomial model or, in RWM’s case, the Poisson-normal mixture model. [See Terza (2010, pp. 555–556) for discussion of this issue.] If the model is reformulated to accommodate heterogeneity, as in

$$\lambda_i = \exp(\mathbf{x}'_i \boldsymbol{\beta} + \delta \text{Income}_i + \varepsilon_i),$$

then  $\text{Income}_i$  will be endogenous if  $u_i$  and  $\varepsilon_i$  are correlated.

A bivariate normal model for  $(u_i, \varepsilon_i)$  with zero means, variances  $\sigma_u^2$  and  $\sigma_\varepsilon^2$  and correlation  $\rho$  provides a convenient (and the usual) platform to operationalize this idea. By projecting  $\varepsilon_i$  on  $u_i$ , we have

$$\varepsilon_i = (\rho\sigma_\varepsilon/\sigma_u)u_i + v_i,$$

where  $v_i$  is normally distributed with mean zero and variance  $\sigma_\varepsilon^2(1 - \rho^2)$ . It will prove convenient to parameterize these based on the regression and the specific parameters as follows:

$$\begin{aligned}\varepsilon_i &= \rho\sigma_\varepsilon(\text{Income}_i - \mathbf{z}'_i \boldsymbol{\gamma})/\sigma_u + v_i, \\ &= \tau[(\text{Income}_i - \mathbf{z}'_i \boldsymbol{\gamma})/\sigma_u] + \theta w_i.\end{aligned}$$

where  $w_i$  will be normally distributed with mean zero and variance one while  $\tau = \rho\sigma_\varepsilon$  and  $\theta^2 = \sigma_\varepsilon^2(1 - \rho^2)$ . Then, combining terms,

$$\varepsilon_i = \tau u_i^* + \theta w_i.$$

With this parameterization, the conditional mean function in the Poisson regression model is

$$\lambda_i = \exp(\mathbf{x}'_i \boldsymbol{\beta} + \delta \text{Income}_i + \tau u_i^* + \theta w_i).$$

The parameters to be estimated are  $\boldsymbol{\beta}$ ,  $\boldsymbol{\gamma}$ ,  $\delta$ ,  $\sigma_\varepsilon$ ,  $\sigma_u$ , and  $\rho$ . There are two ways to proceed. A two-step method can be based on the fact that  $\boldsymbol{\gamma}$  and  $\sigma_u$  can consistently be estimated by linear regression of  $\text{Income}$  on  $\mathbf{z}$ . After this first step, we can compute values of  $u_i^*$  and formulate the Poisson regression model in terms of

$$\hat{\lambda}_i(w_i) = \exp[\mathbf{x}'_i \boldsymbol{\beta} + \delta \text{Income}_i + \tau \hat{u}_i + \theta w_i].$$

The log-likelihood to be maximized at the second step is

$$\ln L(\boldsymbol{\beta}, \delta, \tau, \theta | \mathbf{w}) = \sum_{i=1}^n -\hat{\lambda}_i(w_i) + y_i \ln \hat{\lambda}_i(w_i) - \ln y_i!.$$

A remaining complication is that the unobserved heterogeneity,  $w_i$  remains in the equation so it must be integrated out of the log-likelihood function. The unconditional log-likelihood function is obtained by integrating the standard normally distributed  $w_i$  out

**828 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

of the conditional densities.

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau, \theta) = \sum_{i=1}^n \ln \left\{ \int_{-\infty}^{\infty} \left[ \frac{\exp(-\hat{\lambda}_i(w_i)) (\hat{\lambda}_i(w_i))^{y_i}}{y_i!} \right] \phi(w_i) dw_i \right\}.$$

The method of Butler and Moffitt or maximum simulated likelihood that we used to fit a probit model in Section 17.4.2 can be used to estimate  $\boldsymbol{\beta}$ ,  $\delta$ ,  $\tau$ , and  $\theta$ . Estimates of  $\rho$  and  $\sigma_\varepsilon$  can be deduced from the last two of these;  $\sigma_\varepsilon^2 = \theta^2 + \tau^2$  and  $\rho = \tau/\sigma_\varepsilon$ . This is the control function method discussed in Section 17.3.5 and is also the “residual inclusion” method discussed by Terza, Basu, and Rathouz (2008).

The full set of parameters can be estimated in a single step using **full information maximum likelihood**. To estimate all parameters simultaneously and efficiently, we would form the log-likelihood from joint density of *DocVis* and *Income* as  $P(DocVis|Income) f(Income)$ . Thus,

$$f(DocVis, Income) = \frac{\exp[-\lambda_i(w_i)] [\lambda_i(w_i)]^{y_i}}{y_i!} \frac{1}{\sigma_u} \phi\left(\frac{Income - \mathbf{z}'_i \boldsymbol{\gamma}}{\sigma_u}\right)$$

$$\lambda_i(w_i) = \exp(\mathbf{x}'_i \boldsymbol{\beta} + \delta Income_i + \tau (Income_i - \mathbf{z}'_i \boldsymbol{\gamma})/\sigma_u + \theta w_i)$$

As before, the unobserved  $w_i$  must be integrated out of the log-likelihood function. Either quadrature or simulation can be used. The parameters to be estimated by maximizing the full log-likelihood are  $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \delta, \sigma_u, \sigma_\varepsilon, \rho)$ . The invariance principle has been used to simplify the estimation a bit by parameterizing the log-likelihood function in terms of  $\tau$  and  $\theta$ . Some additional simplification can also be obtained by using the Olsen (1978) [and Tobin (1958)] transformations,  $\eta = 1/\sigma_u$  and  $\omega = (1/\sigma_u)\boldsymbol{\gamma}$ .

An endogenous binary variable, such as *Public* or *AddOn* in our *DocVis* example is handled similarly but is a bit simpler. The structural equations of the model are

$$T^* = \mathbf{z}'_i \boldsymbol{\gamma} + u_i, \quad u \sim N[0, 1],$$

$$T = 1(T^* > 0),$$

$$\lambda = \exp(\mathbf{x}' \boldsymbol{\beta} + \delta T + \varepsilon) \quad \varepsilon \sim N[0, \sigma_\varepsilon^2],$$

with  $\text{Cov}(u, \varepsilon) = \rho\sigma_\varepsilon$ . The endogeneity of  $T$  is implied by a nonzero  $\rho$ . We use the bivariate normal result

$$u = (\rho/\sigma_\varepsilon)\varepsilon + v$$

where  $v$  is normally distributed with mean zero and variance  $1 - \rho^2$ . Then, using our earlier results for the probit model (Section 17.2),

$$P(T|\varepsilon) = \Phi \left[ (2T - 1) \left( \frac{\mathbf{z}' \boldsymbol{\gamma} + (\rho/\sigma_\varepsilon)\varepsilon}{\sqrt{1 - \rho^2}} \right) \right], \quad T = 0, 1.$$

It will be convenient once again to write  $\varepsilon = \sigma_\varepsilon w$  where  $w \sim N[0, 1]$ . Making the substitution, we have

$$P(T|w) = \Phi \left[ (2T - 1) \left( \frac{\mathbf{z}' \boldsymbol{\gamma} + \rho w}{\sqrt{1 - \rho^2}} \right) \right], \quad T = 0, 1.$$

**CHAPTER 18 ♦ Discrete Choices and Event Counts 829**

The probability density function for  $y|T, w$  is Poisson with  $\lambda(w) = \exp(\mathbf{x}'\boldsymbol{\beta} + \delta T + \sigma_\varepsilon w)$ . Combining terms,

$$P(y, T|w) = \frac{\exp[-\lambda(w)] [\lambda(w)]^y}{y!} \Phi \left[ (2T - 1) \left( \frac{\mathbf{z}'\boldsymbol{\gamma} + \rho w}{\sqrt{1 - \rho^2}} \right) \right].$$

This last result provides the terms that enter the log-likelihood for  $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \delta, \rho, \sigma_\varepsilon)$ . As before, the unobserved heterogeneity,  $w$ , must be integrated out of the log-likelihood, so either the quadrature or simulation method discussed in Chapter 17 is used to obtain the parameter estimates. Note that this model may also be estimated in two steps, with  $\boldsymbol{\gamma}$  obtained in the first-step probit. The two-step method will not be appreciably simpler, since the second term in the density must remain to identify  $\rho$ . The residual inclusion method is not feasible here since  $T^*$  is not observed.

This same set of methods is used to allow for endogeneity of the participation equation in the hurdle model in Section 18.4.8. Mechanically, the hurdle model with endogenous participation is essentially the same as the endogenous binary variable. [See Greene (2005, 2007).]

## 18.5 SUMMARY AND CONCLUSIONS

The analysis of individual decisions in microeconomics is largely about discrete decisions such as whether to participate in an activity or not, whether to make a purchase or not, or what brand of product to buy. This chapter and Chapter 17 have developed the four essential models used in that type of analysis. Random utility, the binary choice model, and regression-style modeling of probabilities developed in Chapter 17 are the three fundamental building blocks of discrete choice modeling. This chapter extended those tools into the three primary areas of choice modeling, unordered choice models, ordered choice models, and models for counts. In each case, we developed a core modeling framework that provides the broad platform and then developed a variety of extensions.

In the analysis of unordered choice models, such as brand or location, the multinomial logit (MNL) model has provided the essential starting point. The MNL works well to provide a basic framework, but as a behavioral model in its own right, it has some important shortcomings. Much of the recent research in this area has focused on relaxing these behavioral assumptions. The most recent research in this area, on the mixed logit model, has produced broadly flexible functional forms that can match behavioral modeling to empirical specification and estimation.

The ordered choice model is a natural extension of the binary choice setting and also a convenient bridge between models of choice between two alternatives and more complex models of choice among multiple alternatives. We began this analysis with the ordered probit and logit model pioneered by Zavoina and McKelvey (1975). Recent developments of this model have produced the same sorts of extensions to panel data and modeling heterogeneity that we considered in Chapter 17 for binary choice. We also examined some multiple-equation specifications. For all its versatility, the familiar ordered choice models have an important shortcoming in the assumed constancy underlying preference behind the rating scale. The current work on differential item

## 830 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

functioning, such as King et al. (2004), has produced significant progress on filling this gap in the theory.

Finally, we examined probability models for counts of events. Here, the Poisson regression model provides the broad framework for the analysis. The Poisson model has two shortcomings that have motivated the current stream of research. The functional form binds the mean of the random variable to its variance, producing an unrealistic regression specification. Second, the basic model has no component that accommodates unmeasured heterogeneity. (This second feature is what produces the first.) Current research has produced a rich variety of models for counts, such as two-part behavioral models that account for many different aspects of the decision-making process and the mechanisms that generate the observed data.

### Key Terms and Concepts

- Bivariate ordered probit
- Censoring
- Choice based sample
- Conditional logit model
- Count data
- Deviance
- Differential item functioning (DIF)
- Event count
- Exposure
- Full information maximum likelihood (FIML)
- Heterogeneity
- Hurdle model
- Identification through functional form
- Inclusive value
- Independence from irrelevant alternatives (IIA)
- Lagrange multiplier test
- Limited information
- Log-odds
- Loglinear model
- Method of simulated moments
- Mixed logit model
- Multinomial choice
- Multinomial logit model
- Multinomial probit model (MNP)
- Negative binomial model
- Negbin 1 (NB1) form
- Negbin 2 (NB2) form
- Negbin P (NBP) model
- Nested logit model
- Nonnested models
- Ordered choice model
- Overdispersion
- Parallel regression assumption
- Poisson regression model
- Random coefficients
- Random parameters logit model (RPL)
- Revealed preference data
- Specification error
- Stated choice experiment
- Subjective well-being
- Unordered choice model
- Willingness to pay space
- Zero inflated Poisson model (ZIP)

### Exercises

1. We are interested in the ordered probit model. Our data consist of 250 observations, of which the responses are

$y$	0	1	2	3	4	
$n$	50	40	45	80	35	

Using the preceding data, obtain maximum likelihood estimates of the unknown parameters of the model. (*Hint:* Consider the probabilities as the unknown parameters.)

2. For the zero-inflated Poisson (ZIP) model in Section 18.4.8, we derived the conditional mean function,  $E[y_i|\mathbf{x}_i, \mathbf{w}_i] = (1 - F_i)\lambda_i$ .
  - a. For the same model, now obtain  $\text{Var}[y_i|\mathbf{x}_i, \mathbf{w}_i]$ . Then, obtain  $\tau_i = \text{Var}[y_i|\mathbf{x}_i, \mathbf{w}_i]/E[y_i|\mathbf{x}_i, \mathbf{w}_i]$ . Does the zero inflation produce overdispersion? (That is, is the ratio greater than one?)
  - b. Obtain the partial effect for a variable  $z_i$  that appears in both  $\mathbf{w}_i$  and  $\mathbf{x}_i$ .

**CHAPTER 18 ♦ Discrete Choices and Event Counts 831**

3. Consider estimation of a Poisson regression model for  $y_i | x_i$ . The data are truncated on the left—these are on-site observations at a recreation site, so zeros do not appear in the data set. The data are censored on the right—any response greater than 5 is recorded as a 5. Construct the log-likelihood for a data set drawn under this sampling scheme.

**Applications**

1. Appendix Table F10.1 provides Fair's (1978) *Redbook Magazine* survey on extramarital affairs. The variables in the data set are as follows:

$id$  = an identification number

$C$  = constant, value = 1

$yrb$  = a constructed measure of time spent in extramarital affairs

$v_1$  = a rating of the marriage, coded 1 to 5

$v_2$  = age, in years, aggregated

$v_3$  = number of years married

$v_4$  = number of children, top coded at 5

$v_5$  = religiosity, 1 to 4, 1 = not, 4 = very

$v_6$  = education, coded 9, 12, 14, 16, 17, 20

$v_7$  = occupation

$v_8$  = husband's occupation

and three other variables that are not used. The sample contains a survey of 6,366 married women. For this exercise, we will analyze, first, the binary variable  $A = 1$  if  $yrb > 0$ , 0 otherwise. The regressors of interest are  $v_1$  to  $v_8$ ; however, not necessarily all of them belong in your model. Use these data to build a bin choice model for  $A$ . Report all computed results for the model. Compute the marginal effects for the variables you choose. Compare the results you obtain for a probit model to those for a logit model. Are there any substantial differences in the results for the two models?

2. Continuing the analysis of the first application, we now consider the self-reported rating, This is a natural candidate for an ordered choice model, because the simple Likert item coding is a censored version of what would be a continuous scale on some subjective satisfaction variable. Analyze this variable using an ordered probit model. What variables appear to explain the response to this survey question? (Note: The variable is coded 1, 2, 3, 4, 5. Some programs accept data for ordered choice modeling in this form, for example, *Stata*, while others require the variable to be coded 0, 1, 2, 3, 4, for example, *LIMDEP*. Be sure to determine which is appropriate for the program you are using and transform the data if necessary.) Can you obtain the partial effects for your model? Report them as well. What do they suggest about the impact of the different independent variables on the reported ratings?
3. Several applications in the preceding chapters using the German health care data have examined the variable Doc Vis, the reported number of visits to the doctor. The data are described in Appendix Table F7.1. A second count variable in that data set that we have not examined is Hosp Vis, the number of visits to hospital. For this application, we will examine this variable. To begin, we treat the full sample (27,326) observations as a cross section.

**832 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

- a. Begin by fitting a Poisson regression model to this variable. The exogenous variables are listed in Appendix Table F7.1. Determine an appropriate specification for the right-hand side of your model. Report the regression results and the partial effects.
- b. Estimate the model using ordinary least squares and compare your least squares results to the partial eff you computed in part a. What do you find?
- c. Is there evidence of overdispersion in the data? Test for overdispersion. Now, reestimate the model using a negative binomial specification. What is the result? Do your results change? Use a likelihood ratio test to test the hypothesis of the negative binomial model against the Poisson.
4. The GSOEP data are an unbalanced panel, with 7,293 groups. Continue your analysis in Application 3 by fitting the Poisson model with fixed and with random effects and compare your results. (Recall, like the linear model, the Poisson fixed effects model may not contain any time-invariant variables.) How do the panel data results compare to the po results?
5. Appendix Table F18.2 contains data on ship accidents reported in McCullagh and Nelder (1983). The data set contains 40 observations on the number of incidents of wave damage for oceangoing ships. Regressors include “aggregate months of service”, and three sets of dummy variables, Type (1, . . . , 5), operation period (1960–1974 or 1975–1979), and construction period (1960–1964, 1965–1969, or 1970–1974). There are six missing values on the dependent variable, leaving 34 usable observations.
  - a. Fit a Poisson model for these data, using the log of service months, four types of dummy variables, two construction period variables, and one operation period dummy variable. Report your results.
  - b. The authors note that the rate of accidents is supposed to be per period, but the exposure (aggregate months) differs by ship. Reestimate your model constraining the coefficient on log of service months to equal one.
  - c. The authors take overdispersion as a given in these data. Do you find evidence of over dispersion? Show your results.

## 19

# LIMITED DEPENDENT VARIABLES—TRUNCATION, CENSORING, AND SAMPLE SELECTION

---

## 19.1 INTRODUCTION

This chapter is concerned with **truncation** and **censoring**. As we saw in Section 18.4.6, these features complicate the analysis of data that might otherwise be amenable to conventional estimation methods such as regression. “Truncation” effects arise when one attempts to make inferences about a larger population from a sample that is drawn from a distinct subpopulation. For example, studies of income based on incomes above or below some poverty line may be of limited usefulness for inference about the whole population. Truncation is essentially a characteristic of the distribution from which the sample data are drawn. Censoring is a more common feature of recent studies. To continue the example, suppose that instead of being unobserved, all incomes below the poverty line are reported as if they were *at* the poverty line. The censoring of a range of values of the variable of interest introduces a distortion into conventional statistical results that is similar to that of truncation. Unlike truncation, however, censoring is essentially a defect in the sample data. Presumably, if they were not censored, the data would be a representative sample from the population of interest. We will also examine a form of truncation called the **sample selection** problem. Although most empirical work in this area involves censoring rather than truncation, we will study the simpler model of truncation first. It provides most of the theoretical tools we need to analyze models of censoring and sample selection.

The discussion will examine the general characteristics of truncation, censoring, and sample selection, and then, in each case, develop a major area of application of the principles. The stochastic frontier model [Aigner, Lovell, and Schmidt (1977), Fried, Lovell, and Schmidt (2008)] is a leading application of results for truncated distributions in empirical models. Censoring appears prominently in the analysis of labor supply and in modeling of duration data. Finally, the sample selection model has appeared in all areas of the social sciences and plays a significant role in the evaluation of treatment effects and program evaluation.

## 19.2 TRUNCATION

In this section, we are concerned with inferring the characteristics of a full population from a sample drawn from a restricted part of that population.

## 834 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

### 19.2.1 TRUNCATED DISTRIBUTIONS

A **truncated distribution** is the part of an untruncated distribution that is above or below some specified value. For instance, in Example 19.2, we are given a characteristic of the distribution of incomes above \$100,000. This subset is a part of the full distribution of incomes which range from zero to (essentially) infinity.

#### THEOREM 19.1 Density of a Truncated Random Variable

If a continuous random variable  $x$  has pdf  $f(x)$  and  $a$  is a constant, then<sup>1</sup>

$$f(x | x > a) = \frac{f(x)}{\text{Prob}(x > a)}.$$

The proof follows from the definition of conditional probability and amounts merely to scaling the density so that it integrates to one over the range above  $a$ . Note that the truncated distribution is a conditional distribution.

Most recent applications based on continuous random variables use the **truncated normal distribution**. If  $x$  has a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ , then

$$\text{Prob}(x > a) = 1 - \Phi\left(\frac{a - \mu}{\sigma}\right) = 1 - \Phi(\alpha),$$

where  $\alpha = (a - \mu)/\sigma$  and  $\Phi(\cdot)$  is the standard normal cdf. The density of the truncated normal distribution is then

$$f(x | x > a) = \frac{f(x)}{1 - \Phi(\alpha)} = \frac{(2\pi\sigma^2)^{-1/2}e^{-(x-\mu)^2/(2\sigma^2)}}{1 - \Phi(\alpha)} = \frac{\frac{1}{\sigma}\phi\left(\frac{x - \mu}{\sigma}\right)}{1 - \Phi(\alpha)},$$

where  $\phi(\cdot)$  is the standard normal pdf. The **truncated standard normal distribution**, with  $\mu = 0$  and  $\sigma = 1$ , is illustrated for  $a = -0.5, 0$ , and  $0.5$  in Figure 19.1. Another truncated distribution that has appeared in the recent literature, this one for a discrete random variable, is the truncated at zero Poisson distribution,

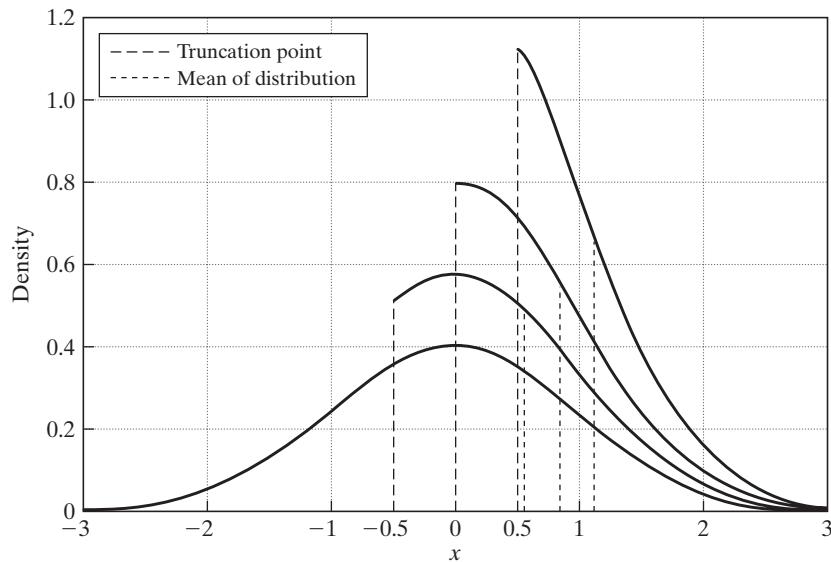
$$\begin{aligned} \text{Prob}[Y = y | y > 0] &= \frac{(e^{-\lambda}\lambda^y)/y!}{\text{Prob}[Y > 0]} = \frac{(e^{-\lambda}\lambda^y)/y!}{1 - \text{Prob}[Y = 0]} \\ &= \frac{(e^{-\lambda}\lambda^y)/y!}{1 - e^{-\lambda}}, \quad \lambda > 0, y = 1, \dots \end{aligned}$$

This distribution is used in models of uses of recreation and other kinds of facilities where observations of zero uses are discarded.<sup>2</sup>

For convenience in what follows, we shall call a random variable whose distribution is truncated a **truncated random variable**.

<sup>1</sup>The case of truncation from above instead of below is handled in an analogous fashion and does not require any new results.

<sup>2</sup>See Shaw (1988). An application of this model appears in Section 18.4.6 and Example 18.8.

**FIGURE 19.1** Truncated Normal Distributions.**19.2.2 MOMENTS OF TRUNCATED DISTRIBUTIONS**

We are usually interested in the mean and variance of the truncated random variable. They would be obtained by the general formula:

$$E[x | x > a] = \int_a^\infty xf(x | x > a) dx$$

for the mean and likewise for the variance.

**Example 19.1 Truncated Uniform Distribution**

If  $x$  has a standard uniform distribution, denoted  $U(0, 1)$ , then

$$f(x) = 1, \quad 0 \leq x \leq 1.$$

The truncated at  $x = \frac{1}{3}$  distribution is also uniform:

$$f\left(x | x > \frac{1}{3}\right) = \frac{f(x)}{\text{Prob}(x > \frac{1}{3})} = \frac{1}{\left(\frac{2}{3}\right)} = \frac{3}{2}, \quad \frac{1}{3} \leq x \leq 1.$$

The expected value is

$$E\left[x | x > \frac{1}{3}\right] = \int_{1/3}^1 x \left(\frac{3}{2}\right) dx = \frac{2}{3}.$$

For a variable distributed uniformly between  $L$  and  $U$ , the variance is  $(U - L)^2 / 12$ . Thus,

$$\text{Var}\left[x | x > \frac{1}{3}\right] = \frac{1}{27}.$$

The mean and variance of the untruncated distribution are  $\frac{1}{2}$  and  $\frac{1}{12}$ , respectively.

## 836 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

Example 19.1 illustrates two results.

1. If the truncation is from below, then the mean of the truncated variable is greater than the mean of the original one. If the truncation is from above, then the mean of the truncated variable is smaller than the mean of the original one.
2. Truncation reduces the variance compared with the variance in the untruncated distribution.

Henceforth, we shall use the terms **truncated mean** and **truncated variance** to refer to the mean and variance of the random variable with a truncated distribution.

For the truncated normal distribution, we have the following theorem:<sup>3</sup>

### THEOREM 19.2 Moments of the Truncated Normal Distribution

If  $x \sim N[\mu, \sigma^2]$  and  $a$  is a constant, then

$$E[x | \text{truncation}] = \mu + \sigma\lambda(\alpha), \quad (19-1)$$

$$\text{Var}[x | \text{truncation}] = \sigma^2[1 - \delta(\alpha)], \quad (19-2)$$

where  $\alpha = (a - \mu)/\sigma$ ,  $\phi(\alpha)$  is the standard normal density and

$$\lambda(\alpha) = \phi(\alpha)/[1 - \Phi(\alpha)] \quad \text{if truncation is } x > a, \quad (19-3a)$$

$$\lambda(\alpha) = -\phi(\alpha)/\Phi(\alpha) \quad \text{if truncation is } x < a, \quad (19-3b)$$

and

$$\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha]. \quad (19-4)$$

An important result is

$$0 < \delta(\alpha) < 1 \quad \text{for all values of } \alpha,$$

which implies point 2 after Example 19.1. A result that we will use at several points below is  $d\phi(\alpha)/d\alpha = -\alpha\phi(\alpha)$ . The function  $\lambda(\alpha)$  is called the **inverse Mills ratio**. The function in (19-3a) is also called the **hazard function** for the standard normal distribution.

#### Example 19.2 A Truncated Lognormal Income Distribution

"The typical 'upper affluent American'... makes \$142,000 per year.... The people surveyed had household income of at least \$100,000."<sup>4</sup> Would this statistic tell us anything about the "typical American"? As it stands, it probably does not (popular impressions notwithstanding). The 1987 article where this appeared went on to state, "If you're in that category, pat yourself on the back—only 2 percent of American households make the grade, according to the survey." Because the **degree of truncation** in the sample is 98 percent, the \$142,000 was probably quite far from the mean in the full population.

Suppose that incomes,  $x$ , in the population were lognormally distributed—see Section B.4.4. Then the log of income,  $y$ , had a normal distribution with, say, mean  $\mu$  and

<sup>3</sup>Details may be found in Johnson, Kotz, and Balakrishnan (1994, pp. 156–158). Proofs appear in Cameron and Trivedi (2005).

<sup>4</sup>See *New York Post* (1987).

**CHAPTER 19 ♦ Limited Dependent Variables 837**

standard deviation,  $\sigma$ . Suppose that the survey was large enough for us to treat the sample average as the true mean. Assuming so, we'll deduce  $\mu$  and  $\sigma$  and then determine the population mean income.

Two useful numbers for this example are  $\ln 100 = 4.605$  and  $\ln 142 = 4.956$ . The article states that

$$\text{Prob}[x \geq 100] = \text{Prob}[\exp(y) \geq 100] = 0.02,$$

or

$$\text{Prob}(y < 4.605) = 0.98.$$

This implies that

$$\text{Prob}[(y - \mu)/\sigma < (4.605 - \mu)/\sigma] = 0.98.$$

Because  $\Phi[(4.605 - \mu)/\sigma] = 0.98$ , we know that

$$\Phi^{-1}(0.98) = 2.054 = (4.605 - \mu)/\sigma,$$

or

$$4.605 = \mu + 2.054\sigma.$$

The article also states that

$$E[x | x > 100] = E[\exp(y) | \exp(y) > 100] = 142,$$

or

$$E[\exp(y) | y > 4.645] = 142.$$

To proceed, we need another result for the lognormal distribution:

$$\text{If } y \sim N[\mu, \sigma^2], \text{ then } E[\exp(y) | y > a] = \exp(\mu + \sigma^2/2) \times \frac{\Phi(\sigma - (a - \mu)/\sigma)}{1 - \Phi((a - \mu)/\sigma)}.$$

[See Johnson, Kotz and Balakrishnan (1995, p. 241).] For our application, we would equate this expression to 142, and  $a$  to  $\ln 100 = 4.605$ . This provides a second equation. To estimate the two parameters, we used the method of moments. We solved the minimization problem

$$\begin{aligned} \text{Minimize}_{\mu, \sigma} & [4.605 - (\mu + 2.054\sigma)]^2 \\ & + [142\Phi((\mu - 4.605)/\sigma) - \exp(\mu + \sigma^2/2)\Phi(\sigma - (4.605 - \mu)/\sigma)]^2. \end{aligned}$$

The two solutions are 2.89372 and 0.83314 for  $\mu$  and  $\sigma$ , respectively. To obtain the mean income, we now use the result that if  $y \sim N[\mu, \sigma^2]$  and  $x = \exp(y)$ , then  $E[x] = \exp(\mu + \sigma^2/2)$ . Inserting our values for  $\mu$  and  $\sigma$  gives  $E[x] = \$25,554$ . The 1987 Statistical Abstract of the United States gives the mean of household incomes across all groups for the United States as about \$25,000. So, the estimate based on surprisingly little information would have been relatively good. These meager data did, indeed, tell us something about the average American.

**19.2.3 THE TRUNCATED REGRESSION MODEL**

In the model of the earlier examples, we now assume that

$$\mu_i = \mathbf{x}'_i \boldsymbol{\beta}$$

is the deterministic part of the classical regression model. Then

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i,$$

**838 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

where

$$\varepsilon_i | \mathbf{x}_i \sim N[0, \sigma^2],$$

so that

$$y_i | \mathbf{x}_i \sim N[\mathbf{x}'_i \boldsymbol{\beta}, \sigma^2]. \quad (19-5)$$

We are interested in the distribution of  $y_i$  given that  $y_i$  is greater than the truncation point  $a$ . This is the result described in Theorem 19.2. It follows that

$$E[y_i | y_i > a] = \mathbf{x}'_i \boldsymbol{\beta} + \sigma \frac{\phi[(a - \mathbf{x}'_i \boldsymbol{\beta})/\sigma]}{1 - \Phi[(a - \mathbf{x}'_i \boldsymbol{\beta})/\sigma]}. \quad (19-6)$$

The conditional mean is therefore a nonlinear function of  $a$ ,  $\sigma$ ,  $\mathbf{x}$ , and  $\boldsymbol{\beta}$ .

The partial effects in this model *in the subpopulation* can be obtained by writing

$$E[y_i | y_i > a] = \mathbf{x}'_i \boldsymbol{\beta} + \sigma \lambda(\alpha_i), \quad (19-7)$$

where now  $\alpha_i = (a - \mathbf{x}'_i \boldsymbol{\beta})/\sigma$ . For convenience, let  $\lambda_i = \lambda(\alpha_i)$  and  $\delta_i = \delta(\alpha_i)$ . Then

$$\begin{aligned} \frac{\partial E[y_i | y_i > a]}{\partial \mathbf{x}_i} &= \boldsymbol{\beta} + \sigma(d\lambda_i/d\alpha_i) \frac{\partial \alpha_i}{\partial \mathbf{x}_i} \\ &= \boldsymbol{\beta} + \sigma(\lambda_i^2 - \alpha_i \lambda_i)(-\boldsymbol{\beta}/\sigma) \\ &= \boldsymbol{\beta}(1 - \lambda_i^2 + \alpha_i \lambda_i) \\ &= \boldsymbol{\beta}(1 - \delta_i). \end{aligned} \quad (19-8)$$

Note the appearance of the scale factor  $1 - \delta_i$  from the truncated variance. Because  $(1 - \delta_i)$  is between zero and one, we conclude that for every element of  $\mathbf{x}_i$ , the marginal effect is less than the corresponding coefficient. There is a similar **attenuation** of the variance. In the subpopulation  $y_i > a$ , the regression variance is not  $\sigma^2$  but

$$\text{Var}[y_i | y_i > a] = \sigma^2(1 - \delta_i). \quad (19-9)$$

Whether the partial effect in (19-7) or the coefficient  $\boldsymbol{\beta}$  itself is of interest depends on the intended inferences of the study. If the analysis is to be confined to the subpopulation, then (19-7) is of interest. If the study is intended to extend to the entire population, however, then it is the coefficients  $\boldsymbol{\beta}$  that are actually of interest.

One's first inclination might be to use ordinary least squares to estimate the parameters of this regression model. For the subpopulation from which the data are drawn, we could write (19-6) in the form

$$y_i | y_i > a = E[y_i | y_i > a] + u_i = \mathbf{x}'_i \boldsymbol{\beta} + \sigma \lambda_i + u_i, \quad (19-10)$$

where  $u_i$  is  $y_i$  minus its conditional expectation. By construction,  $u_i$  has a zero mean, but it is heteroscedastic:

$$\text{Var}[u_i] = \sigma^2(1 - \lambda_i^2 + \lambda_i \alpha_i) = \sigma^2(1 - \delta_i),$$

which is a function of  $\mathbf{x}_i$ . If we estimate (19-10) by ordinary least squares regression of  $\mathbf{y}$  on  $\mathbf{X}$ , then we have omitted a variable, the nonlinear term  $\lambda_i$ . All the biases that arise because of an omitted variable can be expected.<sup>5</sup>

Without some knowledge of the distribution of  $\mathbf{x}$ , it is not possible to determine how serious the bias is likely to be. A result obtained by Chung and Goldberger (1984) is broadly suggestive. If  $E[\mathbf{x} | \mathbf{y}]$  in the full population is a linear function of  $\mathbf{y}$ , then  $\text{plim } \mathbf{b} = \boldsymbol{\beta}\tau$  for some proportionality constant  $\tau$ . This result is consistent with the widely observed (albeit rather rough) proportionality relationship between least squares estimates of this model and maximum likelihood estimates.<sup>6</sup> The proportionality result appears to be quite general. In applications, it is usually found that, compared with consistent maximum likelihood estimates, the OLS estimates are biased toward zero. (See Example 19.5.)

#### 19.2.4 THE STOCHASTIC FRONTIER MODEL

A lengthy literature commencing with theoretical work by Knight (1933), Debreu (1951), and Farrell (1957) and the pioneering empirical study by Aigner, Lovell, and Schmidt (ALS, 1977) has been directed at models of production that specifically account for the textbook proposition that a production function is a theoretical ideal.<sup>7</sup> If  $y = f(\mathbf{x})$  defines a production relationship between inputs,  $\mathbf{x}$ , and an output,  $y$ , then for any given  $\mathbf{x}$ , the observed value of  $y$  must be less than or equal to  $f(\mathbf{x})$ . The implication for an empirical regression model is that in a formulation such as  $y = h(\mathbf{x}, \boldsymbol{\beta}) + u$ ,  $u$  must be negative. Because the theoretical production function is an ideal—the frontier of efficient production—any nonzero disturbance must be interpreted as the result of inefficiency. A strictly orthodox interpretation embedded in a Cobb–Douglas production model might produce an empirical frontier production model such as

$$\ln y = \beta_1 + \sum_k \beta_k \ln x_k - u, \quad u \geq 0.$$

The gamma model described in Example 4.7 was an application. One-sided disturbances such as this one present a particularly difficult estimation problem. The primary theoretical problem is that any measurement error in  $\ln y$  must be embedded in the disturbance. The practical problem is that the entire estimated function becomes a slave to any single errantly measured data point.

Aigner, Lovell, and Schmidt proposed instead a formulation within which observed deviations from the production function could arise from two sources: (1) productive inefficiency, as we have defined it earlier and that would necessarily be negative, and (2) idiosyncratic effects that are specific to the firm and that could enter the model with either sign. The end result was what they labeled the **stochastic frontier**:

$$\begin{aligned} \ln y &= \beta_1 + \sum_k \beta_k \ln x_k - u + v, \quad u \geq 0, \quad v \sim N[0, \sigma_v^2]. \\ &= \beta_1 + \sum_k \beta_k \ln x_k + \varepsilon. \end{aligned}$$

<sup>5</sup>See Heckman (1979) who formulates this as a “specification error.”

<sup>6</sup>See the appendix in Hausman and Wise (1977) and Greene (1983) as well.

<sup>7</sup>A survey by Greene (2008a) appears in Fried, Lovell, and Schmidt (2008). Kumbhakar and Lovell (2000) is a comprehensive reference on the subject.

**840 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**

The frontier for any particular firm is  $h(\mathbf{x}, \boldsymbol{\beta}) + v$ , hence the name *stochastic frontier*. The inefficiency term is  $u$ , a random variable of particular interest in this setting. Because the data are in log terms,  $u$  is a measure of the percentage by which the particular observation fails to achieve the frontier, ideal production rate.

To complete the specification, they suggested two possible distributions for the inefficiency term: the absolute value of a normally distributed variable, which has the truncated at zero distribution shown in Figure 19.1, and an exponentially distributed variable. The density functions for these two compound variables are given by Aigner, Lovell, and Schmidt; let  $\varepsilon = v - u$ ,  $\lambda = \sigma_u/\sigma_v$ ,  $\sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$ , and  $\Phi(z) =$  the probability to the left of  $z$  in the standard normal distribution (see Section B.4.1). For the “half-normal” model,

$$\ln h(\varepsilon_i | \boldsymbol{\beta}, \lambda, \sigma) = \left[ -\ln \sigma + \left( \frac{1}{2} \right) \ln \frac{2}{\pi} - \frac{1}{2} \left( \frac{\varepsilon_i}{\sigma} \right)^2 + \ln \Phi \left( \frac{-\varepsilon_i \lambda}{\sigma} \right) \right],$$

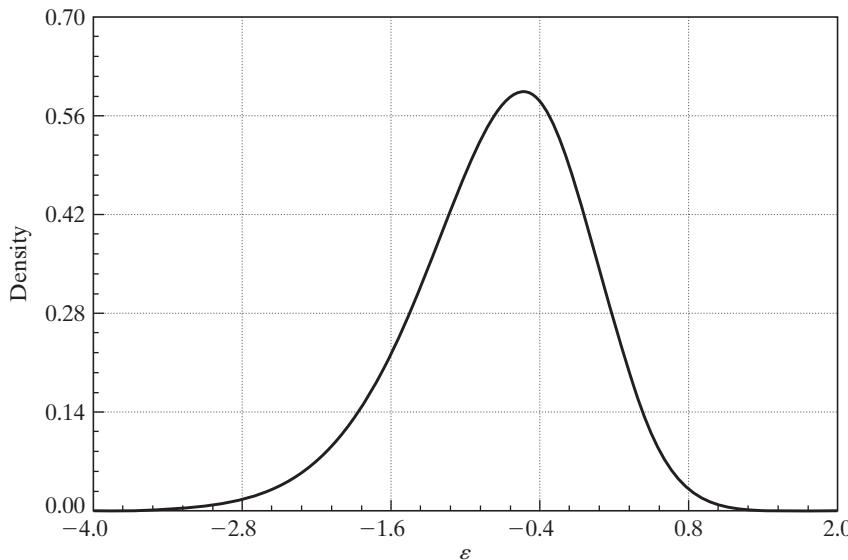
whereas for the exponential model

$$\ln h(\varepsilon_i | \boldsymbol{\beta}, \theta, \sigma_v) = \left[ \ln \theta + \frac{1}{2} \theta^2 \sigma_v^2 + \theta \varepsilon_i + \ln \Phi \left( -\frac{\varepsilon_i}{\sigma_v} - \theta \sigma_v \right) \right].$$

Both these distributions are asymmetric. We thus have a regression model with a nonnormal distribution specified for the disturbance. The disturbance,  $\varepsilon$ , has a nonzero mean as well;  $E[\varepsilon] = -\sigma_u(2/\pi)^{1/2}$  for the half-normal model and  $-1/\theta$  for the exponential model. Figure 19.2 illustrates the density for the half-normal model with  $\sigma = 1$  and  $\lambda = 2$ . By writing  $\beta_0 = \beta_1 + E[\varepsilon]$  and  $\varepsilon^* = \varepsilon - E[\varepsilon]$ , we obtain a more conventional formulation

$$\ln y = \beta_0 + \sum_k \beta_k \ln x_k + \varepsilon^*,$$

**FIGURE 19.2** Density for the Disturbance in the Stochastic Frontier Model.



CHAPTER 19 ♦ Limited Dependent Variables **841**

which does have a disturbance with a zero mean but an asymmetric, nonnormal distribution. The asymmetry of the distribution of  $\varepsilon^*$  does not negate our basic results for least squares in this classical regression model. This model satisfies the assumptions of the Gauss–Markov theorem, so least squares is unbiased and consistent (save for the constant term) and efficient among linear unbiased estimators. In this model, however, the maximum likelihood estimator is not linear, and it is more efficient than least squares.

The log-likelihood function for the half normal model is given in ALS (1977):

$$\ln L = -n \ln \sigma + \frac{n}{2} \ln \frac{2}{\pi} - \frac{1}{2} \sum_{i=1}^n \left( \frac{\varepsilon_i}{\sigma} \right)^2 + \sum_{i=1}^n \ln \Phi \left( \frac{-\varepsilon_i \lambda}{\sigma} \right). \quad (19-11)$$

Maximization programs for this model are built into modern software packages such as *Stata*, *NLOGIT*, and *TSP*. The log-likelihood is simple enough that it can also be readily adapted to the generic optimization routines in, for example, *MatLab* or *Gauss*. Some treatments in the literature use the parameterization employed by Battese and Coelli (1992) and Coelli (1996),  $\gamma = \sigma_u^2/\sigma^2$ . This is a one-to-one transformation of  $\lambda$ ;  $\lambda = (\gamma/(1-\gamma))^{1/2}$ , so which parameterization is employed is a matter of convenience; the empirical results will be the same. The log-likelihood function for the exponential model can be built up from the density given earlier. For the half-normal model, we would also rely on the invariance of maximum likelihood estimators to recover estimates of the structural variance parameters,  $\sigma_v^2 = \sigma^2/(1+\lambda^2)$  and  $\sigma_u^2 = \sigma^2\lambda^2/(1+\lambda^2)$ .<sup>8</sup> (Note, the variance of the truncated variable,  $u_i$ , is not  $\sigma_u^2$ ; using (19-2), it reduces to  $(1-2/\pi)\sigma_u^2$ .) In addition, a structural parameter of interest is the proportion of the total variance of  $\varepsilon$  that is due to the inefficiency term. For the half-normal model,  $\text{Var}[\varepsilon] = \text{Var}[u] + \text{Var}[v] = (1-2/\pi)\sigma_u^2 + \sigma_v^2$  whereas for the exponential model, the counterpart is  $1/\theta^2 + \sigma_v^2$ .

Modeling in the stochastic frontier setting is rather unlike what we are accustomed to up to this point, in that the disturbance, specifically  $u_i$ , not the model parameters, is the central focus of the analysis. The reason is that in this context, the disturbance,  $u_i$ , rather than being the catchall for the unknown and unknowable factors omitted from the equation, has a particular interpretation—it is the firm-specific inefficiency. Ideally, we would like to estimate  $u_i$  for each firm in the sample to compare them on the basis of their productive efficiency. Unfortunately, the data do not permit a direct estimate, because with estimates of  $\beta$  in hand, we are only able to compute a direct estimate of  $\varepsilon_i = y_i - \mathbf{x}'_i \beta$ . Jondrow et al. (1982), however, have derived a useful approximation that is now the standard measure in these settings,

$$E[u_i | \varepsilon_i] = \frac{\sigma \lambda}{1 + \lambda^2} \left[ \frac{\phi(z_i)}{1 - \Phi(z_i)} - z_i \right], \quad z_i = \frac{\varepsilon_i \lambda}{\sigma}$$

for the half-normal model, and

$$E[u_i | \varepsilon_i] = z_i + \sigma_v \frac{\phi(z_i/\sigma_v)}{\Phi(z_i/\sigma_v)}, \quad z_i = -(\varepsilon_i + \theta \sigma_v^2)$$

for the exponential model. These values can be computed using the maximum likelihood estimates of the structural parameters in the model. In some cases in which researchers

<sup>8</sup>A vexing problem for estimation of the model is that if the ordinary least squares residuals are skewed in the positive (wrong) direction (See Figure 19.2), OLS with  $\hat{\lambda} = 0$  will be the MLE. OLS residuals with a positive skew are apparently inconsistent with a model in which, in theory, they should have a negative skew. [See Waldman (1982) for theoretical development of this result.]

## 842 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

are interested in discovering best practice [e.g., WHO (2000), Tandon et al. (2000)], the estimated values are sorted and the ranks of the individuals in the sample become of interest.

Research in this area since the methodological developments beginning in the 1930s and the building of the empirical foundations in 1977 and 1982 has proceeded in several directions. Most theoretical treatments of “inefficiency” as envisioned here attribute it to aspects of management of the firm. It remains to establish a firm theoretical connection between the theory of firm behavior and the stochastic frontier model as a device for measurement of inefficiency.

In the context of the model, many studies have developed alternative, more flexible functional forms that (it is hoped) can provide a more realistic model for inefficiency. Two that are relevant in this chapter are Stevenson’s (1980) truncated normal model and the normal-gamma frontier. One intuitively appealing form of the truncated normal model is

$$U_i \sim N[\mu + \mathbf{z}'_i \boldsymbol{\alpha}, \sigma_u^2], \\ u_i = |U_i|.$$

The original normal-half-normal model results if  $\mu$  equals zero and  $\boldsymbol{\alpha}$  equals zero. This is a device by which the environmental variables noted in the next paragraph can enter the model of inefficiency. A truncated normal model is presented in Example 19.3. The half-normal, truncated normal, and exponential models all take the form of distribution shown in Figure 19.1. The gamma model,

$$f(u) = [\theta^P / \Gamma(P)] \exp(-\theta u) u^{P-1},$$

is a flexible model that presents the advantage that the distribution of inefficiency can move away from zero. If  $P$  is greater than one, then the density at  $u = 0$  equals zero and the entire distribution moves away from the origin. The implication is that the distribution of inefficiency among firms can move away from zero. The gamma model is estimated by simulation methods—either Bayesian MCMC [Huang (2003) and Tsionas (2002)] or maximum simulated likelihood [Greene (2003)]. Many other functional forms have been proposed. [See Greene (2008) for a survey.]

There are usually elements in the environment in which the firm operates that impact the firm’s output and/or costs but are not, themselves, outputs, inputs, or input prices. In example 19.3, the costs of the Swiss railroads are affected by three variables; track width, long tunnels, and curvature. It is not yet specified how such factors should be incorporated into the model; four candidates are in the mean and variance of  $u_i$ , directly in the function, or in the variance of  $v_i$ . [See Hadri, Guermat, and Whittaker (2003) and Kumbhakar (1997c).] All of these can be found in the received studies. This aspect of the model was prominent in the discussion of the famous World Health Organization efficiency study of world health systems [WHO (2000), Tandon, Murray, Lauer, and Evans (2000), and Greene (2004)]. In Example 19.3, we have placed the environmental factors in the mean of the inefficiency distribution. This produces a rather extreme set of results for the JLMS estimates of inefficiency—many railroads are estimated to be extremely inefficient. An alternative formulation would be a “heteroscedastic” model in which  $\sigma_{u,i} = \sigma_u \exp(z'_i \boldsymbol{\delta})$  or  $\sigma_{v,i} = \sigma_v \exp(\mathbf{z}'_i \boldsymbol{\eta})$ , or both. We can see from the JLMS formula that the term heteroscedastic is actually a bit misleading, since both

CHAPTER 19 ♦ Limited Dependent Variables **843**

standard deviations enter (now)  $\lambda_i$ , which is, in turn, a crucial parameter in the mean of inefficiency.

How should inefficiency be modeled in panel data, such as in our example? It might be tempting to treat it as a time-invariant “effect” [as in Schmidt and Sickles (1984) and Pitt and Lee (1984) in two pioneering papers]. Greene (2004) argued that a preferable approach would be to allow inefficiency to vary freely over time in a panel, and to the extent that there is a common time-invariant effect in the model, that should be treated as unobserved heterogeneity, not inefficiency. A string of studies, including Battese and Coelli (1992, 1995), Cuesta (2000), Kumbhakar (1997a) Kumbhakar and Orea (2004), and many others have proposed hybrid forms that treat the core random part of inefficiency as a time-invariant firm-specific effect that is modified over time by a deterministic, possibly firm-specific, function. The Battese-Coelli form,

$$u_{it} = \exp[-\eta(t - T)]|U_i| \text{ where } U_i \sim N[0, \sigma_u^2],$$

has been used in a number of applications. Cuesta (2000) suggests allowing  $\eta$  to vary across firms, producing a model that bears some relationship to a fixed-effects specification. This thread of the literature is one of the most active ongoing pursuits.

Is it reasonable to use a possibly restrictive parametric approach to modeling inefficiency? Sickles (2005) and Kumbhakar, Simar, Park, and Tsionas (2007) are among numerous studies that have explored less parametric approaches to efficiency analysis. Proponents of **data envelopment analysis** [see, e.g., Simar and Wilson (2000, 2007)] have developed methods that impose absolutely no parametric structure on the production function. Among the costs of this high degree of flexibility is a difficulty to include environmental effects anywhere in the analysis, and the uncomfortable implication that any unmeasured heterogeneity of any sort is necessarily included in the measure of inefficiency. That is, data envelopment analysis returns to the deterministic frontier approach where this section began.

#### **Example 19.3 Stochastic Cost Frontier for Swiss Railroads**

Farsi, Filippini, and Greene (2005) analyzed the cost efficiency of Swiss railroads. In order to use the stochastic frontier approach to analyze costs of production, rather than production, we rely on the fundamental duality of production and cost [see Samuelson (1938), Shephard (1953), and Kumbhakar and Lovell (2000)]. An appropriate cost frontier model for a firm that produces more than one output—the Swiss railroads carry both freight and passengers—will appear as the following:

$$\ln(C/P_K) = \alpha + \sum_{k=1}^{K-1} \beta_k \ln(P_k/P_K) + \sum_{m=1}^M \gamma_m \ln Q_m + v + u.$$

The requirement that the cost function be homogeneous of degree one in the input prices has been imposed by normalizing total cost,  $C$ , and the first  $K - 1$  prices by the  $K$ th input price. In this application, the three factors are labor, capital, and electricity—the third is used as the numeraire in the cost function. Notice that the inefficiency term,  $u$ , enters the cost function positively; actual cost is above the frontier cost. [The MLE is modified simply by replacing  $\varepsilon_i$  with  $-\varepsilon_i$  in (19-11).] In analyzing costs of production, we recognize that there is an additional source of inefficiency that is absent when we analyze production. On the production side, inefficiency measures the difference between output and frontier output, which arises because of technical inefficiency. By construction, if output fails to reach the efficient level for the given input usage, then costs must be higher than frontier costs. However, costs can be excessive even if the firm is technically efficient if it is “allocatively inefficient.” That is, the firm can be technically efficient while not using inputs in the cost minimizing mix (equating

## 844 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics

the ratio of marginal products to the input price ratios). It follows that on the cost side, “ $u$ ” can contain both elements of inefficiency while on the production side, we would expect to measure only technical inefficiency. [See Kumbhakar (1997b).]

The data for this study are an unbalanced panel of 50 railroads with  $T_i$  ranging from 1 to 13. (Thirty-seven of the firms are observed 13 times, 8 are observed 12 times, and the remaining 5 are observed 10, 7, 7, 3, and 1 times.) The variables we will use here are

- $CT$ : Total costs adjusted for inflation (1,000 Swiss franc)
- $QP$ : Total passenger-output in passenger-kilometers
- $QF$ : Total goods-output in ton-kilometers
- $PL$ : Labor price adjusted for inflation (in Swiss Francs per person per year)
- $PK$ : Capital price with capital stock proxied by total number of seats
- $PE$ : Price of electricity (Swiss franc per kWh)

Logs of costs and prices ( $\ln CT$ ,  $\ln PK$ ,  $\ln PL$ ) are normalized by  $PE$ . We will also use these environmental variables:

- $NARROW\_T$ : Dummy for the networks with narrow track (1 m wide) The usual width is 1.435m.
- $TUNNEL$ : Dummy for networks that have tunnels with an average length of more than 300 meters.
- $VIRAGE$ : Dummy for the networks whose minimum radius of curvature is 100 meters or less.

The full data set is given in Appendix Table F19.1. Several other variables not used here are presented in the appendix table. In what follows, we will ignore the panel data aspect of the data set. This would be a focal point of a more extensive study.

There have been dozens of models proposed for the inefficiency component of the stochastic frontier model. Table 19.1 presents several different forms. The basic half-normal model is given in the first column. The estimated cost function parameters across the different

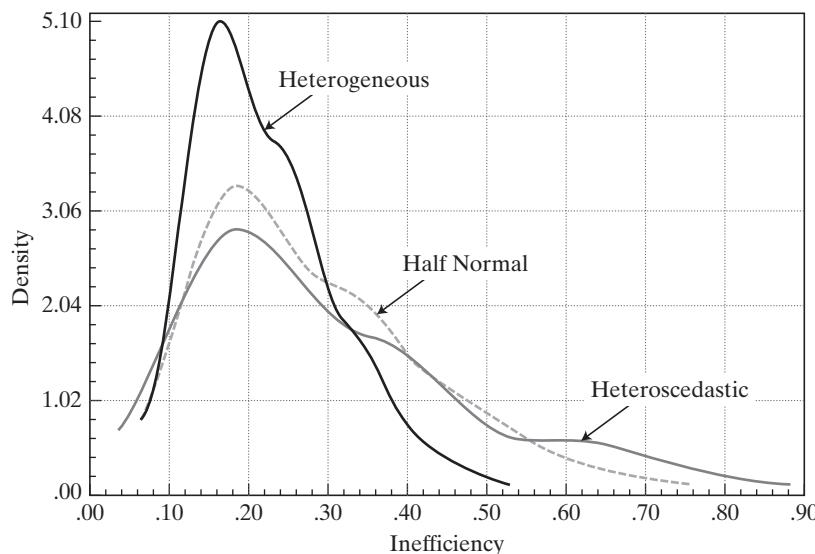
**TABLE 19.1** Estimated Stochastic Frontier Cost Functions<sup>a</sup>

<i>Variable</i>	<i>Model</i>					
	<i>Half Normal</i>	<i>Truncated Normal</i>	<i>Exponential</i>	<i>Gamma</i>	<i>Heterosced</i>	<i>Heterogen</i>
Constant	-10.0799	-9.80624	-10.1838	-10.1944	-9.82189	-10.2891
$\ln QP$	0.64220	0.62573	0.64403	0.64401	0.61976	0.63576
$\ln QF$	0.06904	0.07708	0.06803	0.06810	0.07970	0.07526
$\ln PK$	0.26005	0.26625	0.25883	0.25886	0.25464	0.25893
$\ln PL$	0.53845	0.50474	0.56138	0.56047	0.53953	0.56036
Constant	0.44116				-2.48218 <sup>b</sup>	
Narrow		0.29881			2.16264 <sup>b</sup>	0.14355
Virage		-0.20738			-1.52964 <sup>b</sup>	-0.10483
Tunnel		0.01118			0.35748 <sup>b</sup>	-0.01914
$\sigma$	0.44240	0.38547	(0.34325)	(0.34288)	0.45392 <sup>c</sup>	0.40597
$\lambda$	1.27944	2.35055				0.91763
$P$			1.0000	1.22920		
$\theta$			13.2922	12.6915		
$\sigma_u$	(0.34857)	(0.35471)	(0.07523)	(0.09685)	0.37480 <sup>c</sup>	0.27448
$\sigma_v$	(0.27244)	(0.15090)	0.33490	0.33197	0.25606	0.29912
Mean $E[u \varepsilon]$	0.27908	0.52858	0.075232	0.096616	0.29499	0.21926
$\ln L$	-210.495	-200.67	-211.42	-211.091	-201.731	-208.349

<sup>a</sup>Estimates in parentheses are derived from other MLEs.

<sup>b</sup>Estimates used in computation of  $\sigma_u$ .

<sup>c</sup>Obtained by averaging  $\lambda = \sigma_{u,i}/\sigma_v$  over observations.



**FIGURE 19.3** Kernel Density Estimator for JLMS Estimates.

forms are broadly similar, as might be expected as  $(\alpha, \beta)$  are consistently estimated in all cases. There are fairly pronounced differences in the implications for the components of  $\varepsilon$ , however.

There is an ambiguity in the model as to whether modifications to the distribution of  $u_i$  will affect the mean of the distribution, the variance, or both. The following results suggest that it is both for these data. The gamma and exponential models appear to remove most of the inefficiency from the data. Note that the estimates of  $\sigma_u$  are considerably smaller under these specifications, and  $\sigma_v$  is correspondingly larger. The second to last row shows the sample averages of the Jondrow estimators—this estimates  $E_\varepsilon E[u|\varepsilon] = E[u]$ . There is substantial difference across the specifications.

The estimates in the rightmost two columns illustrate two different placements of the measured heterogeneity: in the variance of  $u_i$  and directly in the cost function. The log-likelihood function appears to favor the first of these. However, the models are not nested and involve the same number of parameters. We used the Vuong test (see Section 14.6.6), instead and obtained a value of  $-2.65$  in favor of the heteroscedasticity model. Figure 19.3 describes the values of  $E[u_i|\varepsilon_i]$  estimated for the sample observations for the half-normal, heteroscedastic and heterogeneous models. The smaller estimate of  $\sigma_u$  for the third of these is evident in the figure, which suggests a somewhat tighter concentration of values than the other two.

### 19.3 CENSORED DATA

A very common problem in microeconomic data is **censoring** of the dependent variable. When the dependent variable is censored, values in a certain range are all transformed to (or reported as) a single value. Some examples that have appeared in the empirical literature are as follows:<sup>9</sup>

1. Household purchases of durable goods [Tobin (1958)]
2. The number of extramarital affairs [Fair (1977, 1978)]

<sup>9</sup>More extensive listings may be found in Amemiya (1984) and Maddala (1983).

## 846 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics

3. The number of hours worked by a woman in the labor force [Quester and Greene (1982)]
4. The number of arrests after release from prison [Witte (1980)]
5. Household expenditure on various commodity groups [Jarque (1987)]
6. Vacation expenditures [Melenberg and van Soest (1996)]

Each of these studies analyzes a dependent variable that is zero for a significant fraction of the observations. Conventional regression methods fail to account for the qualitative difference between *limit* (zero) observations and *nonlimit* (continuous) observations.

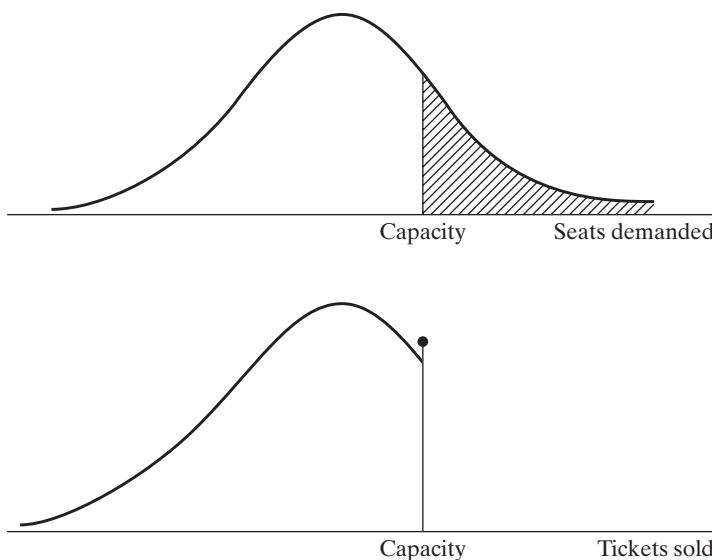
### 19.3.1 THE CENSORED NORMAL DISTRIBUTION

The relevant distribution theory for a **censored variable** is similar to that for a truncated one. Once again, we begin with the normal distribution, as much of the received work has been based on an assumption of normality. We also assume that the censoring point is zero, although this is only a convenient normalization. In a truncated distribution, only the part of distribution above  $y = 0$  is relevant to our computations. To make the distribution integrate to one, we scale it up by the probability that an observation in the untruncated population falls in the range that interests us. When data are censored, the distribution *that applies to the sample data* is a mixture of discrete and continuous distributions. Figure 19.4 illustrates the effects.

To analyze this distribution, we define a new random variable  $y$  transformed from the original one,  $y^*$ , by

$$\begin{aligned} y &= 0 && \text{if } y^* \leq 0, \\ y &= y^* && \text{if } y^* > 0. \end{aligned}$$

**FIGURE 19.4** Partially Censored Distribution.



CHAPTER 19 ♦ Limited Dependent Variables **847**

The distribution that applies if  $y^* \sim N[\mu, \sigma^2]$  is  $\text{Prob}(y = 0) = \text{Prob}(y^* \leq 0) = \Phi(-\mu/\sigma) = 1 - \Phi(\mu/\sigma)$ , and if  $y^* > 0$ , then  $y$  has the density of  $y^*$ .

This distribution is a mixture of discrete and continuous parts. The total probability is one, as required, but instead of scaling the second part, we simply assign the full probability in the censored region to the censoring point, in this case, zero.

**THEOREM 19.3 Moments of the Censored Normal Variable**

If  $y^* \sim N[\mu, \sigma^2]$  and  $y = a$  if  $y^* \leq a$  or else  $y = y^*$ , then

$$E[y] = \Phi a + (1 - \Phi)(\mu + \sigma\lambda),$$

and

$$\text{Var}[y] = \sigma^2(1 - \Phi)[(1 - \delta) + (\alpha - \lambda)^2\Phi],$$

where

$$\Phi[(a - \mu)/\sigma] = \Phi(\alpha) = \text{Prob}(y^* \leq a) = \Phi, \quad \lambda = \phi/(1 - \Phi),$$

and

$$\delta = \lambda^2 - \lambda\alpha.$$

**Proof:** For the mean,

$$\begin{aligned} E[y] &= \text{Prob}(y = a) \times E[y | y = a] + \text{Prob}(y > a) \times E[y | y > a] \\ &= \text{Prob}(y^* \leq a) \times a + \text{Prob}(y^* > a) \times E[y^* | y^* > a] \\ &= \Phi a + (1 - \Phi)(\mu + \sigma\lambda) \end{aligned}$$

using Theorem 19.2. For the variance, we use a counterpart to the decomposition in (B-69), that is,  $\text{Var}[y] = E[\text{conditional variance}] + \text{Var}[\text{conditional mean}]$ , and Theorem 19.2.

For the special case of  $a = 0$ , the mean simplifies to

$$E[y | a = 0] = \Phi(\mu/\sigma)(\mu + \sigma\lambda), \quad \text{where } \lambda = \frac{\phi(\mu/\sigma)}{\Phi(\mu/\sigma)}.$$

For censoring of the upper part of the distribution instead of the lower, it is only necessary to reverse the role of  $\Phi$  and  $1 - \Phi$  and redefine  $\lambda$  as in Theorem 19.2.

**Example 19.4 Censored Random Variable**

We are interested in the number of tickets *demanded* for events at a certain arena. Our only measure is the number actually *sold*. Whenever an event sells out, however, we know that the actual number demanded is larger than the number sold. The number of tickets demanded is censored when it is transformed to obtain the number sold. Suppose that the arena in question has 20,000 seats and, in a recent season, sold out 25 percent of the time. If the average attendance, including sellouts, was 18,000, then what are the mean and standard deviation of the demand for seats? According to Theorem 19.3, the 18,000 is an estimate of

$$E[\text{sales}] = 20,000(1 - \Phi) + [\mu + \sigma\lambda]\Phi.$$

Because this is censoring from above, rather than below,  $\lambda = -\phi(\alpha)/\Phi(\alpha)$ . The argument of  $\Phi$ ,  $\phi$ , and  $\lambda$  is  $\alpha = (20,000 - \mu)/\sigma$ . If 25 percent of the events are sellouts, then  $\Phi = 0.75$ . Inverting the standard normal at 0.75 gives  $\alpha = 0.675$ . In addition, if  $\alpha = 0.675$ ,

## 848 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

then  $-\phi(0.675)/0.75 = \lambda = -0.424$ . This result provides two equations in  $\mu$  and  $\sigma$ , (a)  $18,000 = 0.25(20,000) + 0.75(\mu - 0.424\sigma)$  and (b)  $0.675\sigma = 20,000 - \mu$ . The solutions are  $\sigma = 2426$  and  $\mu = 18,362$ .

For comparison, suppose that we were told that the mean of 18,000 applies only to the events that were *not* sold out and that, on average, the arena sells out 25 percent of the time. Now our estimates would be obtained from the equations (a)  $18,000 = \mu - 0.424\sigma$  and (b)  $0.675\sigma = 20,000 - \mu$ . The solutions are  $\sigma = 1820$  and  $\mu = 18,772$ .

### 19.3.2 THE CENSORED REGRESSION (TOBIT) MODEL

The regression model based on the preceding discussion is referred to as the **censored regression model** or the **tobit model** [in reference to Tobin (1958), where the model was first proposed]. The regression is obtained by making the mean in the preceding correspond to a classical regression model. The general formulation is usually given in terms of an index function,

$$\begin{aligned} y_i^* &= \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i, \\ y_i &= 0 \quad \text{if } y_i^* \leq 0, \\ y_i &= y_i^* \quad \text{if } y_i^* > 0. \end{aligned}$$

There are potentially three conditional mean functions to consider, depending on the purpose of the study. For the index variable, sometimes called the *latent variable*,  $E[y_i^* | \mathbf{x}_i]$  is  $\mathbf{x}_i' \boldsymbol{\beta}$ . If the data are always censored, however, then this result will usually not be useful. Consistent with Theorem 19.3, for an observation randomly drawn from the population, which may or may not be censored,

$$E[y_i | \mathbf{x}_i] = \Phi\left(\frac{\mathbf{x}_i' \boldsymbol{\beta}}{\sigma}\right)(\mathbf{x}_i' \boldsymbol{\beta} + \sigma \lambda_i),$$

where

$$\lambda_i = \frac{\phi[(0 - \mathbf{x}_i' \boldsymbol{\beta})/\sigma]}{1 - \Phi[(0 - \mathbf{x}_i' \boldsymbol{\beta})/\sigma]} = \frac{\phi(\mathbf{x}_i' \boldsymbol{\beta}/\sigma)}{\Phi(\mathbf{x}_i' \boldsymbol{\beta}/\sigma)}. \quad (19-12)$$

Finally, if we intend to confine our attention to uncensored observations, then the results for the truncated regression model apply. The limit observations should not be discarded, however, because the truncated regression model is no more amenable to least squares than the censored data model. It is an unresolved question which of these functions should be used for computing predicted values from this model. Intuition suggests that  $E[y_i | \mathbf{x}_i]$  is correct, but authors differ on this point. For the setting in Example 19.4, for predicting the number of tickets sold, say, to plan for an upcoming event, the censored mean is obviously the relevant quantity. On the other hand, if the objective is to study the need for a new facility, then the mean of the latent variable  $y_i^*$  would be more interesting.

There are differences in the partial effects as well. For the index variable,

$$\frac{\partial E[y_i^* | \mathbf{x}_i]}{\partial \mathbf{x}_i} = \boldsymbol{\beta}.$$

But this result is not what will usually be of interest, because  $y_i^*$  is unobserved. For the observed data,  $y_i$ , the following general result will be useful:<sup>10</sup>

<sup>10</sup>See Greene (1999) for the general result and Rosett and Nelson (1975) and Nakamura and Nakamura (1983) for applications based on the normal distribution.

**THEOREM 19.4 Partial Effects in the Censored Regression Model**

In the censored regression model with latent regression  $y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$  and observed dependent variable,  $y = a$  if  $y^* \leq a$ ,  $y = b$  if  $y^* \geq b$ , and  $y = y^*$  otherwise, where  $a$  and  $b$  are constants, let  $f(\varepsilon)$  and  $F(\varepsilon)$  denote the density and cdf of  $\varepsilon$ . Assume that  $\varepsilon$  is a continuous random variable with mean 0 and variance  $\sigma^2$ , and  $f(\varepsilon | \mathbf{x}) = f(\varepsilon)$ . Then

$$\frac{\partial E[y | \mathbf{x}]}{\partial \mathbf{x}} = \boldsymbol{\beta} \times \text{Prob}[a < y^* < b].$$

 **Proof:** By definition,

$$\begin{aligned} E[y | \mathbf{x}] &= a \text{Prob}[y^* \leq a | \mathbf{x}] + b \text{Prob}[y^* \geq b | \mathbf{x}] \\ &\quad + \text{Prob}[a < y^* < b | \mathbf{x}] E[y^* | a < y^* < b | \mathbf{x}]. \end{aligned}$$

Let  $\alpha_j = (j - \mathbf{x}'\boldsymbol{\beta})/\sigma$ ,  $F_j = F(\alpha_j)$ ,  $f_j = f(\alpha_j)$ , and  $j = a, b$ . Then

$$E[y | \mathbf{x}] = aF_a + b(1 - F_b) + (F_b - F_a)E[y^* | a < y^* < b, \mathbf{x}].$$

Because  $y^* = \mathbf{x}'\boldsymbol{\beta} + \sigma[(y^* - \boldsymbol{\beta}'\mathbf{x})/\sigma]$ , the conditional mean may be written

$$\begin{aligned} E[y^* | a < y^* < b, \mathbf{x}] &= \mathbf{x}'\boldsymbol{\beta} + \sigma E\left[\frac{y^* - \mathbf{x}'\boldsymbol{\beta}}{\sigma} \middle| \frac{a - \mathbf{x}'\boldsymbol{\beta}}{\sigma} < \frac{y^* - \mathbf{x}'\boldsymbol{\beta}}{\sigma} < \frac{b - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right] \\ &= \mathbf{x}'\boldsymbol{\beta} + \sigma \int_{\alpha_a}^{\alpha_b} \frac{(\varepsilon/\sigma)f(\varepsilon/\sigma)}{F_b - F_a} d\left(\frac{\varepsilon}{\sigma}\right). \end{aligned}$$

Collecting terms, we have

$$E[y | \mathbf{x}] = aF_a + b(1 - F_b) + (F_b - F_a)\boldsymbol{\beta}'\mathbf{x} + \sigma \int_{\alpha_a}^{\alpha_b} \left(\frac{\varepsilon}{\sigma}\right) f\left(\frac{\varepsilon}{\sigma}\right) d\left(\frac{\varepsilon}{\sigma}\right).$$

Now, differentiate with respect to  $\mathbf{x}$ . The only complication is the last term, for which the differentiation is with respect to the limits of integration. We use Leibnitz's theorem and use the assumption that  $f(\varepsilon)$  does not involve  $\mathbf{x}$ . Thus,

$$\begin{aligned} \frac{\partial E[y | \mathbf{x}]}{\partial \mathbf{x}} &= \left(\frac{-\boldsymbol{\beta}}{\sigma}\right) a f_a - \left(\frac{-\boldsymbol{\beta}}{\sigma}\right) b f_b + (F_b - F_a)\boldsymbol{\beta} + (\mathbf{x}'\boldsymbol{\beta})(f_b - f_a)\left(\frac{-\boldsymbol{\beta}}{\sigma}\right) \\ &\quad + \sigma[\alpha_b f_b - \alpha_a f_a]\left(\frac{-\boldsymbol{\beta}}{\sigma}\right). \end{aligned}$$

After inserting the definitions of  $\alpha_a$  and  $\alpha_b$ , and collecting terms, we find all terms sum to zero save for the desired result,

$$\frac{\partial E[y | \mathbf{x}]}{\partial \mathbf{x}} = (F_b - F_a)\boldsymbol{\beta} = \boldsymbol{\beta} \times \text{Prob}[a < y_i^* < b].$$

## 850 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

Note that this general result includes censoring in either or both tails of the distribution, and it does not assume that  $\varepsilon$  is normally distributed. For the standard case with censoring at zero and normally distributed disturbances, the result specializes to

$$\frac{\partial E[y_i | \mathbf{x}_i]}{\partial \mathbf{x}_i} = \boldsymbol{\beta} \Phi\left(\frac{\mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right).$$

Although not a formal result, this does suggest a reason why, in general, least squares estimates of the coefficients in a tobit model usually resemble the MLEs times the proportion of nonlimit observations in the sample.

McDonald and Moffitt (1980) suggested a useful decomposition of  $\partial E[y_i | \mathbf{x}_i]/\partial \mathbf{x}_i$ ,

$$\frac{\partial E[y_i | \mathbf{x}_i]}{\partial \mathbf{x}_i} = \boldsymbol{\beta} \times \{ \Phi_i[1 - \lambda_i(\alpha_i + \lambda_i)] + \phi_i(\alpha_i + \lambda_i) \},$$

where  $\alpha_i = \mathbf{x}'_i \boldsymbol{\beta}/\sigma$ ,  $\Phi_i = \Phi(\alpha_i)$  and  $\lambda_i = \phi_i/\Phi_i$ . Taking the two parts separately, this result decomposes the slope vector into

$$\frac{\partial E[y_i | \mathbf{x}_i]}{\partial \mathbf{x}_i} = \text{Prob}[y_i > 0] \frac{\partial E[y_i | \mathbf{x}_i, y_i > 0]}{\partial \mathbf{x}_i} + E[y_i | \mathbf{x}_i, y_i > 0] \frac{\partial \text{Prob}[y_i > 0]}{\partial \mathbf{x}_i}.$$

Thus, a change in  $\mathbf{x}_i$  has two effects: It affects the conditional mean of  $y_i^*$  in the positive part of the distribution, and it affects the probability that the observation will fall in that part of the distribution.

### 19.3.3 ESTIMATION

The tobit model has become so routine and been incorporated in so many computer packages that despite formidable obstacles in years past, estimation is now essentially on the level of ordinary linear regression. The log-likelihood for the censored regression model is

$$\ln L = \sum_{y_i > 0} -\frac{1}{2} \left[ \log(2\pi) + \ln \sigma^2 + \frac{(y_i - \mathbf{x}'_i \boldsymbol{\beta})^2}{\sigma^2} \right] + \sum_{y_i=0} \ln \left[ 1 - \Phi\left(\frac{\mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right) \right]. \quad (19-13)$$

The two parts correspond to the classical regression for the nonlimit observations and the relevant probabilities for the limit observations, respectively. This likelihood is a nonstandard type, because it is a mixture of discrete and continuous distributions. In a seminal paper, Amemiya (1973) showed that despite the complications, proceeding in the usual fashion to maximize  $\ln L$  would produce an estimator with all the familiar desirable properties attained by MLEs.

The log-likelihood function is fairly involved, but Olsen's (1978) **reparameterization** simplifies things considerably. With  $\boldsymbol{\gamma} = \boldsymbol{\beta}/\sigma$  and  $\theta = 1/\sigma$ , the log-likelihood is

$$\ln L = \sum_{y_i > 0} -\frac{1}{2} [\ln(2\pi) - \ln \theta^2 + (\theta y_i - \mathbf{x}'_i \boldsymbol{\gamma})^2] + \sum_{y_i=0} \ln [1 - \Phi(\mathbf{x}'_i \boldsymbol{\gamma})]. \quad (19-14)$$

The results in this setting are now very similar to those for the truncated regression. The Hessian is always negative definite, so Newton's method is simple to use and usually converges quickly. After convergence, the original parameters can be recovered using  $\sigma = 1/\theta$  and  $\boldsymbol{\beta} = \boldsymbol{\gamma}/\theta$ . The asymptotic covariance matrix for these estimates can be obtained from that for the estimates of  $[\boldsymbol{\gamma}, \theta]$  using the **delta method**:

CHAPTER 19 ♦ Limited Dependent Variables **851****TABLE 19.2** Tobit Estimates of an Hours Worked Equation

	<i>White Wives</i>		<i>Black Wives</i>		<i>Least</i>	<i>Scaled</i>
	<i>Coefficient</i>	<i>Slope</i>	<i>Coefficient</i>	<i>Slope</i>	<i>Squares</i>	<i>OLS</i>
Constant	-1803.13 (-8.64)		-2753.87 (-9.68)			
Small kids	-1324.84 (-19.78)	-385.89	-824.19 (-10.14)	-376.53	-352.63	-766.56
Education difference	-48.08 (-4.77)	-14.00	22.59 (1.96)	10.32	11.47	24.93
Relative wage	312.07 (5.71)	90.90	286.39 (3.32)	130.93	123.95	269.46
Second marriage	175.85 (3.47)	51.51	25.33 (0.41)	11.57	13.14	28.57
Mean divorce probability	417.39 (6.52)	121.58	481.02 (5.28)	219.75	219.22	476.57
High divorce probability	670.22 (8.40)	195.22	578.66 (5.33)	264.36	244.17	530.80
$\sigma$	1559	618	1511	826		
Sample size		7459		2798		
Proportion working		0.29		0.46		

Est. Asy. Var[ $\hat{\beta}, \hat{\sigma}$ ] =  $\hat{\mathbf{J}}$  Asy. Var[ $\hat{\gamma}, \hat{\theta}$ ]  $\hat{\mathbf{J}}'$ , where

$$\mathbf{J} = \begin{bmatrix} \partial\beta/\partial\gamma' & \partial\beta/\partial\theta \\ \partial\sigma/\partial\gamma' & \partial\sigma/\partial\theta \end{bmatrix} = \begin{bmatrix} (1/\theta)\mathbf{I} & (-1/\theta^2)\gamma \\ \mathbf{0}' & (-1/\theta^2) \end{bmatrix}.$$

Researchers often compute ordinary least squares estimates despite their inconsistency. Almost without exception, it is found that the OLS estimates are smaller in absolute value than the MLEs. A striking empirical regularity is that the maximum likelihood estimates can often be approximated by dividing the OLS estimates by the proportion of nonlimit observations in the sample.<sup>11</sup> The effect is illustrated in the last two columns of Table 19.2. Another strategy is to discard the limit observations, but we now see that just trades the censoring problem for the truncation problem.

#### **Example 19.5 Estimated Tobit Equations for Hours Worked**

In their study of the number of hours worked in a survey year by a large sample of wives, Quester and Greene (1982) were interested in whether wives whose marriages were statistically more likely to dissolve hedged against that possibility by spending, on average, more time working. They reported the tobit estimates given in Table 19.2. The last figure in the table implies that a very large proportion of the women reported zero hours, so least squares regression would be inappropriate.

The figures in parentheses are the ratio of the coefficient estimate to the estimated asymptotic standard error. The dependent variable is hours worked in the survey year. "Small kids" is a dummy variable indicating whether there were children in the household. The "education difference" and "relative wage" variables compare husband and wife on these two dimensions. The wage rate used for wives was predicted using a previously estimated regression model and is thus available for all individuals, whether working or not. "Second marriage" is a

<sup>11</sup>This concept is explored further in Greene (1980b), Goldberger (1981), and Chung and Goldberger (1984).

## 852 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

dummy variable. Divorce probabilities were produced by a large microsimulation model presented in another study [Orcutt, Caldwell, and Wertheimer (1976)]. The variables used here were dummy variables indicating “mean” if the predicted probability was between 0.01 and 0.03 and “high” if it was greater than 0.03. The “slopes” are the marginal effects described earlier.

Note the marginal effects compared with the tobit coefficients. Likewise, the estimate of  $\sigma$  is quite misleading as an estimate of the standard deviation of hours worked.

The effects of the divorce probability variables were as expected and were quite large. One of the questions raised in connection with this study was whether the divorce probabilities could reasonably be treated as independent variables. It might be that for these individuals, the number of hours worked was a significant determinant of the probability.

### 19.3.4 TWO-PART MODELS AND CORNER SOLUTIONS

The tobit model contains a restriction that might be unreasonable in an economic setting. Consider a behavioral outcome,  $y$  = charitable donation. Two implications of the tobit model are that

$$\text{Prob}(y > 0 | \mathbf{x}) = \text{Prob}(\mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0 | \mathbf{x}) = \Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)$$

and [from (19-7)]

$$E[y | y > 0, \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta} + \sigma\phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)/\Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma).$$

Differentiating both of these, we find from (17-11) and (19-8),

$$\partial \text{Prob}(y > 0 | \mathbf{x})/\partial \mathbf{x} = [\phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)/\sigma]\boldsymbol{\beta} = \text{a positive multiple of } \boldsymbol{\beta},$$

$$\partial E[y | y > 0, \mathbf{x}]/\partial \mathbf{x} = \{[1 - \delta(\mathbf{x}'\boldsymbol{\beta}/\sigma)]/\sigma\}\boldsymbol{\beta} = \text{a positive multiple of } \boldsymbol{\beta}.$$

Thus, any variable that appears in the model affects the participation probability and the intensity equation with the same sign. In the case suggested, for example, it is conceivable that age might affect participation and intensity in different directions. Fin and Schmidt (1984) suggest another application, loss due to fire in buildings; older buildings might be more likely to have fires but, because of the greater value of newer buildings, the actual damage might be greater in newer buildings. This fact would require the coefficient on age to have different signs in the two functions, which is impossible in the tobit model because they are the same coefficient.

In an early study in this literature, Cragg (1971) proposed a somewhat more general model in which the probability of a limit observation is independent of the regression model for the nonlimit data. One can imagine, for instance, the decision of whether or not to purchase a car as being different from the decision of how much to spend on the car, having decided to buy one.

A more general model that accommodates these objections is as follows:

#### 1. Participation equation

$$\begin{aligned} \text{Prob}[y_i^* > 0] &= \Phi(\mathbf{x}_i'\boldsymbol{\gamma}), & d_i &= 1 \text{ if } y_i^* > 0, \\ \text{Prob}[y_i^* \leq 0] &= 1 - \Phi(\mathbf{x}_i'\boldsymbol{\gamma}), & d_i &= 0 \text{ if } y_i^* \leq 0. \end{aligned} \tag{19-15}$$

#### 2. Intensity equation for nonlimit observations

$$E[y_i | d_i = 1] = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\lambda_i,$$

CHAPTER 19 ♦ Limited Dependent Variables **853**

according to Theorem 19.2. This two-part model is a combination of the truncated regression model of Section 19.2 and the univariate probit model of Section 17.3, which suggests a method of analyzing it. Note that it is precisely the same approach we considered in Section 18.4.8 and Example 18.12 where we used a hurdle model to model doctor visits. The tobit model returns if  $\gamma = \beta/\sigma$ . The parameters of the regression (intensity) equation can be estimated independently using the truncated regression model of Section 19.2. An application is Melenberg and van Soest (1996).

Lin and Schmidt (1984) considered testing the restriction of the tobit model. Based only on the tobit model, they devised a Lagrange multiplier statistic that, although a bit cumbersome algebraically, can be computed without great difficulty. If one is able to estimate the truncated regression model, the tobit model, and the probit model separately, then there is a simpler way to test the hypothesis. The tobit log-likelihood is the sum of the log-likelihoods for the truncated regression and probit models. To show this result, add and subtract  $\sum_{y_i=1} \ln \Phi(\mathbf{x}'_i \boldsymbol{\beta})$  in (19-13). This produces the log-likelihood for the truncated regression model (considered in the exercises) plus (17-20) for the probit model. Therefore, a likelihood ratio statistic can be computed using

$$\lambda = -2[\ln LT - (\ln LP + \ln LTR)],$$

where

$LT$  = likelihood for the tobit model in (19-13), with the same coefficients

$LP$  = likelihood for the probit model in (17-17), fit separately

$LTR$  = likelihood for the truncated regression model, fit separately

The two-part model just considered extends the tobit model, but it stops a bit short of the generality we might achieve. In the preceding hurdle model, we have assumed that the same regressors appear in both equations. Although this produces a convenient way to retreat to the tobit model as a parametric restriction, it couples the two decisions perhaps unreasonably. In our example to follow, where we model extramarital affairs, the decision whether or not to spend any time in an affair may well be an entirely different decision from how much time to spend having once made that commitment. The obvious way to proceed is to reformulate the hurdle model as

### 1. Participation equation

$$\begin{aligned} \text{Prob}[d_i^* > 0] &= \Phi(\mathbf{z}'_i \boldsymbol{\gamma}), & d_i = 1 \text{ if } d_i^* > 0, \\ \text{Prob}[d_i^* \leq 0] &= 1 - \Phi(\mathbf{z}'_i \boldsymbol{\gamma}), & d_i = 0 \text{ if } d_i^* \leq 0. \end{aligned} \quad (19-16)$$

### 2. Intensity equation for nonlimit observations

$$E[y_i | d_i = 1] = \mathbf{x}'_i \boldsymbol{\beta} + \sigma \lambda_i.$$

This extension, however, omits an important element; it seems unlikely that the two decisions would be uncorrelated; that is, the implicit disturbances in the equations should be correlated. The combination of these produces what has been labeled a **type-II tobit model**. [Amemiya (1985) identified five possible permutations of the model specification and observation mechanism. The familiar tobit model is type I; this is type-II.] The full model is

**854 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**
**1. Participation equation**

$$\begin{aligned} d_i^* &= \mathbf{z}'_i \boldsymbol{\gamma} + u_i, & u_i &\sim N[0, 1] \\ d_i &= 1 \text{ if } d_i^* > 0, & 0 &\text{ otherwise.} \end{aligned}$$

**2. Intensity equation**

$$y_i^* = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N[0, \sigma^2].$$

**3. Observation mechanism**

- (a)  $y_i^* = 0$  if  $d_i = 0$  and  $y_i = y_i^*$  if  $d_i = 1$ .
- (b)  $y_i = y_i^*$  if  $d_i = 1$  and  $y_i$  is unobserved if  $d_i = 0$ .

**4. Endogeneity**

$$(u_i, \varepsilon_i) \sim \text{bivariate normal with correlation } \rho.$$

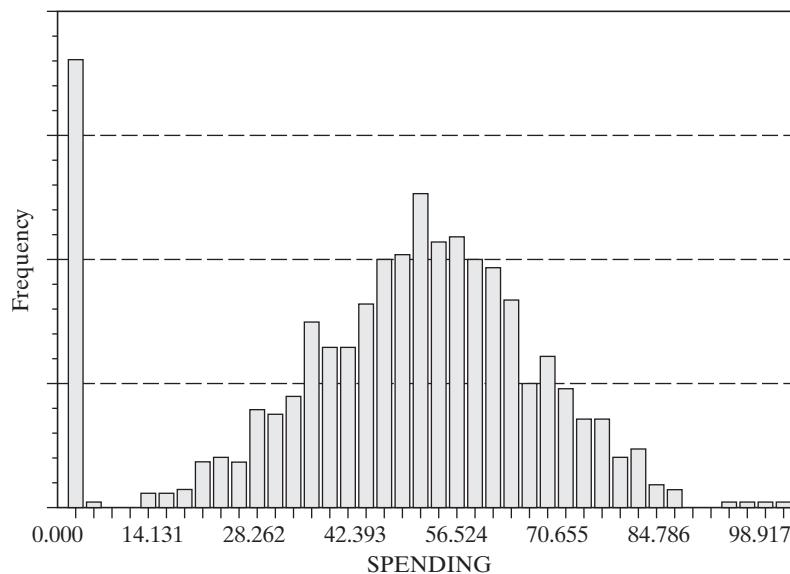
Mechanism (a) produces Amemiya's type II model. (Amemiya blends these two interpretations. In the statement of the model, he presents (a), but in the subsequent discussion, assumes (b). The difference is substantive if  $\mathbf{x}_i$  is observed in case (b). Otherwise, they are the same, and " $y_i = 0$ " is not actually meaningful. Amemiya notes, " $y_i^* = 0$  merely signifies the event  $d_i^* \leq 0$ ." If  $\mathbf{x}_i$  is observed when  $d_i = 0$ , then these observations will contribute to the likelihood for the full sample. If not, then they will not. We will develop this idea later when we consider Heckman's selection model [which is case (b) without observed  $\mathbf{x}_i$  when  $d_i = 0$ ].

There are two estimation strategies that can be used to fit the type II model. A two-step method can proceed as follows: The probit model for  $d_i$  can be estimated using maximum likelihood as shown in Section 17.3. For the second step, we make use of our theorems on truncation (and Theorem 19.5 that will appear later) to write

$$\begin{aligned} E[y_i | d_i = 1, \mathbf{x}_i, \mathbf{z}_i] &= \mathbf{x}'_i \boldsymbol{\beta} + E[\varepsilon_i | d_i = 1, \mathbf{x}_i, \mathbf{z}_i] \\ &= \mathbf{x}'_i \boldsymbol{\beta} + \rho \sigma \frac{\phi(\mathbf{z}'_i \boldsymbol{\gamma})}{\Phi(\mathbf{z}'_i \boldsymbol{\gamma})} \\ &= \mathbf{x}'_i \boldsymbol{\beta} + \rho \sigma \lambda_i. \end{aligned} \tag{19-17}$$

Since we have estimated  $\gamma$  at step 1, we can compute  $\hat{\lambda}_i = \phi(\mathbf{z}'_i \hat{\boldsymbol{\gamma}})/\Phi(\mathbf{z}'_i \hat{\boldsymbol{\gamma}})$  using the first-step estimates, and we can estimate  $\beta$  and  $\theta = (\rho\sigma)$  by least squares regression of  $y_i$  on  $\mathbf{x}_i$  and  $\hat{\lambda}_i$ . It will be necessary to correct the asymptotic covariance matrix that is computed for  $(\hat{\boldsymbol{\beta}}, \hat{\theta})$ . This is a template application of the Murphy and Topel (2002) results that appear in Section 14.7. The second approach is full information maximum likelihood, estimating all the parameters in both equations simultaneously. We will return to the details of estimation of the type II tobit model in Section 19.5 where we examine Heckman's model of "sample selection" model (which is the type II tobit model).

Many of the applications of the tobit model in the received literature are constructed not to accommodate censoring of the underlying data, but, rather, to model the appearance of a large cluster of zeros. Cragg's application is clearly related to this phenomenon. Consider, for example, survey data on purchases of consumer durables, firm expenditure on research and development, or consumer savings. In each case, the

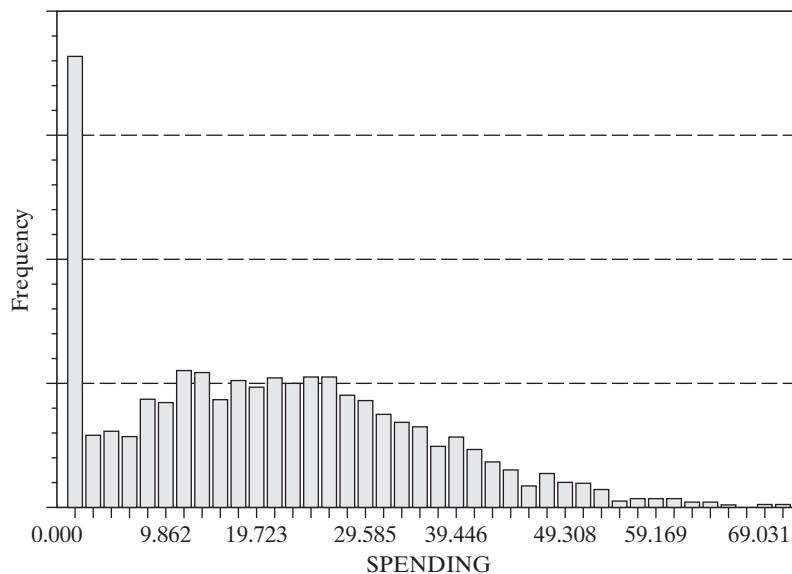


**FIGURE 19.5** Hypothetical Spending Data.

observed data will consist of zero or some positive amount. Arguably, there are two decisions at work in these scenarios: First, whether to engage in the activity or not, and second, given that the answer to the first question is yes, how intensively to engage in it—how much to spend, for example. This is precisely the motivation behind the hurdle model. This specification has been labeled a “**corner solution model**”; see Wooldridge (2002a, pp. 518–519).

In practical terms, the difference between the **hurdle model** and the tobit model should be evident in the data. Often overlooked in tobit analyses is that the model predicts not only a cluster of zeros (or limit observations), but also a grouping of observations *near zero* (or the limit point). For example, the tobit model is surely misspecified for the sort of (hypothetical) spending data shown in Figure 19.5 for a sample of 1,000 observations. Neglecting for the moment the earlier point about the underlying decision process, Figure 19.6 shows the characteristic appearance of a (substantively) censored variable. The implication for the model builder is that an appropriate specification would consist of two equations, one for the “participation decision,” and one for the distribution of the positive dependent variable. Formally, we might, continuing the development of Cragg’s specification, model the first decision with a binary choice (e.g., probit or logit model). The second equation is a model for  $y | y > 0$ , for which the truncated regression model of Section 19.2.3 is a natural candidate. As we will see, this is essentially the model behind the sample selection treatment developed in Section 19.5.

Two practical issues frequently intervene at this point. First, one might well have a model in mind for the intensity (regression) equation, but none for the participation equation. This is the usual backdrop for the uses of the tobit model, which produces the considerations in the previous section. The second issue concerns the appropriateness

**856 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**

**FIGURE 19.6** Hypothetical Censored Data.

of the truncation or censoring model to data such as those in Figure 19.6. If we consider only the nonlimit observations in Figure 19.5, the underlying distribution does not appear to be truncated at all. The truncated regression model in Section 19.2.3 fit to these data will not depart significantly from ordinary least squares [because the underlying probability in the denominator of (19-6) will equal one and the numerator will equal zero]. But, this is not the case of a tobit model forced on these same data. Forcing the model in (19-13) on data such as these will significantly distort the estimator—all else equal, it will significantly attenuate the coefficients, the more so the larger is the proportion of limit observations in the sample. Once again, this stands as a caveat for the model builder. The tobit model is manifestly misspecified for data such as those in Figure 19.5.

**Example 19.6 Two-Part Model for Extramarital Affairs**

In Example 18.9, we examined Fair's (1977) *Psychology Today* survey data on extramarital affairs. The 601 observations in the data set are mostly zero—451 of the 601. This feature of the data motivated the author to use a tobit model to analyze these data. In our example, we reconsidered the model, since the nonzero observations were a count, not a continuous variable. Another data set in Fair's study was the *Redbook Magazine* survey of 6,366 married women. Once again, the outcome variable of interest was extramarital affairs. However, in this instance, the outcome data were transformed to a measure of time spent, which, being continuous, lends itself more naturally to the tobit model we are studying here. The variables in the data set are as follows (excluding three unidentified and not used):

- id* = Identification number
- C* = Constant, value = 1
- yrb* = Constructed measure of time spent in extramarital affairs
- v<sub>1</sub>* = Rating of the marriage, coded 1 to 4
- v<sub>2</sub>* = Age, in years, aggregated
- v<sub>3</sub>* = Number of years married

CHAPTER 19 ♦ Limited Dependent Variables **857****TABLE 19.3** Estimated Censored Regression Models (*t*-ratios in parentheses)

	Model						
	Linear OLS	Tobit	Truncated Regression	Probit	Tobit/ $\sigma$	Hurdle Participation	Hurdle Intensity
Constant	3.62346 (13.63)	7.83653 (10.98)	8.89449 (2.90)	2.21010 (12.60)	1.74189	1.56419 (17.75)	4.84602 (5.87)
RateMarr	-0.42053 (-14.79)	-1.53071 (-20.85)	-0.44303 (-1.45)	-0.42874 (-23.40)	-0.34024	-0.42582 (-23.61)	-0.24603 (-.46)
Age	-0.01457 (-1.59)	-0.10514 (-4.24)	-0.22394 (-1.83)	-0.03542 (-5.87)	-0.02337		-0.01903 (-.77)
YrsMarr	-0.01599 (-1.62)	0.12829 (4.86)	-0.94437 (-7.27)	0.06563 (10.18)	0.02852		-0.16822 (-6.52)
NumKids	-0.01705 (-.57)	-0.02777 (-0.36)	-0.02280 (-0.06)	-0.00394 (-0.21)	-0.00617	0.14024 (11.55)	-0.28365 (-1.49)
Religious	-0.24374 (-7.83)	-0.94350 (-11.11)	-0.50490 (-1.29)	-0.22281 (-10.88)	-0.20972	-0.21466 (-10.64)	-0.05452 (-0.19)
Education	-0.01743 (-1.24)	-0.08598 (-2.28)	-0.06406 (-0.38)	-0.02373 (-2.60)	-0.01911		0.00338 (0.09)
Wife Occ.	0.06577 (2.10)	0.31284 (3.82)	0.00805 (0.02)	0.09539 (4.75)	0.06954		0.01505 (0.19)
Hus. Occ.	0.00405 (0.19)	0.01421 (0.26)	-0.09946 (-0.41)	0.00659 (0.49)	0.00316		-0.02911 (-0.53)
$\sigma$	2.14351	4.49887	5.46846				3.43748
ln $L$	R <sup>2</sup> = 0.05479	-7804.38	-3463.71	-3469.58			

$v_4$  = Number of children, top coded at 5

$v_5$  = Religiosity, 1 to 4, 1 = not, 4 = very

$v_6$  = Education, coded 9, 12, 14, 16, 17, 20

$v_7$  = Wife's Occupation—Hollingshead scale

$v_8$  = Husband's occupation—Hollingshead scale

This is a cross section of 6,366 observations with 4,313 zeros and 2,053 positive values.

Table 19.3 presents estimates of various models for  $yrb$ . The leftmost column presents the OLS estimates. The least squares estimator is inconsistent in this model. The empirical regularity that the OLS estimator appears to be biased toward zero, the more so is the smaller the proportion of limit observations. Here, the ratio, based on the tobit estimates in the second column, appears to be about 4 or 5 to 1. Likewise, the OLS estimator of  $\sigma$  appears to be greatly underestimated. This would be expected, as the OLS estimator is treating the limit observations, which have no variation in the dependent variable, as if they were nonlimit observations. The third set of results is the truncated regression estimator. In principle, the truncated regression estimator is also consistent. However, it will be less efficient as it is based on less information. In our example, this estimator seems to be quite erratic, again compared to the tobit estimator. Note, for example, the coefficient on years married, which, although it is "significant" in both cases, changes sign. The *t* ratio on Religiousness falls from -11.11 to -1.29 in the truncation model. The probit estimator based on  $yrb > 0$  appears next. As a rough check on the corner solution aspect of our model, we would expect the normalized tobit coefficients ( $\beta/\sigma$ ) to approximate the probit coefficients, which they appear to. However, the likelihood ratio statistic for testing the internal consistency based on the three estimated models is  $2[7804.38 - 3463.71 - 3469.58] = 1742.18$  with nine degrees of freedom. The hypothesis of parameter constancy implied by the tobit model is rejected. The last two sets of results are for a hurdle model in which the intensity equation is fit by the two-step method.

## 858 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

### 19.3.5 SOME ISSUES IN SPECIFICATION

Two issues that commonly arise in microeconomic data, heteroscedasticity and nonnormality, have been analyzed at length in the tobit setting.<sup>12</sup>

#### 19.3.5.a Heteroscedasticity

Maddala and Nelson (1975), Hurd (1979), Arabmazar and Schmidt (1982a,b), and Brown and Moffitt (1982) all have varying degrees of pessimism regarding how inconsistent the maximum likelihood estimator will be when **heteroscedasticity** occurs. Not surprisingly, the degree of censoring is the primary determinant. Unfortunately, all the analyses have been carried out in the setting of very specific models—for example, involving only a single dummy variable or one with groupwise heteroscedasticity—so the primary lesson is the very general conclusion that heteroscedasticity emerges as an obviously serious problem.

One can approach the heteroscedasticity problem directly. Petersen and Waldman (1981) present the computations needed to estimate a tobit model with heteroscedasticity of several types. Replacing  $\sigma$  with  $\sigma_i$  in the log-likelihood function and including  $\sigma_i^2$  in the summations produces the needed generality. Specification of a particular model for  $\sigma_i$  provides the empirical model for estimation.

#### *Example 19.7 Multiplicative Heteroscedasticity in the Tobit Model*

Petersen and Waldman (1981) analyzed the volume of short interest in a cross section of common stocks. The regressors included a measure of the market component of heterogeneous expectations as measured by the firm's *BETA* coefficient; a company-specific measure of heterogeneous expectations, *NONMARKET*; the *NUMBER* of analysts making earnings forecasts for the company; the number of common shares to be issued for the acquisition of another firm, *MERGER*; and a dummy variable for the existence of *OPTIONS*. They report the results listed in Table 19.4 for a model in which the variance is assumed to be of the form  $\sigma_i^2 = \exp(\mathbf{x}_i'\boldsymbol{\alpha})$ . The values in parentheses are the ratio of the coefficient to the estimated asymptotic standard error.

The effect of heteroscedasticity on the estimates is extremely large. We do note, however, a common misconception in the literature. The change in the coefficients is often misleading. The marginal effects in the heteroscedasticity model will generally be very similar to those computed from the model which assumes homoscedasticity. (The calculation is pursued in the exercises.)

A test of the hypothesis that  $\boldsymbol{\alpha} = \mathbf{0}$  (except for the constant term) can be based on the likelihood ratio statistic. For these results, the statistic is  $-2[-547.3 - (-466.27)] = 162.06$ . This statistic has a limiting chi-squared distribution with five degrees of freedom. The sample value exceeds the critical value in the table of 11.07, so the hypothesis can be rejected.

In the preceding example, we carried out a likelihood ratio test against the hypothesis of homoscedasticity. It would be desirable to be able to carry out the test without having to estimate the unrestricted model. A **Lagrange multiplier test** can be used for

---

<sup>12</sup>Two symposia that contain numerous results on these subjects are Blundell (1987) and Duncan (1986b). An application that explores these two issues in detail is Melenberg and van Soest (1996). Developing specification tests for the tobit model has been a popular enterprise. A sampling of the received literature includes Nelson (1981); Bera, Jarque, and Lee (1982); Chesher and Irish (1987); Chesher, Lancaster, and Irish (1985); Gourieroux et al. (1984, 1987); Newey (1986); Rivers and Vuong (1988); Horowitz and Neumann (1989); and Pagan and Vella (1989). Newey (1985a,b) are useful references on the general subject of conditional moment testing. More general treatments of specification testing are Godfrey (1988) and Ruud (1984).

**TABLE 19.4** Estimates of a Tobit Model (standard errors in parentheses)

	<i>Homoscedastic</i>		<i>Heteroscedastic</i>	
	$\beta$		$\beta$	$\alpha$
Constant	-18.28 (5.10)		-4.11 (3.28)	-0.47 (0.60)
Beta	10.97 (3.61)		2.22 (2.00)	1.20 (1.81)
Nonmarket	0.65 (7.41)		0.12 (1.90)	0.08 (7.55)
Number	0.75 (5.74)		0.33 (4.50)	0.15 (4.58)
Merger	0.50 (5.90)		0.24 (3.00)	0.06 (4.17)
Option	2.56 (1.51)		2.96 (2.99)	0.83 (1.70)
ln $L$	-547.30		-466.27	
Sample size	200		200	

that purpose. Consider the heteroscedastic tobit model in which we specify that

$$\sigma_i^2 = \sigma^2 [\exp(\mathbf{w}'_i \boldsymbol{\alpha})]^2. \quad (19-18)$$

This model is a fairly general specification that includes many familiar ones as special cases. The null hypothesis of homoscedasticity is  $\boldsymbol{\alpha} = \mathbf{0}$ . (We used this specification in the probit model in Section 17.3.7 and in the linear regression model in Section 9.7.1) Using the BHHH estimator of the Hessian as usual, we can produce a Lagrange multiplier statistic as follows: Let  $z_i = 1$  if  $y_i$  is positive and 0 otherwise,

$$\begin{aligned} a_i &= z_i \left( \frac{\varepsilon_i}{\sigma^2} \right) + (1 - z_i) \left( \frac{(-1)\lambda_i}{\sigma} \right), \\ b_i &= z_i \left( \frac{(\varepsilon_i^2/\sigma^2 - 1)}{2\sigma^2} \right) + (1 - z_i) \left( \frac{(\mathbf{x}'_i \boldsymbol{\beta})\lambda_i}{2\sigma^3} \right), \\ \lambda_i &= \frac{\phi(\mathbf{x}'_i \boldsymbol{\beta}/\sigma)}{1 - \Phi(\mathbf{x}'_i \boldsymbol{\beta}/\sigma)}. \end{aligned} \quad (19-19)$$

The data vector is  $\mathbf{g}_i = [a_i \mathbf{x}'_i, b_i, b_i \mathbf{w}'_i]'$ . The sums are taken over all observations, and all functions involving unknown parameters ( $\varepsilon_i, \phi_i, \Phi_i, \mathbf{x}'_i \boldsymbol{\beta}, \sigma, \lambda_i$ ) are evaluated at the restricted (homoscedastic) maximum likelihood estimates. Then,

$$LM = \mathbf{i}' \mathbf{G} [\mathbf{G}' \mathbf{G}]^{-1} \mathbf{G}' \mathbf{i} = n R^2 \quad (19-20)$$

in the regression of a column of ones on the  $K + 1 + P$  derivatives of the log-likelihood function for the model with multiplicative heteroscedasticity, evaluated at the estimates from the restricted model. (If there were no limit observations, then it would reduce to the Breusch-Pagan statistic discussed in Section 9.5.2.) Given the maximum likelihood estimates of the tobit model coefficients, it is quite simple to compute. The statistic has a limiting chi-squared distribution with degrees of freedom equal to the number of variables in  $\mathbf{w}_i$ .

### 19.3.5.b Nonnormality

Nonnormality is an especially difficult problem in this setting. It has been shown that if the underlying disturbances are not normally distributed, then the estimator based

## 860 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

on (19-13) is inconsistent. Research is ongoing both on alternative estimators and on methods for testing for this type of misspecification.<sup>13</sup>

One approach to the estimation is to use an alternative distribution. Kalbfleisch and Prentice (2002) present a unifying treatment that includes several distributions such as the exponential, lognormal, and Weibull. (Their primary focus is on survival analysis in a medical statistics setting, which is an interesting convergence of the techniques in very different disciplines.) Of course, assuming some other specific distribution does not necessarily solve the problem and may make it worse. A preferable alternative would be to devise an estimator that is robust to changes in the distribution. Powell's (1981, 1984) least absolute deviations (LAD) estimator appears to offer some promise.<sup>14</sup> The main drawback to its use is its computational complexity. An extensive application of the LAD estimator is Melenberg and van Soest (1996). Although estimation in the nonnormal case is relatively difficult, testing for this failure of the model is worthwhile to assess the estimates obtained by the conventional methods. Among the tests that have been developed are Hausman tests, Lagrange multiplier tests [Bera and Jarque (1981, 1982), Bera, Jarque, and Lee (1982)], and **conditional moment tests** [Nelson (1981)].

### 19.3.6 PANEL DATA APPLICATIONS

Extension of the familiar panel data results to the tobit model parallel the probit model, with the attendant problems. The random effects or random parameters models discussed in Chapter 17 can be adapted to the censored regression model using simulation or quadrature. The same reservations with respect to the orthogonality of the effects and the regressors will apply here, as will the applicability of the Mundlak (1978) correction to accommodate it.

Most of the attention in the theoretical literature on panel data methods for the tobit model has been focused on fixed effects. The departure point would be the maximum likelihood estimator for the static fixed effects model,

$$\begin{aligned} y_{it}^* &= \alpha_i + x'_{it}\beta + \varepsilon_{it}, \quad \varepsilon_{it} \sim N[0, \sigma^2], \\ y_{it} &= \text{Max}(0, y_{it}). \end{aligned}$$

However, there are no firm theoretical results on the behavior of the MLE in this model. Intuition might suggest, based on the findings for the binary probit model, that the MLE would be biased in the same fashion, away from zero. Perhaps surprisingly, the results in Greene (2004) persistently found that not to be the case in a variety of model specifications. Rather, the incidental parameters, such as it is, manifests in a downward bias in the estimator of  $\sigma$ , not an upward (or downward) bias in the MLE of  $\beta$ . However, this is less surprising when the tobit estimator is juxtaposed with the MLE in the linear regression model with fixed effects. In that model, the MLE is the within-groups (LSDV) estimator which is unbiased and consistent. But, the ML estimator of the disturbance variance in the linear regression model is  $\mathbf{e}'_{\text{LSDV}} \mathbf{e}_{\text{LSDV}} / (nT)$ , which is biased downward

<sup>13</sup>See Duncan (1983, 1986b), Goldberger (1983), Pagan and Vella (1989), Lee (1996), and Fernandez (1986).

<sup>14</sup>See Duncan (1986a,b) for a symposium on the subject and Amemiya (1984). Additional references are Newey, Powell, and Walker (1990); Lee (1996); and Robinson (1988).

by a factor of  $(T - 1)/T$ . [This is the result found in the original source on the incidental parameters problem, Neyman and Scott (1948).] So, what evidence there is suggests that unconditional estimation of the tobit model behaves essentially like that for the linear regression model. That does not settle the problem, however; if the evidence is correct, then it implies that although consistent estimation of  $\beta$  is possible, appropriate statistical inference is not. The bias in the estimation of  $\sigma$  shows up in any estimator of the asymptotic covariance of the MLE of  $\beta$ .

Unfortunately, there is no conditional estimator of  $\beta$  for the tobit (or truncated regression) model. First differencing or taking group mean deviations does not preserve the model. Because the latent variable is censored before observation, these transformations are not meaningful. Some progress has been made on theoretical, **semiparametric estimators** for this model. See, for example, Honoré and Kyriazidou (2000) for a survey. Much of the theoretical development has also been directed at dynamic models where the benign result of the previous paragraph (such as it is) is lost once again. Arellano (2001) contains some general results. Hahn and Kuersteiner (2004) have characterized the bias of the MLE, and suggested methods of reducing the bias of the estimators in dynamic binary choice and censored regression models.

## 19.4 MODELS FOR DURATION

The leading application of the censoring models we examined in Section 19.3 is models for durations and events. We consider the time until some kind of transition as the duration, and the transition, itself, as the event. The length of a spell of unemployment (until rehire or exit from the market), the duration of a strike, the amount of time until a patient ends a health-related spell in connection with a disease or operation, and the length of time between origination and termination (via prepayment, default, or some other mechanism) of a mortgage are all examples of durations and transitions. The role that censoring plays in these scenarios is that in almost all cases in which we as analysts study duration data, some or even many of the spells we observe do not end in transitions. For example, in studying the lengths of unemployment spells, many of the individuals in the sample may still be unemployed at the time the study ends—the analyst observes (or believes) that the spell will end some time after the observation window closes. These data on spell lengths are, by construction, censored. Models of duration will generally account explicitly for censoring of the duration data.

This section is concerned with models of duration. In some aspects, the regression-like models we have studied, such as the discrete choice models, are the appropriate tools. As in the previous two chapters, however, the models are nonlinear, and the familiar regression methods are not appropriate. Most of this analysis focuses on maximum likelihood estimators. In modeling duration, although an underlying regression model is, in fact, at work, it is generally not the conditional mean function that is of interest. More likely, as we will explore next, the objects of estimation are certain probabilities of events, for example in the conditional probability of a transition in a given interval given that the spell has lasted up to the point of interest. These are known as “hazard models”—the probability is labeled the hazard function—and are a central focus of this type of analysis.

**862 PART IV ♦ Cross Sections, Panel Data, and Microeometrics****19.4.1 MODELS FOR DURATION DATA<sup>15</sup>**

Intuition might suggest that the longer a strike persists, the more likely it is that it will end within, say, the next week. Or is it? It seems equally plausible to suggest that the longer a strike has lasted, the more difficult must be the problems that led to it in the first place, and hence the *less* likely it is that it will end in the next short time interval. A similar kind of reasoning could be applied to spells of unemployment or the interval between conceptions. In each of these cases, it is not only the duration of the event, per se, that is interesting, but also the likelihood that the event will end in “the next period” given that it has lasted as long as it has.

Analysis of the length of *time until failure* has interested engineers for decades. For example, the models discussed in this section were applied to the durability of electric and electronic components long before economists discovered their usefulness. Likewise, the analysis of *survival times*—for example, the length of survival after the onset of a disease or after an operation such as a heart transplant—has long been a staple of biomedical research. Social scientists have recently applied the same body of techniques to strike duration, length of unemployment spells, intervals between conception, time until business failure, length of time between arrests, length of time from purchase until a warranty claim is made, intervals between purchases, and so on.

This section will give a brief introduction to the econometric analysis of duration data. As usual, we will restrict our attention to a few straightforward, relatively uncomplicated techniques and applications, primarily to introduce terms and concepts. The reader can then wade into the literature to find the extensions and variations. We will concentrate primarily on what are known as **parametric models**. These apply familiar inference techniques and provide a convenient departure point. Alternative approaches are considered at the end of the discussion.

**19.4.2 DURATION DATA**

The variable of interest in the analysis of duration is the length of time that elapses from the beginning of some event either until its end or until the measurement is taken, which may precede termination. Observations will typically consist of a cross section of durations,  $t_1, t_2, \dots, t_n$ . The process being observed may have begun at different points in calendar time for the different individuals in the sample. For example, the strike duration data examined in Example 19.8 are drawn from nine different years.

Censoring is a pervasive and usually unavoidable problem in the analysis of duration data. The common cause is that the measurement is made while the process is ongoing. An obvious example can be drawn from medical research. Consider analyzing the survival times of heart transplant patients. Although the beginning times may be known with precision, at the time of the measurement, observations on any individuals who are still alive are necessarily censored. Likewise, samples of spells of unemployment drawn from surveys will probably include some individuals who are still unemployed at the time the survey is taken. For these individuals, duration, or survival, is at least the

<sup>15</sup>There are a large number of highly technical articles on this topic, but relatively few accessible sources for the uninitiated. A particularly useful introductory survey is Kiefer (1988), upon which we have drawn heavily for this section. Other useful sources are Kalbfleisch and Prentice (2002), Heckman and Singer (1984a), Lancaster (1990), Florens, Fougere, and Mouchart (1996) and Cameron and Trivedi (2005, Chapters 17–19).

CHAPTER 19 ♦ Limited Dependent Variables **863**

observed  $t_i$ , but not equal to it. Estimation must account for the censored nature of the data for the same reasons as considered in Section 19.3. The consequences of ignoring censoring in duration data are similar to those that arise in regression analysis.

In a conventional regression model that characterizes the conditional mean and variance of a distribution, the regressors can be taken as fixed characteristics at the point in time or for the individual for which the measurement is taken. When measuring duration, the observation is implicitly on a process that has been under way for an interval of time from zero to  $t$ . If the analysis is conditioned on a set of covariates (the counterparts to regressors)  $\mathbf{x}_t$ , then the duration is implicitly a function of the entire time path of the variable  $\mathbf{x}(t)$ ,  $t = (0, t)$ , which may have changed during the interval. For example, the observed duration of employment in a job may be a function of the individual's rank in the firm. But their rank may have changed several times between the time they were hired and when the observation was made. As such, observed rank at the end of the job tenure is not necessarily a complete description of the individual's rank *while they were employed*. Likewise, marital status, family size, and amount of education are all variables that can change during the duration of unemployment and that one would like to account for in the duration model. The treatment of **time-varying covariates** is a considerable complication.<sup>16</sup>

#### **19.4.3 A REGRESSION-LIKE APPROACH: PARAMETRIC MODELS OF DURATION**

We will use the term *spell* as a catchall for the different duration variables we might measure. Spell length is represented by the random variable  $T$ . A simple approach to duration analysis would be to apply regression analysis to the sample of observed spells. By this device, we could characterize the expected duration, perhaps conditioned on a set of covariates whose values were measured at the end of the period. We could also assume that conditioned on an  $\mathbf{x}$  that has remained fixed from  $T = 0$  to  $T = t$ ,  $t$  has a normal distribution, as we commonly do in regression. We could then characterize the probability distribution of observed duration times. But, normality turns out not to be particularly attractive in this setting for a number of reasons, not least of which is that duration is positive by construction, while a normally distributed variable can take negative values. (*Lognormality* turns out to be a palatable alternative, but it is only one among a long list of candidates.)

##### **19.4.3.a Theoretical Background**

Suppose that the random variable  $T$  has a continuous probability distribution  $f(t)$ , where  $t$  is a realization of  $T$ . The cumulative probability is

$$F(t) = \int_0^t f(s) ds = \text{Prob}(T \leq t).$$

We will usually be more interested in the probability that the spell is of length *at least*  $t$ , which is given by the **survival function**,

$$S(t) = 1 - F(t) = \text{Prob}(T \geq t).$$

---

<sup>16</sup>See Petersen (1986) for one approach to this problem.

## 864 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

Consider the question raised in the introduction: Given that the spell has lasted until time  $t$ , what is the probability that it will end in the next short interval of time, say,  $\Delta t$ ? It is

$$l(t, \Delta t) = \text{Prob}(t \leq T \leq t + \Delta t | T \geq t).$$

A useful function for characterizing this aspect of the distribution is the **hazard rate**,

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{\text{Prob}(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{F(t + \Delta t) - F(t)}{\Delta t S(t)} = \frac{f(t)}{S(t)}.$$

Roughly, the hazard rate is the rate at which spells are completed after duration  $t$ , given that they last at least until  $t$ . As such, the hazard function gives an answer to our original question.

The hazard function, the density, the CDF, and the survival function are all related. The hazard function is

$$\lambda(t) = \frac{-d \ln S(t)}{dt},$$

so

$$f(t) = S(t)\lambda(t).$$

Another useful function is the **integrated hazard function**

$$\Lambda(t) = \int_0^t \lambda(s) ds,$$

for which

$$S(t) = e^{-\Lambda(t)},$$

so

$$\Lambda(t) = -\ln S(t).$$

The integrated hazard function is **generalized residual** in this setting. [See Chesher and Irish (1987) and Example 19.8.]

### 19.4.3.b Models of the Hazard Function

For present purposes, the hazard function is more interesting than the survival rate or the density. Based on the previous results, one might consider modeling the hazard function itself, rather than, say, modeling the survival function and then obtaining the density and the hazard. For example, the base case for many analyses is a hazard rate that does not vary over time. That is,  $\lambda(t)$  is a constant  $\lambda$ . This is characteristic of a process that has no memory; the *conditional* probability of “failure” in a given short interval is the same regardless of when the observation is made. Thus,

$$\lambda(t) = \lambda.$$

From the earlier definition, we obtain the simple differential equation,

$$\frac{-d \ln S(t)}{dt} = \lambda.$$

The solution is

$$\ln S(t) = k - \lambda t,$$

CHAPTER 19 ♦ Limited Dependent Variables **865**

or

$$S(t) = Ke^{-\lambda t},$$

where  $K$  is the constant of integration. The terminal condition that  $S(0) = 1$  implies that  $K = 1$ , and the solution is

$$S(t) = e^{-\lambda t}.$$

This solution is the **exponential** distribution, which has been used to model the time until failure of electronic components. Estimation of  $\lambda$  is simple, because with an exponential distribution,  $E[t] = 1/\lambda$ . The maximum likelihood estimator of  $\lambda$  would be the reciprocal of the sample mean.

A natural extension might be to model the hazard rate as a linear function,  $\lambda(t) = \alpha + \beta t$ . Then  $\Lambda(t) = \alpha t + \frac{1}{2}\beta t^2$  and  $f(t) = \lambda(t)S(t) = \lambda(t)\exp[-\Lambda(t)]$ . To avoid a negative hazard function, one might depart from  $\lambda(t) = \exp[g(t, \theta)]$ , where  $\theta$  is a vector of parameters to be estimated. With an observed sample of durations, estimation of  $\alpha$  and  $\beta$  is, at least in principle, a straightforward problem in maximum likelihood. [Kennan (1985) used a similar approach.]

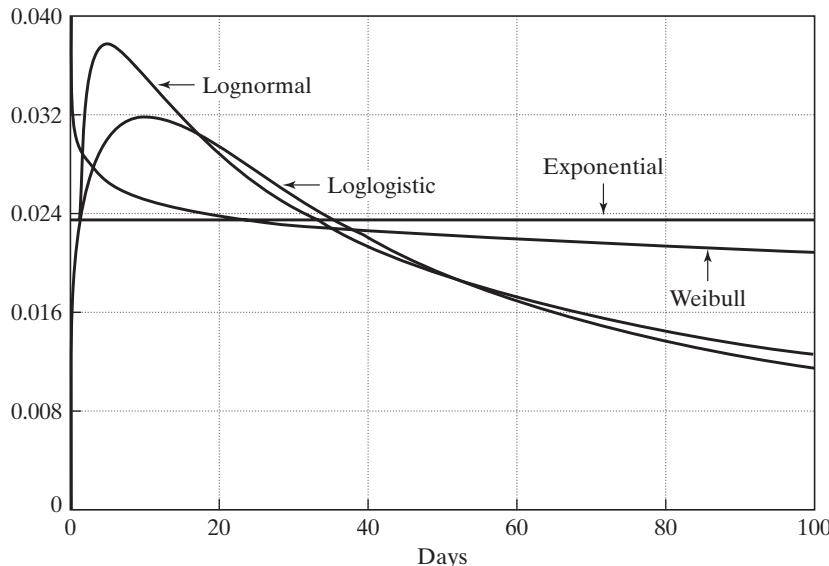
A distribution whose hazard function slopes upward is said to have **positive duration dependence**. For such distributions, the likelihood of failure at time  $t$ , conditional upon duration up to time  $t$ , is increasing in  $t$ . The opposite case is that of decreasing hazard or **negative duration dependence**. Our question in the introduction about whether the strike is more or less likely to end at time  $t$  given that it has lasted until time  $t$  can be framed in terms of positive or negative duration dependence. The assumed distribution has a considerable bearing on the answer. If one is unsure at the outset of the analysis whether the data can be characterized by positive or negative duration dependence, then it is counterproductive to assume a distribution that displays one characteristic or the other over the entire range of  $t$ . Thus, the exponential distribution and our suggested extension could be problematic. The literature contains a cornucopia of choices for duration models: normal, inverse normal [inverse Gaussian; see Lancaster (1990)], lognormal,  $F$ , gamma, Weibull (which is a popular choice), and many others.<sup>17</sup> To illustrate the differences, we will examine a few of the simpler ones. Table 19.5 lists the hazard functions and survival functions for four commonly used distributions. Each involves two parameters, a location parameter  $\lambda$ , and a scale parameter,  $p$ . [Note that in the benchmark case of the exponential distribution,  $\lambda$  is the hazard function. In all other cases, the hazard function is a function of  $\lambda$ ,  $p$ , and, where there is duration dependence,  $t$  as well. Different authors, for example, Kiefer (1988), use different parameterizations of these models. We follow the convention of Kalbfleisch and Prentice (2002).]

All these are distributions for a nonnegative random variable. Their hazard functions display very different behaviors, as can be seen in Figure 19.7. The hazard function for the exponential distribution is constant, that for the Weibull is monotonically increasing or decreasing depending on  $p$ , and the hazards for lognormal and loglogistic distributions first increase and then decrease. Which among these or the many alternatives is likely to be best in any application is uncertain.

<sup>17</sup>Three sources that contain numerous specifications are Kalbfleisch and Prentice (2002), Cox and Oakes (1985), and Lancaster (1990).

**866 PART IV ♦ Cross Sections, Panel Data, and Microeconometrics**
**TABLE 19.5** Survival Distributions

<i>Distribution</i>	<i>Hazard Function, <math>\lambda(t)</math></i>	<i>Survival Function, <math>S(t)</math></i>
Exponential	$\lambda$ ,	$S(t) = e^{-\lambda t}$
Weibull	$\lambda p(\lambda t)^{p-1}$ ,	$S(t) = e^{-(\lambda t)^p}$
Lognormal	$f(t) = (p/t)\phi[p \ln(\lambda t)]$ [ $\ln t$ is normally distributed with mean $-\ln \lambda$ and standard deviation $1/p$ .]	$S(t) = \Phi[-p \ln(\lambda t)]$
Loglogistic	$\lambda(t) = \lambda p(\lambda t)^{p-1}/[1 + (\lambda t)^p]$ , [ $\ln t$ has a logistic distribution with mean $-\ln \lambda$ and variance $\pi^2/(3p^2)$ .]	$S(t) = 1/[1 + (\lambda t)^p]$

**FIGURE 19.7** Parametric Hazard Functions.
**19.4.3.c Maximum Likelihood Estimation**

The parameters  $\lambda$  and  $p$  of these models can be estimated by maximum likelihood. For observed duration data,  $t_1, t_2, \dots, t_n$ , the log-likelihood function can be formulated and maximized in the ways we have become familiar with in earlier chapters. Censored observations can be incorporated as in Section 19.3 for the tobit model. [See (19-13).] As such,

$$\ln L(\theta) = \sum_{\text{uncensored observations}} \ln f(t | \theta) + \sum_{\text{censored observations}} \ln S(t | \theta),$$

where  $\theta = (\lambda, p)$ . For some distributions, it is convenient to formulate the log-likelihood function in terms of  $f(t) = \lambda(t)S(t)$  so that

$$\ln L = \sum_{\text{uncensored observations}} \ln \lambda(t | \theta) + \sum_{\text{all observations}} \ln S(t | \theta).$$

CHAPTER 19 ♦ Limited Dependent Variables **867**

Inference about the parameters can be done in the usual way. Either the BHHH estimator or actual second derivatives can be used to estimate asymptotic standard errors for the estimates. The transformation  $w = p(\ln t + \ln \lambda)$  for these distributions greatly facilitates maximum likelihood estimation. For example, for the Weibull model, by defining  $w = p(\ln t + \ln \lambda)$ , we obtain the very simple density  $f(w) = \exp[w - \exp(w)]$  and survival function  $S(w) = \exp(-\exp(w))$ .<sup>18</sup> Therefore, by using  $\ln t$  instead of  $t$ , we greatly simplify the log-likelihood function. Details for these and several other distributions may be found in Kalbfleisch and Prentice (2002, pp. 68–70). The Weibull distribution is examined in detail in the next section.

**19.4.3.d Exogenous Variables**

One limitation of the models given earlier is that external factors are not given a role in the survival distribution. The addition of “covariates” to duration models is fairly straightforward, although the interpretation of the coefficients in the model is less so. Consider, for example, the Weibull model. (The extension to other distributions will be similar.) Let

$$\lambda_i = e^{-\mathbf{x}'_i \boldsymbol{\beta}},$$

where  $\mathbf{x}_i$  is a constant term and a set of variables that are assumed not to change from time  $T = 0$  until the “failure time,”  $T = t_i$ . Making  $\lambda_i$  a function of a set of regressors is equivalent to changing the units of measurement on the time axis. For this reason, these models are sometimes called **accelerated failure time models**. Note as well that in all the models listed (and generally), the regressors do not bear on the question of duration dependence, which is a function of  $p$ .

Let  $\sigma = 1/p$  and let  $\delta_i = 1$  if the spell is completed and  $\delta_i = 0$  if it is censored. As before, let

$$w_i = p \ln(\lambda_i t_i) = \frac{(\ln t_i - \mathbf{x}'_i \boldsymbol{\beta})}{\sigma},$$

and denote the density and survival functions  $f(w_i)$  and  $S(w_i)$ . The observed random variable is

$$\ln t_i = \sigma w_i + \mathbf{x}'_i \boldsymbol{\beta}.$$

The Jacobian of the transformation from  $w_i$  to  $\ln t_i$  is  $d\ln t_i/d\ln w_i = 1/\sigma$ , so the density and survival functions for  $\ln t_i$  are

$$f(\ln t_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma) = \frac{1}{\sigma} f\left(\frac{\ln t_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right), \quad \text{and} \quad S(\ln t_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma) = S\left(\frac{\ln t_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right).$$

The log-likelihood for the observed data is

$$\ln L(\boldsymbol{\beta}, \sigma | \text{data}) = \sum_{i=1}^n [\delta_i \ln f(\ln t_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma) + (1 - \delta_i) \ln S(\ln t_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma)].$$

<sup>18</sup>The transformation is  $\exp(w) = (\lambda t)^p$  so  $t = (1/\lambda)[\exp(w)]^{1/p}$ . The Jacobian of the transformation is  $dt/dw = [\exp(w)]^{1/p}/(\lambda p)$ . The density in Table 19.5 is  $\lambda p[\exp(w)]^{-(1/p)-1}[\exp(-\exp(w))]$ . Multiplying by the Jacobian produces the result,  $f(w) = \exp[w - \exp(w)]$ . The survival function is the antiderivative,  $[\exp(-\exp(w))]$ .

## 868 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

For the **Weibull model**, for example (see footnote 18),

$$f(w_i) = \exp(w_i - e^{w_i}),$$

and

$$S(w_i) = \exp(-e^{w_i}).$$

Making the transformation to  $\ln t_i$  and collecting terms reduces the log-likelihood to

$$\ln L(\beta, \sigma | \text{data}) = \sum_i \left[ \delta_i \left( \frac{\ln t_i - \mathbf{x}'_i \beta}{\sigma} - \ln \sigma \right) - \exp \left( \frac{\ln t_i - \mathbf{x}'_i \beta}{\sigma} \right) \right].$$

(Many other distributions, including the others in Table 19.5, simplify in the same way. The exponential model is obtained by setting  $\sigma$  to one.) The derivatives can be equated to zero using the methods described in Section E.3. The individual terms can also be used to form the BHHH estimator of the asymptotic covariance matrix for the estimator.<sup>19</sup> The Hessian is also simple to derive, so Newton's method could be used instead.<sup>20</sup>

Note that the hazard function generally depends on  $t$ ,  $p$ , and  $\mathbf{x}$ . The sign of an estimated coefficient suggests the direction of the effect of the variable on the hazard function when the hazard is monotonic. But in those cases, such as the loglogistic, in which the hazard is nonmonotonic, even this may be ambiguous. The magnitudes of the effects may also be difficult to interpret in terms of the hazard function. In a few cases, we do get a regression-like interpretation. In the Weibull and exponential models,  $E[t | \mathbf{x}_i] = \exp(\mathbf{x}'_i \beta) \Gamma[(1/p) + 1]$ , whereas for the lognormal and loglogistic models,  $E[\ln t | \mathbf{x}_i] = \mathbf{x}'_i \beta$ . In these cases,  $\beta_k$  is the derivative (or a multiple of the derivative) of this conditional mean. For some other distributions, the conditional median of  $t$  is easily obtained. Numerous cases are discussed by Kiefer (1988), Kalbfleisch and Prentice (2002), and Lancaster (1990).

### 19.4.3.e Heterogeneity

The problem of **heterogeneity** in duration models can be viewed essentially as the result of an incomplete specification. Individual specific covariates are intended to incorporate observation specific effects. But if the model specification is incomplete and if systematic individual differences in the distribution remain after the observed effects are accounted for, then inference based on the improperly specified model is likely to be problematic. We have already encountered several settings in which the possibility of heterogeneity mandated a change in the model specification; the fixed and random effects regression, logit, and probit models all incorporate observation-specific effects. Indeed, all the failures of the linear regression model discussed in the preceding chapters can be interpreted as a consequence of heterogeneity arising from an incomplete specification.

There are a number of ways of extending duration models to account for heterogeneity. The strictly nonparametric approach of the Kaplan–Meier estimator (see Section 19.4.4) is largely immune to the problem, but it is also rather limited in how

<sup>19</sup> Note that the log-likelihood function has the same form as that for the tobit model in Section 19.3.2. By just reinterpreting the nonlimit observations in a tobit setting, we can, therefore, use this framework to apply a wide range of distributions to the tobit model. [See Greene (1995a) and references given therein.]

<sup>20</sup> See Kalbfleisch and Prentice (2002) for numerous other examples.

CHAPTER 19 ♦ Limited Dependent Variables **869**

much information can be culled from it. One direct approach is to model heterogeneity in the parametric model. Suppose that we posit a survival function conditioned on the individual specific effect  $v_i$ . We treat the survival function as  $S(t_i|v_i)$ . Then add to that a model for the unobserved heterogeneity  $f(v_i)$ . (Note that this is a counterpart to the incorporation of a disturbance in a regression model and follows the same procedures that we used in the Poisson model with random effects.) Then

$$S(t) = E_v[S(t|v)] = \int_v S(t|v) f(v) dv.$$

The gamma distribution is frequently used for this purpose.<sup>21</sup> Consider, for example, using this device to incorporate heterogeneity into the Weibull model we used earlier. As is typical, we assume that  $v$  has a gamma distribution with mean 1 and variance  $\theta = 1/k$ . Then

$$f(v) = \frac{k^k}{\Gamma(k)} e^{-kv} v^{k-1},$$

and

$$S(t|v) = e^{-(v\lambda t)^p}.$$

After a bit of manipulation, we obtain the unconditional distribution,

$$S(t) = \int_0^\infty S(t|v) f(v) dv = [1 + \theta(\lambda t)^p]^{-1/\theta}.$$

The limiting value, with  $\theta = 0$ , is the **Weibull survival model**, so  $\theta = 0$  corresponds to  $\text{Var}[v] = 0$ , or no heterogeneity.<sup>22</sup> The hazard function for this model is

$$\lambda(t) = \lambda p(\lambda t)^{p-1} [S(t)]^\theta,$$

which shows the relationship to the Weibull model.

This approach is common in parametric modeling of heterogeneity. In an important paper on this subject, Heckman and Singer (1984b) argued that this approach tends to overparameterize the survival distribution and can lead to rather serious errors in inference. They gave some dramatic examples to make the point. They also expressed some concern that researchers tend to choose the distribution of heterogeneity more on the basis of mathematical convenience than on any sensible economic basis.

#### 19.4.4 NONPARAMETRIC AND SEMIPARAMETRIC APPROACHES

The parametric models are attractive for their simplicity. But by imposing as much structure on the data as they do, the models may distort the estimated hazard rates. It may be that a more accurate representation can be obtained by imposing fewer restrictions.

---

<sup>21</sup>See, for example, Hausman, Hall, and Griliches (1984), who use it to incorporate heterogeneity in the Poisson regression model. The application is developed in Section 18.4.4.

<sup>22</sup>For the strike data analyzed in Figure 19.7, the maximum likelihood estimate of  $\theta$  is 0.0004, which suggests that at least in the context of the Weibull model, latent heterogeneity does not appear to be a feature of these data.

**870 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

The Kaplan–Meier (1958) **product limit estimator** is a strictly empirical, nonparametric approach to survival and hazard function estimation. Assume that the observations on duration are sorted in ascending order so that  $t_1 \leq t_2$  and so on and, for now, that no observations are censored. Suppose as well that there are  $K$  distinct survival times in the data, denoted  $T_k$ ;  $K$  will equal  $n$  unless there are ties. Let  $n_k$  denote the number of individuals whose observed duration is at least  $T_k$ . The set of individuals whose duration is at least  $T_k$  is called the **risk set** at this duration. (We borrow, once again, from biostatistics, where the risk set is those individuals still “at risk” at time  $T_k$ ). Thus,  $n_k$  is the size of the risk set at time  $T_k$ . Let  $h_k$  denote the number of observed spells completed at time  $T_k$ . A strictly empirical estimate of the survivor function would be

$$\hat{S}(T_k) = \prod_{i=1}^k \frac{n_i - h_i}{n_i} = \frac{n_i - h_i}{n_1}.$$

The estimator of the hazard rate is

$$\hat{\lambda}(T_k) = \frac{h_k}{n_k}. \quad (19-21)$$

Corrections are necessary for observations that are censored. Lawless (1982), Kalbfleisch and Prentice (2002), Kiefer (1988), and Greene (1995a) give details. Susin (2001) points out a fundamental ambiguity in this calculation (one which he argues appears in the 1958 source). The estimator in (19-21) is not a “rate” as such, as the width of the time window is undefined, and could be very different at different points in the chain of calculations. Because many intervals, particularly those late in the observation period, might have zeros, the failure to acknowledge these intervals should impart an upward bias to the estimator. His proposed alternative computes the counterpart to (19-21) over a mesh of defined intervals as follows:

$$\hat{\lambda}(I_a^b) = \frac{\sum_{j=a}^b h_j}{\sum_{j=a}^b n_j b_j},$$

where the interval is from  $t = a$  to  $t = b$ ,  $h_j$  is the number of failures in each period in this interval,  $n_j$  is the number of individuals at risk in that period and  $b_j$  is the width of the period. Thus, an interval  $(a, b)$  is likely to include several “periods.”

Cox’s (1972) approach to the **proportional hazard** model is another popular, **semi-parametric** method of analyzing the effect of covariates on the hazard rate. The model specifies that

$$\lambda(t_i) = \exp(\mathbf{x}'_i \boldsymbol{\beta}) \lambda_0(t_i)$$

The function  $\lambda_0$  is the “baseline” hazard, which is the individual heterogeneity. In principle, this hazard is a parameter for each observation that must be estimated. Cox’s **partial likelihood** estimator provides a method of estimating  $\boldsymbol{\beta}$  without requiring estimation of  $\lambda_0$ . The estimator is somewhat similar to Chamberlain’s estimator for the logit model with panel data in that a conditioning operation is used to remove the heterogeneity. (See Section 17.4.4.) Suppose that the sample contains  $K$  distinct exit times,  $T_1, \dots, T_K$ . For any time  $T_k$ , the risk set, denoted  $R_k$ , is all individuals whose exit time is at least  $T_k$ . The risk set is defined with respect to any moment in time  $T$  as the set of individuals who

CHAPTER 19 ♦ Limited Dependent Variables **871**

have not yet exited just prior to that time. For every individual  $i$  in risk set  $R_k$ ,  $t_i \geq T_k$ . The probability that an individual exits at time  $T_k$  given that exactly one individual exits at this time (which is the counterpart to the conditioning in the binary logit model in Chapter 17) is

$$\text{Prob}[t_i = T_k | \text{risk set}_k] = \frac{e^{\mathbf{x}'_i \boldsymbol{\beta}}}{\sum_{j \in R_k} e^{\mathbf{x}'_j \boldsymbol{\beta}}}.$$

Thus, the conditioning sweeps out the baseline hazard functions. For the simplest case in which exactly one individual exits at each distinct exit time and there are no censored observations, the partial log-likelihood is

$$\ln L = \sum_{k=1}^K \left[ \mathbf{x}'_k \boldsymbol{\beta} - \ln \sum_{j \in R_k} e^{\mathbf{x}'_j \boldsymbol{\beta}} \right].$$

If  $m_k$  individuals exit at time  $T_k$ , then the contribution to the log-likelihood is the sum of the terms for each of these individuals.

The proportional hazard model is a common choice for modeling durations because it is a reasonable compromise between the Kaplan–Meier estimator and the possibly excessively structured parametric models. Hausman and Han (1990) and Meyer (1988), among others, have devised other, “semiparametric” specifications for hazard models.

**Example 19.8 Survival Models for Strike Duration**

The strike duration data given in Kennan (1985, pp. 14–16) have become a familiar standard for the demonstration of hazard models. Appendix Table F19.2 lists the durations, in days, of 62 strikes that commenced in June of the years 1968 to 1976. Each involved at least 1,000 workers and began at the expiration or reopening of a contract. Kennan reported the actual duration. In his survey, Kiefer (1985), using the same observations, censored the data at 80 days to demonstrate the effects of censoring. We have kept the data in their original form; the interested reader is referred to Kiefer for further analysis of the censoring problem.<sup>23</sup>

Parameter estimates for the four duration models are given in Table 19.6. The estimate of the median of the survival distribution is obtained by solving the equation  $S(t) = 0.5$ . For example, for the Weibull model,

$$S(M) = 0.5 = \exp[-(\lambda M)^p],$$

or

$$M = [(\ln 2)^{1/p}] / \lambda.$$

For the exponential model,  $p = 1$ . For the lognormal and loglogistic models,  $M = 1/\lambda$ . The delta method is then used to estimate the standard error of this function of the parameter estimates. (See Section 4.4.4.) All these distributions are skewed to the right. As such,  $E[t]$  is greater than the median. For the exponential and Weibull models,  $E[t] = [1/\lambda]\Gamma[(1/p) + 1]$ ; for the normal,  $E[t] = (1/\lambda)[\exp(1/p^2)]^{1/2}$ . The implied hazard functions are shown in Figure 19.7.

The variable  $x$  reported with the strike duration data is a measure of unanticipated aggregate industrial production net of seasonal and trend components. It is computed as the residual in a regression of the log of industrial production in manufacturing on time, time squared, and monthly dummy variables. With the industrial production variable included as

<sup>23</sup>Our statistical results are nearly the same as Kiefer's despite the censoring.

**872 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**
**TABLE 19.6** Estimated Duration Models (estimated standard errors in parentheses)

	$\lambda$	$p$	<i>Median Duration</i>
Exponential	0.02344 (0.00298)	1.00000 (0.00000)	29.571 (3.522)
Weibull	0.02439 (0.00354)	0.92083 (0.11086)	27.543 (3.997)
Loglogistic	0.04153 (0.00707)	1.33148 (0.17201)	24.079 (4.102)
Lognormal	0.04514 (0.00806)	0.77206 (0.08865)	22.152 (3.954)

a covariate, the estimated Weibull model is

$$-\ln \lambda = 3.7772 - 9.3515x, \quad p = 1.00288, \\ (0.1394) (2.973) \quad (0.1217),$$

$$\text{median strike length} = 27.35(3.667) \text{ days}, E[t] = 39.83 \text{ days}.$$

Note that the Weibull model is now almost identical to the exponential model ( $p = 1$ ). Because the hazard conditioned on  $x$  is approximately equal to  $\lambda_i$ , it follows that the hazard function is increasing in “unexpected” industrial production. A 1 percent increase in  $x$  leads to a 9.35 percent increase in  $\lambda$ , which because  $p \approx 1$  translates into a 9.35 percent decrease in the median strike length or about 2.6 days. (Note that  $M = \ln 2/\lambda$ .)

The proportional hazard model does not have a constant term. (The baseline hazard is an individual specific constant.) The estimate of  $\beta$  is  $-9.0726$ , with an estimated standard error of 3.225. This is very similar to the estimate obtained for the Weibull model.

## 19.5 INCIDENTAL TRUNCATION AND SAMPLE SELECTION

The topic of sample selection, or **incidental truncation**, has been the subject of an enormous recent literature, both theoretical and applied.<sup>24</sup> This analysis combines both of the previous topics.

### *Example 19.9 Incidental Truncation*

In the high-income survey discussed in Example 19.2, respondents were also included in the survey if their net worth, not including their homes, was at least \$500,000. Suppose that the survey of incomes was based *only* on people whose net worth was at least \$500,000. This selection is a form of truncation, but not quite the same as in Section 19.2. This selection criterion does not necessarily exclude individuals whose incomes at the time might be quite low. Still, one would expect that, on average, individuals with a high net worth would have a high income as well. Thus, the average income in this subpopulation would in all likelihood also be misleading as an indication of the income of the typical American. The data in such a survey would be nonrandomly selected or incidentally truncated.

Econometric studies of nonrandom sampling have analyzed the deleterious effects of sample selection on the properties of conventional estimators such as least squares; have produced a variety of alternative estimation techniques; and, in the process, have

<sup>24</sup>A large proportion of the analysis in this framework has been in the area of labor economics. See, for example, Vella (1998), which is an extensive survey for practitioners. The results, however, have been applied in many other fields, including, for example, long series of stock market returns by financial economists (“survivorship bias”) and medical treatment and response in long-term studies by clinical researchers (“attrition bias”). Some studies that comment on methodological issues are Heckman (1990), Manski (1989, 1990, 1992), and Newey, Powell, and Walker (1990).

**CHAPTER 19 ♦ Limited Dependent Variables 873**

yielded a rich crop of empirical models. In some cases, the analysis has led to a reinterpretation of earlier results.

**19.5.1 INCIDENTAL TRUNCATION IN A BIVARIATE DISTRIBUTION**

Suppose that  $y$  and  $z$  have a bivariate distribution with correlation  $\rho$ . We are interested in the distribution of  $y$  given that  $z$  exceeds a particular value. Intuition suggests that if  $y$  and  $z$  are positively correlated, then the truncation of  $z$  should push the distribution of  $y$  to the right. As before, we are interested in (1) the form of the incidentally truncated distribution and (2) the mean and variance of the incidentally truncated random variable. Because it has dominated the empirical literature, we will focus first on the bivariate normal distribution.

The truncated *joint* density of  $y$  and  $z$  is

$$f(y, z | z > a) = \frac{f(y, z)}{\text{Prob}(z > a)}.$$

To obtain the incidentally truncated marginal density for  $y$ , we would then integrate  $z$  out of this expression. The moments of the incidentally truncated normal distribution are given in Theorem 19.5.<sup>25</sup>

**THEOREM 19.5 Moments of the Incidentally Truncated Bivariate Normal Distribution**

*If  $y$  and  $z$  have a bivariate normal distribution with means  $\mu_y$  and  $\mu_z$ , standard deviations  $\sigma_y$  and  $\sigma_z$ , and correlation  $\rho$ , then*

$$\begin{aligned} E[y | z > a] &= \mu_y + \rho\sigma_y\lambda(\alpha_z), \\ \text{Var}[y | z > a] &= \sigma_y^2[1 - \rho^2\delta(\alpha_z)], \end{aligned}$$

where

$$\alpha_z = (a - \mu_z)/\sigma_z, \lambda(\alpha_z) = \phi(\alpha_z)/[1 - \Phi(\alpha_z)], \text{ and } \delta(\alpha_z) = \lambda(\alpha_z)[\lambda(\alpha_z) - \alpha_z].$$

Note that the expressions involving  $z$  are analogous to the moments of the truncated distribution of  $x$  given in Theorem 19.2. If the truncation is  $z < a$ , then we make the replacement  $\lambda(\alpha_z) = -\phi(\alpha_z)/\Phi(\alpha_z)$ .

As expected, the truncated mean is pushed in the direction of the correlation if the truncation is from below and in the opposite direction if it is from above. In addition, the incidental truncation reduces the variance, because both  $\delta(\alpha)$  and  $\rho^2$  are between zero and one.

**19.5.2 REGRESSION IN A MODEL OF SELECTION**

To motivate a regression model that corresponds to the results in Theorem 19.5, we consider the following example.

---

<sup>25</sup>Much more general forms of the result that apply to multivariate distributions are given in Johnson and Kotz (1974). See also Maddala (1983, pp. 266–267).

## 874 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

### Example 19.10 A Model of Labor Supply

A simple model of female labor supply that has been examined in many studies consists of two equations:<sup>26</sup>

1. *Wage equation.* The difference between a person's *market wage*, what she could command in the labor market, and her *reservation wage*, the wage rate necessary to make her choose to participate in the labor market, is a function of characteristics such as age and education as well as, for example, number of children and where a person lives.
2. *Hours equation.* The desired number of labor hours supplied depends on the wage, home characteristics such as whether there are small children present, marital status, and so on.

The problem of truncation surfaces when we consider that the second equation describes desired hours, but an actual figure is observed only if the individual is working. (In most such studies, only a *participation equation*, that is, whether hours are positive or zero, is observable.) We infer from this that the market wage exceeds the reservation wage. Thus, the hours variable in the second equation is incidentally truncated.

To put the preceding examples in a general framework, let the equation that determines the sample selection be

$$z_i^* = \mathbf{w}'_i \boldsymbol{\gamma} + u_i,$$

and let the equation of primary interest be

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i.$$

The sample rule is that  $y_i$  is observed only when  $z_i^*$  is greater than zero. Suppose as well that  $\varepsilon_i$  and  $u_i$  have a bivariate normal distribution with zero means and correlation  $\rho$ . Then we may insert these in Theorem 19.5 to obtain the model *that applies to the observations in our sample*:

$$\begin{aligned} E[y_i | y_i \text{ is observed}] &= E[y_i | z_i^* > 0] \\ &= E[y_i | u_i > -\mathbf{w}'_i \boldsymbol{\gamma}] \\ &= \mathbf{x}'_i \boldsymbol{\beta} + E[\varepsilon_i | u_i > -\mathbf{w}'_i \boldsymbol{\gamma}] \\ &= \mathbf{x}'_i \boldsymbol{\beta} + \rho \sigma_\varepsilon \lambda_i(\alpha_u) \\ &= \mathbf{x}'_i \boldsymbol{\beta} + \beta_\lambda \lambda_i(\alpha_u), \end{aligned}$$

where  $\alpha_u = -\mathbf{w}'_i \boldsymbol{\gamma} / \sigma_u$  and  $\lambda(\alpha_u) = \phi(\mathbf{w}'_i \boldsymbol{\gamma} / \sigma_u) / \Phi(\mathbf{w}'_i \boldsymbol{\gamma} / \sigma_u)$ . So,

$$\begin{aligned} y_i | z_i^* > 0 &= E[y_i | z_i^* > 0] + v_i \\ &= \mathbf{x}'_i \boldsymbol{\beta} + \beta_\lambda \lambda_i(\alpha_u) + v_i. \end{aligned}$$

Least squares regression using the observed data—for instance, OLS regression of hours on its determinants, using only data for women who are working—produces inconsistent estimates of  $\boldsymbol{\beta}$ . Once again, we can view the problem as an omitted variable. Least squares regression of  $y$  on  $\mathbf{x}$  and  $\lambda$  would be a consistent estimator, but if  $\lambda$  is omitted, then the **specification error** of an omitted variable is committed. Finally, note that the second part of Theorem 19.5 implies that even if  $\lambda_i$  were observed, then least squares would be inefficient. The disturbance  $v_i$  is heteroscedastic.

<sup>26</sup>See, for example, Heckman (1976). This strand of literature begins with an exchange by Gronau (1974) and Lewis (1974).

CHAPTER 19 ♦ Limited Dependent Variables **875**

The marginal effect of the regressors on  $y_i$  in the observed sample consists of two components. There is the direct effect on the mean of  $y_i$ , which is  $\beta$ . In addition, for a particular independent variable, if it appears in the probability that  $z_i^*$  is positive, then it will influence  $y_i$  through its presence in  $\lambda_i$ . The full effect of changes in a regressor that appears in both  $\mathbf{x}_i$  and  $\mathbf{w}_i$  on  $y$  is

$$\frac{\partial E[y_i | z_i^* > 0]}{\partial x_{ik}} = \beta_k - \gamma_k \left( \frac{\rho \sigma_\varepsilon}{\sigma_u} \right) \delta_i(\alpha_u),$$

where<sup>27</sup>

$$\delta_i = \lambda_i^2 - \alpha_i \lambda_i.$$

Suppose that  $\rho$  is positive and  $E[y_i]$  is greater when  $z_i^*$  is positive than when it is negative. Because  $0 < \delta_i < 1$ , the additional term serves to reduce the marginal effect. The change in the probability affects the mean of  $y_i$  in that the mean in the group  $z_i^* > 0$  is higher. The second term in the derivative compensates for this effect, leaving only the marginal effect of a change given that  $z_i^* > 0$  to begin with. Consider Example 19.12, and suppose that education affects both the probability of migration and the income in either state. If we suppose that the income of migrants is higher than that of otherwise identical people who do not migrate, then the marginal effect of education has two parts, one due to its influence in increasing the probability of the individual's entering a higher-income group and one due to its influence on income within the group. As such, the coefficient on education in the regression overstates the marginal effect of the education of migrants and understates it for nonmigrants. The sizes of the various parts depend on the setting. It is quite possible that the magnitude, sign, and statistical significance of the effect might all be different from those of the estimate of  $\beta$ , a point that appears frequently to be overlooked in empirical studies.

In most cases, the selection variable  $z^*$  is not observed. Rather, we observe only its sign. To consider our two examples, we typically observe only whether a woman is working or not working or whether an individual migrated or not. We can infer the sign of  $z^*$ , but not its magnitude, from such information. Because there is no information on the scale of  $z^*$ , the disturbance variance in the selection equation cannot be estimated. (We encountered this problem in Chapter 17 in connection with the probit model.) Thus, we reformulate the model as follows:

$$\begin{aligned} \text{selection mechanism: } z_i^* &= \mathbf{w}'_i \boldsymbol{\gamma} + u_i, z_i = 1 \text{ if } z_i^* > 0 \text{ and } 0 \text{ otherwise;} \\ \text{Prob}(z_i = 1 | \mathbf{w}_i) &= \Phi(\mathbf{w}'_i \boldsymbol{\gamma}); \text{ and} \\ \text{Prob}(z_i = 0 | \mathbf{w}_i) &= 1 - \Phi(\mathbf{w}'_i \boldsymbol{\gamma}). \end{aligned} \tag{19-22}$$

regression model:  $y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i$  observed only if  $z_i = 1$ ,

$$(u_i, \varepsilon_i) \sim \text{bivariate normal } [0, 0, 1, \sigma_\varepsilon, \rho].$$

Suppose that, as in many of these studies,  $z_i$  and  $\mathbf{w}_i$  are observed for a random sample of individuals but  $y_i$  is observed only when  $z_i = 1$ . This model is precisely the one we

<sup>27</sup>We have reversed the sign of  $\alpha_u$  in (Theorem 19.5) because  $a = 0$ , and  $\alpha = \mathbf{w}' \boldsymbol{\gamma} / \sigma_M$  is somewhat more convenient. Also, as such,  $\partial \lambda / \partial \alpha = -\delta$ .

## 876 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

examined earlier, with

$$E[y_i | z_i = 1, \mathbf{x}_i, \mathbf{w}_i] = \mathbf{x}'_i \boldsymbol{\beta} + \rho \sigma_\varepsilon \lambda(\mathbf{w}'_i \boldsymbol{\gamma}).$$

### 19.5.3 TWO-STEP AND MAXIMUM LIKELIHOOD ESTIMATION

The parameters of the sample selection model can be estimated by maximum likelihood.<sup>28</sup> However, Heckman's (1979) **two-step estimation** procedure is usually used instead. Heckman's method is as follows:<sup>29</sup>

1. Estimate the probit equation by maximum likelihood to obtain estimates of  $\gamma$ . For each observation in the selected sample, compute  $\hat{\lambda}_i = \phi(\mathbf{w}'_i \hat{\boldsymbol{\gamma}})/\Phi(\mathbf{w}'_i \hat{\boldsymbol{\gamma}})$  and  $\hat{\delta}_i = \hat{\lambda}_i(\hat{\lambda}_i + \mathbf{w}'_i \hat{\boldsymbol{\gamma}})$ .
2. Estimate  $\boldsymbol{\beta}$  and  $\beta_\lambda = \rho \sigma_\varepsilon$  by least squares regression of  $y$  on  $\mathbf{x}$  and  $\hat{\lambda}$ .

It is possible also to construct consistent estimators of the individual parameters  $\rho$  and  $\sigma_\varepsilon$ . At each observation, the true conditional variance of the disturbance would be

$$\sigma_i^2 = \sigma_\varepsilon^2(1 - \rho^2 \delta_i).$$

The average conditional variance for the sample would converge to

$$\text{plim} \frac{1}{n} \sum_{i=1}^n \sigma_i^2 = \sigma_\varepsilon^2(1 - \rho^2 \bar{\delta}),$$

which is what is estimated by the least squares residual variance  $\mathbf{e}'\mathbf{e}/n$ . For the square of the coefficient on  $\lambda$ , we have

$$\text{plim } b_\lambda^2 = \rho^2 \sigma_\varepsilon^2,$$

whereas based on the probit results we have

$$\text{plim} \frac{1}{n} \sum_{i=1}^n \hat{\delta}_i = \bar{\delta}.$$

We can then obtain a consistent estimator of  $\sigma_\varepsilon^2$  using

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n} \mathbf{e}'\mathbf{e} + \hat{\delta} b_\lambda^2.$$

Finally, an estimator of  $\rho^2$  is

$$\hat{\rho}^2 = \frac{b_\lambda^2}{\hat{\sigma}_\varepsilon^2}, \tag{19-23}$$

which provides a complete set of estimators of the model's parameters.<sup>30</sup>

To test hypotheses, an estimate of the asymptotic covariance matrix of  $[\mathbf{b}', b_\lambda]$  is needed. We have two problems to contend with. First, we can see in Theorem 19.5 that

<sup>28</sup>See Greene (1995a).

<sup>29</sup>Perhaps in a mimicry of the "tobit" estimator described earlier, this procedure has come to be known as the "Heckit" estimator.

<sup>30</sup>Note that  $\hat{\rho}^2$  is not a sample correlation and, as such, is not limited to  $[0, 1]$ . See Greene (1981) for discussion.

CHAPTER 19 ♦ Limited Dependent Variables **877**

the disturbance term in

$$(y_i | z_i = 1, \mathbf{x}_i, \mathbf{w}_i) = \mathbf{x}'_i \boldsymbol{\beta} + \rho \sigma_\varepsilon \lambda_i + v_i \quad (19-24)$$

is heteroscedastic;

$$\text{Var}[v_i | z_i = 1, \mathbf{x}_i, \mathbf{w}_i] = \sigma_\varepsilon^2 (1 - \rho^2 \delta_i).$$

Second, there are unknown parameters in  $\lambda_i$ . Suppose that we assume for the moment that  $\lambda_i$  and  $\delta_i$  are known (i.e., we do not have to estimate  $\gamma$ ). For convenience, let  $\mathbf{x}_i^* = [\mathbf{x}_i, \lambda_i]$ , and let  $\mathbf{b}^*$  be the least squares coefficient vector in the regression of  $y$  on  $\mathbf{x}^*$  in the selected data. Then, using the appropriate form of the variance of ordinary least squares in a heteroscedastic model from Chapter 9, we would have to estimate

$$\begin{aligned} \text{Var}[\mathbf{b}^*] &= \sigma_\varepsilon^2 [\mathbf{X}'_* \mathbf{X}_*]^{-1} \left[ \sum_{i=1}^n (1 - \rho^2 \delta_i) \mathbf{x}_i^* \mathbf{x}_i^{*\prime} \right] [\mathbf{X}'_* \mathbf{X}_*]^{-1} \\ &= \sigma_\varepsilon^2 [\mathbf{X}'_* \mathbf{X}_*]^{-1} [\mathbf{X}'_* (\mathbf{I} - \rho^2 \Delta) \mathbf{X}_*] [\mathbf{X}'_* \mathbf{X}_*]^{-1}, \end{aligned}$$

where  $\mathbf{I} - \rho^2 \Delta$  is a diagonal matrix with  $(1 - \rho^2 \delta_i)$  on the diagonal. Without any other complications, this result could be computed fairly easily using  $\mathbf{X}$ , the sample estimates of  $\sigma_\varepsilon^2$  and  $\rho^2$ , and the assumed known values of  $\lambda_i$  and  $\delta_i$ .

The parameters in  $\gamma$  do have to be estimated using the probit equation. Rewrite (19-24) as

$$(y_i | z_i = 1, \mathbf{x}_i, \mathbf{w}_i) = \mathbf{x}'_i \boldsymbol{\beta} + \beta_\lambda \hat{\lambda}_i + v_i - \beta_\lambda (\hat{\lambda}_i - \lambda_i).$$

In this form, we see that in the preceding expression we have ignored both an additional source of variation in the compound disturbance and correlation across observations; the same estimate of  $\gamma$  is used to compute  $\hat{\lambda}_i$  for every observation. Heckman has shown that the earlier covariance matrix can be appropriately corrected by adding a term inside the brackets,

$$\mathbf{Q} = \hat{\rho}^2 (\mathbf{X}'_* \hat{\Delta} \mathbf{W}) \text{Est. Asy. Var}[\hat{y}] (\mathbf{W}' \hat{\Delta} \mathbf{X}_*) = \hat{\rho}^2 \hat{\mathbf{F}} \hat{\mathbf{V}} \hat{\mathbf{F}}',$$

where  $\hat{\mathbf{V}} = \text{Est. Asy. Var}[\hat{y}]$ , the estimator of the asymptotic covariance of the probit coefficients. Any of the estimators in (17-22) to (17-24) may be used to compute  $\hat{\mathbf{V}}$ . The complete expression is<sup>31</sup>

$$\text{Est. Asy. Var}[\mathbf{b}, b_\lambda] = \hat{\sigma}_\varepsilon^2 [\mathbf{X}'_* \mathbf{X}_*]^{-1} [\mathbf{X}'_* (\mathbf{I} - \hat{\rho}^2 \hat{\Delta}) \mathbf{X}_* + \mathbf{Q}] [\mathbf{X}'_* \mathbf{X}_*]^{-1}.$$

The sample selection model can also be estimated by maximum likelihood. The full log-likelihood function for the data is built up from

$$\text{Prob(selection)} \times \text{density} | \text{selection for observations with } z_i = 1,$$

and

$$\text{Prob}(nonselection) \text{ for observations with } z_i = 0.$$

<sup>31</sup>This matrix formulation is derived in Greene (1981). Note that the Murphy and Topel (1985) results for two-step estimators given in Theorem 14.8 would apply here as well. Asymptotically, this method would give the same answer. The Heckman formulation has become standard in the literature.

**878 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

Combining the parts produces the full log-likelihood function,

$$\ln L = \sum_{z=1} \ln \left[ \frac{\exp(-(1/2)\varepsilon_i^2/\sigma_\varepsilon^2)}{\sigma_\varepsilon \sqrt{2\pi}} \Phi \left( \frac{\rho\varepsilon_i/\sigma_\varepsilon + \mathbf{w}'_i \boldsymbol{\gamma}}{\sqrt{1-\rho^2}} \right) \right] + \sum_{z=0} [1 - \ln \Phi(\mathbf{w}'_i \boldsymbol{\gamma})],$$

where  $\varepsilon_i = y_i - \mathbf{x}'_i \boldsymbol{\beta}$ . Note, the FIML estimator with its assumption of bivariate normality is not less robust than the two-step estimator, because the latter also requires bivariate normality to form the conditional mean for the regression.

Two virtues of the FIML estimator will be the greater efficiency brought by using the likelihood function rather than the method of moments and, second, the estimation of  $\rho$  subject to the constraint  $-1 < \rho < 1$ . (This is typically done by reparameterizing the model in terms of the monotonic inverse hyperbolic tangent,  $\tau = (1/2) \ln [(1+\rho)/(1-\rho)] = \text{atanh}(\rho)$ . The transformed parameter,  $\tau$ , is unrestricted. The inverse transformation is  $\rho = [\exp(2\tau) - 1]/[\exp(2\tau) + 1]$  which is bounded between zero and one.) One possible drawback (it might be argued) could be the complexity of the likelihood function that would make estimation more difficult than the two-step estimator. However, the MLE for the selection model appears as a built-in procedure in modern software such as *Stata* and *NLOGIT*, and it is straightforward to implement in *Gauss* and *MatLab*, so this might be a moot point. Surprisingly, the MLE is by far less common than the two-step estimator in the received applications. The estimation of  $\rho$  is the difficult part of the estimation process (this is often the case). It is quite common for the method of moments estimator and the FIML estimator to be very different—our application in Example 19.11 is a case. Perhaps surprisingly so, the moment-based estimator of  $\rho$  in (19-23) is not bounded by zero and one. [See Greene (1981).] This would seem to recommend the MLE.

The fully parametric bivariate normality assumption of the model has been viewed as a potential drawback. However, relatively little progress has been made on devising informative semi- and nonparametric estimators—see, for one example, Gallant and Nychka (1987). The obstacle here is that, ultimately, the model hangs on a parameterization of the correlation of the unobservables in the two equations. So, method of moment estimators or kernel-based estimators must still incorporate this feature of a bivariate distribution. Some results have been obtained using the method of copula functions. [See Smith (2003, 2005) and Trivedi and Zimmer (2007).]

**Example 19.11 Female Labor Supply**

Examples 17.1 and 17.8 proposed a labor force participation model for a sample of 753 married women in a sample analyzed by Mroz (1987). The data set contains wage and hours information for the 428 women who participated in the formal market (*LFP*=1). Following Mroz, we suppose that for these 428 individuals, the offered wage exceeded the reservation wage and, moreover, the unobserved effects in the two wage equations are correlated. As such, a wage equation based on the market data should account for the sample selection problem. We specify a simple wage model:

$$\text{Wage} = \beta_1 + \beta_2 \text{Exper} + \beta_3 \text{Exper}^2 + \beta_4 \text{Education} + \beta_5 \text{City} + \varepsilon$$

where *Exper* is labor market experience and *City* is a dummy variable indicating that the individual lived in a large urban area. Maximum likelihood, Heckman two-step, and ordinary least squares estimates of the wage equation are shown in Table 19.7. The maximum likelihood estimates are FIML estimates—the labor force participation equation is reestimated at the same time. Only the parameters of the wage equation are shown next. Note as well that the two-step estimator estimates the single coefficient on  $\lambda_i$  and the structural parameters  $\sigma$

**TABLE 19.7** Estimated Selection Corrected Wage Equation

	<i>Two-Step</i>		<i>Maximum Likelihood</i>		<i>Least Squares</i>	
	<i>Estimate</i>	<i>Std. Err.</i>	<i>Estimate</i>	<i>Std. Err.</i>	<i>Estimate</i>	<i>Std. Err.</i>
$\beta_1$	-0.971	(2.06)	-1.963	(1.684)	-2.56	(0.929)
$\beta_2$	0.021	(0.0625)	0.0279	(0.0756)	0.0325	(0.0616)
$\beta_3$	0.000137	(0.00188)	-0.0001	(0.00234)	-0.000260	(0.00184)
$\beta_4$	0.417	(0.100)	0.457	(0.0964)	0.481	(0.0669)
$\beta_5$	0.444	(0.316)	0.447	(0.427)	(0.449)	0.318
$(\rho\sigma)$	-1.098	(1.266)				
$\rho$	-0.343		-0.132	(0.224)	0.000	
$\sigma$	3.200		3.108	(0.0837)	3.111	

and  $\rho$  are deduced by the method of moments. The maximum likelihood estimator computes estimates of these parameters directly. [Details on maximum likelihood estimation may be found in Maddala (1983).]

The differences between the two-step and maximum likelihood estimates in Table 19.7 are surprisingly large. The difference is even more striking in the marginal effects. The effect for education is estimated as  $0.417 + 0.0641$  for the two-step estimators and 0.480 in total for the maximum likelihood estimates. For the kids variable, the marginal effect is  $-.293$  for the two-step estimates and only  $-.11003$  for the MLEs. Surprisingly, the direct test for a selection effect in the maximum likelihood estimates, a nonzero  $\rho$ , fails to reject the hypothesis that  $\rho$  equals zero.

In some settings, the selection process is a nonrandom sorting of individuals into two or more groups. The mover-stayer model in the next example is a familiar case.

**Example 19.12 A Mover-Stayer Model for Migration**

The model of migration analyzed by Nakosteen and Zimmer (1980) fits into the framework described in this section. The equations of the model are

$$\begin{aligned} \text{net benefit of moving: } M_i^* &= \mathbf{w}'_i \boldsymbol{\gamma} + u_i, \\ \text{income if moves: } I_{i1} &= \mathbf{x}'_{i1} \boldsymbol{\beta}_1 + \varepsilon_{i1}, \\ \text{income if stays: } I_{i0} &= \mathbf{x}'_{i0} \boldsymbol{\beta}_0 + \varepsilon_{i0}. \end{aligned}$$

One component of the net benefit is the market wage individuals could achieve if they move, compared with what they could obtain if they stay. Therefore, among the determinants of the net benefit are factors that also affect the income received in either place. An analysis of income in a sample of migrants must account for the incidental truncation of the mover's income on a positive net benefit. Likewise, the income of the stayer is incidentally truncated on a nonpositive net benefit. The model implies an income after moving for all observations, but we observe it only for those who actually do move. Nakosteen and Zimmer (1980) applied the selectivity model to a sample of 9,223 individuals with data for two years (1971 and 1973) sampled from the Social Security Administration's Continuous Work History Sample. Over the period, 1,078 individuals migrated and the remaining 8,145 did not. The independent variables in the migration equation were as follows:

$SE$  = self-employment dummy variable; 1 if yes

$\Delta EMP$  = rate of growth of state employment

$\Delta PCI$  = growth of state per capita income

$x$  = age, race (nonwhite= 1), sex (female= 1)

$\Delta SIC$  = 1 if individual changes industry

**880 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**
**TABLE 19.8** Estimated Earnings Equations

	<i>Migration</i>	<i>Migrant Earnings</i>	<i>Nonmigrant Earnings</i>
Constant	-1.509	9.041	8.593
SE	-0.708 (-5.72)	-4.104 (-9.54)	-4.161 (-57.71)
$\Delta EMP$	-1.488 (-2.60)	—	—
$\Delta PCI$	1.455 (3.14)	—	—
Age	-0.008 (-5.29)	—	—
Race	-0.065 (-1.17)	—	—
Sex	-0.082 (-2.14)	—	—
$\Delta SIC$	0.948 (24.15)	-0.790 (-2.24)	-0.927 (-9.35)
$\lambda$	—	0.212 (0.50)	0.863 (2.84)

The earnings equations included  $\Delta S/C$  and  $SE$ . The authors reported the results given in Table 19.8. The figures in parentheses are asymptotic  $t$  ratios.

#### 19.5.4 SAMPLE SELECTION IN NONLINEAR MODELS

The preceding analysis has focused on an extension of the linear regression (or the estimation of simple averages of the data). The method of analysis changes in nonlinear models. To begin, it is not necessarily obvious what the impact of the sample selection is on the response variable, or how it can be accommodated in a model. Consider the model analyzed by Boyes, Hoffman, and Lowe (1989):

$$\begin{aligned}y_{i1} &= 1 \text{ if individual } i \text{ defaults on a loan, 0 otherwise,} \\y_{i2} &= 1 \text{ if the individual is granted a loan, 0 otherwise.}\end{aligned}$$

Wynand and van Praag (1981) also used this framework to analyze consumer insurance purchases in the first application of the selection methodology in a nonlinear model. Greene (1992) applied the same model to  $y_{i1}$  = default on credit card loans, in which  $y_{i2}$  denotes whether an application for the card was accepted or not. [Mohanty (2002) also used this model to analyze teen employment in California.] For a given individual,  $y_{i1}$  is not observed unless  $y_{i2} = 1$ . Following the lead of the linear regression case in Section 19.5.3, a natural approach might seem to be to fit the second (selection) equation using a univariate probit model, compute the inverse Mills ratio,  $\lambda_i$ , and add it to the first equation as an additional “control” variable to accommodate the selection effect. [This is the approach used by Wynand and van Praag (1981) and Greene (1994).] The problems with this control function approach are, first, it is unclear what in the model is being “controlled” and, second, assuming the first model is correct, the appropriate model conditioned on the sample selection is unlikely to contain an inverse Mills ratio anywhere in it. [See Terza (2010) for discussion.] That result is specific to the linear model, where it arises as  $E[\epsilon_i | \text{selection}]$ . What would seem to be the apparent counterpart for this probit model,

$$\text{Prob}(y_{i1} = 1 | \text{selection on } y_{i2} = 1) = \Phi(\mathbf{x}'_{i1} \boldsymbol{\beta}_1 + \theta \lambda_i),$$

is not, in fact, the appropriate conditional mean, or probability. For this particular application, the appropriate conditional probability (extending the bivariate probit model

of Section 17.5) would be

$$\text{Prob}[y_{i1} = 1 | y_{i2} = 1] = \frac{\Phi_2(\mathbf{x}'_{i1}\boldsymbol{\beta}_1, \mathbf{x}'_{i2}\boldsymbol{\beta}_2, \rho)}{\Phi(\mathbf{x}'_{i2}\boldsymbol{\beta}_2)}.$$

We would use this result to build up the likelihood function for the three observed outcomes, as follows: The three types of observations in the sample, with their unconditional probabilities, are

$$\begin{aligned} y_{i2} = 0: \text{Prob}(y_{i2} = 0 &| \mathbf{x}_{i1}, \mathbf{x}_{i2}) = 1 - \Phi(\mathbf{x}'_{i2}\boldsymbol{\beta}_2), \\ y_{i1} = 0, y_{i2} = 1: \text{Prob}(y_{i1} = 0, y_{i2} = 1 | \mathbf{x}_{i1}, \mathbf{x}_{i2}) &= \Phi_2(-\mathbf{x}'_{i1}\boldsymbol{\beta}_1, \mathbf{x}'_{i2}\boldsymbol{\beta}_2, -\rho), \quad (19-25) \\ y_{i1} = 1, y_{i2} = 1: \text{Prob}(y_{i1} = 1, y_{i2} = 1 | \mathbf{x}_{i1}, \mathbf{x}_{i2}) &= \Phi_2(\mathbf{x}'_{i1}\boldsymbol{\beta}_1, \mathbf{x}'_{i2}\boldsymbol{\beta}_2, \rho). \end{aligned}$$

The log-likelihood function is based on these probabilities.<sup>32</sup> An application appears in Section 17.5.6.

**Example 19.13 Doctor Visits and Insurance**

Continuing our analysis of the utilization of the German health care system, we observe that the data set contains an indicator of whether the individual subscribes to the “Public” health insurance or not. Roughly 87 percent of the observations in the sample do. We might ask whether the selection on public insurance reveals any substantive difference in visits to the physician. We estimated a logit specification for this model in Example 17.4. Using (19-25) as the framework, we define  $y_{i2}$  to be presence of insurance and  $y_{i1}$  to be the binary variable defined to equal 1 if the individual makes at least one visit to the doctor in the survey year.

The estimation results are given in Table 19.9. Based on these results, there does appear to be a very strong relationship. The coefficients do change somewhat in the conditional model. A Wald test for the presence of the selection effect against the null hypothesis that  $\rho$  equals zero produces a test statistic of  $(-7.188)^2 = 51.667$ , which is larger than the critical value of 3.84. Thus, the hypothesis is rejected. A likelihood ratio statistic is computed as the difference between the log-likelihood for the full model and the sum of the two separate log-likelihoods for the independent probit models when  $\rho$  equals zero. The result is

$$\lambda_{LR} = 2[-23969.58 - (-15536.39 + (-8471.508))] = 77.796$$

The hypothesis is rejected once again. Partial effects were computed using the results in Section 17.5.3.

The large correlation coefficient can be misleading. The estimated  $-0.9299$  does not state that the presence of insurance makes it much less likely to go to the doctor. This is the correlation among the unobserved factors in each equation. The factors that make it more likely to purchase insurance make it less likely to use a physician. To obtain a simple correlation between the two variables, we might use the tetrachoric correlation defined in Example 17.18. This would be computed by fitting a bivariate probit model for the two binary variables without any other variables. The estimated value is 0.120.

More general cases are typically much less straightforward. Greene (2005, 2006, 2010) and Terza (1998, 2010) present sample selection models for nonlinear specifications based on the underlying logic of the Heckman model in Section 19.5.3, that the influence of the incidental truncation acts on the unobservable variables in the model. (That is the source of the “selection bias” in conventional estimators.) The modeling extension introduces the unobservables into the model in a natural fashion that parallels the regression model. Terza (2010) presents a survey of the general results.

<sup>32</sup>Extensions of the bivariate probit model to other types of censoring are discussed in Poirier (1980) and Abowd and Farber (1982).

**882 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**
**TABLE 19.9** Estimated Probit Equations for Doctor Visits

<i>Variable</i>	<i>Independent: No Selection</i>			<i>Sample Selection Model</i>		
	<i>Estimate</i>	<i>Standard Error</i>	<i>Partial Effect</i>	<i>Estimate</i>	<i>Standard Error</i>	<i>Partial Effect</i>
Constant	0.05588	0.06564		-9.4366	0.06760	
Age	0.01331	0.0008399	0.004971	0.01284	0.0008131	0.005042
Income	-0.1034	0.05089	-0.03860	-0.1030	0.04582	-0.04060
Kids	-0.1349	0.01947	-0.05059	-0.1264	0.01790	-0.04979
Education	-0.01920	0.004254	-0.007170	0.03660	0.004744	0.002703
Married	0.03586	0.02172	0.01343	0.03564	0.02016	0.01404
ln L		-15536.39				
Constant	3.3585	0.06959		3.2699	0.06916	
Age	0.0001868	0.0009744		-0.0002679	0.001036	
Education	-0.1854	0.003941		-0.1807	0.003936	
Female	0.1150	0.02186	0.0000 <sup>a</sup>	0.2230	0.02101	0.01446 <sup>a</sup>
ln L		-8471.508				
$\rho$	0.0000	0.0000		-0.9299	0.1294	
ln L		-24007.90			-23969.58	

<sup>a</sup>Indirect effect from second equation.

The generic model will take the form

1. Probit selection equation:

$$\begin{aligned} z_i^* &= \mathbf{w}'_i \boldsymbol{\alpha} + u_i \text{ in which } u_i \sim N[0, 1], \\ z_i &= 1 \text{ if } z_i^* > 0, 0 \text{ otherwise.} \end{aligned} \quad (19-26)$$

2. Nonlinear index function model with unobserved heterogeneity and sample selection:

$$\begin{aligned} \mu_i | \varepsilon_i &= \mathbf{x}'_i \boldsymbol{\beta} + \sigma \varepsilon_i, \varepsilon_i \sim N[0, 1], \\ y_i | \mathbf{x}_i, \varepsilon_i &\sim \text{density } g(y_i | \mathbf{x}_i, \varepsilon_i) = f(y_i | \mathbf{x}'_i \boldsymbol{\beta} + \sigma \varepsilon_i), \\ y_i, \mathbf{x}_i &\text{ are observed only when } z_i = 1, \\ [u_i, \varepsilon_i] &\sim N[(0, 1), (1, \rho, 1)]. \end{aligned} \quad (19-27)$$

For example, in a Poisson regression model, the conditional mean function becomes  $E(y_i | \mathbf{x}_i) = \lambda_i = \exp(\mathbf{x}'_i \boldsymbol{\beta} + \sigma \varepsilon_i) = \exp(\mu_i)$ . (We used this specification of the model in Chapter 18 to introduce random effects in the Poisson regression model for panel data.)

The log-likelihood function for the full model is the joint density for the observed data. When  $z_i$  equals one,  $(y_i, \mathbf{x}_i, z_i, \mathbf{w}_i)$  are all observed. To obtain the joint density  $p(y_i, z_i = 1 | \mathbf{x}_i, \mathbf{w}_i)$ , we proceed as follows:

$$p(y_i, z_i = 1 | \mathbf{x}_i, \mathbf{w}_i) = \int_{-\infty}^{\infty} p(y_i, z_i = 1 | \mathbf{x}_i, \mathbf{w}_i, \varepsilon_i) f(\varepsilon_i) d\varepsilon_i.$$

Conditioned on  $\varepsilon_i$ ,  $z_i$  and  $y_i$  are independent. Therefore, the joint density is the product,

$$p(y_i, z_i = 1 | \mathbf{x}_i, \mathbf{w}_i, \varepsilon_i) = f(y_i | \mathbf{x}'_i \boldsymbol{\beta} + \sigma \varepsilon_i) \text{Prob}(z_i = 1 | \mathbf{w}_i, \varepsilon_i).$$

CHAPTER 19 ♦ Limited Dependent Variables **883**

The first part,  $f(y_i | \mathbf{x}'_i \boldsymbol{\beta} + \sigma \varepsilon_i)$  is the conditional index function model in (19-27). By joint normality,  $f(u_i | \varepsilon_i) = N[\rho \varepsilon_i, (1 - \rho^2)]$ , so  $u_i | \varepsilon_i = \rho \varepsilon_i + (u_i - \rho \varepsilon_i) = \rho \varepsilon_i + v_i$  where  $E[v_i] = 0$  and  $\text{Var}[v_i] = (1 - \rho^2)$ . Therefore,

$$\text{Prob}(z_i = 1 | \mathbf{w}_i, \varepsilon_i) = \Phi\left(\frac{\mathbf{w}'_i \boldsymbol{\alpha} + \rho \varepsilon_i}{\sqrt{1 - \rho^2}}\right).$$

Combining terms and using the earlier approach, the unconditional joint density is

$$p(y_i, z_i = 1 | \mathbf{x}_i, \mathbf{w}_i) = \int_{-\infty}^{\infty} f(y_i | \mathbf{x}'_i \boldsymbol{\beta} + \sigma \varepsilon_i) \Phi\left(\frac{\mathbf{w}'_i \boldsymbol{\alpha} + \rho \varepsilon_i}{\sqrt{1 - \rho^2}}\right) \frac{\exp(-\varepsilon_i^2/2)}{\sqrt{2\pi}} d\varepsilon_i. \quad (19-28)$$

The other part of the likelihood function for the observations with  $z_i = 0$  will be

$$\begin{aligned} \text{Prob}(z_i = 0 | \mathbf{w}_i) &= \int_{-\infty}^{\infty} \text{Prob}(z_i = 0 | \mathbf{w}_i, \varepsilon_i) f(\varepsilon_i) d\varepsilon_i. \\ &= \int_{-\infty}^{\infty} \left[ 1 - \Phi\left(\frac{\mathbf{w}'_i \boldsymbol{\alpha} + \rho \varepsilon_i}{\sqrt{1 - \rho^2}}\right) \right] f(\varepsilon_i) d\varepsilon_i \\ &= \int_{-\infty}^{\infty} \Phi\left(\frac{-(\mathbf{w}'_i \boldsymbol{\alpha} + \rho \varepsilon_i)}{\sqrt{1 - \rho^2}}\right) \frac{\exp(-\varepsilon_i^2/2)}{\sqrt{2\pi}} d\varepsilon_i. \end{aligned} \quad (19-29)$$

For convenience, we can use the invariance principle to reparameterize the likelihood function in terms of  $\gamma = \boldsymbol{\alpha}/\sqrt{1 - \rho^2}$  and  $\tau = \rho/\sqrt{1 - \rho^2}$ . Combining all the preceding terms, the log-likelihood function to be maximized is

$$\ln L = \sum_{i=1}^n \ln \int_{-\infty}^{\infty} [(1 - z_i) + z_i f(y_i | \mathbf{x}'_i \boldsymbol{\beta} + \sigma \varepsilon_i)] \Phi[(2z_i - 1)(\mathbf{w}'_i \gamma + \tau \varepsilon_i)] \phi(\varepsilon_i) d\varepsilon_i. \quad (19-30)$$

This can be maximized with respect to  $(\boldsymbol{\beta}, \sigma, \gamma, \tau)$  using quadrature or simulation. When done,  $\rho$  can be recovered from  $\rho = \tau / (1 + \tau^2)^{1/2}$  and  $\boldsymbol{\alpha} = (1 - \rho^2)^{1/2} \gamma$ . All that differs from one model to another is the specification of  $f(y_i | \mathbf{x}'_i \boldsymbol{\beta} + \sigma \varepsilon_i)$ . This is the specification used in Terza (1998) and Terza and Kenkel (2001). (In these two papers, the authors also analyzed  $E[y_i | z_i = 1]$ . This estimator was based on nonlinear least squares, but as earlier, it is necessary to integrate the unobserved heterogeneity out of the conditional mean function.) Greene (2010) applies the method to a stochastic frontier model.

#### 19.5.5 PANEL DATA APPLICATIONS OF SAMPLE SELECTION MODELS

The development of methods for extending sample selection models to panel data settings parallels the literature on cross-section methods. It begins with Hausman and Wise (1979) who devised a maximum likelihood estimator for a two-period model with attrition—the “selection equation” was a formal model for attrition from the sample. Subsequent research has drawn the analogy between attrition and sample selection in a variety of applications, such as Keane et al. (1988) and Verbeek and Nijman (1992), and produced theoretical developments including Wooldridge (2002a, b).

The direct extension of panel data methods to sample selection brings several new issues for the modeler. An immediate question arises concerning the nature of the

## 884 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

selection itself. Although much of the theoretical literature [e.g., Kyriazidou (1997, 2001)] treats the panel as if the selection mechanism is run anew in every period, in practice, the selection process often comes in two very different forms. First, selection may take the form of selection of the entire group of observations into the panel data set. Thus, the selection mechanism operates once, perhaps even before the observation window opens. Consider the entry (or not) of eligible candidates for a job training program. In this case, it is not appropriate to build the model to allow entry, exit, and then reentry. Second, for most applications, selection comes in the form of attrition or retention. Once an observation is “deselected,” it does not return. Leading examples would include “survivorship” in time-series–cross-section models of firm performance and attrition in medical trials and in panel data applications involving large national survey data bases, such as Contoyannis et al. (2004). Each of these cases suggests the utility of a more structured approach to the selection mechanism.

### 19.5.5.a Common Effects in Sample Selection Models

A formal “effects” treatment for sample selection was first suggested in complete form by Verbeek (1990), who formulated a random effects model for the probit equation and a fixed effects approach for the main regression. Zabel (1992) criticized the specification for its asymmetry in the treatment of the effects in the two equations. He also argued that the likelihood function that neglected correlation between the effects and regressors in the probit model would render the FIML estimator inconsistent. His proposal involved fixed effects in both equations. Recognizing the difficulty of fitting such a model, he then proposed using the Mundlak correction. The full model is

$$\begin{aligned} y_{it}^* &= \eta_i + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}, \quad \eta_i = \bar{\mathbf{x}}'_i\boldsymbol{\pi} + \tau w_i, w_i \sim N[0, 1], \\ d_{it}^* &= \theta_i + \mathbf{z}'_{it}\boldsymbol{\alpha} + u_{it}, \quad \theta_i = \bar{\mathbf{z}}'_i\boldsymbol{\delta} + \omega v_i, v_i \sim N[0, 1], \\ (\varepsilon_{it}, u_{it}) &\sim N_2[(0, 0), (\sigma^2, \mathbf{1}, \rho\sigma)]. \end{aligned} \tag{19-31}$$

The “selectivity” in the model is carried through the correlation between  $\varepsilon_{it}$  and  $u_{it}$ . The resulting log-likelihood is built up from the contribution of individual  $i$ ,

$$\begin{aligned} L_i &= \int_{-\infty}^{\infty} \prod_{d_{it}=0} \Phi[-\mathbf{z}'_{it}\boldsymbol{\alpha} - \bar{\mathbf{z}}'_i\boldsymbol{\delta} - \omega v_i] \phi(v_i) dv_i \\ &\quad \times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \prod_{d_{it}=1} \Phi \left[ \frac{\mathbf{z}'_{it}\boldsymbol{\alpha} + \bar{\mathbf{z}}'_i\boldsymbol{\delta} + \omega v_i + (\rho/\sigma)\varepsilon_{it}}{\sqrt{1 - \rho^2}} \right] \\ &\quad \times \frac{1}{\sigma} \phi \left( \frac{\varepsilon_{it}}{\sigma} \right) \phi_2(v_i, w_i) dv_i dw_i, \\ \varepsilon_{it} &= y_{it} - \mathbf{x}'_{it}\boldsymbol{\beta} - \bar{\mathbf{x}}'_i\boldsymbol{\pi} - \tau w_i. \end{aligned} \tag{19-32}$$

The log-likelihood is then  $\ln L = \sum_i \ln L_i$ .

The log-likelihood requires integration in two dimensions for any selected observations. Vella (1998) suggested two-step procedures to avoid the integration. However, the bivariate normal integration is actually the product of two univariate normals, because in the preceding specification,  $v_i$  and  $w_i$  are assumed to be uncorrelated. As such, the likelihood function in (19-32) can be readily evaluated using familiar simulation or quadrature techniques. [See Sections 14.9.6.c and 15.6. Vella and Verbeek (1999)

## CHAPTER 19 ♦ Limited Dependent Variables 885

suggest this in a footnote, but do not pursue it.] To show this, note that the first line in the log-likelihood is of the form  $E_v[\prod_{d=0} \Phi(\dots)]$  and the second line is of the form  $E_w[E_v[\Phi(\dots)\phi(\dots)/\sigma]]$ . Either of these expectations can be satisfactorily approximated with the average of a sufficient number of draws from the standard normal populations that generate  $w_i$  and  $v_i$ . The term in the simulated likelihood that follows this prescription is

$$\begin{aligned} L_i^S &= \frac{1}{R} \sum_{r=1}^R \prod_{d_i=0} \Phi[-\mathbf{z}'_i \boldsymbol{\alpha} - \bar{\mathbf{z}}'_i \boldsymbol{\delta} - \omega v_{i,r}] \\ &\quad \times \frac{1}{R} \sum_{r=1}^R \prod_{d_i=1} \Phi \left[ \frac{\mathbf{z}'_i \boldsymbol{\alpha} + \bar{\mathbf{z}}'_i \boldsymbol{\delta} + \omega v_{i,r} + (\rho/\sigma) \varepsilon_{it,r}}{\sqrt{1-\rho^2}} \right] \frac{1}{\sigma} \phi \left( \frac{\varepsilon_{it,r}}{\sigma} \right), \quad (19-33) \\ \varepsilon_{it,r} &= y_{it} - \mathbf{x}'_i \boldsymbol{\beta} - \bar{\mathbf{x}}'_i \boldsymbol{\pi} - \tau w_{i,r}. \end{aligned}$$

Maximization of this log-likelihood with respect to  $(\boldsymbol{\beta}, \sigma, \rho, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\pi}, \tau, \omega)$  by conventional gradient methods is quite feasible. Indeed, this formulation provides a means by which the likely correlation between  $v_i$  and  $w_i$  can be accommodated in the model. Suppose that  $w_i$  and  $v_i$  are bivariate standard normal with correlation  $\rho_{vw}$ . We can project  $w_i$  on  $v_i$  and write

$$w_i = \rho_{vw} v_i + (1 - \rho_{vw}^2)^{1/2} h_i,$$

where  $h_i$  has a standard normal distribution. To allow the correlation, we now simply substitute this expression for  $w_i$  in the simulated (or original) log-likelihood and add  $\rho_{vw}$  to the list of parameters to be estimated. The simulation is still over independent normal variates,  $v_i$  and  $h_i$ .

Notwithstanding the preceding derivation, much of the recent attention has focused on simpler two-step estimators. Building on Ridder and Wansbeek (1990) and Verbeek and Nijman (1992) [see Vella (1998) for numerous additional references], Vella and Verbeek (1999) propose a two-step methodology that involves a random effects framework similar to the one in (19-31). As they note, there is some loss in efficiency by not using the FIML estimator. But, with the sample sizes typical in contemporary panel data sets, that efficiency loss may not be large. As they note, their two-step template encompasses a variety of models including the tobit model examined in the preceding sections and the mover-stayer model noted earlier.

The Vella and Verbeek model requires some fairly intricate maximum likelihood procedures. Wooldridge (1995) proposes an estimator that, with a few probably—but not necessarily—innocent assumptions, can be based on straightforward applications of conventional, everyday methods. We depart from a fixed effects specification,

$$\begin{aligned} y_{it}^* &= \eta_i + \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_{it}, \\ d_{it}^* &= \theta_i + \mathbf{z}'_i \boldsymbol{\alpha} + u_{it}, \\ (\varepsilon_{it}, u_{it}) &\sim N_2[(0, 0), (\sigma^2, 1, \rho\sigma)]. \end{aligned}$$

Under the **mean independence assumption**  $E[\varepsilon_{it} | \eta_i, \theta_i, \mathbf{z}_{i1}, \dots, \mathbf{z}_{it}, v_{i1}, \dots, v_{it}, d_{i1}, \dots, d_{it}] = \rho u_{it}$ , it will follow that

$$E[y_{it} | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}, \eta_i, \theta_i, \mathbf{z}_{i1}, \dots, \mathbf{z}_{it}, v_{i1}, \dots, v_{it}, d_{i1}, \dots, d_{it}] = \eta_i + \mathbf{x}'_i \boldsymbol{\beta} + \rho u_{it}.$$

**886 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

This suggests an approach to estimating the model parameters; however, it requires computation of  $u_{it}$ . That would require estimation of  $\theta_i$ , which cannot be done, at least not consistently—and that precludes simple estimation of  $u_{it}$ . To escape the dilemma, Wooldridge (2002c) suggests Chamberlain's approach to the fixed effects model,

$$\theta_i = f_0 + \mathbf{z}'_{i1}\mathbf{f}_1 + \mathbf{z}'_{i2}\mathbf{f}_2 + \cdots + \mathbf{z}'_{iT}\mathbf{f}_T + h_i.$$

With this substitution,

$$\begin{aligned} d_{it}^* &= \mathbf{z}'_{it}\boldsymbol{\alpha} + f_0 + \mathbf{z}'_{i1}\mathbf{f}_1 + \mathbf{z}'_{i2}\mathbf{f}_2 + \cdots + \mathbf{z}'_{iT}\mathbf{f}_T + h_i + u_{it} \\ &= \mathbf{z}'_{it}\boldsymbol{\alpha} + f_0 + \mathbf{z}'_{i1}\mathbf{f}_1 + \mathbf{z}'_{i2}\mathbf{f}_2 + \cdots + \mathbf{z}'_{iT}\mathbf{f}_T + w_{it}, \end{aligned}$$

where  $w_{it}$  is independent of  $\mathbf{z}_{it}$ ,  $t = 1, \dots, T$ . This now implies that

$$\begin{aligned} E[y_{it} | \mathbf{x}_{i1}, \dots, \mathbf{x}_{it}, \eta_i, \theta_i, \mathbf{z}_{i1}, \dots, \mathbf{z}_{it}, v_{i1}, \dots, v_{it}, d_{i1}, \dots, d_{it}] &= \eta_i + \mathbf{x}'_{it}\boldsymbol{\beta} + \rho(w_{it} - h_i) \\ &= (\eta_i - \rho h_i) + \mathbf{x}'_{it}\boldsymbol{\beta} + \rho w_{it}. \end{aligned}$$

To complete the estimation procedure, we now compute  $T$  cross-sectional probit models (reestimating  $f_0, \mathbf{f}_1, \dots$  each time) and compute  $\hat{\lambda}_{it}$  from each one. The resulting equation,

$$y_{it} = a_i + \mathbf{x}'_{it}\boldsymbol{\beta} + \rho\hat{\lambda}_{it} + v_{it},$$

now forms the basis for estimation of  $\boldsymbol{\beta}$  and  $\rho$  by using a conventional fixed effects linear regression with the observed data.

#### 19.5.5.b Attrition

The recent literature or sample selection contains numerous analyses of two-period models, such as Kyriazidou (1997, 2001). They generally focus on non- and semiparametric analyses. An early parametric contribution of Hausman and Wise (1979) is also a two-period model of attrition, which would seem to characterize many of the studies suggested in the current literature. The model formulation is a two-period random effects specification:

$$\begin{aligned} y_{i1} &= \mathbf{x}'_{i1}\boldsymbol{\beta} + \varepsilon_{i1} + u_i \quad (\text{first period regression}), \\ y_{i2} &= \mathbf{x}'_{i2}\boldsymbol{\beta} + \varepsilon_{i2} + u_i \quad (\text{second period regression}). \end{aligned}$$

Attrition is likely in the second period (to begin the study, the individual must have been observed in the first period). The authors suggest that the probability that an observation is made in the second period varies with the value of  $y_{i2}$  as well as some other variables,

$$z_{i2}^* = \delta y_{i2} + \mathbf{x}'_{i2}\boldsymbol{\theta} + \mathbf{w}'_{i2}\boldsymbol{\alpha} + v_{i2}.$$

Attrition occurs if  $z_{i2}^* \leq 0$ , which produces a probit model,

$$z_{i2} = 1(z_{i2}^* > 0) \quad (\text{attrition indicator observed in period 2}).$$

An observation is made in the second period if  $z_{i2} = 1$ , which makes this an early version of the familiar sample selection model. The reduced form of the observation

equation is

$$\begin{aligned} z_{i2}^* &= \mathbf{x}'_{i2}(\delta\beta + \theta) + \mathbf{w}'_{i2}\alpha + \delta\varepsilon_{i2} + v_{i2} \\ &= \mathbf{x}'_{i2}\pi + \mathbf{w}'_{i2}\alpha + h_{i2} \\ &= \mathbf{r}'_{i2}\gamma + h_{i2}. \end{aligned}$$

The variables in the probit equation are all those in the second period regression plus any additional ones dictated by the application. The estimable parameters in this model are  $\beta$ ,  $\gamma$ ,  $\sigma^2 = \text{Var}[\varepsilon_{it} + u_i]$ , and two correlation coefficients,

$$\rho_{12} = \text{Corr}[\varepsilon_{i1} + u_i, \varepsilon_{i2} + u_i] = \text{Var}[u_i]/\sigma^2,$$

and

$$\rho_{23} = \text{Corr}[h_{i2}, \varepsilon_{i2} + u_i].$$

All disturbances are assumed to be normally distributed. (Readers are referred to the paper for motivation and details on this specification.)

The authors propose a full information maximum likelihood estimator. Estimation can be simplified somewhat by using two steps. The parameters of the probit model can be estimated first by maximum likelihood. Then the remaining parameters are estimated by maximum likelihood, conditionally on these first-step estimates. The Murphy and Topel adjustment is made after the second step. [See Greene (2007a).]

The Hausman and Wise model covers the case of two periods in which there is a formal mechanism in the model for retention in the second period. It is unclear how the procedure could be extended to a multiple-period application such as that in Contoyannis et al. (2004), which involved a panel data set with eight waves. In addition, in that study, the variables in the main equations were counts of hospital visits and physician visits, which complicates the use of linear regression. A workable solution to the problem of attrition in a multiperiod panel is the **inverse probability weighted estimator** [Wooldridge (2002a, 2006b) and Rotnitzky and Robins (2005).] In the Contoyannis application, there are eight waves in the panel. Attrition is taken to be “ignorable” so that the unobservables in the attrition equation and in the main equation(s) of interest are uncorrelated. (Note that Hausman and Wise do not make this assumption.) This enables Contoyannis et al. to fit a “retention” probit equation for each observation present at wave 1, for waves 2–8, using characteristics observed at the entry to the panel. (This defines, then, “selection (retention) on observables.”) Defining  $d_{it}$  to be the indicator for presence ( $d_{it} = 1$ ) or absence ( $d_{it} = 0$ ) of observation  $i$  in wave  $t$ , it will follow that the sequence of observations will begin at 1 and either stay at 1 or change to 0 for the remaining waves. Let  $\hat{p}_{it}$  denote the predicted probability from the probit estimator at wave  $t$ . Then, their full log-likelihood is constructed as

$$\ln L = \sum_{i=1}^n \sum_{t=1}^T \frac{d_{it}}{\hat{p}_{it}} \ln L_{it}.$$

Wooldridge (2002b) presents the underlying theory for the properties of this weighted maximum likelihood estimator. [Further details on the use of the inverse probability weighted estimator in the Contoyannis et al. (2004) study appear in Jones, Koolman, and Rice (2006) and in Section 17.4.9.]

**888 PART IV ♦ Cross Sections, Panel Data, and Microeometrics****19.6 EVALUATING TREATMENT EFFECTS**

The leading recent application of models of selection and endogeneity is the evaluation of “**treatment effects**.” The central focus is on analysis of the effect of participation in a treatment,  $T$ , on an outcome variable,  $y$ —examples include job training programs [LaLonde (1986), Business Week (2009; Example 19.14)] and education [e.g., test scores, Angrist and Lavy (1999), Van der Klaauw (2002)]. Wooldridge and Imbens (2009, pp. 22–23) cite a number of labor market applications. Recent more narrow examples include Munkin and Trivedi’s (2007) analysis of the effect of dental insurance and Jones and Rice’s (2010) survey that notes a variety of techniques and applications in health economics.

***Example 19.14 German Labor Market Interventions***

“Germany long had the highest ratio of unfilled jobs to unemployed people in Europe. Then, in 2003, Berlin launched the so-called Hartz reforms, ending generous unemployment benefits that went on indefinitely. Now payouts for most recipients drop sharply after a year, spurring people to look for work. From 12.7% in 2005, unemployment fell to 7.1% last November. Even now, after a year of recession, Germany’s jobless rate has risen to just 8.6%.

At the same time, lawmakers introduced various programs intended to make it easier for people to learn new skills. One initiative instructed the Federal Labor Agency, which had traditionally pushed the long-term unemployed into government-funded make-work positions, to cooperate more closely with private employers to create jobs. That program last year paid Dutch staffing agency Randstad to teach 15,000 Germans information technology, business English, and other skills. And at a Daimler truck factory in Wörth, 55 miles west of Stuttgart, several dozen short-term employees at risk of being laid off got government help to continue working for the company as mechanic trainees.

Under a second initiative, Berlin pays part of the wages of workers hired from the ranks of the jobless. Such payments make employers more willing to take on the costs of training new workers. That extra training, in turn, helps those workers keep their jobs after the aid expires, a study by the government-funded Institute for Employment Research found. Café Nenninger in the city of Kassel, for instance, used the program to train an unemployed single mother. Co-owner Verena Nenninger says she was willing to take a chance on her in part because the government picked up about a third of her salary the first year. ‘It was very helpful, because you never know what’s going to happen,’ Nenninger says” [Business Week (2009)].

Empirical measurement of treatment effects, such as the impact of going to college or participating in a job training program, presents a large variety of econometric complications. The natural, ultimate objective of an analysis of a “treatment” or intervention would be the “effect of treatment on the treated.” For example, what is the effect of a college education on the lifetime income of someone who goes to college? Measuring this effect econometrically encounters at least two compelling computations:

**Endogeneity of the treatment:** The analyst risks attributing to the treatment causal effects that should be attributed to factors that motivate both the treatment and the outcome. In our example, the individual who goes to college might well have succeeded (more) in life than their counterpart who did not go to college even if they (themselves) did not attend college.

**Missing counterfactual:** The preceding thought experiment is not actually the effect we wish to measure. In order to measure the impact of college attendance on lifetime earnings in a pure sense, we would have to run an individual’s lifetime twice, once with

CHAPTER 19 ♦ Limited Dependent Variables **889**

college attendance and once without. Any individual is observed in only one of the two states, so the pure measurement is impossible.

Accommodating these two problems forms the focal point of this enormous and still growing literature. **Rubin's causal model** (1974, 1978) provides a useful framework for the analysis. Every individual in a population has a potential outcome,  $y$  and can be exposed to the treatment,  $C$ . We will denote by  $C_i$  the indicator whether or not the individual receives the treatment. Thus, the potential outcomes are  $y_i | (C_i = 1) = y_{i1}$  and  $y_i | (C_i = 0) = y_{i0}$ . The **average treatment effect**, averaged across the entire population is

$$\text{ATE} = E[y_{i1} - y_{i0}].$$

The compelling complication is that the individual will exist in only one of the two states, so it is not possible to estimate ATE without further assumptions. More specifically, what the researcher would prefer see is the **average treatment effect on the treated**,

$$\text{ATET} = E[y_{i1} - y_{i0} | C_i = 1]$$

and note that the second term is the missing counterfactual.

One of the major themes of the recent research is to devise robust methods of estimation that do not rely heavily on fragile assumptions such as identification by functional form (e.g., relying on bivariate normality) and identification by exclusion restrictions (e.g., relying on basic instrumental variable estimators). This is a challenging exercise—we have relied heavily on these assumptions in most of the work in this book up to this point. For purposes of the general specification, we will denote by  $\mathbf{x}$  the exogenous information that will be brought to bear on this estimation problem. The vector  $\mathbf{x}$  may (usually will) be a set of variables that will appear in a regression model, but it is useful to think more generally than that and consider  $\mathbf{x}$  rather to be an information set. Certain minimal assumptions are necessary to make any headway at all. The following appear at different points in the analysis.

**Conditional independence:** Receiving the treatment,  $C_i$ , does not depend on the outcome variable once the effect of  $\mathbf{x}$  on the outcome is accounted for. If assignment to the treatment group is completely random, then we would omit the effect of  $\mathbf{x}$  in this assumption. This assumption is extended for regression approaches with the **conditional mean assumption**:  $E[y_{i0} | \mathbf{x}_i, C_i = 1] = E[y_{i0} | \mathbf{x}_i, C_i = 0] = E[y_{i0} | \mathbf{x}]$ . This states that the outcome in the untreated state does not affect the participation.

**Distribution of potential outcomes:** The model that is used for the outcomes is the same for treated and nontreated,  $f(y | \mathbf{x}, T = 1) = f(y | \mathbf{x}, T = 0)$ . In a regression context, this would mean that the same regression applies in both states and that the disturbance is uncorrelated with  $T$ , or that  $T$  is exogenous. This is a very strong assumption that we will relax later. For the present, it removes one of the complications noted previously, so a step in the model-building exercise will be to relax this assumption.

**Overlap assumption:** For any value of  $\mathbf{x}$ ,  $0 < \text{Prob}(C_i = 1 | \mathbf{x}) < 1$ . The strict inequality in this assumption means that for any  $\mathbf{x}$ , the population will contain a mix of treated and nontreated individuals. The usefulness of the overlap assumption is that with it, we can expect to find, for any treated individual, an individual who looks like them but is not treated. This assumption will be useful for regression approaches.

## 890 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

The following sections will describe three major parts of the research agenda on treatment effects: regression analysis with control functions in Section 19.6.1, propensity score matching in Section 19.6.2, and regression discontinuity design in Section 19.6.3. A fourth area, instrumental variable estimation, was developed in Chapter 8. As noted, this is a huge and rapidly growing literature. For example, Imbens and Wooldridge's (2009) survey paper runs 85 pages and includes nearly 300 references, most of them since 2000. Our purpose here is to provide some of the vocabulary and a superficial introduction to methods. The survey papers by Imbens and Wooldridge (2009) and Jones and Rice (2010) provide greater detail. The conference volume by Millment, Smith, and Vytlacil (2008) contains many theoretical contributions and empirical applications.<sup>33</sup> A *Journal of Business and Economic Statistics* symposium [Angrist (2001)] raised many of the important questions on whether and how it is possible to measure treatment effects.

### 19.6.1 REGRESSION ANALYSIS OF TREATMENT EFFECTS

The basic model of selectivity outlined earlier has been extended in an impressive variety of directions. An interesting application that has found wide use is the measurement of **treatment effects** and program effectiveness.

An earnings equation that accounts for the value of a college education is

$$\text{earnings}_i = \mathbf{x}'_i \boldsymbol{\beta} + \delta C_i + \varepsilon_i,$$

where  $C_i$  is a dummy variable indicating whether or not the individual attended college. The same format has been used in any number of other analyses of programs, experiments, and treatments. The question is: Does  $\delta$  measure the value of a college education (assuming that the rest of the regression model is correctly specified)? The answer is no if the typical individual who chooses to go to college would have relatively high earnings whether or not he or she went to college. The problem is one of self-selection. If our observation is correct, then least squares estimates of  $\delta$  will actually overestimate the treatment effect. The same observation applies to estimates of the treatment effects in other settings in which the individuals themselves decide whether or not they will receive the treatment.

To put this in a more familiar context, suppose that we model program participation (e.g., whether or not the individual goes to college) as

$$\begin{aligned} C_i^* &= \mathbf{w}'_i \boldsymbol{\gamma} + u_i, \\ C_i &= 1 \text{ if } C_i^* > 0, 0 \text{ otherwise.} \end{aligned}$$

We also suppose that, consistent with our previous conjecture,  $u_i$  and  $\varepsilon_i$  are correlated. Coupled with our earnings equation, we find that

$$\begin{aligned} E[y_i | C_i = 1, \mathbf{x}_i, \mathbf{w}_i] &= \mathbf{x}'_i \boldsymbol{\beta} + \delta + E[\varepsilon_i | C_i = 1, \mathbf{x}_i, \mathbf{w}_i] \\ &= \mathbf{x}'_i \boldsymbol{\beta} + \delta + \rho \sigma_\varepsilon \lambda(-\mathbf{w}'_i \boldsymbol{\gamma}) \end{aligned} \tag{19-34}$$

once again. [See (19-24).] Evidently, a viable strategy for estimating this model is to use the two-step estimator discussed earlier. The net result will be a different estimate of  $\delta$

---

<sup>33</sup>In the initial essay in the volume, Goldberger (2008) reproduces Goldberger (1972) in which the author explores the endogeneity issue in detail with specific reference to the Head Start program of the 1960s.

CHAPTER 19 ♦ Limited Dependent Variables **891**

that will account for the self-selected nature of program participation. For nonparticipants, the counterpart to (19-34) is

$$E[y_i | C_i = 0, \mathbf{x}_i, \mathbf{w}_i] = \mathbf{x}'_i \boldsymbol{\beta} + \rho \sigma_\varepsilon \left[ \frac{-\phi(\mathbf{w}'_i \boldsymbol{\gamma})}{1 - \Phi(\mathbf{w}'_i \boldsymbol{\gamma})} \right]. \quad (19-35)$$

The difference in expected earnings between participants and nonparticipants is, then,

$$E[y_i | C_i = 1, \mathbf{x}_i, \mathbf{w}_i] - E[y_i | C_i = 0, \mathbf{x}_i, \mathbf{w}_i] = \delta + \rho \sigma_\varepsilon \left[ \frac{\phi_i}{\Phi_i(1 - \Phi_i)} \right]. \quad (19-36)$$

If the selectivity correction  $\lambda_i$  is omitted from the least squares regression, then this difference is what is estimated by the least squares coefficient on the treatment dummy variable. But because (by assumption) all terms are positive, we see that least squares overestimates the treatment effect. Note, finally, that simply estimating separate equations for participants and nonparticipants does not solve the problem. In fact, doing so would be equivalent to estimating the two regressions of Example 19.12 by least squares, which, as we have seen, would lead to inconsistent estimates of both sets of parameters.

To describe the problem created by **selection on the unobservables**, we will drop the independence assumptions. The model with endogenous participation and different outcome equations would be

$$\begin{aligned} C_i^* &= \mathbf{w}'_i \boldsymbol{\gamma} + u_i, \quad C_i = 1 \text{ if } C_i^* > 0 \text{ and 0 otherwise,} \\ y_{i0} &= \mathbf{x}'_i \boldsymbol{\beta}_0 + \varepsilon_{i0}, \\ y_{i1} &= \mathbf{x}'_i \boldsymbol{\beta}_1 + \varepsilon_{i1}. \end{aligned}$$

It is useful to combine the second and third equations in

$$y_{ij} = C_i(\mathbf{x}'_i \boldsymbol{\beta}_1 + \varepsilon_{i1}) + (1 - C_i)(\mathbf{x}'_i \boldsymbol{\beta}_0 + \varepsilon_{i0}), \quad j = 0, 1.$$

We assume joint normality for the three disturbances;

$$\begin{pmatrix} u_i \\ \varepsilon_{i0} \\ \varepsilon_{i1} \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_0 \theta_0 & \rho_1 \theta_1 \\ \rho_0 \theta_0 & \theta_0^2 & \theta_{01} \\ \rho_1 \theta_1 & \theta_{01} & \theta_1^2 \end{pmatrix} \right].$$

The variance in the participation equation is normalized to one for a binary outcome, as described earlier (Section 17.2). Endogeneity of the participation is implied by the nonzero values of the correlations  $\rho_0$  and  $\rho_1$ . The familiar problem of the missing counterfactual appears here in our inability to estimate  $\theta_{01}$ . The data will never contain information on both states simultaneously, so it will be impossible to estimate a covariance of  $y_{i0}$  and  $y_{i1}$  (conditioned on  $\mathbf{x}_i$  or otherwise). Thus, the parameter  $\theta_{01}$  is not identified (estimable)—we normalize it to zero. The parameters of this model after the two normalizations can be estimated by two-step least squares as suggested in Section 19.XX, or by full information maximum likelihood. The average treatment effect on the treated would be

$$\begin{aligned} \text{ATET} &= E[y_{i1} | C_i = 1, \mathbf{x}_i, \mathbf{w}_i] - E[y_{i0} | C_i = 1, \mathbf{x}_i, \mathbf{w}_i] \\ &= \mathbf{x}'_i (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + (\rho_1 \theta_1 - \rho_0 \theta_0) \frac{\phi(\mathbf{w}'_i \boldsymbol{\gamma})}{\Phi(\mathbf{w}'_i \boldsymbol{\gamma})}. \end{aligned}$$

## 892 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

[See (19-34).] If the treatment assignment is completely random, then  $\rho_1 = \rho_0 = 0$ , and we are left with the first term. But, of course, it is the nonrandomness of the treatment assignment that brought us to this point. Finally, if the two coefficient vectors differ only in their constant terms,  $\beta_{0,0}$  and  $\beta_{1,0}$ , then we are left with the same  $\delta$  that appears in (19-36)—the ATET would be  $\beta_{0,1} + C_i(\beta_{1,0} - \beta_{0,0})$ .

There are many variations of this model in the empirical literature. They have been applied to the analysis of education,<sup>34</sup> the Head Start program,<sup>35</sup> and a host of other settings.<sup>36</sup> This strand of literature is particularly important because the use of dummy variable models to analyze treatment effects and program participation has a long history in empirical economics. This analysis has called into question the interpretation of a number of received studies.

### 19.6.1.a The Normality Assumption

Some research has cast some skepticism on the selection model based on the normal distribution. [See Goldberger (1983) for an early salvo in this literature.] Among the findings are that the parameter estimates are surprisingly sensitive to the distributional assumption that underlies the model. Of course, this fact in itself does not invalidate the normality assumption, but it does call its generality into question. On the other hand, the received evidence is convincing that sample selection, in the abstract, raises serious problems, distributional questions aside. The literature—for example, Duncan (1986b), Manski (1989, 1990), and Heckman (1990)—has suggested some promising approaches based on robust and nonparametric estimators. These approaches obviously have the virtue of greater generality. Unfortunately, the cost is that they generally are quite limited in the breadth of the models they can accommodate. That is, one might gain the robustness of a nonparametric estimator at the cost of being unable to make use of the rich set of accompanying variables usually present in the panels to which selectivity models are often applied. For example, the nonparametric bounds approach of Manski (1990) is defined for two regressors. Other methods [e.g., Duncan (1986b)] allow more elaborate specifications.

Recent research includes specific attempts to move away from the normality assumption.<sup>37</sup> An example is Martins (2001), building on Newey (1991), which takes the core specification as given in (19-22) as the platform but constructs an alternative to the assumption of bivariate normality. Martins's specification modifies the Heckman model by employing an equation of the form

$$E[y_i | z_i = 1, \mathbf{x}_i, \mathbf{w}_i] = \mathbf{x}'_i \boldsymbol{\beta} + \mu(\mathbf{w}'_i \boldsymbol{\gamma})$$

where the latter “selectivity correction” is not the inverse Mills ratio, but some other result from a different model. The correction term is estimated using the Klein and Spady model discussed in Section 23.6.1. This is labeled a “semiparametric” approach.

 Whether the conditional mean in the selected sample should even remain a linear index function remains to be settled. Not surprisingly, Martins's results, based on two-step

<sup>34</sup> Willis and Rosen (1979).

<sup>35</sup> Goldberger (1972, 2008).

<sup>36</sup> A useful summary of the issues is Barnow, Cain, and Goldberger (1981). See, also, Imbens and Wooldridge (2009).

<sup>37</sup> Again, Angrist (2001) is an important contribution to this literature.

CHAPTER 19 ♦ Limited Dependent Variables **893**

least squares differ only slightly from the conventional results based on normality. This approach is arguably only a fairly small step away from the tight parameterization of the Heckman model. Other non- and semiparametric specifications, for example, Honore and Kyriazidou (1997, 2000) represent more substantial departures from the normal model, but are much less operational.<sup>38</sup> The upshot is that the issue remains unsettled. For better or worse, the empirical literature on the subject continues to be dominated by Heckman's original model built around the joint normal distribution.

**19.6.1.b Estimating the Effect of Treatment on the Treated**

Consider a regression approach to analyzing treatment effects in a two-period setting,

$$y_{it} = \theta_t + \mathbf{x}'_{it}\boldsymbol{\beta} + \gamma C_i + u_i + \varepsilon_{it}, \quad t = 0, 1,$$

where  $C_i$  is the treatment dummy variable and  $u_i$  is the unobserved individual effect. The setting is the pre- and posttreatment analysis of the sort considered in this section, where we examine the impact of a job training program on post training earnings. Because there are two periods, a natural approach to the analysis is to examine the changes,

$$\Delta y_i = (\theta_1 - \theta_0) + \gamma \Delta C_i + (\Delta \mathbf{x}'_{it})'\boldsymbol{\beta} + \Delta \varepsilon_{it}$$

where  $\Delta C_i = 1$  for the treated and 0 for the nontreated individuals, and the first differences eliminate the unobserved individual effects. In the absence of controls (regressors,  $\mathbf{x}_{it}$ ), or assuming that the controls are unchanged, the estimator of the effect of the treatment will be

$$\hat{\gamma} = \frac{\overline{\Delta y}}{\overline{\Delta C_i}} = \frac{\overline{\Delta y} | (\Delta C_i = 1)}{\overline{\Delta y} | (C_i = 0)},$$



which is the **difference in differences** estimator. This simplifies the problem considerably but has several shortcomings. Most important, by using the simple differences, we have lost our ability to discern what induced the change, whether it was the program or something else, presumably in  $\mathbf{x}_{it}$ .

Even without the normality assumption, the preceding regression approach is more tightly structured than many are comfortable with. A considerable amount of research has focused on what assumptions are needed to reach that model and whether they are likely to be appropriate in a given setting.<sup>39</sup> The overall objective of the analysis of the preceding two sections is to evaluate the effect of a treatment,  $C_i$ , on the individual treated. The implicit counterfactual is an observation on what the “response” (dependent variable) of the treated individual would have been had they not been treated. But, of course, an individual will be in one state or the other, not both. Denote by  $y_0$  the random variable that is the outcome variable in the absence of the treatment and by  $y_1$  the outcome when the treatment has taken place. The **average treatment effect**,

<sup>38</sup>This particular work considers selection in a “panel” (mainly two periods). But, the panel data setting for sample selection models is more involved than a cross-section analysis. In a panel data set, the “selection” is likely to be a decision at the beginning of Period 1 to be in the data set for all subsequent periods. As such, something more intricate than the model we have considered here is called for.

<sup>39</sup>A sampling of the more important parts of the literature on this issue includes Heckman (1992, 1997), Imbens and Angrist (1994), Manski (1996), and Wooldridge (2002a, Chapter 18).

## 894 PART IV ♦ Cross Sections, Panel Data, and Microeometrics

averaged over the entire population is

$$ATE = E[y_1 - y_0].$$

This is the impact of the treatment on an individual drawn at random from the entire population. However, the desired quantity is not necessarily the *ATE*, but the **average treatment effect on the treated**, which would be

$$ATE | T = E[y_1 - y_0 | C = 1].$$

The difficulty of measuring this is, once again, the counterfactual,  $E[y_0 | C = 1]$ . Whether these two measures will be the same is at the center of the much of the discussion on this subject. If treatment is completely randomly assigned, then  $E[y_j | C = 1] = E[y_j | C = 0] = E[y_j | C = j]$ ,  $j = 0, 1$ . This means that with completely random treatment assignment

$$ATE = E[y_1 | C = 1] - E[y_0 | C = 0].$$

To put this in our example, if college attendance were completely randomly distributed throughout the population, then the impact of college attendance on income (neglecting other covariates at this point) could be measured simply by averaging the incomes of college attendees and subtracting the average income of nonattendees. The preceding theory might work for the treatment “having brown eyes,” but it is unlikely to work for college attendance. Not only is the college attendance treatment not randomly distributed, but the treatment “assignment” is surely related to expectations about  $y_1$  versus  $y_0$ , and, at a minimum,  $y_0$  itself. (College is expensive.) More generally, the researcher faces the difficulty in calculating treatment effects that assignment to the treatment might not be exogenous.

The **control function** approach that we used in (19-34)–(19-36) is used to account for the endogeneity of the treatment assignment in the regression context. The very specific assumptions of the bivariate normal distribution of the unobservables somewhat simplifies the estimation, because they make explicit what control function ( $\lambda_i$ ) is appropriate to use in the regression. As Wooldridge (2002a, p. 622) points out, however, the binary variable in the treatment effects regression represents simply an endogenous variable in a linear equation, amenable to **instrumental variable estimation** (assuming suitable instruments are available). Barnow, Cain, and Goldberger (1981) proposed a two-stage least squares estimator, with instrumental variable equal to the predicted probability from the probit treatment assignment model. This is slightly less **parametric** than (19-36) because, in principle, its validity does not rely on joint normality of the disturbances. [Wooldridge (2002a, pp. 621–633) discusses the underlying assumptions.]

### 19.6.2 PROPENSITY SCORE MATCHING

If the treatment assignment is “completely ignorable,” then, as noted, estimation of the treatment effects is greatly simplified. Suppose, as well, that there are observable variables that influence both the outcome and the treatment assignment. Suppose it is possible to obtain pairs of individuals matched by a common  $\mathbf{x}_i$ , one with  $C_i = 0$ , the other with  $C_i = 1$ . If done with a sufficient number of pairs so as to average

CHAPTER 19 ♦ Limited Dependent Variables **895**

over the population of  $\mathbf{x}_i$ 's, then a **matching estimator**, the average value of  $(y_i | C_i = 1) - (y_i | C_i = 0)$ , would estimate  $E[y_1 - y_0]$ , which is what we seek. Of course, it is optimistic to hope to find a large sample of such matched pairs, both because the sample overall is finite and because there may be many regressors, and the "cells" in the distribution of  $\mathbf{x}_i$  are likely to be thinly populated. This will be worse when the regressors are continuous, for example, with a "family income" variable. Rosenbaum and Rubin (1983) and others<sup>40</sup> suggested, instead, matching on the **propensity score**,  $F(\mathbf{x}_i) = \text{Prob}(C_i = 1 | \mathbf{x}_i)$ . Individuals with similar propensity scores are paired and the average treatment effect is then estimated by the differences in outcomes. Various strategies are suggested by the authors for obtaining the necessary subsamples and for verifying the conditions under which the procedures will be valid. [See, e.g., Becker and Ichino (2002) and Greene (2007c).]

**Example 19.15 Treatment Effects on Earnings**

LaLonde (1986) analyzed the results of a labor market experiment, The National Supported Work Demonstration, in which a group of disadvantaged workers lacking basic job skills were given work experience and counseling in a sheltered environment. Qualified applicants were assigned to training positions randomly. The treatment group received the benefits of the program. Those in the control group "were left to fend for themselves." [The demonstration was run in numerous cities in the mid-1970s. See LaLonde (1986, pp. 605–609) for details on the NSW experiments.] The training period was 1976–1977; the outcome of interest for the sample examined here was posttraining 1978 earnings. LaLonde reports a large variety of estimates of the treatment effect, for different subgroups and using different estimation methods. Nonparametric estimates for the group in our sample are roughly \$900 for the income increment in the posttraining year. (See LaLonde, p. 609.) Similar results are reported from a two-step regression-based estimator similar to (19-34) to (19-36). (See LaLonde's footnote to Table 6, p. 616.)

LaLonde's data are fairly well traveled, having been used in replications and extensions in, for example, Dehejia and Wahba (1999), Becker and Ichino (2002), and Greene (2007b, c). We have reestimated the matching estimates reported in Becker and Ichino. The data in the file used there (and here) contain 2,490 control observations and 185 treatment observations on the following variables:

*t* = treatment dummy variable  
*age* = age in years  
*educ* = education in years  
*marr* = dummy variable for married  
*black* = dummy variable for black  
*hisp* = dummy variable for Hispanic  
*nodegree* = dummy for no degree (not used)  
*re74* = real earnings in 1974  
*re75* = real earnings in 1975  
*re78* = real earnings in 1978



<sup>40</sup>Other important references in this literature are Becker and Ichino (1999), Dehejia and Wahba (1999), LaLonde (1986), Heckman, Ichimura, and Todd (1997, 1998), Heckman, Ichimura, Smith, and Todd (1998), Heckman, LaLonde, and Smith (1999), Heckman, Tobias, and Vytlacil (2003), and Heckman and Vytlacil (2000).

**896 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

Transformed variables added to the equation are

$$\text{age}^2 = \text{age squared}$$

$$\text{educ}^2 = \text{educ squared}$$

$$\text{re74}^2 = \text{re74 squared}$$

$$\text{re75}^2 = \text{re75 squared}$$

$$\text{black}_{74} = \text{black times } 1(\text{re74} = 0)$$



We also scaled all earnings variables by 10,000 before beginning the analysis. (See Appendix Table F19.3. The data are downloaded from the website <http://www.nber.org/%7Erdehejia/nswdata.html>. The two specific subsamples are in <http://www.nber.org/%7Erdehejia//psid-controls.txt> and [http://www.nber.org/%7Erdehejia/nswre74\\_treated.txt](http://www.nber.org/%7Erdehejia/nswre74_treated.txt).) (We note that Becker and Ichino report they were unable to replicate Dehejia and Wahba's results, although they could come reasonably close. We, in turn, were not able to replicate either set of results, though we, likewise, obtained quite similar results.)

The analysis proceeded as follows: A logit model in which the included variables were a constant, age,  $\text{age}^2$ , education,  $\text{education}^2$ , marr, black, hisp, re74, re75, re742, re752, and black74 was computed for the treatment assignment. The fitted probabilities are used for the propensity scores. By means of an iterative search, the range of propensity scores was partitioned into eight regions within which, by a simple *F* test, the mean scores of the treatments and controls were not statistically different. The partitioning is shown in Table 19.10. The 1,347 observations are all the treated observations and the 1,162 control observations are those whose propensity scores fell within the range of the scores for the treated observations.

Within each interval, each treated observation is paired with a small number of the nearest control observations. We found the average difference between treated observation and control to equal \$1,574.35. Becker and Ichino reported \$1,537.94.

As an experiment, we refit the propensity score equation using a probit model, retaining the fitted probabilities. We then used the two-step estimator described earlier to fit (19-34) and (19-35) using the entire sample. The estimates of  $\delta$ ,  $\rho$ , and  $\sigma$  were  $-1.01437$ ,  $0.35519$ ,  $1.38426$ . Using the results from the probit model, we averaged the result in (19-36) for the entire sample, obtaining an estimated treatment effect of \$1,476.30.

**TABLE 19.10** Empirical Distribution of Propensity Scores

Percent	Lower	Upper	Lower	Upper	# Obs
0–5	0.000591	0.000783		Sample size = 1,347	
5–10	0.000787	0.001061		Average score = 0.137238	
10–15	0.001065	0.001377		Std. Dev score = 0.274079	
15–20	0.001378	0.001748			
20–25	0.001760	0.002321			
25–30	0.002340	0.002956	1	0.000591	0.098016
30–35	0.002974	0.004057	2	0.098016	0.195440
35–40	0.004059	0.005272	3	0.195440	0.390289
40–45	0.005278	0.007486	4	0.390289	0.585138
45–50	0.007557	0.010451	5	0.585138	0.779986
50–55	0.010563	0.014643	6	0.779986	0.877411
55–60	0.014686	0.022462	7	0.877411	0.926123
60–65	0.022621	0.035060	8	0.926123	0.974835
65–70	0.035075	0.051415			
70–75	0.051415	0.076188			
75–80	0.076376	0.134189			
80–85	0.134238	0.320638			
85–90	0.321233	0.616002			
90–95	0.624407	0.949418			
95–100	0.949418	0.974835			

### 19.6.3 REGRESSION DISCONTINUITY

There are many situations in which there is no possibility of randomized assignment of treatments. Examples include student outcomes and policy interventions in schools. Angrist and Lavy (1999), for example, studied the effect of class sizes on test scores. Van der Klaauw studied financial aid offers that were tied to SAT scores and grade point averages. In these cases, the natural experiment approach advocated by Angrist and Pischke (2009) is an appealing way to proceed, when it is feasible. The **regression discontinuity design** presents an alternative strategy. The conditions under which the approach can be effective are when (1) the outcome,  $y$ , is a continuous variable; (2) the outcome varies smoothly with an assignment variable,  $A$ , and (3) treatment is “sharply” assigned based on the value of  $A$ , specifically  $C = 1(A > A^*)$  where  $A^*$  is a fixed threshold or cutoff value. [A “**fuzzy design** is based on  $\text{Prob}(C = 1 | A) = F(A)$ . The identification problems with fuzzy design are much more complicated than with sharp design. Readers are referred to Van der Klaauw (2002) for further discussion of fuzzy design.] We assume, then, that

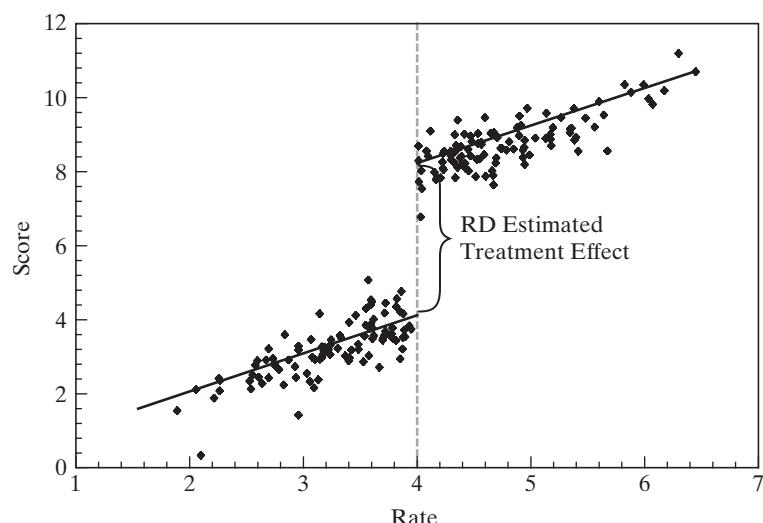
$$y = f(A, C) + \varepsilon.$$

Suppose, for example, the outcome variable is a test score, and that an administrative treatment such as a special education program is funded based on the poverty rates of certain communities. The ideal conditions for a regression discontinuity design based on these assumptions is shown in Figure 19.8. The logic of the calculation is that the points near the threshold value, which have “essentially” the same stimulus value, constitute a nearly random sample of observations which are segmented by the treatment.

The method requires that  $E[\varepsilon | A, C] = E[\varepsilon | A]$ —the assignment variable—be exogenous to the experiment. The result in Figure 19.8 is consistent with

$$y = f(A) + \alpha C + \varepsilon,$$

**FIGURE 19.8** Regression Discontinuity.



**898 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

where  $\alpha$  will be the treatment effect to be estimated. The specification  $f(A)$  can be problematic; assuming a linear function when something more general  will bias the estimate of  $\alpha$ . For this reason, nonparametric methods, such as the LOWESS regression (see Section 12.3.5) might be attractive. This is likely to enable the analyst to make fuller use of the observations that are more distant from the cutoff point. [See Van der Klaauw (2002).] Identification of the treatment effect begins with the assumption that  $f(A)$  is continuous at  $A^*$ , so that

$$\lim_{A \uparrow A^*} f(A) = \lim_{A \downarrow A^*} f(A) = f(A^*).$$

Then

$$\begin{aligned} \lim_{A \downarrow A^*} E[y | A] - \lim_{A \uparrow A^*} E[y | A] &= f(A^*) + \alpha + \lim_{A \downarrow A^*} E[\varepsilon | A] - f(A^*) - \lim_{A \uparrow A^*} E[\varepsilon | A] \\ &= \alpha. \end{aligned}$$

With this in place, the treatment effect can be estimated by the difference of the average outcomes for those individuals “close” to the threshold value,  $A^*$ . Details on regression discontinuity design are provided by Trochim (1984, 2000) and Van der Klaauw (2002).

## 19.7 SUMMARY AND CONCLUSIONS

This chapter has examined settings in which, in principle, the linear regression model of Chapter 2 would apply, but the data generating mechanism produces a nonlinear form: truncation, censoring, and sample selection or endogenous sampling. For each case, we develop the basic theory of the effect and then use the results in a major area of research in econometrics.

In the truncated regression model, the range of the dependent variable is restricted substantively. Certainly all economic data are restricted in this way—aggregate income data cannot be negative, for example. But when data are truncated so that plausible values of the dependent variable are precluded, for example, when zero values for expenditure are discarded, the data that remain are analyzed with models that explicitly account for the truncation. The stochastic frontier model is based on a composite disturbance in which one part follows the assumptions of the familiar regression model while the second component is built on a platform of the truncated regression.

When data are censored, values of the dependent variable that could in principle be observed are masked. Ranges of values of the true variable being studied are observed as a single value. The basic problem this presents for model building is that in such a case, we observe variation of the independent variables without the corresponding variation in the dependent variable that might be expected. Consistent estimation, and useful interpretation of estimation results are based on maximum likelihood or some other technique that explicitly accounts for the censoring mechanism. The most common case of censoring in observed data arises in the context of duration analysis, or survival functions (which borrows a term from medical statistics where this style of model building originated). It is useful to think of duration, or survival data, as the measurement of time between transitions or changes of state. We examined three modeling approaches that correspond to the description in Chapter 12; nonparametric (survival tables), semiparametric (the proportional hazard models), and parametric (various forms such as the Weibull model).

**CHAPTER 19 ♦ Limited Dependent Variables 899**

Finally, the issue of sample selection arises when the observed data are not drawn randomly from the population of interest. Failure to account for this nonrandom sampling produces a model that describes only the nonrandom subsample, not the larger population. In each case, we examined the model specification and estimation techniques which are appropriate for these variations of the regression model. Maximum likelihood is usually the method of choice, but for the third case, a two-step estimator has become more common. The leading contemporary application of selection methods and endogenous sampling is in the measure of treatment effects. We considered three approaches to analysis of treatment effects; regression methods, propensity score matching, and regression discontinuity.

**Key Terms and Concepts**

- Accelerated failure time model
- Attenuation
- Average treatment effect
- Average treatment effect on the treated
- Censored regression model
- Censored variable
- Censoring
- Conditional mean assumption
- Conditional moment test
- Control function
- Corner solution model
- Data envelopment analysis
- Degree of truncation
- Delta method
- Difference in differences
- Duration model
- Exponential
- Exponential model
- Fuzzy design
- Generalized residual
- Hazard function
- Hazard rate
- Heterogeneity
- Heteroscedasticity
- Hurdle model
- Incidental truncation
- Instrumental variable estimation
- Integrated hazard function
- Inverse probability weighted estimator
- Inverse Mills ratio
- Lagrange multiplier test
- Matching estimator
- Mean independence assumption
- Missing counterfactual
- Negative duration dependence
- Olsen's reparameterization
- Parametric
- Parametric model
- Partial likelihood
- Positive duration dependence
- Product limit estimator
- Propensity score
- Proportional hazard
- Regression discontinuity design
- Risk set
- Rubin causal model
- Sample selection
- Selection on observables
- Selection on unobservables
- Semiparametric estimator
- Semiparametric model
- Specification error
- Stochastic frontier model
- Survival function
- Time-varying covariate
- Tobit model
- Treatment effect
- Truncated distribution
- Truncated mean
- Truncated normal distribution
- Truncated random variable
- Truncated standard normal distribution
- Truncated variance
- Truncation
- Two-step estimation
- Type II tobit model
- Weibull model
- Weibull survival model

**Exercises**

1. The following 20 observations are drawn from a censored normal distribution:

3.8396	7.2040	0.00000	0.00000	4.4132	8.0230
5.7971	7.0828	0.00000	0.80260	13.0670	4.3211
0.00000	8.6801	5.4571	0.00000	8.1021	0.00000
1.2526	5.6016				

**900 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

The applicable model is

$$\begin{aligned}y_i^* &= \mu + \varepsilon_i, \\y_i &= y_i^* \quad \text{if } \mu + \varepsilon_i > 0, 0 \text{ otherwise,} \\&\varepsilon_i \sim N[0, \sigma^2].\end{aligned}$$

Exercises 1 through 4 in this section are based on the preceding information. The OLS estimator of  $\mu$  in the context of this tobit model is simply the sample mean. Compute the mean of all 20 observations. Would you expect this estimator to over- or underestimate  $\mu$ ? If we consider only the nonzero observations, then the truncated regression model applies. The sample mean of the nonlimit observations is the least squares estimator in this context. Compute it and then comment on whether this sample mean should be an overestimate or an underestimate of the true mean.

2. We now consider the tobit model that applies to the full data set.
  - a. Formulate the log-likelihood for this very simple tobit model.
  - b. Reformulate the log-likelihood in terms of  $\theta = 1/\sigma$  and  $\gamma = \mu/\sigma$ . Then derive the necessary conditions for maximizing the log-likelihood with respect to  $\theta$  and  $\gamma$ .
  - c. Discuss how you would obtain the values of  $\theta$  and  $\gamma$  to solve the problem in part b.
  - d. Compute the maximum likelihood estimates of  $\mu$  and  $\sigma$ .
3. Using only the nonlimit observations, repeat Exercise 2 in the context of the truncated regression model. Estimate  $\mu$  and  $\sigma$  by using the method of moments estimator outlined in Example 19.2. Compare your results with those in the previous exercises.
4. Continuing to use the data in Exercise 1, consider once again only the nonzero observations. Suppose that the sampling mechanism is as follows:  $y^*$  and another normally distributed random variable  $z$  have population correlation 0.7. The two variables,  $y^*$  and  $z$ , are sampled jointly. When  $z$  is greater than zero,  $y$  is reported. When  $z$  is less than zero, both  $z$  and  $y$  are discarded. Exactly 35 draws were required to obtain the preceding sample. Estimate  $\mu$  and  $\sigma$ . (*Hint:* Use Theorem 19.5.)
5. Derive the partial effects for the tobit model with heteroscedasticity that is described in Section 19.3.5.a.
6. Prove that the Hessian for the tobit model in (19-14) is negative definite after Olsen's transformation is applied to the parameters.

## Applications

1. We examined Ray Fair's famous analysis (*Journal of Political Economy*, 1978) of a *Psychology Today* survey on extramarital affairs in Example 18.9 using a Poisson regression model. Although the dependent variable used in that study was a count, Fair (1978) used the tobit model as the platform for his study. You can reproduce the tobit estimates in Fair's paper easily with any software package that contains a tobit estimator—most do. The data appear in Appendix Table F18.1. Reproduce

CHAPTER 19 ♦ Limited Dependent Variables **901**

Fair's least squares and tobit estimates. Compute the partial effects for the model and interpret all results.

2. Fair's original study also included but did not analyze a second data set that was a similar survey conducted by *Redbook* magazine. The data are reproduced in Appendix Table F17.2. (Our thanks to Ray Fair for providing these data.) This sample contains observations on 6,366 women and the following variables:

$id$  = an identification number

$C$  = constant, value = 1

$yrb$  = a constructed measure of time spent in extramarital affairs

$v_1$  = a rating of the marriage, coded 1 to 4

$v_2$  = age, in years, aggregated

$v_3$  = number of years married

$v_4$  = number of children, top coded at 5

$v_5$  = religiosity, 1 to 4, 1 = not, 4 = very

$v_6$  = education, coded 9, 12, 14, 16, 17, 20

$v_7$  = occupation

$v_8$  = husband's occupation

Three other variables were not used. Details on the variables in the model are given in Fair's (1978) *Journal of Political Economy* paper. Using these data, conduct a parallel study to the *Psychology Today* study that was done in Fair (1978). Are the results consistent? Report all results, including partial effects and relevant diagnostic statistics.

3. Continuing the analysis of the previous application, note that these data conform precisely to the description of "corner solutions" in Section 19.3.4. The dependent variable is not censored in the fashion usually assumed for a tobit model. To investigate whether the dependent variable is determined by a two-part decision process (yes/no and, if yes, how much), specify and estimate a two-part model in which the first equation analyzes the binary  $A = 1$  if  $yrb > 0$  and 0 otherwise and the second equation analyzes  $yrb | yrb > 0$ . What is the appropriate model? What do you find? Report all results. (Note: If you analyze the second dependent variable using the truncated regression, you should remove some extreme observations from your sample. The truncated regression estimator refuses to converge with the full data set but works nicely for the example if you omit observations with  $yrb > 5$ .)

4. **StochasticFrontier Model.** Section 10.5.1 presents estimates of a Cobb-Douglas cost function using Nerlove's 1955 data on the U.S. electric power industry. Christensen and Greene's 1976 update of this study used 1970 data for this industry. The Christensen and Greene data are given in Appendix Table F4.3. These data have provided a standard test data set for estimating different forms of production and cost functions, including the stochastic frontier model discussed in Section 19.2.4. It has been suggested that one explanation for the apparent finding of economies of

**902 PART IV ♦ Cross Sections, Panel Data, and Microeometrics**

scale in these data is that the smaller firms were inefficient for other reasons. The stochastic frontier might allow one to disentangle these effects. Use these data to fit a frontier cost function which includes a quadratic term in log output in addition to the linear term and the factor prices. Then examine the estimated Jondrow et al. residuals to see if they do indeed vary negatively with output, as suggested. (This will require either some programming on your part or specialized software. The stochastic frontier model is provided as an option in Stata, TSP, and LIMDEP. Or, the likelihood function can be programmed fairly easily for RATS, MatLab, or GAUSS. (*Note:* For a cost frontier as opposed to a production frontier, it is necessary to reverse the sign on the argument in the  $\Phi$  function that appears in the log-likelihood.)

# 20

## SERIAL CORRELATION

---

### 20.1 INTRODUCTION

Time-series data often display **autocorrelation**, or serial correlation of the disturbances across periods. Consider, for example, the plot of the least squares residuals in the following example.

**Example 20.1 Money Demand Equation**

Appendix Table F5.2 contains quarterly data from 1950.1 to 2000.4 on the U.S. money stock ( $M1$ ) and output (real GDP) and the price level (CPI-U). Consider a simple (extremely) model of money demand.<sup>1</sup>

$$\ln M1_t = \beta_1 + \beta_2 \ln GDP_t + \beta_3 \ln CPI_t + \varepsilon_t.$$

A plot of the least squares residuals is shown in Figure 20.1. The pattern in the residuals suggests that knowledge of the sign of a residual in one period is a good indicator of the sign of the residual in the next period. This knowledge suggests that the effect of a given disturbance is carried, at least in part, across periods. This sort of “memory” in the disturbances creates the long, slow swings from positive values to negative ones that is evident in Figure 20.1. One might argue that this pattern is the result of an obviously naive model, but that is one of the important points in this discussion. Patterns such as this usually do not arise spontaneously; to a large extent, they are, indeed, a result of an incomplete or flawed model specification.

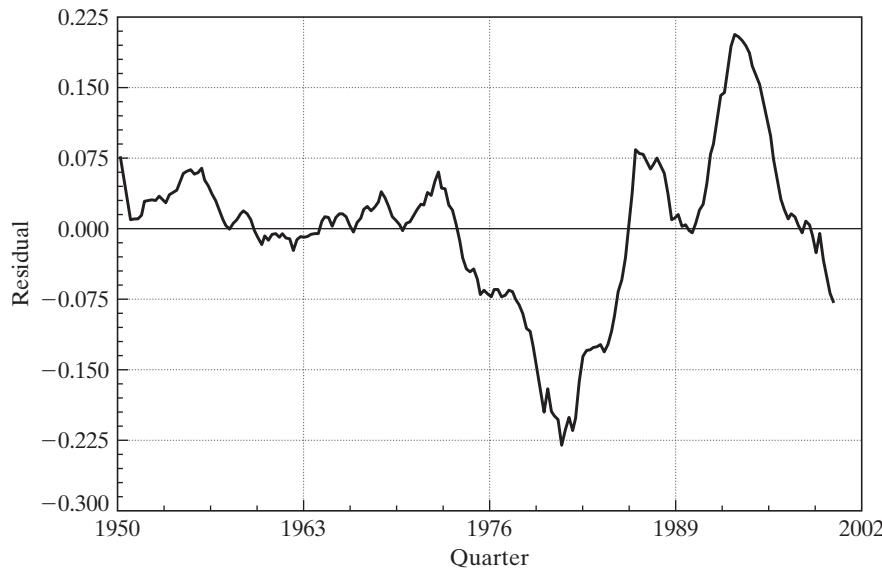
One explanation for autocorrelation is that relevant factors omitted from the time-series regression, like those included, are correlated across periods. This fact may be due to serial correlation in factors that should be in the regression model. It is easy to see why this situation would arise. Example 20.2 shows an obvious case.

**Example 20.2 Autocorrelation Induced by Misspecification  
the Model**

In Examples 2.3 and 6.7, we examined yearly time-series data on the U.S. gasoline market from 1953 to 2004. The evidence in the examples was convincing that a regression model of variation in  $\ln G/Pop$  should include, at a minimum, a constant,  $\ln P_G$  and  $\ln \text{income}/\text{Pop}$ . Other price variables and a time trend also provide significant explanatory power, but these two are a bare minimum. Moreover, we also found on the basis of a Chow test of structural change that apparently this market changed structurally after 1974. Figure 20.2 displays plots of four sets of least squares residuals. Parts (a) through (c) show clearly that as the specification of the regression is expanded, the autocorrelation in the “residuals” diminishes. Part (c) shows the effect of forcing the coefficients in the equation to be the same both before and after the structural shift. In part (d), the residuals in the two subperiods 1953 to 1974 and 1975 to 2004 are produced by separate unrestricted regressions. This latter set of residuals is almost nonautocorrelated. (Note also that the range of variation of the residuals falls as

---

<sup>1</sup>Because this chapter deals exclusively with time-series data, we shall use the index  $t$  for observations and  $T$  for the sample size throughout.

**904 PART V ♦ Time Series and Microeometrics**


**FIGURE 20.1** Autocorrelated Least Squares Residuals.

the model is improved, i.e., as its fit improves.) The full equation is

$$\begin{aligned} \ln \frac{G_t}{Pop_t} = & \beta_1 + \beta_2 \ln P_{Gt} + \beta_3 \ln \frac{I_t}{Pop_t} + \beta_4 \ln P_{Nct} + \beta_5 \ln P_{Uct} \\ & + \beta_6 \ln P_{PTt} + \beta_7 \ln P_{Nt} + \beta_8 \ln P_{Dt} + \beta_9 \ln P_{St} + \beta_{10} t + \varepsilon_t. \end{aligned}$$

Finally, we consider an example in which serial correlation is an anticipated part of the model.

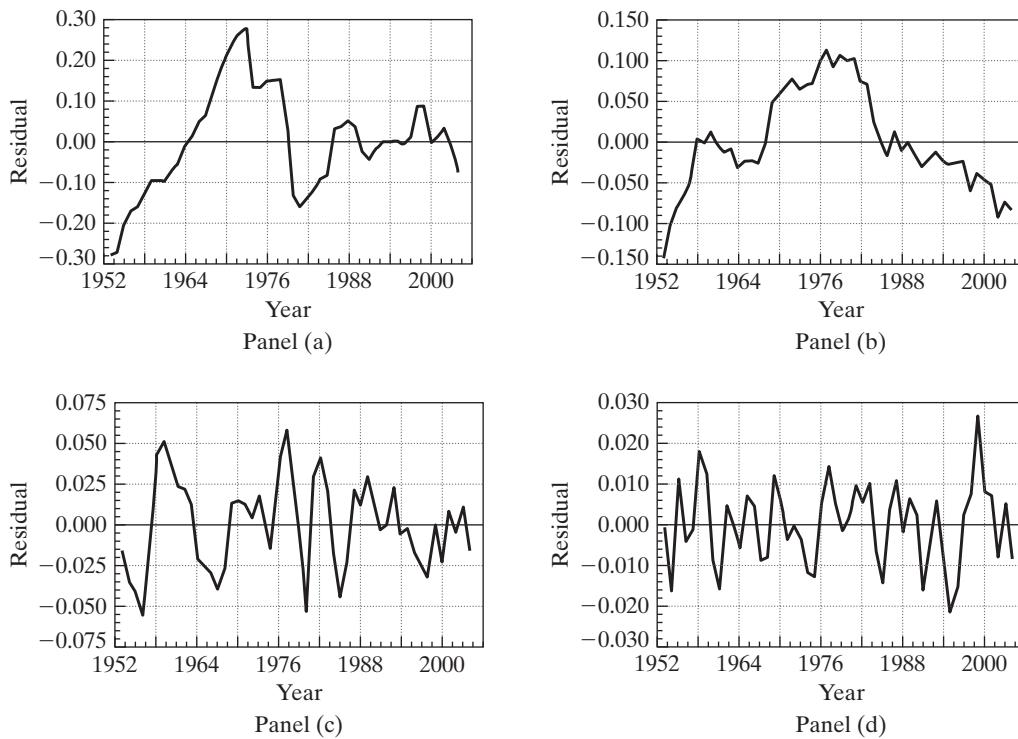
**Example 20.3 Negative Autocorrelation in the Phillips Curve**

The Phillips curve [Phillips (1957)] has been one of the most intensively studied relationships in the macroeconomics literature. As originally proposed, the model specifies a negative relationship between wage inflation and unemployment in the United Kingdom over a period of 100 years. Recent research has documented a similar relationship between unemployment and price inflation. It is difficult to justify the model when cast in simple levels; labor market theories of the relationship rely on an uncomfortable proposition that markets persistently fall victim to money illusion, even when the inflation can be anticipated. Current research [e.g., Staiger et al. (1996)] has reformulated a short-run (disequilibrium) “expectations augmented Phillips curve” in terms of unexpected inflation and unemployment that deviates from a long-run equilibrium or “natural rate.” The **expectations-augmented Phillips curve** can be written as

$$\Delta p_t - E[\Delta p_t | \Psi_{t-1}] = \beta[u_t - u^*] + \varepsilon_t$$

where  $\Delta p_t$  is the rate of inflation in year  $t$ ,  $E[\Delta p_t | \Psi_{t-1}]$  is the forecast of  $\Delta p_t$  made in period  $t - 1$  based on information available at time  $t - 1$ ,  $\Psi_{t-1}$ ,  $u_t$  is the unemployment rate and  $u^*$  is the natural, or equilibrium rate. (Whether  $u^*$  can be treated as an unchanging parameter, as we are about to do, is controversial.) By construction,  $[u_t - u^*]$  is disequilibrium, or cyclical unemployment. In this formulation,  $\varepsilon_t$  would be the supply shock (i.e., the stimulus that produces the disequilibrium situation). To complete the model, we require a model for the expected inflation. We will revisit this in some detail in Example 21.2. For the present, we'll

## CHAPTER 20 ♦ Serial Correlation 905



**FIGURE 20.2** Unstandardized Residuals (Bars mark mean res. and  $\pm 2s(\epsilon)$ ).

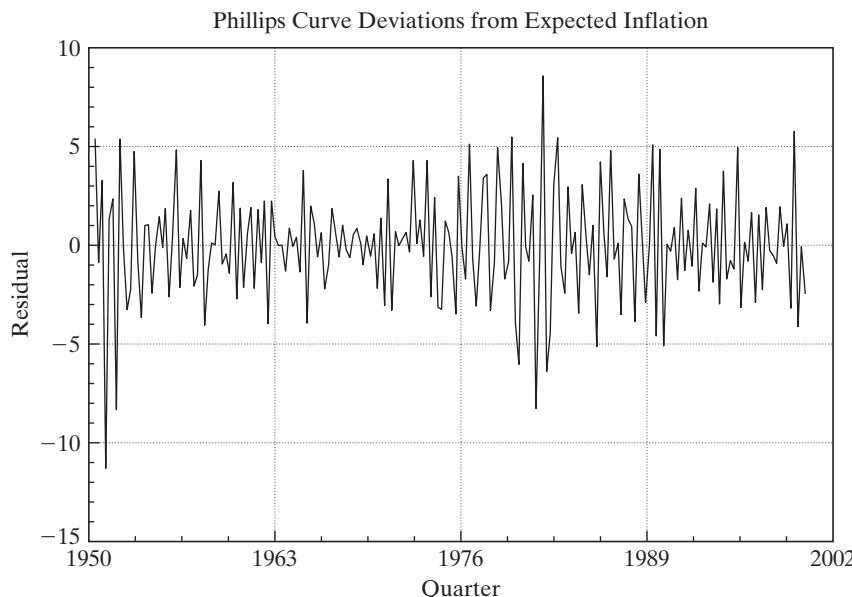
assume that economic agents are rank empiricists. The forecast of next year's inflation is simply this year's value. This produces the estimating equation

$$\Delta p_t - \Delta p_{t-1} = \beta_1 + \beta_2 u_t + \varepsilon_t$$

where  $\beta_2 = \beta$  and  $\beta_1 = -\beta u^*$ . Note that there is an implied estimate of the natural rate of unemployment embedded in the equation. After estimation,  $u^*$  can be estimated by  $-b_1/b_2$ . The equation was estimated with the 1950.1–2000.4 data in Appendix Table F5.2 that were used in Example 20.1 (minus two quarters for the change in the rate of inflation). Least squares estimates (with standard errors in parentheses) are as follows:

$$\begin{aligned} \Delta p_t - \Delta p_{t-1} &= 0.49189 - 0.090136 u_t + \varepsilon_t \\ (0.7405) \quad (0.1257) \quad R^2 &= 0.002561, \quad T = 202. \end{aligned}$$

The implied estimate of the natural rate of unemployment is 5.46 percent, which is in line with other recent estimates. The estimated asymptotic covariance of  $b_1$  and  $b_2$  is  $-0.08973$ . Using the delta method, we obtain a standard error of 2.2062 for this estimate, so a confidence interval for the natural rate is  $5.46 \text{ percent} \pm 1.96 (2.21 \text{ percent}) = (1.13 \text{ percent}, 9.79 \text{ percent})$  (which seems fairly wide, but, again, whether it is reasonable to treat this as a parameter is at least questionable). The regression of the least squares residuals on their past values gives a slope of  $-0.4263$  with a highly significant  $t$  ratio of  $-6.725$ . We thus conclude that the residuals (and, apparently, the disturbances) in this model are highly negatively autocorrelated. This is consistent with the striking pattern in Figure 20.3.

**906 PART V ♦ Time Series and Microeometrics**


**FIGURE 20.3** Negatively Autocorrelated Residuals.

The problems for estimation and inference caused by autocorrelation are similar to (although, unfortunately, more involved than) those caused by heteroscedasticity. As before, least squares is inefficient, and inference based on the least squares estimates is adversely affected. Depending on the underlying process, however, GLS and FGLS estimators can be devised that circumvent these problems. There is one qualitative difference to be noted. In Chapter 18, we examined models in which the generalized regression model can be viewed as an extension of the regression model to the conditional second moment of the dependent variable. In the case of autocorrelation, the phenomenon arises in almost all cases from a misspecification of the model. Views differ on how one should react to this failure of the classical assumptions, from a pragmatic one that treats it as another “problem” in the data to an orthodox methodological view that it represents a major specification issue—see, for example, “A Simple Message to Autocorrelation Correctors: Don’t” [Mizon (1995).]

We should emphasize that the models we shall examine here are quite far removed from the classical regression. The exact or small-sample properties of the estimators are rarely known, and only their asymptotic properties have been derived.

## 20.2 THE ANALYSIS OF TIME-SERIES DATA

The treatment in this chapter will be the first structured analysis of time-series data in the text. Time-series analysis requires some revision of the interpretation of both data generation and sampling that we have maintained thus far.

## CHAPTER 20 ♦ Serial Correlation 907

A time-series model will typically describe the path of a variable  $y_t$  in terms of contemporaneous (and perhaps lagged) factors  $\mathbf{x}_t$ , disturbances (**innovations**),  $\varepsilon_t$ , and its own past,  $y_{t-1}, \dots$ . For example,

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + \varepsilon_t.$$

The time series is a single occurrence of a random event. For example, the quarterly series on real output in the United States from 1950 to 2000 that we examined in Example 20.1 is a single realization of a process,  $GDP_t$ . The entire history over this period constitutes a realization of the process. At least in economics, the process could not be repeated. There is no counterpart to repeated sampling in a cross section or replication of an experiment involving a time-series process in physics or engineering. Nonetheless, were circumstances different at the end of World War II, the observed history *could* have been different. In principle, a completely different realization of the entire series might have occurred. The sequence of observations,  $\{y_t\}_{t=-\infty}^{\infty}$  is a **time-series process**, which is characterized by its time ordering and its systematic correlation between observations in the sequence. The signature characteristic of a time-series process is that empirically, the data generating mechanism produces exactly one realization of the sequence. Statistical results based on sampling characteristics concern not random sampling from a population, but from distributions of statistics constructed from sets of observations taken from this realization in a **time window**,  $t = 1, \dots, T$ . Asymptotic distribution theory in this context concerns behavior of statistics constructed from an increasingly long window in this sequence.

The properties of  $y_t$  as a random variable in a cross section are straightforward and are conveniently summarized in a statement about its mean and variance or the probability distribution generating  $y_t$ . The statement is less obvious here. It is common to assume that innovations are generated independently from one period to the next, with the familiar assumptions

$$E[\varepsilon_t] = 0,$$

$$\text{Var}[\varepsilon_t] = \sigma_\varepsilon^2,$$

and

$$\text{Cov}[\varepsilon_t, \varepsilon_s] = 0 \quad \text{for } t \neq s.$$

In the current context, this distribution of  $\varepsilon_t$  is said to be **covariance stationary** or **weakly stationary**. Thus, although the substantive notion of “random sampling” must be extended for the time series  $\varepsilon_t$ , the mathematical results based on that notion apply here. It can be said, for example, that  $\varepsilon_t$  is generated by a time-series process whose mean and variance are not changing over time. As such, by the method we will discuss in this chapter, we could, at least in principle, obtain sample information and use it to characterize the distribution of  $\varepsilon_t$ . Could the same be said of  $y_t$ ? There is an obvious difference between the series  $\varepsilon_t$  and  $y_t$ ; observations on  $y_t$  at different points in time are necessarily correlated. Suppose that the  $y_t$  series is weakly stationary and that, for the moment,  $\beta_2 = 0$ . Then we could say that

$$E[y_t] = \beta_1 + \beta_3 E[y_{t-1}] + E[\varepsilon_t] = \beta_1 / (1 - \beta_3)$$

and

$$\text{Var}[y_t] = \beta_3^2 \text{Var}[y_{t-1}] + \text{Var}[\varepsilon_t],$$

**908 PART V ♦ Time Series and Microeometrics**

or

$$\gamma_0 = \beta_3^2 \gamma_0 + \sigma_\varepsilon^2,$$

so that

$$\gamma_0 = \frac{\sigma_\varepsilon^2}{1 - \beta_3^2}.$$

Thus,  $\gamma_0$ , the variance of  $y_t$ , is a fixed characteristic of the process generating  $y_t$ . Note how the stationarity assumption, which apparently includes  $|\beta_3| < 1$ , has been used. The assumption that  $|\beta_3| < 1$  is needed to ensure a finite and positive variance.<sup>2</sup> Finally, the same results can be obtained for nonzero  $\beta_2$  if it is further assumed that  $x_t$  is a weakly stationary series.<sup>3</sup>

Alternatively, consider simply repeated substitution of lagged values into the expression for  $y_t$ :

$$y_t = \beta_1 + \beta_3(\beta_1 + \beta_3 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \quad (20-1)$$

and so on. We see that, in fact, the current  $y_t$  is an accumulation of the entire history of the innovations,  $\varepsilon_t$ . So if we wish to characterize the distribution of  $y_t$ , then we might do so in terms of sums of random variables. By continuing to substitute for  $y_{t-2}$ , then  $y_{t-3}, \dots$  in (20-1), we obtain an explicit representation of this idea,

$$y_t = \sum_{i=0}^{\infty} \beta_3^i (\beta_1 + \varepsilon_{t-i}).$$

Do sums that reach back into infinite past make any sense? We might view the process as having begun generating data at some remote, effectively “infinite” past. As long as distant observations become progressively less important, the extension to an infinite past is merely a mathematical convenience. The diminishing importance of past observations is implied by  $|\beta_3| < 1$ . Notice that, not coincidentally, this requirement is the same as that needed to solve for  $\gamma_0$  in the preceding paragraphs. A second possibility is to assume that the *observation of this time series* begins at some time 0 [with  $(x_0, \varepsilon_0)$  called the **initial conditions**], by which time the underlying process has reached a state such that the mean and variance of  $y_t$  are not (or are no longer) changing over time. The mathematics are slightly different, but we are led to the same characterization of the random process generating  $y_t$ . In fact, the same weak stationarity assumption ensures both of them.

Except in very special cases, we would expect all the elements in the  $T$  component random vector  $(y_1, \dots, y_T)$  to be correlated. In this instance, said correlation is called “autocorrelation.” As such, the results pertaining to estimation with independent or uncorrelated observations that we used in the previous chapters are no longer usable. In point of fact, we have a sample of but one observation on the multivariate random variable  $[y_t, t = 1, \dots, T]$ . There is a counterpart to the cross-sectional notion of parameter estimation, but only under assumptions (e.g., weak stationarity) that establish that parameters in the familiar sense even exist. Even with stationarity, it will emerge

<sup>2</sup>The current literature in macroeconomics and time series analysis is dominated by analysis of cases in which  $\beta_3 = 1$  (or counterparts in different models). We will return to this subject in Chapter 23.

<sup>3</sup>See Section 20.4.1 on the stationarity assumption.

that for estimation and inference, none of our earlier finite-sample results are usable. Consistency and asymptotic normality of estimators are somewhat more difficult to establish in time-series settings because results that require independent observations, such as the central limit theorems, are no longer usable. Nonetheless, counterparts to our earlier results have been established for most of the estimation problems we consider here and in Chapters 21 and 22.

## 20.3 DISTURBANCE PROCESSES

The preceding section has introduced a bit of the vocabulary and aspects of time-series specification. To obtain the theoretical results, we need to draw some conclusions about autocorrelation and add some details to that discussion.

### 20.3.1 CHARACTERISTICS OF DISTURBANCE PROCESSES

In the usual time-series setting, the disturbances are assumed to be homoscedastic but correlated across observations, so that

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}] = \sigma^2 \boldsymbol{\Omega},$$

where  $\sigma^2 \boldsymbol{\Omega}$  is a full, positive definite matrix with a constant  $\sigma^2 = \text{Var}[\varepsilon_t | \mathbf{X}]$  on the diagonal. As will be clear in the following discussion, we shall also assume that  $\boldsymbol{\Omega}_{ts}$  is a function of  $|t - s|$ , but not of  $t$  or  $s$  alone, which is a **stationarity** assumption. (See the preceding section.) It implies that the covariance between observations  $t$  and  $s$  is a function only of  $|t - s|$ , the distance apart in time of the observations. Because  $\sigma^2$  is not restricted, we normalize  $\boldsymbol{\Omega}_{tt} = 1$ . We define the **autocovariances**:

$$\text{Cov}[\varepsilon_t, \varepsilon_{t-s} | \mathbf{X}] = \text{Cov}[\varepsilon_{t+s}, \varepsilon_t | \mathbf{X}] = \sigma^2 \boldsymbol{\Omega}_{t,t-s} = \gamma_s = \gamma_{-s}.$$

Note that  $\sigma^2 \boldsymbol{\Omega}_{tt} = \gamma_0$ . The correlation between  $\varepsilon_t$  and  $\varepsilon_{t-s}$  is their autocorrelation,

$$\text{Corr}[\varepsilon_t, \varepsilon_{t-s} | \mathbf{X}] = \frac{\text{Cov}[\varepsilon_t, \varepsilon_{t-s} | \mathbf{X}]}{\sqrt{\text{Var}[\varepsilon_t | \mathbf{X}] \text{Var}[\varepsilon_{t-s} | \mathbf{X}]}} = \frac{\gamma_s}{\gamma_0} = \rho_s = \rho_{-s}.$$

We can then write

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}] = \boldsymbol{\Gamma} = \gamma_0 \mathbf{R},$$

where  $\boldsymbol{\Gamma}$  is an **autocovariance matrix** and  $\mathbf{R}$  is an **autocorrelation matrix**—the  $ts$  element is an **autocorrelation coefficient**

$$\rho_{ts} = \frac{\gamma_{|t-s|}}{\gamma_0}.$$

(Note that the matrix  $\boldsymbol{\Gamma} = \gamma_0 \mathbf{R}$  is the same as  $\sigma^2 \boldsymbol{\Omega}$ . The name change conforms to standard usage in the literature.) We will usually use the abbreviation  $\rho_s$  to denote the autocorrelation between observations  $s$  periods apart.

Different types of processes imply different patterns in  $\mathbf{R}$ . For example, the most frequently analyzed process is a **first-order autoregression** or **AR(1)** process,

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t,$$

## 910 PART V ♦ Time Series and Microeconometrics

where  $u_t$  is a stationary, nonautocorrelated (**white noise**) process and  $\rho$  is a parameter. We will verify later that for this process,  $\rho_s = \rho^s$ . Higher-order **autoregressive processes** of the form

$$\varepsilon_t = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_p \varepsilon_{t-p} + u_t$$

imply more involved patterns, including, for some values of the parameters, cyclical behavior of the autocorrelations.<sup>4</sup> Stationary autoregressions are structured so that the influence of a given disturbance fades as it recedes into the more distant past but vanishes only asymptotically. For example, for the AR(1),  $\text{Cov}[\varepsilon_t, \varepsilon_{t-s}]$  is never zero, but it does become negligible if  $|\rho|$  is less than 1. **Moving-average processes**, conversely, have a short memory. For the MA(1) process,

$$\varepsilon_t = u_t - \lambda u_{t-1},$$

the memory in the process is only one period:  $\gamma_0 = \sigma_u^2(1 + \lambda^2)$ ,  $\gamma_1 = -\lambda\sigma_u^2$ , but  $\gamma_s = 0$  if  $s > 1$ .

### 20.3.2 AR(1) DISTURBANCES

Time-series processes such as the ones listed here can be characterized by their order, the values of their parameters, and the behavior of their autocorrelations.<sup>5</sup> We shall consider various forms at different points. The received empirical literature is overwhelmingly dominated by the AR(1) model, which is partly a matter of convenience. Processes more involved than this model are usually extremely difficult to analyze. There is, however, a more practical reason. It is very optimistic to expect to know precisely the correct form of the appropriate model for the disturbance in any given situation. The first-order autoregression has withstood the test of time and experimentation as a reasonable *model* for underlying processes that probably, in truth, are impenetrably complex. AR(1) works as a first pass—higher-order models are often constructed as a refinement—as in the following example.

The first-order autoregressive disturbance, or AR(1) process, is represented in the **autoregressive form** as

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t, \tag{20-2}$$

where

$$E[u_t | \mathbf{X}] = 0,$$

$$E[u_t^2 | \mathbf{X}] = \sigma_u^2,$$

and

$$\text{Cov}[u_t, u_s | \mathbf{X}] = 0 \quad \text{if } t \neq s.$$

Because  $u_t$  is white noise, the conditional moments equal the unconditional moments. Thus  $E[\varepsilon_t | \mathbf{X}] = E[\varepsilon_t]$  and so on.

<sup>4</sup>This model is considered in more detail in Chapter 22.

<sup>5</sup>See Box and Jenkins (1984) for an authoritative study.

## CHAPTER 20 ♦ Serial Correlation 911

By repeated substitution, we have

$$\varepsilon_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \dots \quad (20-3)$$

From the preceding **moving-average form**, it is evident that each disturbance  $\varepsilon_t$  embodies the entire past history of the  $u$ 's, with the most recent observations receiving greater weight than those in the distant past. Depending on the sign of  $\rho$ , the series will exhibit clusters of positive and then negative observations or, if  $\rho$  is negative, regular oscillations of sign (as in Example 20.3).

Because the successive values of  $u_t$  are uncorrelated, the variance of  $\varepsilon_t$  is the variance of the right-hand side of (20-3):

$$\text{Var}[\varepsilon_t] = \sigma_u^2 + \rho^2 \sigma_u^2 + \rho^4 \sigma_u^2 + \dots \quad (20-4)$$

To proceed, a restriction must be placed on  $\rho$ ,

$$|\rho| < 1, \quad (20-5)$$

because otherwise, the right-hand side of (20-4) will become infinite. This result is the stationarity assumption discussed earlier. With (20-5), which implies that  $\lim_{s \rightarrow \infty} \rho^s = 0$ ,  $E[\varepsilon_t] = 0$  and

$$\text{Var}[\varepsilon_t] = \frac{\sigma_u^2}{1 - \rho^2} = \sigma_\varepsilon^2. \quad (20-6)$$

With the stationarity assumption, there is an easier way to obtain the variance

$$\text{Var}[\varepsilon_t] = \rho^2 \text{Var}[\varepsilon_{t-1}] + \sigma_u^2$$

because  $\text{Cov}[u_t, \varepsilon_s] = 0$  if  $t > s$ . With stationarity,  $\text{Var}[\varepsilon_{t-1}] = \text{Var}[\varepsilon_t]$ , which implies (20-6). Proceeding in the same fashion,

$$\text{Cov}[\varepsilon_t, \varepsilon_{t-1}] = E[\varepsilon_t \varepsilon_{t-1}] = E[\varepsilon_{t-1}(\rho \varepsilon_{t-1} + u_t)] = \rho \text{Var}[\varepsilon_{t-1}] = \frac{\rho \sigma_u^2}{1 - \rho^2}. \quad (20-7)$$

By repeated substitution in (20-2), we see that for any  $s$ ,

$$\varepsilon_t = \rho^s \varepsilon_{t-s} + \sum_{i=0}^{s-1} \rho^i u_{t-i}$$

(e.g.,  $\varepsilon_t = \rho^3 \varepsilon_{t-3} + \rho^2 u_{t-2} + \rho u_{t-1} + u_t$ ). Therefore, because  $\varepsilon_s$  is not correlated with any  $u_t$  for which  $t > s$  (i.e., any subsequent  $u_t$ ), it follows that

$$\text{Cov}[\varepsilon_t, \varepsilon_{t-s}] = E[\varepsilon_t \varepsilon_{t-s}] = \frac{\rho^s \sigma_u^2}{1 - \rho^2}. \quad (20-8)$$

Dividing by  $\gamma_0 = \sigma_u^2 / (1 - \rho^2)$  provides the autocorrelations:

$$\text{Corr}[\varepsilon_t, \varepsilon_{t-s}] = \rho_s = \rho^s. \quad (20-9)$$

With the stationarity assumption, the autocorrelations fade over time. Depending on the sign of  $\rho$ , they will either be declining in geometric progression or alternating in

## 912 PART V ♦ Time Series and Microeconometrics

sign if  $\rho$  is negative. Collecting terms, we have

$$\sigma^2 \boldsymbol{\Omega} = \frac{\sigma_u^2}{1 - \rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \rho \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & \rho & 1 \end{bmatrix}. \quad (20-10)$$

### 20.4 SOME ASYMPTOTIC RESULTS FOR ANALYZING TIME-SERIES DATA

Because  $\boldsymbol{\Omega}$  is not equal to  $\mathbf{I}$ , the now-familiar complications will arise in establishing the properties of estimators of  $\beta$ , in particular of the least squares estimator. The finite sample properties of the OLS and GLS estimators remain intact. Least squares will continue to be unbiased; The earlier general proof allows for autocorrelated disturbances. The Aitken theorem (Theorem 9.4) and the distributional results for normally distributed disturbances can still be established conditionally on  $\mathbf{X}$ . (However, even these will be complicated when  $\mathbf{X}$  contains lagged values of the dependent variable.) But, finite sample properties are of very limited usefulness in time-series contexts. Nearly all that can be said about estimators involving time-series data is based on their asymptotic properties.

As we saw in our analysis of heteroscedasticity, whether least squares is consistent or not, depends on the matrices

$$\mathbf{Q}_T = (1/T)\mathbf{X}'\mathbf{X},$$

and

$$\mathbf{Q}_T^* = (1/T)\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}.$$

In our earlier analyses, we were able to argue for convergence of  $\mathbf{Q}_T$  to a positive definite matrix of constants,  $\mathbf{Q}$ , by invoking laws of large numbers. But, these theorems assume that the observations in the sums are independent, which as suggested in Section 20.2, is surely not the case here. Thus, we require a different tool for this result. We can expand the matrix  $\mathbf{Q}_T^*$  as

$$\mathbf{Q}_T^* = \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T \rho_{ts} \mathbf{x}_t' \mathbf{x}_s', \quad (20-11)$$

where  $\mathbf{x}_t'$  and  $\mathbf{x}_s'$  are rows of  $\mathbf{X}$  and  $\rho_{ts}$  is the autocorrelation between  $\varepsilon_t$  and  $\varepsilon_s$ . Sufficient conditions for this matrix to converge are that  $\mathbf{Q}_T$  converge and that the correlations between disturbances die off reasonably rapidly as the observations become further apart in time. For example, if the disturbances follow the AR(1) process described earlier, then  $\rho_{ts} = \rho^{|t-s|}$  and if  $\mathbf{x}_t$  is sufficiently well behaved,  $\mathbf{Q}_T^*$  will converge to a positive definite matrix  $\mathbf{Q}^*$  as  $T \rightarrow \infty$ .

**Asymptotic normality** of the least squares and GLS estimators will depend on the behavior of sums such as

$$\sqrt{T} \bar{\mathbf{w}}_T = \sqrt{T} \left( \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \varepsilon_t \right) = \sqrt{T} \left( \frac{1}{T} \mathbf{X}' \boldsymbol{\varepsilon} \right).$$

Asymptotic normality of least squares is difficult to establish for this general model. The central limit theorems we have relied on thus far do not extend to sums of *dependent* observations. The results of Amemiya (1985), Mann and Wald (1943), and Anderson (1971) do carry over to most of the familiar types of autocorrelated disturbances, including those that interest us here, so we shall ultimately conclude that ordinary least squares, GLS, and instrumental variables continue to be consistent and asymptotically normally distributed, and, in the case of OLS, inefficient. This section will provide a brief introduction to some of the underlying principles that are used to reach these conclusions.

#### 20.4.1 CONVERGENCE OF MOMENTS—THE ERGODIC THEOREM

The discussion thus far has suggested (appropriately) that stationarity (or its absence) is an important characteristic of a process. The points at which we have encountered this notion concerned requirements that certain sums converge to finite values. In particular, for the AR(1) model,  $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$ , for the variance of the process to be finite, we require  $|\rho| < 1$ , which is a sufficient condition. However, this result is only a byproduct. Stationarity (at least, the weak stationarity we have examined) is only a characteristic of the sequence of moments of a distribution.

#### **DEFINITION 20.1 Strong Stationarity**

A time-series process,  $\{z_t\}_{t=-\infty}^{t=\infty}$  is strongly stationary, or “stationary,” if the joint probability distribution of any set of  $k$  observations in the sequence  $[z_t, z_{t+1}, \dots, z_{t+k-1}]$  is the same regardless of the origin,  $t$ , in the time scale.

For example, in (20-2), if we add  $u_t \sim N[0, \sigma_u^2]$ , then the resulting process  $\{\varepsilon_t\}_{t=-\infty}^{t=\infty}$  can easily be shown to be strongly stationary.

#### **DEFINITION 20.2 Weak Stationarity**

A time-series process,  $\{z_t\}_{t=-\infty}^{t=\infty}$  is weakly stationary (or covariance stationary) if  $E[z_t]$  is finite and is the same for all  $t$  and if the covariances between any two observations (labeled their autocovariance),  $\text{Cov}[z_t, z_{t-k}]$ , is a finite function only of model parameters and their distance apart in time,  $k$ , but not of the absolute location of either observation on the time scale.

Weak stationary is obviously implied by strong stationary, although it requires less because the distribution can, at least in principle, be changing on the time axis. The distinction is rarely necessary in applied work. In general, save for narrow theoretical examples, it will be difficult to come up with a process that is weakly but not strongly stationary. The reason for the distinction is that in much of our work, only weak stationary is required, and, as always, when possible, econometricians will dispense with unnecessary assumptions.

## 914 PART V ♦ Time Series and Microeconometrics

As we will discover shortly, stationarity is a crucial characteristic at this point in the analysis. If we are going to proceed to parameter estimation in this context, we will also require another characteristic of a time series, **ergodicity**. There are various ways to delineate this characteristic, none of them particularly intuitive. We borrow one definition from Davidson and MacKinnon (1993, p. 132) which comes close:

### DEFINITION 20.3 Ergodicity

*A strongly stationary time-series process,  $\{z_t\}_{t=-\infty}^{t=\infty}$ , is ergodic if for any two bounded functions that map vectors in the  $a$  and  $b$  dimensional real vector spaces to real scalars,  $f: \mathbf{R}^a \rightarrow \mathbf{R}^1$  and  $g: \mathbf{R}^b \rightarrow \mathbf{R}^1$ ,*

$$\begin{aligned} & \lim_{k \rightarrow \infty} |E[f(z_t, z_{t+1}, \dots, z_{t+a})g(z_{t+k}, z_{t+k+1}, \dots, z_{t+k+b})]| \\ &= |E[f(z_t, z_{t+1}, \dots, z_{t+a-1})]| |E[g(z_{t+k}, z_{t+k+1}, \dots, z_{t+k+b-1})]|. \end{aligned}$$

The definition states essentially that if events are separated far enough in time, then they are “asymptotically independent.” An implication is that in a time series, every observation will contain at least some unique information. Ergodicity is a crucial element of our theory of estimation. When a time series has this property (with stationarity), then we can consider estimation of parameters in a meaningful sense.<sup>6</sup> The analysis relies heavily on the following theorem:

### THEOREM 20.1 The Ergodic Theorem

*If  $\{z_t\}_{t=-\infty}^{t=\infty}$  is a time-series process that is strongly stationary and ergodic and  $E[|z_t|]$  is a finite constant, and if  $\bar{z}_T = (1/T) \sum_{t=1}^T z_t$ , then  $\bar{z}_T \xrightarrow{a.s.} \mu$ , where  $\mu = E[z_t]$ . Note that the convergence is almost surely not in probability (which is implied) or in mean square (which is also implied). [See White (2001, p. 44) and Davidson and MacKinnon (1993, p. 133).]*

What we have in the ergodic theorem is, for sums of dependent observations, a counterpart to the laws of large numbers that we have used at many points in the preceding chapters. Note, once again, the need for this extension is that to this point, our laws of large numbers have required sums of independent observations. But, in this context, by design, observations are distinctly not independent.

<sup>6</sup>Much of the analysis in later chapters will encounter nonstationary series, which are the focus of most of the current literature—tests for nonstationarity largely dominate the recent study in time-series analysis. Ergodicity is a much more subtle and difficult concept. For any process that we will consider, ergodicity will have to be a given, at least at this level. A classic reference on the subject is Doob (1953). Another authoritative treatise is Billingsley (1995). White (2001) provides a concise analysis of many of these concepts as used in econometrics, and some useful commentary.

For this result to be useful, we will require an extension.

### THEOREM 20.2 Ergodicity of Functions

If  $\{z_t\}_{t=-\infty}^{\infty}$  is a time-series process that is strongly stationary and ergodic and if  $y_t = f\{z_t\}$  is a measurable function in the probability space that defines  $z_t$ , then  $y_t$  is also stationary and ergodic. Let  $\{\mathbf{z}_t\}_{t=-\infty}^{\infty}$  define a  $K \times 1$  vector valued stochastic process—each element of the vector is an ergodic and stationary series, and the characteristics of ergodicity and stationarity apply to the joint distribution of the elements of  $\{\mathbf{z}_t\}_{t=-\infty}^{\infty}$ . Then, the ergodic theorem applies to functions of  $\{\mathbf{z}_t\}_{t=-\infty}^{\infty}$ . [See White (2001, pp. 44–45) for discussion.]

Theorem 20.2 produces the results we need to characterize the least squares (and other) estimators. In particular, our minimal assumptions about the data are

**ASSUMPTION 20.1. Ergodic Data Series:** In the regression model,  $y_t = \mathbf{x}'_t \boldsymbol{\beta} + \varepsilon_t$ ,  $[\mathbf{x}_t, \varepsilon_t]_{t=-\infty}^{\infty}$  is a jointly stationary and ergodic process.

By analyzing terms element by element we can use these results directly to assert that averages of  $\mathbf{w}_t = \mathbf{x}_t \varepsilon_t$ ,  $\mathbf{Q}_{tt} = \mathbf{x}_t \mathbf{x}'_t$ , and  $\mathbf{Q}_{tt}^* = \varepsilon_t^2 \mathbf{x}_t \mathbf{x}'_t$  will converge to their population counterparts,  $\mathbf{0}$ ,  $\mathbf{Q}$  and  $\mathbf{Q}^*$ .

#### 20.4.2 CONVERGENCE TO NORMALITY—A CENTRAL LIMIT THEOREM

To form a distribution theory for least squares, GLS, ML, and GMM, we will need a counterpart to the central limit theorem. In particular, we need to establish a large sample distribution theory for quantities of the form

$$\sqrt{T} \left( \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \varepsilon_t \right) = \sqrt{T} \bar{\mathbf{w}}.$$

As noted earlier, we cannot invoke the familiar central limit theorems (Lindeberg–Levy, Lindeberg–Feller, Liapounov) because the observations in the sum are not independent. But, with the assumptions already made, we do have an alternative result. Some needed preliminaries are as follows:

### DEFINITION 20.4 Martingale Sequence

A vector sequence  $\mathbf{z}_t$  is a martingale sequence if  $E[\mathbf{z}_t | \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \dots] = \mathbf{z}_{t-1}$ .

An important example of a martingale sequence is the **random walk**,

$$z_t = z_{t-1} + u_t,$$

where  $\text{Cov}[u_t, u_s] = 0$  for all  $t \neq s$ . Then

$$E[z_t | z_{t-1}, z_{t-2}, \dots] = E[z_{t-1} | z_{t-1}, z_{t-2}, \dots] + E[u_t | z_{t-1}, z_{t-2}, \dots] = z_{t-1} + 0 = z_{t-1}.$$

**916 PART V ♦ Time Series and Microeconometrics**
**DEFINITION 20.5 Martingale Difference Sequence**

*A vector sequence  $\mathbf{z}_t$  is a martingale difference sequence if  $E[\mathbf{z}_t | \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \dots] = \mathbf{0}$ .*

With Definition 20.5, we have the following broadly encompassing result:

**THEOREM 20.3 Martingale Difference Central Limit Theorem**

*If  $\mathbf{z}_t$  is a vector valued stationary and ergodic martingale difference sequence, with  $E[\mathbf{z}_t \mathbf{z}'_t] = \Sigma$ , where  $\Sigma$  is a finite positive definite matrix, and if  $\bar{\mathbf{z}}_T = (1/T) \sum_{t=1}^T \mathbf{z}_t$ , then  $\sqrt{T} \bar{\mathbf{z}}_T \xrightarrow{d} N[\mathbf{0}, \Sigma]$ . [For discussion, see Davidson and MacKinnon (1993, Sections 4.7 and 4.8).]<sup>7</sup>*

Theorem 20.3 is a generalization of the Lindeberg–Levy central limit theorem. It is not yet broad enough to cover cases of autocorrelation, but it does go beyond Lindeberg–Levy, for example, in extending to the GARCH model of Section 20.13.3. [Forms of the theorem that surpass Lindeberg–Feller (D.19) and Liapounov (Theorem D.20) by allowing for different variances at each time,  $t$ , appear in Ruud (2000, p. 479) and White (2001, p. 133). These variants extend beyond our requirements in this treatment.] But, looking ahead, this result encompasses what will be a very important application. Suppose in the classical linear regression model,  $\{\mathbf{x}_t\}_{t=-\infty}^{t=\infty}$  is a stationary and ergodic multivariate stochastic process and  $\{\varepsilon_t\}_{t=-\infty}^{t=\infty}$  is an i.i.d. process—that is, not autocorrelated and not heteroscedastic. Then, this is the most general case of the classical model that still maintains the assumptions about  $\varepsilon_t$  that we made in Chapter 2. In this case, the process  $\{\mathbf{w}_t\}_{t=-\infty}^{t=\infty} = \{\mathbf{x}_t \varepsilon_t\}_{t=-\infty}^{t=\infty}$  is a martingale difference sequence, so that with sufficient assumptions on the moments of  $\mathbf{x}_t$  we could use this result to establish consistency and asymptotic normality of the least squares estimator. [See, e.g., Hamilton (1994, pp. 208–212).]

We now consider a central limit theorem that is broad enough to include the case that interested us at the outset, stochastically dependent observations on  $\mathbf{x}_t$  and autocorrelation in  $\varepsilon_t$ .<sup>8</sup> Suppose as before that  $\{\mathbf{z}_t\}_{t=-\infty}^{t=\infty}$  is a stationary and ergodic stochastic process. We consider  $\sqrt{T} \bar{\mathbf{z}}_T$ . The following conditions are assumed.<sup>9</sup>

1. **Asymptotic uncorrelatedness:**  $E[\mathbf{z}_t | \mathbf{z}_{t-k}, \mathbf{z}_{t-k-1}, \dots]$  converges in mean square to zero as  $k \rightarrow \infty$ . Note that is similar to the condition for ergodicity. White (2001) demonstrates that a (nonobvious) implication of this assumption is  $E[\mathbf{z}_t] = \mathbf{0}$ .

<sup>7</sup>For convenience, we are bypassing a step in this discussion—establishing multivariate normality requires that the result first be established for the marginal normal distribution of each component, then that every linear combination of the variables also be normally distributed (See Theorems D.17 and D.18A.). Our interest at this point is merely to collect the useful end results. Interested users may find the detailed discussions of the many subtleties and narrower points in White (2001) and Davidson and MacKinnon (1993, Chapter 4).

<sup>8</sup>Detailed analysis of this case is quite intricate and well beyond the scope of this book. Some fairly terse analysis may be found in White (2001, pp. 122–133) and Hayashi (2000).

<sup>9</sup>See Hayashi (2000, p. 405) who attributes the results to Gordin (1969).

**2. Summability of autocovariances:** With dependent observations,

$$\lim_{T \rightarrow \infty} \text{Var}[\sqrt{T} \bar{\mathbf{z}}_T] = \sum_{t=0}^{\infty} \sum_{s=0}^{\infty} \text{Cov}[\mathbf{z}_t, \mathbf{z}'_s] = \sum_{k=-\infty}^{\infty} \boldsymbol{\Gamma}_k = \boldsymbol{\Gamma}^*.$$

To begin, we will need to assume that this matrix is finite, a condition called **summability**. Note this is the condition needed for convergence of  $\mathbf{Q}_T^*$  in (20-11). If the sum is to be finite, then the  $k = 0$  term must be finite, which gives us a necessary condition

$$E[\mathbf{z}_t \mathbf{z}'_t] = \boldsymbol{\Gamma}_0, \text{ a finite matrix.}$$

**3. Asymptotic negligibility of innovations:** Let

$$\mathbf{r}_{tk} = E[\mathbf{z}_t | \mathbf{z}_{t-k}, \mathbf{z}_{t-k-1}, \dots] - E[\mathbf{z}_t | \mathbf{z}_{t-k-1}, \mathbf{z}_{t-k-2}, \dots].$$

An observation  $\mathbf{z}_t$  may be viewed as the accumulated information that has entered the process since it began up to time  $t$ . Thus, it can be shown that

$$\mathbf{z}_t = \sum_{s=0}^{\infty} \mathbf{r}_{ts}.$$

The vector  $\mathbf{r}_{tk}$  can be viewed as the information in this accumulated sum that entered the process at time  $t - k$ . The condition imposed on the process is that  $\sum_{s=0}^{\infty} \sqrt{E[\mathbf{r}'_{ts} \mathbf{r}_{ts}]}$  be finite. In words, condition (3) states that information eventually becomes negligible as it fades far back in time from the current observation. The AR(1) model (as usual) helps to illustrate this point. If  $z_t = \rho z_{t-1} + u_t$ , then

$$\begin{aligned} r_{t0} &= E[z_t | z_t, z_{t-1}, \dots] - E[z_t | z_{t-1}, z_{t-2}, \dots] = z_t - \rho z_{t-1} = u_t \\ r_{t1} &= E[z_t | z_{t-1}, z_{t-2}, \dots] - E[z_t | z_{t-2}, z_{t-3}, \dots] \\ &= E[\rho z_{t-1} + u_t | z_{t-1}, z_{t-2}, \dots] - E[\rho(\rho z_{t-2} + u_{t-1}) + u_t | z_{t-2}, z_{t-3}, \dots] \\ &= \rho(z_{t-1} - \rho z_{t-2}) \\ &= \rho u_{t-1}. \end{aligned}$$

By a similar construction,  $r_{tk} = \rho^k u_{t-k}$  from which it follows that  $z_t = \sum_{s=0}^{\infty} \rho^s u_{t-s}$ , which we saw earlier in (20-3). You can verify that if  $|\rho| < 1$ , the negligibility condition will be met.

With all this machinery in place, we now have the theorem we will need:

**THEOREM 20.4 Gordin's Central Limit Theorem**

If  $\mathbf{z}_t$  is strongly stationary and ergodic and if conditions (1)–(3) are met, then  $\sqrt{T} \bar{\mathbf{z}}_T \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Gamma}^*]$ .

We will be able to employ these tools when we consider the least squares, IV, and GLS estimators in the discussion to follow.

**918 PART V ♦ Time Series and Microeconometrics**
**20.5 LEAST SQUARES ESTIMATION**

The least squares estimator is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + \left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1} \left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{T}\right).$$

Unbiasedness follows from the results in Chapter 4—no modification is needed. We know from Chapter 9 that the Gauss–Markov theorem has been lost—assuming it exists (that remains to be established), the GLS estimator is efficient and OLS is not. How much information is lost by using least squares instead of GLS depends on the data. Broadly, least squares fares better in data that have long periods and little cyclical variation, such as aggregate output series. As might be expected, the greater is the auto-correlation in  $\boldsymbol{\varepsilon}$ , the greater will be the benefit to using generalized least squares (when this is possible). Even if the disturbances are normally distributed, the usual  $F$  and  $t$  statistics do not have those distributions. So, not much remains of the finite sample properties we obtained in Chapter 4. The asymptotic properties remain to be established.

**20.5.1 ASYMPTOTIC PROPERTIES OF LEAST SQUARES**

The asymptotic properties of  $\mathbf{b}$  are straightforward to establish given our earlier results. If we assume that the process generating  $\mathbf{x}_t$  is stationary and ergodic, then by Theorems 20.1 and 20.2,  $(1/T)(\mathbf{X}'\mathbf{X})$  converges to  $\mathbf{Q}$  and we can apply the Slutsky theorem to the inverse. If  $\boldsymbol{\varepsilon}_t$  is not serially correlated, then  $\mathbf{w}_t = \mathbf{x}_t \boldsymbol{\varepsilon}_t$  is a martingale difference sequence, so  $(1/T)(\mathbf{X}'\boldsymbol{\varepsilon})$  converges to zero. This establishes consistency for the simple case. On the other hand, if  $[\mathbf{x}_t, \boldsymbol{\varepsilon}_t]$  are jointly stationary and ergodic, then we can invoke the ergodic theorems 20.1 and 20.2 for both moment matrices and establish consistency. Asymptotic normality is a bit more subtle. For the case without serial correlation in  $\boldsymbol{\varepsilon}_t$ , we can employ Theorem 20.3 for  $\sqrt{T}\bar{\mathbf{w}}$ . The involved case is the one that interested us at the outset of this discussion, that is, where there is autocorrelation in  $\boldsymbol{\varepsilon}_t$  and dependence in  $\mathbf{x}_t$ . Theorem 20.4 is in place for this case. Once again, the conditions described in the preceding section must apply and, moreover, the assumptions needed will have to be established both for  $\mathbf{x}_t$  and  $\boldsymbol{\varepsilon}_t$ . Commentary on these cases may be found in Davidson and MacKinnon (1993), Hamilton (1994), White (2001), and Hayashi (2000). Formal presentation extends beyond the scope of this text, so at this point, we will proceed, and assume that the conditions underlying Theorem 20.4 are met. The results suggested here are quite general, albeit only sketched for the general case. For the remainder of our examination, at least in this chapter, we will confine attention to fairly simple processes in which the necessary conditions for the asymptotic distribution theory will be fairly evident.

There is an important exception to the results in the preceding paragraph. If the regression contains any lagged values of the dependent variable, then least squares will no longer be unbiased or consistent. To take the simplest case, suppose that

$$\begin{aligned} y_t &= \beta y_{t-1} + \boldsymbol{\varepsilon}_t, \\ \boldsymbol{\varepsilon}_t &= \rho \boldsymbol{\varepsilon}_{t-1} + u_t, \end{aligned} \tag{20-12}$$

## CHAPTER 20 ♦ Serial Correlation 919

and assume  $|\beta| < 1$ ,  $|\rho| < 1$ . In this model, the regressor and the disturbance are correlated. There are various ways to approach the analysis. One useful way is to rearrange (20-12) by subtracting  $\rho y_{t-1}$  from  $y_t$ . Then,

$$y_t = (\beta + \rho)y_{t-1} - \beta\rho y_{t-2} + u_t, \quad (20-13)$$

which is a classical regression with stochastic regressors. Because  $u_t$  is an innovation in period  $t$ , it is uncorrelated with both regressors, and least squares regression of  $y_t$  on  $(y_{t-1}, y_{t-2})$  estimates  $\rho_1 = (\beta + \rho)$  and  $\rho_2 = -\beta\rho$ . What is estimated by regression of  $y_t$  on  $y_{t-1}$  alone? Let  $\gamma_k = \text{Cov}[y_t, y_{t-k}] = \text{Cov}[y_t, y_{t+k}]$ . By stationarity,  $\text{Var}[y_t] = \text{Var}[y_{t-1}]$ , and  $\text{Cov}[y_t, y_{t-1}] = \text{Cov}[y_{t-1}, y_{t-2}]$ , and so on. These and (20-13) imply the following relationships:

$$\begin{aligned}\gamma_0 &= \rho_1\gamma_1 + \rho_2\gamma_2 + \sigma_u^2, \\ \gamma_1 &= \rho_1\gamma_0 + \rho_2\gamma_1, \\ \gamma_2 &= \rho_1\gamma_1 + \rho_2\gamma_0.\end{aligned}\quad (20-14)$$

(These are the **Yule–Walker equations** for this model. See Section 22.2.3.) The slope in the simple regression estimates  $\gamma_1/\gamma_0$ , which can be found in the solutions to these three equations. (An alternative approach is to use the left-out variable formula, which is a useful way to interpret this estimator.) In this case, we see that the slope in the short regression is an estimator of  $(\beta + \rho) - \beta\rho(\gamma_1/\gamma_0)$ . In either case, solving the three equations in (20-14) for  $\gamma_0$ ,  $\gamma_1$ , and  $\gamma_2$  in terms of  $\rho_1$ ,  $\rho_2$ , and  $\sigma_u^2$  produces

$$\text{plim } b = \frac{\beta + \rho}{1 + \beta\rho}. \quad (20-15)$$

This result is between  $\beta$  (when  $\rho = 0$ ) and 1 (when both  $\beta$  and  $\rho = 1$ ). Therefore, least squares is inconsistent unless  $\rho$  equals zero. The more general case that includes regressors,  $\mathbf{x}_t$ , involves more complicated algebra but gives essentially the same result. This is a general result; When the equation contains a lagged dependent variable in the presence of autocorrelation, OLS and GLS are inconsistent. The problem can be viewed as one of an omitted variable.

#### 20.5.2 ESTIMATING THE VARIANCE OF THE LEAST SQUARES ESTIMATOR

As usual,  $s^2(\mathbf{X}'\mathbf{X})^{-1}$  is an inappropriate estimator of  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Omega\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$ , both because  $s^2$  is a biased estimator of  $\sigma^2$  and because the matrix is incorrect. Generalities are scarce, but in general, for economic time series that are positively related to their past values, the standard errors conventionally *estimated* by least squares are likely to be too small. For slowly changing, trending aggregates such as output and consumption, this is probably the norm. For highly variable data such as inflation, exchange rates, and market returns, the situation is less clear. Nonetheless, as a general proposition, one would normally not want to rely on  $s^2(\mathbf{X}'\mathbf{X})^{-1}$  as an estimator of the asymptotic covariance matrix of the least squares estimator.

In view of this situation, if one is going to use least squares, then it is desirable to have an appropriate estimator of the covariance matrix of the least squares estimator. There are two approaches. If the form of the autocorrelation is known, then one can estimate the parameters of  $\Omega$  directly and compute a consistent estimator. Of course,

## 920 PART V ♦ Time Series and Microeconometrics

if so, then it would be more sensible to use feasible generalized least squares instead and not waste the sample information on an inefficient estimator. The second approach parallels the use of the White estimator for heteroscedasticity.

The extension of White's result to the more general case of autocorrelation is much more difficult than in the heteroscedasticity case. The natural counterpart for estimating

$$\mathbf{Q}_* = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \sigma_{ij} \mathbf{x}_i \mathbf{x}'_j \quad (20-16)$$

in Section 9.2.3 would be

$$\hat{\mathbf{Q}}_* = \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T e_t e_s \mathbf{x}_t \mathbf{x}'_s.$$

But there are two problems with this estimator, one theoretical, which applies to  $\mathbf{Q}_*$  as well, and one practical, which is specific to the latter.

Unlike the heteroscedasticity case, the matrix in (20-16) is  $1/T$  times a sum of  $T^2$  terms, so it is difficult to conclude yet that it will converge to anything at all. This application is most likely to arise in a time-series setting. To obtain convergence, it is necessary to assume that the terms involving unequal subscripts in (20-16) diminish in importance as  $T$  grows. A sufficient condition is that terms with subscript pairs  $|t - s|$  grow smaller as the distance between them grows larger. In practical terms, observation pairs are progressively less correlated as their separation in time grows. Intuitively, if one can think of weights with the diagonal elements getting a weight of 1.0, then in the sum, the weights in the sum grow smaller as we move away from the diagonal. If we think of the sum of the weights rather than just the number of terms, then this sum falls off sufficiently rapidly that as  $n$  grows large, the sum is of order  $T$  rather than  $T^2$ . Thus, we achieve convergence of  $\mathbf{Q}_*$  by assuming that the rows of  $\mathbf{X}$  are well behaved [4.7.6] that the correlations diminish with increasing separation in time. (See Sections 4.7.6 and 22.2.5 for a more formal statement of this condition.)

The practical problem is that  $\hat{\mathbf{Q}}_*$  need not be positive definite. Newey and West (1987a) have devised an estimator that overcomes this difficulty:

$$\begin{aligned} \hat{\mathbf{Q}}_* &= \mathbf{S}_0 + \frac{1}{T} \sum_{l=1}^L \sum_{t=l+1}^T w_l e_t e_{t-l} (\mathbf{x}_t \mathbf{x}'_{t-l} + \mathbf{x}_{t-l} \mathbf{x}'_t), \\ w_l &= 1 - \frac{l}{(L+1)}. \end{aligned} \quad (20-17)$$

[See (9-26).] The **Newey-West autocorrelation consistent covariance estimator** is surprisingly simple and relatively easy to implement.<sup>10</sup> There is a final problem to be solved. It must be determined in advance how large  $L$  is to be. In general, there is little theoretical guidance. Current practice specifies  $L \approx T^{1/4}$ . Unfortunately, the result is not quite as crisp as that for the heteroscedasticity consistent estimator.

We have the result that  $\mathbf{b}$  and  $\mathbf{b}_{IV}$  are asymptotically normally distributed, and we have an appropriate estimator for the asymptotic covariance matrix. We have not

<sup>10</sup>Both estimators are now standard features in modern econometrics computer programs. Further results on different weighting schemes may be found in Hayashi (2000, pp. 406–410).

**TABLE 20.1** Robust Covariance Estimation

<i>Variable</i>	<i>OLS Estimate</i>	<i>OLS SE</i>	<i>Corrected SE</i>
Constant	-1.6331	0.2286	0.3335
In Output	0.2871	0.04738	0.07806
In CPI	0.9718	0.03377	0.06585
$R^2 = 0.98952, d = 0.02477, r = 0.98762.$			

specified the distribution of the disturbances, however. Thus, for inference purposes, the  $F$  statistic is approximate at best. Moreover, for more involved hypotheses, the likelihood ratio and Lagrange multiplier tests are unavailable. That leaves the Wald statistic, including asymptotic “ $t$  ratios,” as the main tool for statistical inference. We will examine a number of applications in the chapters to follow.

The White and Newey–West estimators are standard in the econometrics literature. We will encounter them at many points in the discussion to follow.

**Example 20.4 Autocorrelation Consistent Covariance Estimation**

For the model shown in Example 20.1, the regression results with the uncorrected standard errors and the Newey–West autocorrelation robust covariance matrix for lags of five quarters are shown in Table 20.1. The effect of the very high degree of autocorrelation is evident.

## 20.6 GMM ESTIMATION

The **GMM estimator** in the regression model with autocorrelated disturbances is produced by the empirical moment equations

$$\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t (y_t - \mathbf{x}'_t \hat{\boldsymbol{\beta}}_{GMM}) = \frac{1}{T} \mathbf{X}' \hat{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\beta}}_{GMM}) = \bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}_{GMM}) = \mathbf{0}. \quad (20-18)$$

The estimator is obtained by minimizing

$$q = \bar{\mathbf{m}}'(\hat{\boldsymbol{\beta}}_{GMM}) \mathbf{W} \bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}_{GMM})$$

where  $\mathbf{W}$  is a positive definite weighting matrix. The optimal weighting matrix would be

$$\mathbf{W} = \left\{ \text{Asy. Var}[\sqrt{T} \bar{\mathbf{m}}(\boldsymbol{\beta})] \right\}^{-1},$$

which is the inverse of

$$\text{Asy. Var}[\sqrt{T} \bar{\mathbf{m}}(\boldsymbol{\beta})] = \text{Asy. Var} \left[ \frac{1}{\sqrt{T}} \sum_{i=1}^n \mathbf{x}_i \varepsilon_i \right] = \text{plim}_{n \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T \sigma^2 \rho_{ts} \mathbf{x}_t \mathbf{x}'_s = \sigma^2 \mathbf{Q}^*.$$

The optimal weighting matrix would be  $[\sigma^2 \mathbf{Q}^*]^{-1}$ . As in the heteroscedasticity case, this minimization problem is an exactly identified case, so, the weighting matrix is actually irrelevant to the solution. *The GMM estimator for the regression model with autocorrelated disturbances is ordinary least squares.* We can use the results in Section 20.5.2 to construct the asymptotic covariance matrix. We will require the assumptions in Section 20.4 to obtain convergence of the moments and asymptotic normality. We will wish to extend this simple result in one instance. In the common case in which  $\mathbf{x}_t$  contains

**922 PART V ♦ Time Series and Microeconometrics**

lagged values of  $y_t$ , we will want to use an instrumental variable estimator. We will return to that estimation problem in Section 20.9.3.

## 20.7 TESTING FOR AUTOCORRELATION

The available tests for autocorrelation are based on the principle that if the true disturbances are autocorrelated, then this fact can be detected through the autocorrelations of the least squares residuals. The simplest indicator is the slope in the artificial regression

$$\begin{aligned} e_t &= r e_{t-1} + v_t, \\ e_t &= y_t - \mathbf{x}'_t \mathbf{b}, \\ r &= \left( \sum_{t=2}^T e_t e_{t-1} \right) \Bigg/ \left( \sum_{t=1}^{T-1} e_t^2 \right). \end{aligned} \tag{20-19}$$

If there is autocorrelation, then the slope in this regression will be an estimator of  $\rho = \text{Corr}[\varepsilon_t, \varepsilon_{t-1}]$ . The complication in the analysis lies in determining a formal means of evaluating when the estimator is “large,” that is, on what statistical basis to reject the null hypothesis that  $\rho$  equals zero. As a first approximation, treating (20-19) as a classical linear model and using a  $t$  or  $F$  (squared  $t$ ) test to test the hypothesis is a valid way to proceed based on the Lagrange multiplier principle. We used this device in Example 20.3. The tests we consider here are refinements of this approach.

### 20.7.1 LAGRANGE MULTIPLIER TEST

The Breusch (1978)–Godfrey (1978) test is a Lagrange multiplier test of  $H_0$ : no autocorrelation versus  $H_1: \varepsilon_t = \text{AR}(P)$  or  $\varepsilon_t = \text{MA}(P)$ . The same test is used for either structure. The test statistic is

$$\text{LM} = T \left( \frac{\mathbf{e}' \mathbf{X}_0 (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{X}'_0 \mathbf{e}}{\mathbf{e}' \mathbf{e}} \right) = TR_0^2, \tag{20-20}$$

where  $\mathbf{X}_0$  is the original  $\mathbf{X}$  matrix augmented by  $P$  additional columns containing the lagged OLS residuals,  $e_{t-1}, \dots, e_{t-P}$ . The test can be carried out simply by regressing the ordinary least squares residuals  $e_t$  on  $\mathbf{x}_{t0}$  (filling in missing values for lagged residuals with zeros) and referring  $TR_0^2$  to the tabled critical value for the chi-squared distribution with  $P$  degrees of freedom.<sup>11</sup> Because  $\mathbf{X}' \mathbf{e} = \mathbf{0}$ , the test is equivalent to regressing  $e_t$  on the part of the lagged residuals that is unexplained by  $\mathbf{X}$ . There is therefore a compelling logic to it; if any fit is found, then it is due to correlation between the current and lagged residuals. The test is a joint test of the first  $P$  autocorrelations of  $\varepsilon_t$ , not just the first.

### 20.7.2 BOX AND PIERCE'S TEST AND LJUNG'S REFINEMENT

An alternative test that is asymptotically equivalent to the LM test when the null hypothesis,  $\rho = 0$ , is true and when  $\mathbf{X}$  does not contain lagged values of  $y$  is due to Box

---

<sup>11</sup>A warning to practitioners: Current software varies on whether the lagged residuals are filled with zeros or the first  $P$  observations are simply dropped when computing this statistic. In the interest of replicability, users should determine which is the case before reporting results.

and Pierce (1970). The ***Q* test** is carried out by referring

$$Q = T \sum_{j=1}^P r_j^2, \quad (20-21)$$

where  $r_j = (\sum_{t=j+1}^T e_t e_{t-j}) / (\sum_{t=1}^T e_t^2)$ , to the critical values of the chi-squared table with  $P$  degrees of freedom. A refinement suggested by Ljung and Box (1979) is

$$Q' = T(T+2) \sum_{j=1}^P \frac{r_j^2}{T-j}. \quad (20-22)$$

The essential difference between the Godfrey–Breusch and the Box–Pierce tests is the use of partial correlations (controlling for  $\mathbf{X}$  and the other variables) in the former and simple correlations in the latter. Under the null hypothesis, there is no autocorrelation in  $\varepsilon_t$ , and no correlation between  $\mathbf{x}_t$  and  $\varepsilon_s$  in any event, so the two tests are asymptotically equivalent. On the other hand, because it does not condition on  $\mathbf{x}_t$ , the Box–Pierce test is less powerful than the LM test when the null hypothesis is false, as intuition might suggest.

### 20.7.3 THE DURBIN–WATSON TEST

The Durbin–Watson statistic<sup>12</sup> was the first formal procedure developed for testing for autocorrelation using the least squares residuals. The test statistic is

$$d = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2} = 2(1 - r) - \frac{e_1^2 + e_T^2}{\sum_{t=1}^T e_t^2}, \quad (20-23)$$

where  $r$  is the same first-order autocorrelation that underlies the preceding two statistics. If the sample is reasonably large, then the last term will be negligible, leaving  $d \approx 2(1-r)$ . The statistic takes this form because the authors were able to determine the exact distribution of this transformation of the autocorrelation and could provide tables of critical values. Usable critical values that depend only on  $T$  and  $K$  are presented in tables such as those at the end of this book. The one-sided test for  $H_0: \rho = 0$  against  $H_1: \rho > 0$  is carried out by comparing  $d$  to values  $d_L(T, K)$  and  $d_U(T, K)$ . If  $d < d_L$ , the null hypothesis is rejected; if  $d > d_U$ , the hypothesis is not rejected. If  $d$  lies between  $d_L$  and  $d_U$ , then no conclusion is drawn.

### 20.7.4 TESTING IN THE PRESENCE OF A LAGGED DEPENDENT VARIABLE

The Durbin–Watson test is not likely to be valid when there is a lagged dependent variable in the equation.<sup>13</sup> The statistic will usually be biased toward a finding of no autocorrelation. Three alternatives have been devised. The LM and Q tests can be used whether or not the regression contains a lagged dependent variable. (In the absence of a lagged dependent variable, they are asymptotically equivalent.) As an alternative to the standard test, Durbin (1970) derived a Lagrange multiplier test that is appropriate

<sup>12</sup>Durbin and Watson (1950, 1951, 1971).

<sup>13</sup>This issue has been studied by Nerlove and Wallis (1966), Durbin (1970), and Dezhbakhsh (1990).

## 924 PART V ♦ Time Series and Microeconometrics

in the presence of a lagged dependent variable. The test may be carried out by referring

$$h = r \sqrt{T/(1 - Ts_c^2)}, \quad (20-24)$$

where  $s_c^2$  is the estimated variance of the least squares regression coefficient on  $y_{t-1}$ , to the standard normal tables. Large values of  $h$  lead to rejection of  $H_0$ . The test has the virtues that it can be used even if the regression contains additional lags of  $y_t$ , and it can be computed using the standard results from the initial regression without any further regressions. If  $s_c^2 > 1/T$ , however, then it cannot be computed. An alternative is to regress  $e_t$  on  $\mathbf{x}_t, y_{t-1}, \dots, e_{t-1}$ , and any additional lags that are appropriate for  $e_t$  and then to test the joint significance of the coefficient(s) on the lagged residual(s) with the standard  $F$  test. This method is a minor modification of the Breusch–Godfrey test. Under  $H_0$ , the coefficients on the remaining variables will be zero, so the tests are the same asymptotically.

### 20.7.5 SUMMARY OF TESTING PROCEDURES

The preceding has examined several testing procedures for locating autocorrelation in the disturbances. In all cases, the procedure examines the least squares residuals. We can summarize the procedures as follows:

**LM test.**  $LM = TR^2$  in a regression of the least squares residuals on  $[\mathbf{x}_t, e_{t-1}, \dots, e_{t-P}]$ . Reject  $H_0$  if  $LM > \chi_*^2[P]$ . This test examines the covariance of the residuals with lagged values, controlling for the intervening effect of the independent variables.

**Q test.**  $Q = T(T+2) \sum_{j=1}^P r_j^2 / (T-j)$ . Reject  $H_0$  if  $Q > \chi_*^2[P]$ . This test examines the raw correlations between the residuals and  $P$  lagged values of the residuals.

**Durbin–Watson test.**  $d = 2(1-r)$ . Reject  $H_0: \rho = 0$  if  $d < d_L^*$ . This test looks directly at the first-order autocorrelation of the residuals.

**Durbin's test.**  $F_D$  = the  $F$  statistic for the joint significance of  $P$  lags of the residuals in the regression of the least squares residuals on  $[\mathbf{x}_t, y_{t-1}, \dots, y_{t-R}, e_{t-1}, \dots, e_{t-P}]$ . Reject  $H_0$  if  $F_D > F_*[P, T-K-P]$ . This test examines the partial correlations between the residuals and the lagged residuals, controlling for the intervening effect of the independent variables and the lagged dependent variable.

The Durbin–Watson test has some major shortcomings. The inconclusive region is large if  $T$  is small or moderate. The bounding distributions, while free of the parameters  $\beta$  and  $\sigma$ , do depend on the data (and assume that  $\mathbf{X}$  is nonstochastic). An exact version based on an algorithm developed by Imhof (1980) avoids the inconclusive region, but is rarely used. The LM and Box–Pierce statistics do not share these shortcomings—their limiting distributions are chi-squared independently of the data and the parameters. For this reason, the LM test has become the standard method in applied research.

## 20.8 EFFICIENT ESTIMATION WHEN $\Omega$ IS KNOWN

As a prelude to deriving feasible estimators for  $\beta$  in this model, we consider full generalized least squares estimation assuming that  $\Omega$  is known. In the next section, we will turn to the more realistic case in which  $\Omega$  must be estimated as well.

## CHAPTER 20 ♦ Serial Correlation 925

If the parameters of  $\Omega$  are known, then the GLS estimator,

$$\hat{\beta} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}(\mathbf{X}'\Omega^{-1}\mathbf{y}), \quad (20-25)$$

and the estimate of its sampling variance,

$$\text{Est. Var}[\hat{\beta}] = \hat{\sigma}_\varepsilon^2 [\mathbf{X}'\Omega^{-1}\mathbf{X}]^{-1}, \quad (20-26)$$

where

$$\hat{\sigma}_\varepsilon^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})'\Omega^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})}{T} \quad (20-27)$$

can be computed in one step. For the AR(1) case, data for the transformed model are

$$\mathbf{y}_* = \begin{bmatrix} \sqrt{1-\rho^2}y_1 \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \\ \vdots \\ y_T - \rho y_{T-1} \end{bmatrix}, \quad \mathbf{X}_* = \begin{bmatrix} \sqrt{1-\rho^2}\mathbf{x}_1 \\ \mathbf{x}_2 - \rho\mathbf{x}_1 \\ \mathbf{x}_3 - \rho\mathbf{x}_2 \\ \vdots \\ \mathbf{x}_T - \rho\mathbf{x}_{T-1} \end{bmatrix}. \quad (20-28)$$

These transformations are variously labeled **partial differences**, **quasi differences**, or **pseudo-differences**. Note that in the transformed model, every observation except the first contains a constant term. What was the column of 1s in  $\mathbf{X}$  is transformed to  $[(1-\rho^2)^{1/2}, (1-\rho), (1-\rho), \dots]$ . Therefore, if the sample is relatively small, then the problems with measures of fit noted in Section 3.5 will reappear.

The variance of the transformed disturbance is

$$\text{Var}[\varepsilon_t - \rho\varepsilon_{t-1}] = \text{Var}[u_t] = \sigma_u^2.$$

The variance of the first disturbance is also  $\sigma_u^2$ ; [see (20-6)]. This can be estimated using  $(1-\rho^2)\hat{\sigma}_\varepsilon^2$ .

Corresponding results have been derived for higher-order autoregressive processes. For the AR(2) model,

$$\varepsilon_t = \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + u_t, \quad (20-29)$$

the transformed data for generalized least squares are obtained by

$$\begin{aligned} \mathbf{z}_{*1} &= \left[ \frac{(1+\theta_2)[(1-\theta_2)^2 - \theta_1^2]}{1-\theta_2} \right]^{1/2} \mathbf{z}_1, \\ \mathbf{z}_{*2} &= (1-\theta_2^2)^{1/2} \mathbf{z}_2 - \frac{\theta_1(1-\theta_2^2)^{1/2}}{1-\theta_2} \mathbf{z}_1, \\ \mathbf{z}_{*t} &= \mathbf{z}_t - \theta_1\mathbf{z}_{t-1} - \theta_2\mathbf{z}_{t-2}, \quad t > 2, \end{aligned} \quad (20-30)$$

where  $\mathbf{z}_t$  is used for  $y_t$  or  $\mathbf{x}_t$ . The transformation becomes progressively more complex for higher-order processes.<sup>14</sup>

<sup>14</sup>See Box and Jenkins (1984) and Fuller (1976).

## 926 PART V ♦ Time Series and Microeconometrics

Note that in both the AR(1) and AR(2) models, the transformation to  $y_*$  and  $\mathbf{X}_*$  involves “starting values” for the processes that depend only on the first one or two observations. We can view the process as having begun in the infinite past. Because the sample contains only  $T$  observations, however, it is convenient to treat the first one or two (or  $P$ ) observations as shown and consider them as “initial values.” Whether we view the process as having begun at time  $t = 1$  or in the infinite past is ultimately immaterial in regard to the asymptotic properties of the estimators.

The asymptotic properties for the GLS estimator are quite straightforward given the apparatus we assembled in Section 20.4. We begin by assuming that  $\{\mathbf{x}_t, \varepsilon_t\}$  are jointly an ergodic, stationary process. Then, after the GLS transformation,  $\{\mathbf{x}_{*t}, \varepsilon_{*t}\}$  is also stationary and ergodic. Moreover,  $\varepsilon_{*t}$  is nonautocorrelated by construction. In the transformed model, then,  $\{\mathbf{w}_{*t}\} = \{\mathbf{x}_{*t} \varepsilon_{*t}\}$  is a stationary and ergodic martingale difference series. We can use the ergodic theorem to establish consistency and the central limit theorem for martingale difference sequences to establish asymptotic normality for GLS in this model. Formal arrangement of the relevant results is left as an exercise.

## 20.9 ESTIMATION WHEN $\Omega$ IS UNKNOWN

For an unknown  $\Omega$ , there are a variety of approaches. A consistent estimator of  $\Omega(\rho)$  will suffice—recall from Theorem (9.5) in Section 9.3.2, all that is needed for efficient estimation of  $\beta$  is a consistent estimator of  $\Omega(\rho)$ . The complication arises, as might be expected, in estimating the autocorrelation parameter(s).

### 20.9.1 AR(1) DISTURBANCES

The AR(1) model is the one most widely used and studied. The most common procedure is to begin FGLS with a natural estimator of  $\rho$ , the autocorrelation of the residuals. Because  $\mathbf{b}$  is consistent, we can use  $r$ . Others that have been suggested include Theil’s (1971) estimator,  $r[(T-K)/(T-1)]$  and Durbin’s (1970), the slope on  $y_{t-1}$  in a regression of  $y_t$  on  $y_{t-1}$ ,  $\mathbf{x}_t$  and  $\mathbf{x}_{t-1}$ . The second step is FGLS based on (20-25)–(20-28). This is the **Prais and Winsten (1954) estimator**. The **Cochrane and Orcutt (1949) estimator** (based on computational ease) omits the first observation.

It is possible to iterate any of these estimators to convergence. Because the estimator is asymptotically efficient at every iteration, nothing is gained by doing so. Unlike the heteroscedastic model, iterating when there is autocorrelation does not produce the maximum likelihood estimator. The iterated FGLS estimator, regardless of the estimator of  $\rho$ , does not account for the term  $(1/2) \ln(1 - \rho^2)$  in the log-likelihood function [see the following (20-31)].

Maximum likelihood estimators can be obtained by maximizing the log-likelihood with respect to  $\beta$ ,  $\sigma_u^2$ , and  $\rho$ . The log-likelihood function may be written

$$\ln L = -\frac{\sum_{t=1}^T u_t^2}{2\sigma_u^2} + \frac{1}{2} \ln(1 - \rho^2) - \frac{T}{2} (\ln 2\pi + \ln \sigma_u^2), \quad (20-31)$$

where, as before, the first observation is computed differently from the others using (20-28). The MLE for this model is developed in Section 14.9.2.b. Based on the MLE,

the standard approximations to the asymptotic variances of the estimators are

$$\begin{aligned}\text{Est. Asy. Var}[\hat{\beta}_{ML}] &= \hat{\sigma}_{\varepsilon, ML}^2 [\mathbf{X}' \hat{\Omega}_{ML}^{-1} \mathbf{X}]^{-1}, \\ \text{Est. Asy. Var}[\hat{\sigma}_{u, ML}^2] &= 2\hat{\sigma}_{u, ML}^4 / T, \\ \text{Est. Asy. Var}[\hat{\rho}_{ML}] &= (1 - \hat{\rho}_{ML}^2) / T.\end{aligned}\tag{20-32}$$

All the foregoing estimators have the same asymptotic properties. The available evidence on their small-sample properties comes from Monte Carlo studies and is, unfortunately, only suggestive. Griliches and Rao (1969) find evidence that if the sample is relatively small and  $\rho$  is not particularly large, say, less than 0.3, then least squares is as good as or better than FGLS. The problem is the additional variation introduced into the sampling variance by the variance of  $r$ . Beyond these, the results are rather mixed. Maximum likelihood seems to perform well in general, but the Prais–Winsten estimator is evidently nearly as efficient. Both estimators have been incorporated in all contemporary software. In practice, the Prais and Winsten (1954) and Beach and MacKinnon (1978a) maximum likelihood estimators are probably the most common choices.

#### 20.9.2 APPLICATION: ESTIMATION OF A MODEL WITH AUTOCORRELATION

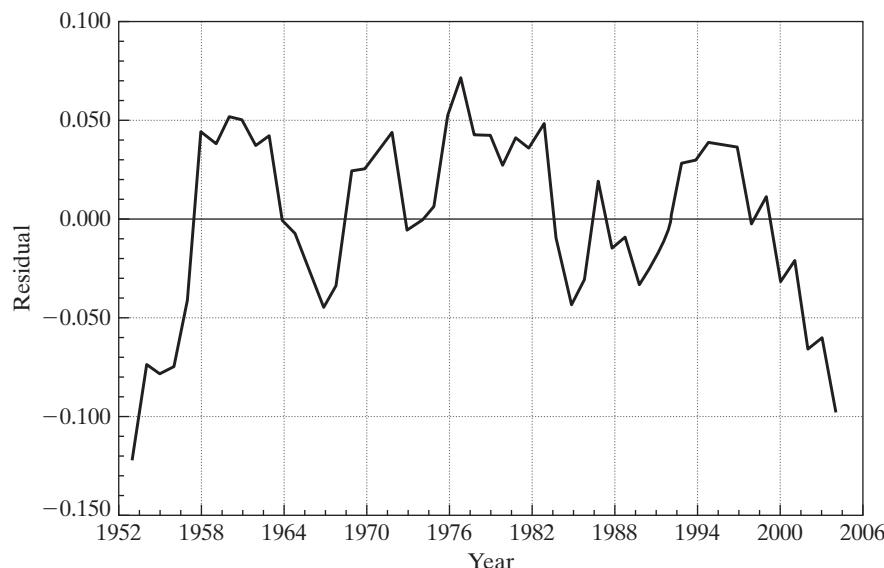
The model of the U.S. gasoline market that appears in Example 6.9 is

$$\ln \frac{G_t}{pop_t} = \beta_1 + \beta_2 \ln \frac{I_t}{pop_t} + \beta_3 \ln P_{G,t} + \beta_4 \ln P_{NC,t} + \beta_5 \ln P_{UC,t} + \beta_6 t + \varepsilon_t.$$

The results in Figure 20.2 suggest that the specification may be incomplete, and, if so, there may be autocorrelation in the disturbances in this specification. Least squares estimates of the parameters using the data in Appendix Table F2.2 appear in the first row of Table 20.2. [The dependent variable is  $\ln(\text{Gas expenditure} / (\text{price} \times \text{population}))$ . These are the OLS results reported in Example 6.9.] The first five autocorrelations of the least squares residuals are 0.667, 0.438, 0.142, -0.018, and -0.198. This produces Box–Pierce and Box–Ljung statistics of 36.217 and 38.789, respectively, both of which are larger than the critical value from the chi-squared table of 11.07. We regressed the least squares residuals on the independent variables and five lags of

**TABLE 20.2** Parameter Estimates (standard errors in parentheses)

	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$	$\beta_6$	$\rho$
OLS	-26.68	1.6250	-0.05392	-0.0834	-0.08467	-0.01393	0.0000
$R^2 = 0.96493$	(2.000)	(0.1952)	(0.04216)	(0.1765)	(0.1024)	(0.00477)	(0.0000)
Prais–Winsten	-18.58	0.7447	-0.1138	-0.1364	-0.008956	0.006689	0.9567
	(1.768)	(0.1761)	(0.03689)	(0.1528)	(0.07213)	(0.004974)	(0.04078)
Cochrane–Orcutt	-18.76	0.7300	-0.1080	-0.06675	0.04190	-0.0001653	0.9695
	(1.382)	(0.1377)	(0.02885)	(0.1201)	(0.05713)	(0.004082)	(0.03434)
Maximum Likelihood	-16.25	0.4690	-0.1387	-0.09682	-0.001485	0.01280	0.9792
	(1.391)	(0.1350)	(0.02794)	(0.1270)	(0.05198)	(0.004427)	(0.02816)
AR(2)	-19.45	0.8116	-0.09538	-0.09099	0.04091	-0.001374	0.8610
	(1.495)	(0.1502)	(0.03117)	(0.1297)	(0.06558)	(0.004227)	(0.07053)

**928 PART V ♦ Time Series and Microeconometrics**


**FIGURE 20.4** Least Squares Residuals.

the residuals. (The missing values in the first five years were filled with zeros.) The coefficients on the lagged residuals and the associated  $t$  statistics are 0.741 (4.635), 0.153 (0.789),  $-0.246$  ( $-1.262$ ), 0.0942(0.472), and  $-0.125$  ( $-0.658$ ). The  $R^2$  in this regression is 0.549086, which produces a chi-squared value of 28.55. This is larger than the critical value of 11.07, so once again, the null hypothesis of zero autocorrelation is rejected. Finally, the Durbin–Watson statistic is 0.425007. For 5 regressors and 52 observations, the critical value of  $d_L$  is 1.36, so on this basis as well, the null hypothesis  $\rho = 0$  would be rejected. The plot of the residuals shown in Figure 20.4 seems consistent with this conclusion.

The Prais and Winsten FGLS estimates appear in the second row of Table 20.2 followed by the Cochrane and Orcutt results then the maximum likelihood estimates. [The autocorrelation coefficient computed using  $(1 - d/2)$  (see Section 20.7.3) is 0.78750. The MLE is computed using the Beach and MacKinnon algorithm. See Section 14.9.2.b.] Finally, we fit the AR(2) model by first regressing the least squares residuals,  $e_t$ , on  $e_{t-1}$  and  $e_{t-2}$  (without a constant term and filling the first two observations with zeros). The two estimates are 0.751941 and  $-0.022464$ , respectively. With the estimates of  $\theta_1$  and  $\theta_2$ , we transformed the data using  $y_t^* = y_t - \theta_1 y_{t-1} - \theta_2 y_{t-2}$  and likewise for each regressor. Two observations are then discarded, so the AR(2) regression uses 50 observations while the Prais–Winsten estimator uses 52 and the Cochrane–Orcutt regression uses 51. In each case, the autocorrelation of the FGLS residuals is computed and reported in the last column of the table.

One might want to examine the residuals after estimation to ascertain whether the AR(1) model is appropriate. In the results just presented, there are two large autocorrelation coefficients listed with the residual based tests, and in computing the LM statistic, we found that the first two coefficients were statistically significant. If the AR(1) model

is appropriate, then one should find that only the coefficient on the first lagged residual is statistically significant in this auxiliary, second-step regression. Another indicator is provided by the FGLS residuals, themselves. After computing the FGLS regression, the estimated residuals,

$$\hat{\varepsilon} = y_t - \mathbf{x}'_t \hat{\beta}$$

will still be autocorrelated. In our results using the Prais–Winsten estimates, the autocorrelation of the FGLS residuals is 0.957. The associated Durbin–Watson statistic is 0.0867. This is to be expected. However, if the model is correct, then the transformed residuals

$$\hat{u}_t = \hat{\varepsilon}_t - \hat{\rho} \hat{\varepsilon}_{t-1}$$

should be at least close to nonautocorrelated. But, for our data, the autocorrelation of these adjusted residuals is only 0.292 with a Durbin–Watson statistic of 1.416. The value of  $d_L$  for one regressor ( $u_{t-1}$ ) and 50 observations is 1.50. It appears on this basis that, in fact, the AR(1) model has largely completed the specification.

### 20.9.3 ESTIMATION WITH A LAGGED DEPENDENT VARIABLE

In Section 20.5.1, we considered the problem of estimation by least squares when the model contains both autocorrelation and lagged dependent variable(s). Because the OLS estimator is inconsistent, the residuals on which an estimator of  $\rho$  would be based are likewise inconsistent. Therefore,  $\hat{\rho}$  will be inconsistent as well. The consequence is that the FGLS estimators described earlier are not usable in this case. There is, however, an alternative way to proceed, based on the method of instrumental variables. The method of instrumental variables was introduced in Section 8.3.2. To review, the general problem is that in the regression model, if

$$\text{plim}(1/T)\mathbf{X}'\boldsymbol{\varepsilon} \neq \mathbf{0},$$

then the least squares estimator is not consistent. A consistent estimator is

$$\mathbf{b}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{y}),$$

where  $\mathbf{Z}$  is a set of  $K$  variables chosen such that  $\text{plim}(1/T)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0}$  but  $\text{plim}(1/T)\mathbf{Z}'\mathbf{X} \neq \mathbf{0}$ . For the purpose of consistency only, any such set of instrumental variables will suffice. The relevance of that here is that the obstacle to consistent FGLS is, at least for the present, the lack of a consistent estimator of  $\rho$ . By using the technique of instrumental variables, we may estimate  $\boldsymbol{\beta}$  consistently, then estimate  $\rho$  and proceed.

Hatanaka (1974, 1976) has devised an efficient two-step estimator based on this principle. To put the estimator in the current context, we consider estimation of the model

$$\begin{aligned} y_t &= \mathbf{x}'_t \boldsymbol{\beta} + \gamma y_{t-1} + \varepsilon_t, \\ \varepsilon_t &= \rho \varepsilon_{t-1} + u_t. \end{aligned}$$

To get to the second step of FGLS, we require a consistent estimator of the slope parameters. These estimates can be obtained using an IV estimator, where the column of  $\mathbf{Z}$  corresponding to  $y_{t-1}$  is the only one that need be different from that of  $\mathbf{X}$ . An appropriate instrument can be obtained by using the fitted values in the regression of

## 930 PART V ♦ Time Series and Microeconometrics

$y_t$  on  $\mathbf{x}_t$  and  $\mathbf{x}_{t-1}$ . The residuals from the IV regression are then used to construct

$$\hat{\rho} = \frac{\sum_{t=3}^T \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}}{\sum_{t=3}^T \hat{\varepsilon}_t^2}, \quad (20-33)$$

where

$$\hat{\varepsilon}_t = y_t - \mathbf{b}'_{IV} \mathbf{x}_t - c_{IV} y_{t-1}.$$

FGLS estimates may now be computed by regressing  $y_{*t} = y_t - \hat{\rho} y_{t-1}$  on

$$\begin{aligned} \mathbf{x}_{*t} &= \mathbf{x}_t - \hat{\rho} \mathbf{x}_{t-1}, \\ y_{*t-1} &= y_{t-1} - \hat{\rho} y_{t-2}, \\ \hat{\varepsilon}_{t-1} &= y_{t-1} - \mathbf{b}'_{IV} \mathbf{x}_{t-1} - c_{IV} y_{t-2}. \end{aligned}$$

Let  $d$  be the coefficient on  $\hat{\varepsilon}_{t-1}$  in this regression. The efficient estimator of  $\rho$  is

$$\hat{\hat{\rho}} = \hat{\rho} + d.$$

Appropriate asymptotic standard errors for the estimators, including  $\hat{\hat{\rho}}$ , are obtained from the  $s^2[\mathbf{X}'_* \mathbf{X}_*]^{-1}$  computed at the second step. Hatanaka shows that these estimators are asymptotically equivalent to maximum likelihood estimators.

### 20.10 AUTOCORRELATION IN PANEL DATA

The extension of the AR(1) model to stationary panel data would mirror the procedures for a single time series. The standard model is

$$y_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + c_i + \varepsilon_{it}, \quad \varepsilon_{it} = \rho \varepsilon_{i,t-1} + v_{it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T_i.$$

[See, e.g., Baltagi (2005, Section 5.2).] The same considerations would apply to the fixed and random effects cases, so we have left the model in generic form. The practical issues are how to obtain an estimate of  $\rho$  and how to carry out an FGLS procedure.

Assuming for the moment that  $\rho$  is known, the Prais and Winsten transformation,

$$\begin{aligned} y_{i1}^* &= (1 - \rho^2)^{1/2} y_{i1}, & \mathbf{x}_{i1}^* &= (1 - \rho^2)^{1/2} \mathbf{x}_{i1}, \\ y_{it}^* &= y_{it} - \rho y_{i,t-1}, & \mathbf{x}_{it}^* &= \mathbf{x}_{it} - \rho \mathbf{x}_{i,t-1}, \end{aligned} \quad (20-34)$$

produces the transformed model

$$y_{it}^* = (\mathbf{x}_{it}^*)' \boldsymbol{\beta} + c_{it}^* + v_{it}, \quad v_{it} | \mathbf{X}_i \sim 0, \sigma_v^2. \quad (20-35)$$

[See (20-28).] This would seem to restore the original panel data model, save for a potentially complicated loose end. The common effect in the transformed model,  $c_{it}^*$ , is treated differently for the first observation from the remaining  $T_i - 1$ , hence the necessity for the double subscript in (20-35). The resulting model is no longer a “common effect” model. Baltagi and Li (1991) have devised a full FGLS treatment for the balanced panel random effects case, including the asymmetric treatment of  $c_i$ . The method is shown in detail in Baltagi (2005, pp. 84–85). The procedure as documented can be generalized to the unbalanced case fairly easily, but overall is quite complicated, again owing to the special treatment of the first observation. FGLS estimation of the fixed effects

model is more complicated yet, because there is no simple transformation comparable to differences from group means that will remove the common effect in (20-35). For least squares estimation of  $\beta$  in (20-35), we would have to use brute force, with  $c_{it}^* = \alpha_i d_{it}$  where  $d_{it} = (1 - \rho^2)^{1/2}$  for individual  $i$  and  $t = 1$  and  $d_{it} = 1 - \rho$  for individual  $i$  and  $t > 1$ . (In principle, the Frisch–Waugh result could be applied group by group to transform the observations for estimation of  $\beta$ . However, the application would involve a different transformation for the first observation and the mean of observations  $2 - T_i$  rather than the full group mean.)

For better or worse, dropping the first observation is a practical compromise that produces a large payoff. The different approaches based on  $T_i - 1$  observations remain consistent in  $n$ , just as in the single time-series case. The question of efficiency might be raised here as it was in an earlier literature in the time-series case [see, e.g., Maeshiro (1979)]. In a given panel,  $T_i$  may well be fairly small. However, with the assumption of common  $\rho$ , this case is likely to be much more favorable than the single time-series case, because the way the model is structured, estimation of  $\beta$  based on the Cochrane–Orcutt transformation becomes analogous to the random effects case with  $\sum_{i=1}^n (T_i - 1)$  observations. If  $n$  is even moderately sized, the efficiency question is likely to be a moot point.

There remains the problem of estimation of  $\rho$ . For either fixed or random effects case, the within (dummy variables) estimator produces a consistent estimator of  $\beta$  and a usable set of residuals,  $e_{it}$  that can be used to estimate  $\rho$  with  $[\sum_{i=1}^n \sum_{t=2}^{T_i} e_{it} e_{i,t-1}] / [\sum_{i=1}^n \sum_{t=2}^{T_i} e_{i,t}^2]$ . [Baltagi and Li (1991a) suggest some alternative estimators that may have better small sample properties.]

#### **Example 20.5 Panel Data Models with Autocorrelation**

Munnell (1990) analyzed the productivity of public capital at the state level using a Cobb–Douglas production function. We will use the data from that study to estimate a log-linear regression model

$$\ln gsp_{it} = \alpha + \beta_1 \ln p_{it} + \beta_2 \ln hwy_{it} + \beta_3 \ln water_{it} \\ + \beta_4 \ln util_{it} + \beta_5 \ln emp_{it} + \beta_6 unemp_{it} + \varepsilon_{it} + u_i,$$

where the variables in the model are

<i>gsp</i>	= gross state product	
<i>p_cap</i>	= public capital	
<i>hwy</i>	= highway capital	
<i>water</i>	= water utility capital	
<i>util</i>	= utility capital	
<i>pc</i>	= private capital	
<i>emp</i>	= employment (labor)	
<i>unemp</i>	= unemployment rate.	

In Example 11.10, we estimated the parameters of the model under the fixed and random effects assumptions. The results are repeated in Table 20.3. Using the fixed effects residuals, the estimate of  $\rho$  is 0.717897, which is quite large. (In computing the estimate, using the preceding result, the sum in the denominator was started at  $t = 1$ , and then the numerator and denominator were divided by  $n(T - 1)$  and  $nT$ , respectively.) We reestimated the two models using the Cochrane–Orcutt transformation. The results are shown at the right in Table 20.3. The estimates of  $\sigma_\varepsilon$  and  $\sigma_u$  in each case are obtained as  $1/(1 - r^2)$  times the estimated variances in the transformed model.

**932 PART V ♦ Time Series and Microeconometrics**
**TABLE 20.3** Estimated Statewide Production Function.

	<b>OLS</b>		<b>Fixed Effects</b>		<b>Random Effects FGLS</b>	
	$\rho = 0$	$AR(1)$	$\rho = 0$	$AR(1)$	$\rho = 0$	$AR(1)$
	<i>Estimate (Std. Err.<sup>a</sup>)</i>	<i>Estimate (Std. Err.)</i>				
$\alpha$	1.9260 (0.2143)	2.1463 (0.01806)			2.1608 (0.1380)	2.8226 (0.1537)
$\beta_1$	0.3120 (0.04678)	0.2615 (0.01338)	0.2350 (0.02621)	0.04041 (0.02383)	0.2755 (0.01972)	0.1342 (0.01943)
$\beta_2$	0.05888 (0.05078)	0.06788 (0.01802)	0.07675 (0.03124)	-0.05831 (0.06101)	0.06167 (0.02168)	0.04585 (0.03044)
$\beta_3$	0.1186 (0.0345)	0.09225 (0.01464)	0.0786 (0.0150)	0.04934 (0.02098)	0.07572 (0.01381)	0.04879 (0.01911)
$\beta_4$	0.00856 (0.0406)	-0.006299 (0.01432)	-0.11478 (0.01814)	-0.07459 (0.02759)	-0.09672 (0.01683)	-0.07977 (0.02273)
$\beta_5$	0.5497 (0.0677)	0.6337 (0.01916)	0.8011 (0.02976)	1.0534 (0.03677)	0.7450 (0.02482)	0.08931 (0.03011)
$\beta_6$	-0.00727 (0.002946)	-0.009327 (0.00083)	-0.005179 (0.000980)	-0.002924 (0.000817)	-0.005963 (0.0008814)	-0.005374 (0.0007513)
$\sigma_\varepsilon$	0.0854228	0.105407	0.0367649	.0302657	.030141 <sup>b</sup>	.0302657 <sup>c</sup>
$\sigma_u$					.0771064 <sup>b</sup>	.074380 <sup>c</sup>

<sup>a</sup>Robust (cluster) standard errors in parentheses

<sup>b</sup>Based on OLS and LSDV residuals

<sup>c</sup>Based on OLS/AR(1) and LSDV/AR(1) residuals

There are several strategies available for testing for serial correlation in a panel data set. [See Baltagi (2005, pp. 93–103).] In general, an obstacle to a simple test against the null hypothesis of no autocorrelation is the possible presence of a time-invariant common effect in the model,

$$y_{it} = c_i + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}.$$

Under the alternative hypothesis,  $\text{Corr}(\varepsilon_{it}, \varepsilon_{i,t-1}) = \rho$ . Many variants of the model, based on AR(1), MA(1), and other specifications have been analyzed. We consider the first, as it is the standard framework in the absence of a specific model that suggests another process. The LM statistic is based on the within-groups (LSDV) residuals from the fixed effects specification,

$$\text{LM} = \left( \frac{NT^2}{T-1} \right) \left( \frac{\sum_{i=1}^n \sum_{t=2}^T e_{it} e_{i,t-1}}{\sum_{i=1}^n \sum_{t=1}^T e_{it}^2} \right)^2.$$

Under the null hypothesis that  $\rho = 0$ , the limiting distribution of LM is chi-squared with one degree of freedom. The Durbin–Watson statistic is obtained by omitting  $[NT^2/(T-1)]$  and replacing  $e_{it} e_{i,t-1}$  with  $(e_{it} - e_{i,t-1})^2$ . Bhargava et al. (1982) showed that the Durbin–Watson version of the test is locally most powerful in the neighborhood of  $\rho = 0$ . In the typical panel, the data set is large enough that the advantage over the simpler

LM statistic is unlikely to be substantial. Both tests will suffer from a loss of power if the model is a random effects model. Baltagi and Li (1991a,b) have devised several test statistics that are appropriate for the random effects specification. Wooldridge (2002a) proposes an alternative approach based on first differences that will be invariant to the presence of fixed or random effects.

## 20.11 COMMON FACTORS

We saw in Example 20.2 that misspecification of an equation could create the appearance of serially correlated disturbances when, in fact, there are none. An orthodox (perhaps somewhat optimistic) purist might argue that autocorrelation is *always* an artifact of misspecification. Although this view might be extreme [see, e.g., Hendry (1980) for a more moderate, but still strident statement], it does suggest a useful point. It might be useful if we could examine the specification of a model statistically with this consideration in mind. The test for **common factors** is such a test. [See, as well, the aforementioned paper by Mizon (1995).]

The assumption that the correctly specified model is

$$y_t = \mathbf{x}'_t \boldsymbol{\beta} + \varepsilon_t, \quad \varepsilon_t = \rho \varepsilon_{t-1} + u_t, \quad t = 1, \dots, T$$

implies the “reduced form,”

$$M_0: y_t = \rho y_{t-1} + (\mathbf{x}_t - \rho \mathbf{x}_{t-1})' \boldsymbol{\beta} + u_t, \quad t = 2, \dots, T,$$

where  $u_t$  is free from serial correlation. The second of these is actually a restriction on the model

$$M_1: y_t = \rho y_{t-1} + \mathbf{x}'_t \boldsymbol{\beta} + \mathbf{x}'_{t-1} \boldsymbol{\alpha} + u_t, \quad t = 2, \dots, T,$$

in which, once again,  $u_t$  is a classical disturbance. The second model contains  $2K + 1$  parameters, but if the model is correct, then  $\boldsymbol{\alpha} = -\rho \boldsymbol{\beta}$  and there are only  $K + 1$  parameters and  $K$  restrictions. Both  $M_0$  and  $M_1$  can be estimated by least squares, although  $M_0$  is a nonlinear model. One might then test the restrictions of  $M_0$  using an  $F$  test. This test will be valid asymptotically, although its exact distribution in finite samples will not be precisely  $F$ . In large samples,  $KF$  will converge to a chi-squared statistic, so we use the  $F$  distribution as usual to be conservative. There is a minor practical complication in implementing this test. Some elements of  $\boldsymbol{\alpha}$  may not be estimable. For example, if  $\mathbf{x}_t$  contains a constant term, then the one in  $\boldsymbol{\alpha}$  is unidentified. If  $\mathbf{x}_t$  contains both current and lagged values of a variable, then the one period lagged value will appear twice in  $M_1$ , once in  $\mathbf{x}_t$  as the lagged value and once in  $\mathbf{x}_{t-1}$  as the current value. There are other combinations that will be problematic, so the actual number of restrictions that appear in the test is reduced to the number of identified parameters in  $\boldsymbol{\alpha}$ .

### Example 20.6 Test for Common Factors

We will reexamine the model estimated in Section 20.9.2. The base model is

$$\ln \frac{G_t}{Pop_t} = \beta_1 + \beta_2 \ln \frac{I_t}{Pop_t} + \beta_3 \ln P_{G,t} + \beta_4 \ln P_{NC,t} + \beta_5 \ln P_{UC,t} + \beta_6 t + \varepsilon_t.$$

### 934 PART V ♦ Time Series and Microeconometrics

If the AR(1) model is appropriate for  $\varepsilon_t$ , that is,  $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$ , then the restricted model,

$$\begin{aligned}\ln \frac{G_t}{Pop_t} &= \rho \left( \ln \frac{G_{t-1}}{Pop_{t-1}} \right) + \beta_1 + \beta_2 \left( \ln \frac{I_t}{Pop_t} - \rho \ln \frac{I_{t-1}}{Pop_{t-1}} \right) + \beta_3 (\ln P_{G,t} - \rho \ln P_{G,t-1}) \\ &\quad + \beta_4 (\ln P_{NC,t} - \rho \ln P_{NC,t-1}) + \beta_5 (\ln P_{UC,t} - \rho \ln P_{UC,t-1}) + \beta_6 [t - \rho(t - 1)] + u_t,\end{aligned}$$

with 7 free coefficients will not significantly degrade the fit of the unrestricted model,

$$\begin{aligned}\ln \frac{G_t}{Pop_t} &= \rho \left( \ln \frac{G_{t-1}}{Pop_{t-1}} \right) + \beta_1 + \alpha_1 + \beta_2 \ln \frac{I_t}{Pop_t} + \alpha_2 \ln \frac{I_{t-1}}{Pop_{t-1}} + \beta_3 \ln P_{G,t} + \alpha_3 \ln P_{G,t-1} \\ &\quad + \beta_4 \ln P_{NC,t} + \alpha_4 \ln P_{NC,t-1} + \beta_5 \ln P_{UC,t} + \alpha_5 \ln P_{UC,t-1} + \beta_6 t + \alpha_6 (t - 1) + u_t,\end{aligned}$$

which has 13 coefficients. Note, however, that  $\alpha_1$  and  $\alpha_6$  are not identified [because  $t = (t - 1) + 1$ ]. Thus, the common factor restriction imposes four restrictions on the model. We fit the unrestricted model [minus one constant and  $(t - 1)$ ] by ordinary least squares and obtained a sum of squared residuals of 0.00737717. We fit the restricted model by nonlinear least squares, using the OLS coefficients from the base model as starting values for  $\beta$  and zero for  $\rho$ . (See Section 7.2.6.) The sum of squared residuals is 0.01084939. This produces an  $F$  statistic of

$$\frac{(0.10184939 - 0.00737717)/4}{0.00737717/(51 - 11)} = 4.707,$$

which is larger than the critical value with 4 and 40 degrees of freedom of 2.606. Thus, we would conclude that the AR(1) model would not be appropriate for this specification and these data.

## 20.12 FORECASTING IN THE PRESENCE OF AUTOCORRELATION

For purposes of forecasting, we refer first to the transformed model,

$$y_{*t} = \mathbf{x}'_{*t} \boldsymbol{\beta} + \varepsilon_{*t}.$$

Suppose that the process generating  $\varepsilon_t$  is an AR(1) and that  $\rho$  is known. This model is a classical regression model, so the results of Section 4.6 may be used. The optimal forecast of  $y_{*T+1}^0$ , given  $\mathbf{x}_{T+1}^0$  and  $\mathbf{x}_T$  (i.e.,  $\mathbf{x}_{*T+1}^0 = \mathbf{x}_{T+1}^0 - \rho \mathbf{x}_T$ ), is

$$\hat{y}_{*T+1}^0 = \mathbf{x}_{*T+1}^{0r} \hat{\boldsymbol{\beta}}.$$

Disassembling  $\hat{y}_{*T+1}^0$ , we find that

$$\hat{y}_{T+1}^0 - \rho y_T = \mathbf{x}_{T+1}^{0r} \hat{\boldsymbol{\beta}} - \rho \mathbf{x}_T' \hat{\boldsymbol{\beta}},$$

or

$$\begin{aligned}\hat{y}_{T+1}^0 &= \mathbf{x}_{T+1}^{0r} \hat{\boldsymbol{\beta}} + \rho(y_T - \mathbf{x}_T' \hat{\boldsymbol{\beta}}) \\ &= \mathbf{x}_{T+1}^{0r} \hat{\boldsymbol{\beta}} + \rho e_T.\end{aligned}\tag{20-36}$$

Thus, we carry forward a proportion  $\rho$  of the estimated disturbance in the preceding period. This step can be justified by reference to

$$E[\varepsilon_{T+1} | \varepsilon_T] = \rho \varepsilon_T.$$

## CHAPTER 20 ♦ Serial Correlation 935

It can also be shown that to forecast  $n$  periods ahead, we would use

$$\hat{y}_{T+n}^0 = \mathbf{x}_{T+n}' \hat{\beta} + \rho^n e_T.$$

The extension to higher-order autoregressions is direct. For a second-order model, for example,

$$\hat{y}_{T+n}^0 = \hat{\beta}' \mathbf{x}_{T+n}^0 + \theta_1 e_{T+n-1} + \theta_2 e_{T+n-2}. \quad (20-37)$$

For residuals that are outside the sample period, we use the recursion

$$e_s = \theta_1 e_{s-1} + \theta_2 e_{s-2}, \quad (20-38)$$

beginning with the last two residuals within the sample.

Moving average models are somewhat simpler, as the autocorrelation lasts for only  $Q$  periods. For an MA(1) model, for the first postsample period,

$$\hat{y}_{T+1}^0 = \mathbf{x}_{T+1}' \hat{\beta} + \hat{\varepsilon}_{T+1},$$

where

$$\hat{\varepsilon}_{T+1} = \hat{u}_{T+1} - \lambda \hat{u}_T.$$

Therefore, a forecast of  $\varepsilon_{T+1}$  will use all previous residuals. One way to proceed is to accumulate  $\hat{\varepsilon}_{T+1}$  from the recursion

$$\hat{u}_t = \hat{\varepsilon}_t + \lambda \hat{u}_{t-1},$$

with  $\hat{u}_{T+1} = \hat{u}_0 = 0$  and  $\hat{\varepsilon}_t = (y_t - \mathbf{x}_t' \hat{\beta})$ . After the first postsample period,

$$\hat{\varepsilon}_{T+n} = \hat{u}_{T+n} - \lambda \hat{u}_{T+n-1} = 0.$$

If the parameters of the disturbance process are known, then  variances for the forecast errors can be computed using the results of Section 5.6. For an AR(1) disturbance, the estimated variance would be

$$s_f^2 = \hat{\sigma}_\varepsilon^2 + (\mathbf{x}_t - \rho \mathbf{x}_{t-1})' \{ \text{Est. Var} [\hat{\beta}] \} (\mathbf{x}_t - \rho \mathbf{x}_{t-1}). \quad (20-39)$$

For a higher-order process, it is only necessary to modify the calculation of  $\mathbf{x}_{*t}$  accordingly. The forecast variances for an MA(1) process are somewhat more involved. Details may be found in Judge et al. (1985) and Hamilton (1994). If the parameters of the disturbance process,  $\rho$ ,  $\lambda$ ,  $\theta_j$ , and so on, are estimated as well, then the forecast variance will be greater. For an AR(1) model, the necessary correction to the forecast variance of the  $n$ -period-ahead forecast error is  $\hat{\sigma}_\varepsilon^2 n^2 \rho^{2(n-1)} / T$ . [For a one-period-ahead forecast, this merely adds a term,  $\hat{\sigma}_\varepsilon^2 / T$  in (20-39)]. Higher-order AR and MA processes are analyzed in Baillie (1979). Finally, if the regressors are stochastic, the expressions become more complex by another order of magnitude.

If  $\rho$  is known, then (20-36) provides the best linear unbiased forecast of  $y_{t+1}$ .<sup>15</sup> If, however,  $\rho$  must be estimated, then this assessment must be modified. There is information about  $\varepsilon_{t+1}$  embodied in  $e_t$ . Having to estimate  $\rho$ , however, implies that some or all the value of this information is offset by the variation introduced into the

<sup>15</sup>See Goldberger (1962).

## 936 PART V ♦ Time Series and Microeometrics

forecast by including the stochastic component  $\hat{\rho}e_t$ .<sup>16</sup> Whether (20-36) is preferable to the obvious expedient  $\hat{y}_{T+n}^0 = \hat{\beta}'\mathbf{x}_{T+n}^0$  in a small sample when  $\rho$  is estimated remains to be settled.

### 20.13 AUTOREGRESSIVE CONDITIONAL HETROSCECDASTICITY

Heteroscedasticity is often associated with cross-sectional data, whereas time series are usually studied in the context of homoscedastic processes. In analyses of macroeconomic data, Engle (1982, 1983) and Cragg (1982) found evidence that for some kinds of data, the disturbance variances in time-series models were less stable than usually assumed. Engle's results suggested that in models of inflation, large and small forecast errors appeared to occur in clusters, suggesting a form of heteroscedasticity in which the variance of the forecast error depends on the size of the previous disturbance. He suggested the autoregressive, conditionally heteroscedastic, or ARCH, model as an alternative to the usual time-series process. More recent studies of financial markets suggest that the phenomenon is quite common. The ARCH model has proven to be useful in studying the volatility of inflation [Coulson and Robins (1985)], the term structure of interest rates [Engle, Hendry, and Trumble (1985)], the volatility of stock market returns [Engle, Lilien, and Robins (1987)], and the behavior of foreign exchange markets [Domowitz and Hakkio (1985) and Bollerslev and Ghysels (1996)], to name but a few. This section will describe specification, estimation, and testing, in the basic formulations of the ARCH model and some extensions.<sup>17</sup>

#### *Example 20.7 Stochastic Volatility*

Figure 20.5 shows Bollerslev and Ghysel's 1974 data on the daily percentage nominal return for the Deutschmark/Pound exchange rate. (These data are given in Appendix Table F20.1.) The variation in the series appears to be fluctuating, with several clusters of large and small movements.

#### 20.13.1 THE ARCH(1) MODEL

The simplest form of this model is the ARCH(1) model,

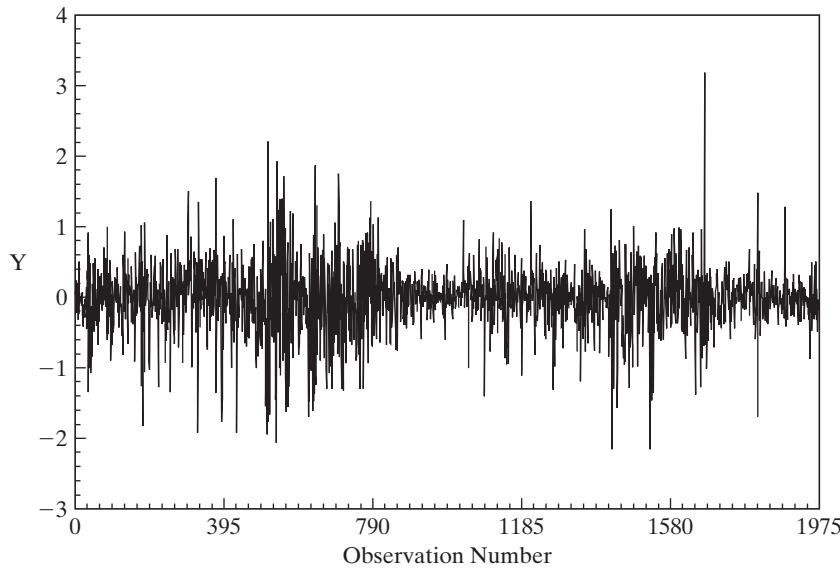
$$\begin{aligned} y_t &= \mathbf{x}_t' \boldsymbol{\beta} + \varepsilon_t, \\ \varepsilon_t &= u_t \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2}, \end{aligned} \tag{20-40}$$

where  $u_t$  is distributed as standard normal.<sup>18</sup> It follows that  $E[\varepsilon_t | \mathbf{x}_t, \varepsilon_{t-1}] = 0$ , so that  $E[\varepsilon_t | \mathbf{x}_t] = 0$  and  $E[y_t | \mathbf{x}_t] = \mathbf{x}_t' \boldsymbol{\beta}$ . Therefore, this model is a classical regression model.

<sup>16</sup>See Baillie (1979).

<sup>17</sup>Engle and Rothschild (1992) give a survey of this literature which describes many extensions. Mills (1993) also presents several applications. See, as well, Bollerslev (1986) and Li, Ling, and McAleer (2001). See McCullough and Renfro (1999) for discussion of estimation of this model.

<sup>18</sup>The assumption that  $u_t$  has unit variance is not a restriction. The scaling implied by any other variance would be absorbed by the other parameters.



**FIGURE 20.5** Nominal Exchange Rate Returns.

But

$$\text{Var}[\varepsilon_t | \varepsilon_{t-1}] = E[\varepsilon_t^2 | \varepsilon_{t-1}] = E[u_t^2] [\alpha_0 + \alpha_1 \varepsilon_{t-1}^2] = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2,$$

so  $\varepsilon_t$  is *conditionally heteroscedastic*, not with respect to  $\mathbf{x}_t$  as we considered in Chapter 18, but with respect to  $\varepsilon_{t-1}$ . The unconditional variance of  $\varepsilon_t$  is

$$\text{Var}[\varepsilon_t] = \text{Var}\{E[\varepsilon_t | \varepsilon_{t-1}]\} + E\{\text{Var}[\varepsilon_t | \varepsilon_{t-1}]\} = \alpha_0 + \alpha_1 E[\varepsilon_{t-1}^2] = \alpha_0 + \alpha_1 \text{Var}[\varepsilon_{t-1}].$$

If the process generating the disturbances is weakly (covariance) stationary (see Definition 19.2),<sup>19</sup> then the unconditional variance is not changing over time so

$$\text{Var}[\varepsilon_t] = \text{Var}[\varepsilon_{t-1}] = \alpha_0 + \alpha_1 \text{Var}[\varepsilon_{t-1}] = \frac{\alpha_0}{1 - \alpha_1}.$$

For this ratio to be finite and positive,  $|\alpha_1|$  must be less than 1. Then, unconditionally,  $\varepsilon_t$  is distributed with mean zero and variance  $\sigma^2 = \alpha_0/(1 - \alpha_1)$ . Therefore, the model obeys the classical assumptions, and ordinary least squares is the most efficient *linear* unbiased estimator of  $\beta$ .

But there is a more efficient *nonlinear* estimator. The log-likelihood function for this model is given by Engle (1982). Conditioned on starting values  $y_0$  and  $\mathbf{x}_0$  (and  $\varepsilon_0$ ), the conditional log-likelihood for observations  $t = 1, \dots, T$  is the one we examined in Section 14.9.2.a for the general heteroscedastic regression model [see (14-52)],

$$\ln L = -\frac{T}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^T \ln(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2) - \frac{1}{2} \sum_{t=1}^T \frac{\varepsilon_t^2}{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2}, \quad \varepsilon_t = y_t - \boldsymbol{\beta}' \mathbf{x}_t.$$

(20-41)

<sup>19</sup>This discussion will draw on the results and terminology of time-series analysis in Section 20.3 and Chapter 22. The reader may wish to peruse this material at this point.

## 938 PART V ♦ Time Series and Microeometrics

Maximization of  $\log L$  can be done with the conventional methods, as discussed in Appendix E.<sup>20</sup>

### 20.13.2 ARCH( $q$ ), ARCH-IN-MEAN, AND GENERALIZED ARCH MODELS

The natural extension of the ARCH(1) model presented before is a more general model with longer lags. The ARCH( $q$ ) process,

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \cdots + \alpha_q \varepsilon_{t-q}^2,$$

is a  $q$ th order moving average [MA( $q$ )] process. (Much of the analysis of the model parallels the results in Chapter 22 for more general time-series models.) [Once again, see Engle (1982).] This section will generalize the ARCH( $q$ ) model, as suggested by Bollerslev (1986), in the direction of the autoregressive-moving average (ARMA) models of Section 22.2.1. The discussion will parallel his development, although many details are omitted for brevity. The reader is referred to that paper for background and for some of the less critical details.

Among the many variants of the capital asset pricing model (CAPM) is an intertemporal formulation by Merton (1980) that suggests an approximate linear relationship between the return and variance of the market portfolio. One of the possible flaws in this model is its assumption of a constant variance of the market portfolio. In this connection, then, the **ARCH-in-Mean**, or ARCH-M, model suggested by Engle, Lilien, and Robins (1987) is a natural extension. The model states that

$$y_t = \beta' \mathbf{x}_t + \delta \sigma_t^2 + \varepsilon_t,$$

$$\text{Var}[\varepsilon_t | \Psi_t] = \text{ARCH}(q).$$

Among the interesting implications of this modification of the standard model is that under certain assumptions,  $\delta$  is the coefficient of relative risk aversion. The ARCH-M model has been applied in a wide variety of studies of volatility in asset returns, including the daily Standard and Poor's Index [French, Schwert, and Stambaugh (1987)] and weekly New York Stock Exchange returns [Chou (1988)]. A lengthy list of applications is given in Bollerslev, Chou, and Kroner (1992).

The ARCH-M model has several noteworthy statistical characteristics. Unlike the standard regression model, misspecification of the variance function does affect the consistency of estimators of the parameters of the mean. [See Pagan and Ullah (1988) for formal analysis of this point.] Recall that in the classical regression setting, weighted least squares is consistent even if the weights are misspecified as long as the weights are uncorrelated with the disturbances. That is not true here. If the ARCH part of the model is misspecified, then conventional estimators of  $\beta$  and  $\delta$  will not be consistent. Bollerslev, Chou, and Kroner (1992) list a large number of studies that called into question the specification of the ARCH-M model, and they subsequently obtained quite different

<sup>20</sup>Engle (1982) and Judge et al. (1985, pp. 441–444) suggest a four-step procedure based on the method of scoring that resembles the two-step method we used for the multiplicative heteroscedasticity model in Section 8.8.1. However, the full MLE is now incorporated in most modern software, so the simple regression based methods, which are difficult to generalize, are less attractive in the current literature. But, see McCullough and Renfro (1999) and Fiorentini, Calzolari, and Panattoni (1996) for commentary and some cautions related to maximum likelihood estimation.

## CHAPTER 20 ♦ Serial Correlation 939

results after respecifying the model. A closely related practical problem is that the mean and variance parameters in this model are no longer uncorrelated. In analysis up to this point, we made quite profitable use of the block diagonality of the Hessian of the log-likelihood function for the model of heteroscedasticity. But the Hessian for the ARCH-M model is not block diagonal. In practical terms, the estimation problem cannot be segmented as we have done previously with the heteroscedastic regression model. All the parameters must be estimated simultaneously.

The model of generalized autoregressive conditional heteroscedasticity (GARCH) is defined as follows.<sup>21</sup> The underlying regression is the usual one in (20-40). *Conditioned on an information set at time t*, denoted  $\Psi_t$ , the distribution of the disturbance is assumed to be

$$\varepsilon_t | \Psi_t \sim N[0, \sigma_t^2],$$

where the conditional variance is

$$\sigma_t^2 = \alpha_0 + \delta_1 \sigma_{t-1}^2 + \delta_2 \sigma_{t-2}^2 + \cdots + \delta_p \sigma_{t-p}^2 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \cdots + \alpha_q \varepsilon_{t-q}^2. \quad (20-42)$$

Define

$$\mathbf{z}_t = [1, \sigma_{t-1}^2, \sigma_{t-2}^2, \dots, \sigma_{t-p}^2, \varepsilon_{t-1}^2, \varepsilon_{t-2}^2, \dots, \varepsilon_{t-q}^2]'$$

and

$$\boldsymbol{\gamma} = [\alpha_0, \delta_1, \delta_2, \dots, \delta_p, \alpha_1, \dots, \alpha_q]' = [\alpha_0, \boldsymbol{\delta}', \boldsymbol{\alpha}']'.$$

Then

$$\sigma_t^2 = \boldsymbol{\gamma}' \mathbf{z}_t.$$

Notice that the conditional variance is defined by an autoregressive-moving average [ARMA ( $p, q$ )] process in the innovations  $\varepsilon_t^2$ , exactly as in Section 22.2.1. The difference here is that the *mean* of the random variable of interest  $y_t$  is described completely by a heteroscedastic, but otherwise ordinary, regression model. The *conditional variance*, however, evolves over time in what might be a very complicated manner, depending on the parameter values and on  $p$  and  $q$ . The model in (20-42) is a GARCH( $p, q$ ) model, where  $p$  refers, as before, to the order of the autoregressive part.<sup>22</sup> As Bollerslev (1986) demonstrates with an example, the virtue of this approach is that a GARCH model with a small number of terms appears to perform as well as or better than an ARCH model with many.

The **stationarity conditions** discussed in Section 22.2.2 are important in this context to ensure that the moments of the normal distribution are finite. The reason is that higher moments of the normal distribution are finite powers of the variance. A normal distribution with variance  $\sigma_t^2$  has fourth moment  $3\sigma_t^4$ , sixth moment  $15\sigma_t^6$ , and so on. [The precise relationship of the even moments of the normal distribution to the variance is  $\mu_{2k} = (\sigma^2)^k (2k)! / (k! 2^k)$ .] Simply ensuring that  $\sigma_t^2$  is stable does not ensure that higher

<sup>21</sup>As have most areas in time-series econometrics, the line of literature on GARCH models has progressed rapidly in recent years and will surely continue to do so. We have presented Bollerslev's model in some detail, despite many recent extensions, not only to introduce the topic as a bridge to the literature, but also because it provides a convenient and interesting setting in which to discuss several related topics such as double-length regression and pseudo-maximum likelihood estimation.

<sup>22</sup>We have changed Bollerslev's notation slightly so as not to conflict with our previous presentation. He used  $\beta$  instead of our  $\delta$  in (20-42) and  $\mathbf{b}$  instead of our  $\boldsymbol{\beta}$  in (20-40).

## 940 PART V ♦ Time Series and Microeometrics

powers are as well.<sup>23</sup> Bollerslev presents a useful figure that shows the conditions needed to ensure stability for moments up to order 12 for a GARCH(1, 1) model and gives some additional discussion. For example, for a GARCH(1, 1) process, for the fourth moment to exist,  $3\alpha_1^2 + 2\alpha_1\delta_1 + \delta_1^2$  must be less than 1.

It is convenient to write (20-42) in terms of polynomials in the lag operator (see Section 21.2.2):

$$\sigma_t^2 = \alpha_0 + D(L)\sigma_t^2 + A(L)\varepsilon_t^2.$$

As discussed in Section 21.2.2, the stationarity condition for such an equation is that the roots of the characteristic equation,  $1 - D(z) = 0$ , must lie outside the unit circle. For the present, we will assume that this case is true for the model we are considering and that  $A(1) + D(1) < 1$ . [This assumption is stronger than that needed to ensure stationarity in a higher-order autoregressive model, which would depend only on  $D(L)$ .] The implication is that the GARCH process is covariance stationary with  $E[\varepsilon_t] = 0$  (unconditionally),  $\text{Var}[\varepsilon_t] = \alpha_0/[1 - A(1) - D(1)]$ , and  $\text{Cov}[\varepsilon_t, \varepsilon_s] = 0$  for all  $t \neq s$ . Thus, unconditionally the model is the classical regression model that we examined in Chapters 2–6.

The usefulness of the GARCH specification is that it allows the variance to evolve over time in a way that is much more general than the simple specification of the ARCH model. The comparison between simple finite-distributed lag models and the dynamic regression model discussed in Chapter 21 is analogous. For the example discussed in his paper, Bollerslev reports that although Engle and Kraft's (1983) ARCH(8) model for the rate of inflation in the GNP deflator appears to remove all ARCH effects, a closer look reveals GARCH effects at several lags. By fitting a GARCH(1, 1) model to the same data, Bollerslev finds that the ARCH effects out to the same eight-period lag as fit by Engle and Kraft and his observed GARCH effects are all satisfactorily accounted for.

### 20.13.3 MAXIMUM LIKELIHOOD ESTIMATION OF THE GARCH MODEL

Bollerslev describes a method of estimation based on the BHHH algorithm. As he shows, the method is relatively simple, although with the line search and first derivative method that he suggests, it probably involves more computation and more iterations than necessary. Following the suggestions of Harvey (1976), it turns out that there is a simpler way to estimate the GARCH model that is also very illuminating. This model is actually very similar to the more conventional model of multiplicative heteroscedasticity that we examined in Section 14.9.2.b.

For normally distributed disturbances, the log-likelihood for a sample of  $T$  observations is<sup>24</sup>

$$\ln L = \sum_{t=1}^T -\frac{1}{2} \left[ \ln(2\pi) + \ln \sigma_t^2 + \frac{\varepsilon_t^2}{\sigma_t^2} \right] = \sum_{t=1}^T \ln f_t(\boldsymbol{\theta}) = \sum_{t=1}^T l_t(\boldsymbol{\theta}),$$

<sup>23</sup>The conditions cannot be imposed a priori. In fact, there is no nonzero set of parameters that guarantees stability of *all* moments, even though the normal distribution has finite moments of all orders. As such, the normality assumption must be viewed as an approximation.

<sup>24</sup>There are three minor errors in Bollerslev's derivation that we note here to avoid the apparent inconsistencies. In his (22),  $\frac{1}{2}h_t$  should be  $\frac{1}{2}h_t^{-1}$ . In (23),  $-2h_t^{-2}$  should be  $-h_t^{-2}$ . In (28),  $h \partial h / \partial \omega$  should, in each case, be  $(1/h) \partial h / \partial \omega$ . [In his (8),  $\alpha_0\alpha_1$  should be  $\alpha_0 + \alpha_1$ , but this has no implications for our derivation.]

## CHAPTER 20 ♦ Serial Correlation 941

where  $\varepsilon_t = y_t - \mathbf{x}'\beta$  and  $\theta = (\beta', \alpha_0, \alpha', \delta')' = (\beta', \gamma')'$ . Derivatives of  $\ln L$  are obtained by summation. Let  $l_t$  denote  $\ln f_t(\theta)$ . The first derivatives with respect to the variance parameters are

$$\frac{\partial l_t}{\partial \gamma} = -\frac{1}{2} \left[ \frac{1}{\sigma_t^2} - \frac{\varepsilon_t^2}{(\sigma_t^2)^2} \right] \frac{\partial \sigma_t^2}{\partial \gamma} = \frac{1}{2} \left( \frac{1}{\sigma_t^2} \right) \frac{\partial \sigma_t^2}{\partial \gamma} \left( \frac{\varepsilon_t^2}{\sigma_t^2} - 1 \right) = \frac{1}{2} \left( \frac{1}{\sigma_t^2} \right) \mathbf{g}_t v_t = \mathbf{b}_t v_t. \quad (20-43)$$

Note that  $E[v_t] = 0$ . Suppose, for now, that there are no regression parameters. Newton's method for estimating the variance parameters would be

$$\hat{\gamma}^{i+1} = \hat{\gamma}^i - \mathbf{H}^{-1} \mathbf{g}, \quad (20-44)$$

where  $\mathbf{H}$  indicates the Hessian and  $\mathbf{g}$  is the first derivatives vector. Following Harvey's suggestion (see Section 14.9.2.a), we will use the method of scoring instead. To do this, we make use of  $E[v_t] = 0$  and  $E[\varepsilon_t^2/\sigma_t^2] = 1$ . After taking expectations in (20-43), the iteration reduces to a linear regression of  $v_{*t} = (1/\sqrt{2})v_t$  on regressors  $\mathbf{w}_{*t} = (1/\sqrt{2})\mathbf{g}_t/\sigma_t^2$ . That is,

$$\hat{\gamma}^{i+1} = \hat{\gamma}^i + [\mathbf{W}'_* \mathbf{W}_*]^{-1} \mathbf{W}'_* \mathbf{v}_* = \hat{\gamma}^i + [\mathbf{W}'_* \mathbf{W}_*]^{-1} \left( \frac{\partial \ln L}{\partial \gamma} \right), \quad (20-45)$$

where row  $t$  of  $\mathbf{W}_*$  is  $\mathbf{w}'_{*t}$ . The iteration has converged when the slope vector is zero, which happens when the first derivative vector is zero. When the iterations are complete, the estimated asymptotic covariance matrix is simply

$$\text{Est. Asy. Var}[\hat{\gamma}] = [\hat{\mathbf{W}}'_* \hat{\mathbf{W}}_*]^{-1}$$

based on the estimated parameters.

The usefulness of the result just given is that  $E[\partial^2 \ln L / \partial \gamma \partial \beta']$  is, in fact, zero. Because the expected Hessian is block diagonal, applying the method of scoring to the full parameter vector can proceed in two parts, exactly as it did in Section 14.9.2.a for the multiplicative heteroscedasticity model. That is, the updates for the mean and variance parameter vectors can be computed separately. Consider then the slope parameters,  $\beta$ . The same type of modified scoring method as used earlier produces the iteration

$$\begin{aligned} \hat{\beta}^{i+1} &= \hat{\beta}^i + \left[ \sum_{t=1}^T \frac{\mathbf{x}_t \mathbf{x}'_t}{\sigma_t^2} + \frac{1}{2} \left( \frac{\mathbf{d}_t}{\sigma_t^2} \right) \left( \frac{\mathbf{d}_t}{\sigma_t^2} \right)' \right]^{-1} \left[ \sum_{t=1}^T \frac{\mathbf{x}_t \varepsilon_t}{\sigma_t^2} + \frac{1}{2} \left( \frac{\mathbf{d}_t}{\sigma_t^2} \right) v_t \right] \\ &= \hat{\beta}^i + \left[ \sum_{t=1}^T \frac{\mathbf{x}_t \mathbf{x}'_t}{\sigma_t^2} + \frac{1}{2} \left( \frac{\mathbf{d}_t}{\sigma_t^2} \right) \left( \frac{\mathbf{d}_t}{\sigma_t^2} \right)' \right]^{-1} \left( \frac{\partial \ln L}{\partial \beta} \right) \\ &= \hat{\beta}^i + \mathbf{h}^i, \end{aligned} \quad (20-46)$$

which has been referred to as a **double-length regression**. [See Orme (1990) and Davidson and MacKinnon (1993, Chapter 14).] The update vector  $\mathbf{h}^i$  is the vector of slopes in an augmented or double-length generalized regression,

$$\mathbf{h}^i = [\mathbf{C}' \boldsymbol{\Omega}^{-1} \mathbf{C}]^{-1} [\mathbf{C}' \boldsymbol{\Omega}^{-1} \mathbf{a}], \quad (20-47)$$

where  $\mathbf{C}$  is a  $2T \times K$  matrix whose first  $T$  rows are the  $\mathbf{X}$  from the original regression model and whose next  $T$  rows are  $(1/\sqrt{2})\mathbf{d}'_t/\sigma_t^2$ ,  $t = 1, \dots, T$ ;  $\mathbf{a}$  is a  $2T \times 1$  vector whose

## 942 PART V ♦ Time Series and Microeometrics

first  $T$  elements are  $\varepsilon_t$  and whose next  $T$  elements are  $(1/\sqrt{2})v_t/\sigma_t^2$ ,  $t = 1, \dots, T$ ; and  $\Omega$  is a diagonal matrix with  $1/\sigma_t^2$  in positions  $1, \dots, T$  and ones below observation  $T$ . At convergence,  $[\mathbf{C}'\Omega^{-1}\mathbf{C}]^{-1}$  provides the asymptotic covariance matrix for the MLE. The resemblance to the familiar result for the generalized regression model is striking, but note that this result is based on the double-length regression.

The iteration is done simply by computing the update vectors to the current parameters as defined earlier.<sup>25</sup> An important consideration is that to apply the scoring method, the estimates of  $\beta$  and  $\gamma$  are updated simultaneously. That is, one does not use the updated estimate of  $\gamma$  in (20-45) to update the weights for the GLS regression to compute the new  $\beta$  in (20-46). The same estimates (the results of the prior iteration) are used on the right-hand sides of both (20-45) and (20-46). The remaining problem is to obtain starting values for the iterations. One obvious choice is  $\mathbf{b}$ , the OLS estimator, for  $\beta$ ,  $\mathbf{e}'\mathbf{e}/T = s^2$  for  $\alpha_0$ , and zero for all the remaining parameters. The OLS slope vector will be consistent under all specifications. A useful alternative in this context would be to start  $\alpha$  at the vector of slopes in the least squares regression of  $e_t^2$ , the squared OLS residual, on a constant and  $q$  lagged values.<sup>26</sup> As discussed later, an LM test for the presence of GARCH effects is then a by-product of the first iteration. In principle, the updated result of the first iteration is an **efficient two-step estimator** of all the parameters. But having gone to the full effort to set up the iterations, nothing is gained by not iterating to convergence. One virtue of allowing the procedure to iterate to convergence is that the resulting log-likelihood function can be used in likelihood ratio tests.

### 20.13.4 TESTING FOR GARCH EFFECTS

The preceding development appears fairly complicated. In fact, it is not, because at each step, nothing more than a linear least squares regression is required. The intricate part of the computation is setting up the derivatives. On the other hand, it does take a fair amount of programming to get this far.<sup>27</sup> As Bollerslev suggests, it might be useful to test for GARCH effects first.

The simplest approach is to examine the squares of the least squares residuals. The autocorrelations (correlations with lagged values) of the squares of the residuals provide evidence about ARCH effects. An LM test of ARCH( $q$ ) against the hypothesis of no ARCH effects [ARCH(0), the classical model] can be carried out by computing  $\chi^2 = TR^2$  in the regression of  $e_t^2$  on a constant and  $q$  lagged values. Under the null hypothesis of no ARCH effects, the statistic has a limiting chi-squared distribution with  $q$  degrees of freedom. Values larger than the critical table value give evidence of the presence of ARCH (or GARCH) effects.

Bollerslev suggests a Lagrange multiplier statistic that is, in fact, surprisingly simple to compute. The LM test for GARCH( $p, 0$ ) against GARCH( $p, q$ ) can be carried out by referring  $T$  times the  $R^2$  in the linear regression defined in (20-42) to the chi-squared

<sup>25</sup>See Fiorentini et al. (1996) on computation of derivatives in GARCH models.

<sup>26</sup>A test for the presence of  $q$  ARCH effects against none can be carried out by carrying  $TR^2$  from this regression into a table of critical values for the chi-squared distribution. But in the presence of GARCH effects, this procedure loses its validity.

<sup>27</sup>Because this procedure is available as a preprogrammed procedure in many computer programs, including TSP, E-Views, Stata, RATS, LIMDEP, and Shazam, this warning might itself be overstated.

**TABLE 20.4** Maximum Likelihood Estimates of a GARCH(1, 1) Model<sup>29</sup>

	$\mu$	$\alpha_0$	$\alpha_1$	$\delta$	$\alpha_0/(1 - \alpha_1 - \delta)$
Estimate	-0.006190	0.01076	0.1531	0.8060	0.2631
Std. Error	0.00873	0.00312	0.0273	0.0302	0.594
t ratio	-0.709	3.445	5.605	26.731	0.443
$\ln L = -1106.61, \ln L_{OLS} = -1311.09, \bar{y} = -0.01642, s^2 = 0.221128$					

critical value with  $q$  degrees of freedom. There is, unfortunately, an indeterminacy in this test procedure. The test for ARCH( $q$ ) against GARCH( $p, q$ ) is exactly the same as that for ARCH( $p$ ) against ARCH( $p + q$ ). For carrying out the test, one can use as starting values a set of estimates that includes  $\delta = 0$  and any consistent estimators for  $\beta$  and  $\alpha$ . Then  $TR^2$  for the regression at the initial iteration provides the test statistic.<sup>28</sup>

A number of recent papers have questioned the use of test statistics based solely on normality. Wooldridge (1991) is a useful summary with several examples.

#### **Example 20.8 GARCH Model for Exchange Rate Volatility**

Bollerslev and Ghysels analyzed the exchange rate data in Example 20.7 using a GARCH(1, 1) model,

$$\begin{aligned} y_t &= \mu + \varepsilon_t, \\ E[\varepsilon_t | \varepsilon_{t-1}] &= 0, \\ \text{Var}[\varepsilon_t | \varepsilon_{t-1}] &= \sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \delta \sigma_{t-1}^2. \end{aligned}$$

The least squares residuals for this model are simply  $e_t = y_t - \bar{y}$ . Regression of the squares of these residuals on a constant and 10 lagged squared values using observations 11–1974 produces an  $R^2 = 0.09795$ . With  $T = 1964$ , the chi-squared statistic is 192.37, which is larger than the critical value from the table of 18.31. We conclude that there is evidence of GARCH effects in these residuals. The maximum likelihood estimates of the GARCH model are given in Table 20.4. Note the resemblance between the OLS unconditional variance (0.221128) and the estimated equilibrium variance from the GARCH model, 0.2631.

#### **20.13.5 PSEUDO-MAXIMUM LIKELIHOOD ESTIMATION**

We now consider an implication of nonnormality of the disturbances. Suppose that the assumption of normality is weakened to only

$$E[\varepsilon_t | \Psi_t] = 0, \quad E\left[\frac{\varepsilon_t^2}{\sigma_t^2} \mid \Psi_t\right] = 1, \quad E\left[\frac{\varepsilon_t^4}{\sigma_t^4} \mid \Psi_t\right] = \kappa < \infty,$$

where  $\sigma_t^2$  is as defined earlier. Now the normal log-likelihood function is inappropriate. In this case, the nonlinear (ordinary or weighted) least squares estimator would have the

<sup>28</sup>Bollerslev argues that in view of the complexity of the computations involved in estimating the GARCH model, it is useful to have a test for GARCH effects. This case is one (as are many other maximum likelihood problems) in which the apparatus for carrying out the test is the same as that for estimating the model. Having computed the LM statistic for GARCH effects, one can proceed to estimate the model just by allowing the program to iterate to convergence. There is no additional cost beyond waiting for the answer.

<sup>29</sup>These data have become a standard data set for the evaluation of software for estimating GARCH models. The values given are the benchmark estimates. Standard errors differ substantially from one method to the next. Those given are the Bollerslev and Wooldridge (1992) results. See McCullough and Renfro (1999).

## 944 PART V ♦ Time Series and Microeconometrics



properties discussed in Chapter 9. It would be more difficult to compute than the MLE discussed earlier, however. It has been shown [see White (1982a) and Weiss (1982)] that the **pseudo-MLE** obtained by maximizing the same log-likelihood as if it were correct produces a consistent estimator despite the misspecification.<sup>30</sup> The asymptotic covariance matrices for the parameter estimators must be adjusted, however.

The general result for cases such as this one [see Gourieroux, Monfort, and Trognon (1984)] is that the appropriate asymptotic covariance matrix for the pseudo-MLE of a parameter vector  $\theta$  would be

$$\text{Asy. Var}[\hat{\theta}] = \mathbf{H}^{-1} \mathbf{F} \mathbf{H}^{-1}, \quad (20-48)$$

where

$$\mathbf{H} = -E \left[ \frac{\partial^2 \ln L}{\partial \theta \partial \theta'} \right],$$

and

$$\mathbf{F} = E \left[ \left( \frac{\partial \ln L}{\partial \theta} \right) \left( \frac{\partial \ln L}{\partial \theta'} \right) \right]$$

(i.e., the BHHH estimator), and  $\ln L$  is the used but inappropriate log-likelihood function. For current purposes,  $\mathbf{H}$  and  $\mathbf{F}$  are still block diagonal, so we can treat the mean and variance parameters separately. In addition,  $E[v_t]$  is still zero, so the second derivative terms in both blocks are quite simple. (The parts involving  $\partial^2 \sigma_t^2 / \partial \beta \partial \beta'$  fall out of the expectation.) Taking expectations and inserting the parts produces the corrected asymptotic covariance matrix for the variance parameters:

$$\text{Asy. Var}[\hat{\gamma}_{\text{PMLE}}] = [\mathbf{W}_*^* \mathbf{W}_*]^{-1} \mathbf{B}' \mathbf{B} [\mathbf{W}_*^* \mathbf{W}_*]^{-1},$$

where the rows of  $\mathbf{W}^*$  are defined in (20-45) and those of  $\mathbf{B}$  are in (20-43). For the slope parameters, the adjusted asymptotic covariance matrix would be

$$\text{Asy. Var}[\hat{\beta}_{\text{PMLE}}] = [\mathbf{C}' \boldsymbol{\Omega}^{-1} \mathbf{C}]^{-1} \left[ \sum_{t=1}^T \mathbf{b}_t \mathbf{b}_t' \right] [\mathbf{C}' \boldsymbol{\Omega}^{-1} \mathbf{C}]^{-1},$$

where the outer matrix is defined in (20-47) and, from the first derivatives given in (20-43) and (20-46),<sup>31</sup>

$$\mathbf{b}_t = \frac{\mathbf{x}_t \varepsilon_t}{\sigma_t^2} + \frac{1}{2} \left( \frac{v_t}{\sigma_t^2} \right) \mathbf{d}_t.$$

<sup>30</sup>White (1982a) gives some additional requirements for the true underlying density of  $\varepsilon_t$ . Gourieroux, Monfort, and Trognon (1984) also consider the issue. Under the assumptions given, the expectations of the matrices in (20-42) and (20-47) remain the same as under normality. The consistency and asymptotic normality of the pseudo-MLE can be argued under the logic of GMM estimators.

<sup>31</sup>McCullough and Renfro (1999) examined several approaches to computing an appropriate asymptotic covariance matrix for the GARCH model, including the conventional Hessian and BHHH estimators and three sandwich style estimators, including the one suggested earlier and two based on the method of scoring suggested by Bollerslev and Wooldridge (1992). None stand out as obviously better, but the Bollerslev and QMLE estimator based on an actual Hessian appears to perform well in Monte Carlo studies.

## 20.14 SUMMARY AND CONCLUSIONS

This chapter has examined the generalized regression model with serial correlation in the disturbances. We began with some general results on analysis of time-series data. When we consider dependent observations and serial correlation, the laws of large numbers and central limit theorems used to analyze independent observations no longer suffice. We presented some useful tools that extend these results to time-series settings. We then considered estimation and testing in the presence of autocorrelation. As usual, OLS is consistent but inefficient. The Newey-West estimator is a robust estimator for the asymptotic covariance matrix of the OLS estimator. This pair of estimators also constitute the GMM estimator for the regression model with autocorrelation. We then considered two-step feasible generalized least squares and maximum likelihood estimation for the special case usually analyzed by practitioners, the AR(1) model. The model with a correction for autocorrelation is a restriction on a more general model with lagged values of both dependent and independent variables. We considered a means of testing this specification as an alternative to “fixing” the problem of autocorrelation. The final section, on ARCH and GARCH effects, describes an extension of the models of autoregression to the conditional variance of  $\varepsilon$  as opposed to the conditional mean. This model embodies elements of both autocorrelation and heteroscedasticity. The set of methods plays a fundamental role in the modern analysis of volatility in financial data.

### Key Terms and Concepts

- AR(1)
- ARCH
- ARCH-in-mean
- Asymptotic negligibility
- Asymptotic normality
- Autocorrelation
- Autocorrelation coefficient
- Autocorrelation matrix
- Autocovariance
- Autocovariance matrix
- Autoregressive form
- Autoregressive processes
- Cochrane–Orcutt estimator
- Common factors
- Common factor model
- Covariance stationarity
- Double-length regression
- Durbin–Watson test
- Efficient two-step estimator
- Ergodicity
- Ergodic theorem
- Expectations-augmented Phillips curve
- First-order autoregression
- GARCH
- GMM estimator
- Initial conditions
- Innovation
- Lagrange multiplier test
- Martingale sequence
- Martingale difference sequence
- Moving average form
- Moving-average process
- Newey–West autocorrelation consistent covariance estimator
- Newey–West robust
- covariance matrix estimator
- Partial difference
- Prais–Winsten estimator
- Pseudo-differences
- Pseudo-MLE
- $Q$  test
- Quasi differences
- Random walk
- Stationarity
- Stationarity conditions
- Summability
- Time-series process
- Time window
- Weakly stationary
- White noise
- Yule–Walker equations

### Exercises

1. Does first differencing reduce autocorrelation? Consider the models  $y_t = \beta' \mathbf{x}_t + \varepsilon_t$ , where  $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$  and  $\varepsilon_t = u_t - \lambda u_{t-1}$ . Compare the autocorrelation of  $\varepsilon_t$  in the original model with that of  $v_t$  in  $y_t - y_{t-1} = \beta' (\mathbf{x}_t - \mathbf{x}_{t-1}) + v_t$ , where  $v_t = \varepsilon_t - \varepsilon_{t-1}$ .

**946 PART V ♦ Time Series and Microeometrics**

2. Derive the disturbance covariance matrix for the model

$$y_t = \beta' \mathbf{x}_t + \varepsilon_t,$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t - \lambda u_{t-1}.$$

What parameter is estimated by the regression of the OLS residuals on their lagged values?

3. The following regression is obtained by ordinary least squares, using 21 observations. (Estimated asymptotic standard errors are shown in parentheses.)

$$y_t = 1.3 + 0.97y_{t-1} + 2.31x_t, \quad D-W = 1.21.$$

$$(0.3) \quad (0.18) \quad (1.04)$$

Test for the presence of autocorrelation in the disturbances.

4. It is commonly asserted that the Durbin–Watson statistic is only appropriate for testing for first-order autoregressive disturbances. What combination of the coefficients of the model is estimated by the Durbin–Watson statistic in each of the following cases: AR(1), AR(2), MA(1)? In each case, assume that the regression model does not contain a lagged dependent variable. Comment on the impact on your results of relaxing this assumption.

### **Applications**

1. The data used to fit the expectations augmented Phillips curve in Example 20.3 are given in Appendix Table F5.2. Using these data, reestimate the model given in the example. Carry out a formal test for first-order autocorrelation using the LM statistic. Then, reestimate the model using an AR(1) model for the disturbance process. Because the sample is large, the Prais–Winsten and Cochrane–Orcutt estimators should give essentially the same answer. Do they? After fitting the model, obtain the transformed residuals and examine them for first-order autocorrelation. Does the AR(1) model appear to have adequately “fixed” the problem?
2. Data for fitting an improved Phillips curve model can be obtained from many sources, including the Bureau of Economic Analysis’s (BEA) own Web site, [www.economagic.com](http://www.economagic.com), and so on. Obtain the necessary data and expand the model of Example 20.3. Does adding additional explanatory variables to the model reduce the extreme pattern of the OLS residuals that appears in Figure 20.3?
3. (This exercise requires appropriate computer software. The computations required can be done with RATS, EViews, Stata, TSP, LIMDEP, and a variety of other software using only preprogrammed procedures.) Quarterly data on the consumer price index for 1950.1 to 2000.4 are given in Appendix Table F5.2. Use these data to fit the model proposed by Engle and Kraft (1983). The model is

$$\pi_t = \beta_0 + \beta_1 \pi_{t-1} + \beta_2 \pi_{t-2} + \beta_3 \pi_{t-3} + \beta_4 \pi_{t-4} + \varepsilon_t,$$

where  $\pi_t = 100 \ln[p_t/p_{t-1}]$  and  $p_t$  is the price index.

- a. Fit the model by ordinary least squares, then use the tests suggested in the text to see if ARCH effects appear to be present.

**CHAPTER 20 ♦ Serial Correlation 947**

- b.** The authors fit an ARCH(8) model with declining weights,

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^8 \left( \frac{9-i}{36} \right) \varepsilon_{t-i}^2.$$

Fit this model. If the software does not allow constraints on the coefficients, you can still do this with a two-step least squares procedure, using the least squares residuals from the first step. What do you find?

- c.** Bollerslev (1986) recomputed this model as a GARCH(1,1). Use the GARCH(1,1) to form and refit your model.

## 21

# MODELS WITH LAGGED VARIABLES

---

## 21.1 INTRODUCTION

This chapter begins our introduction to the analysis of economic time series. By most views, this field has become synonymous with empirical macroeconomics and the analysis of financial markets.<sup>1</sup> In this and the next chapter, we will consider a number of models and topics in which time and relationships through time play an explicit part in the formulation. Consider the **dynamic regression model**

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 x_{t-1} + \gamma y_{t-1} + \varepsilon_t. \quad (21-1)$$

Models of this form specifically include as right-hand-side variables previous as well as contemporaneous values of the regressors. It is also in this context that lagged values of the dependent variable appear as a consequence of the theoretical basis of the model rather than as a computational means of removing autocorrelation. There are several reasons lagged effects might appear in an empirical model:

- In modeling the response of economic variables to policy stimuli, it is expected that there will be possibly long lags between policy changes and their impacts. The length of lag between changes in monetary policy and its impact on important economic variables such as output and investment has been a subject of analysis for several decades.
- Either the dependent variable or one of the independent variables is based on expectations. **Expectations** about economic events are usually formed by aggregating new information and past experience. Thus, we might write the expectation of a future value of variable  $x$ , formed this period, as

$$x_t = E_t[x_{t+1}^* | z_t, x_{t-1}, x_{t-2}, \dots] = g(z_t, x_{t-1}, x_{t-2}, \dots).$$

<sup>1</sup>The literature in this area has grown at an impressive rate, and, more so than in any other area, it has become impossible to provide comprehensive surveys in general textbooks such as this one. Fortunately, specialized volumes have been produced that can fill this need at any level. Harvey (1990) has been in wide use for some time. Among the many other books, three very useful works are Enders (2003), which presents the basics of time-series analysis at an introductory level with several very detailed applications; Hamilton (1994), which gives a relatively technical but quite comprehensive survey of the field; and Lutkepohl (2005), which provides an extremely detailed treatment of the topics presented at the end of this chapter. Hamilton also surveys a number of the applications in the contemporary literature. Two references that are focused on financial econometrics are Mills (1993) and Tsay (2005). There are also a number of important references that are primarily limited to forecasting, including Diebold (1998a, 2003) and Granger and Newbold (1996). A survey of research in many areas of time-series analysis is Engle and McFadden (1994). An extensive, fairly advanced treatise that analyzes in great depth all the issues we touch on in this chapter is Hendry (1995). Finally, Patterson (2000) surveys most of the practical issues in time series and presents a large variety of useful and very detailed applications.

## CHAPTER 21 ♦ Models with Lagged Variables 949

For example, forecasts of prices and income enter demand equations and consumption equations. (See Example 13.1 for an influential application.)

- Certain economic decisions are explicitly driven by a history of related activities. For example, energy demand by individuals is clearly a function not only of current prices and income, but also the accumulated stocks of energy using capital. Even energy demand in the macroeconomy behaves in this fashion—the stock of automobiles and its attendant demand for gasoline is clearly driven by past prices of gasoline and automobiles. Other classic examples are the dynamic relationship between investment decisions and past appropriation decisions and the consumption of addictive goods such as cigarettes and theater performances.

We begin with a general discussion of models containing **lagged variables**. In Section 21.2, we consider some methodological issues in the specification of dynamic regressions. In Sections 21.3 and 21.4, we describe a general dynamic model that encompasses some of the extensions and more formal models for time-series data that are presented in Chapter 22. Section 21.5 takes a closer look at some of issues in model specification. Finally, Section 21.6 considers systems of dynamic equations. This chapter is generally not about methods of estimation. OLS and GMM estimation are usually routine in this context. Because we are examining time-series data, conventional assumptions including ergodicity and stationarity will be made at the outset. In particular, in the general framework, we will assume that the multivariate stochastic process  $(y_t, \mathbf{x}_t, \varepsilon_t)$  are a **stationary** and ergodic process. As such, without further analysis, we will invoke the theorems discussed in Chapters 4, 13, 14, and 20 that support least squares and GMM as appropriate estimate techniques in this context. In most of what follows, in fact, in practical terms, the dynamic regression model can be treated as a linear regression model and estimated by conventional methods (e.g., ordinary least squares or instrumental variables if  $\varepsilon_t$  is autocorrelated). As noted, we will generally not return to the issue of estimation and inference theory except where new results are needed, such as in the discussion of nonstationary processes.

## 21.2 DYNAMIC REGRESSION MODELS

In some settings, economic agents respond not only to current values of independent variables but to past values as well. When effects persist over time, an appropriate model will include lagged variables. Example 21.1 illustrates a familiar case.

### **Example 21.1 A Structural Model of the Demand for Gasoline**

Drivers demand gasoline not for direct consumption, but as fuel for cars to provide a source of energy for transportation. Per capita demand for gasoline in any period,  $G/Pop$ , is determined partly by the current price,  $P_g$ , and per capita income,  $Y/Pop$ , which influence how intensively the existing stock of gasoline using “capital,”  $K$ , is used and partly by the size and composition of the stock of cars and other vehicles. The capital stock is determined, in turn, by income,  $Y/Pop$ ; prices of the equipment such as new and used cars,  $P_{nc}$  and  $P_{uc}$ ; the price of alternative modes of transportation such as public transportation,  $P_{pt}$ ; and past prices of gasoline as they influence forecasts of future gasoline prices. A structural model of

## 950 PART V ♦ Time Series and Macroeconometrics

these effects might appear as follows:

$$\text{per capita demand: } G_t/Pop_t = \alpha + \beta Pg_t + \delta Y_t/Pop_t + \gamma K_t + u_t$$

$$\text{stock of vehicles: } K_t = (1 - \Delta)K_{t-1} + I_t, \Delta = \text{depreciation rate}$$

$$\text{investment in new vehicles: } I_t = \theta Y_t/Pop_t + \phi E_t[Pg_{t+1}] + \lambda_1 Pnc_t + \lambda_2 Puc_t + \lambda_3 Ppt_t$$

$$\text{expected price of gasoline: } E_t[Pg_{t+1}] = w_0 Pg_t + w_1 Pg_{t-1} + w_2 Pg_{t-2}$$

The capital stock is the sum of all past investments, so it is evident that not only current income and prices, but all past values, play a role in determining  $K$ . When income or the price of gasoline changes, the immediate effect will be to cause drivers to use their vehicles more or less intensively. But, over time, vehicles are added to the capital stock, and some cars are replaced with more or less efficient ones. These changes take some time, so the full impact of income and price changes will not be felt for several periods. Two episodes in the recent history have shown this effect clearly. For well over a decade following the 1973 oil shock, drivers gradually replaced their large, fuel-inefficient cars with smaller, less-fuel-intensive models. In the late 1990s in the United States, this process has visibly worked in reverse. As American drivers have become accustomed to steadily rising incomes and steadily falling real gasoline prices, the downsized, efficient coupes and sedans of the 1980s have yielded the highways to a tide of ever-larger, six- and eight-cylinder sport utility vehicles, whose size and power can reasonably be characterized as astonishing.

### 21.2.1 LAGGED EFFECTS IN A DYNAMIC MODEL

The general form of a dynamic regression model is

$$y_t = \alpha + \sum_{i=0}^{\infty} \beta_i x_{t-i} + \varepsilon_t. \quad (21-2)$$

In this model, a one-time change in  $x$  at any point in time will affect  $E[y_s | x_t, x_{t-1}, \dots]$  in every period thereafter. When it is believed that the duration of the lagged effects is extremely long—for example, in the analysis of monetary policy—**infinite lag** models that have effects that gradually fade over time are quite common. But models are often constructed in which changes in  $x$  cease to have any influence after a fairly small number of periods. We shall consider these **finite lag** models first.

Marginal effects in the static classical regression model are one-time events. The response of  $y$  to a change in  $x$  is assumed to be immediate and to be complete at the end of the period of measurement. In a dynamic model, the counterpart to a marginal effect is the effect of a one-time change in  $x_t$  on the **equilibrium** of  $y_t$ . If the level of  $x_t$  has been unchanged from, say,  $\bar{x}$  for many periods prior to time  $t$ , then the equilibrium value of  $E[y_t | x_t, x_{t-1}, \dots]$  (assuming that it exists) will be

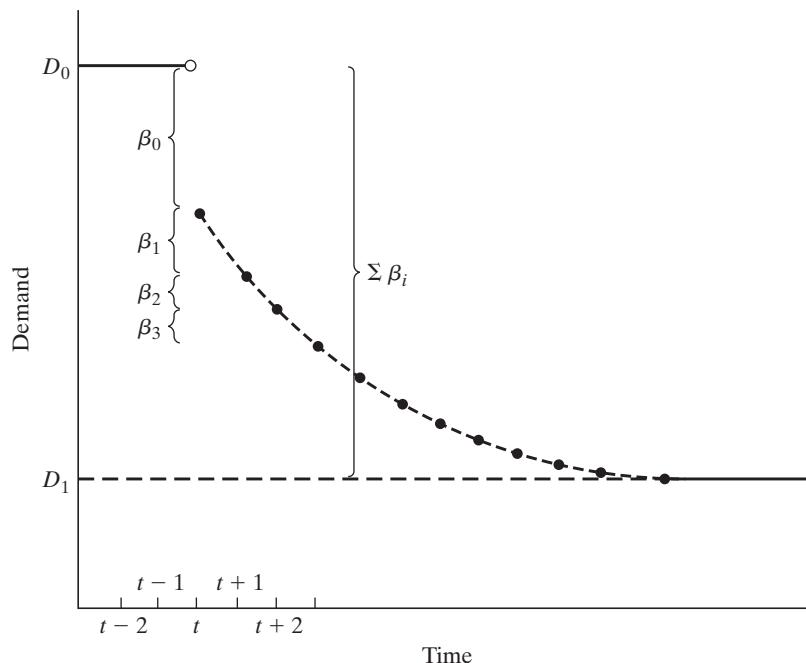
$$\bar{y} = \alpha + \sum_{i=0}^{\infty} \beta_i \bar{x} = \alpha + \bar{x} \sum_{i=0}^{\infty} \beta_i, \quad (21-3)$$

where  $\bar{x}$  is the permanent value of  $x_t$ . For this value to be finite, we require that

$$\left| \sum_{i=0}^{\infty} \beta_i \right| < \infty. \quad (21-4)$$

Consider the effect of a unit change in  $\bar{x}$  occurring in period  $s$ . To focus ideas, consider the earlier example of demand for gasoline and suppose that  $x_t$  is the unit price. Prior to the oil shock, demand had reached an equilibrium consistent with accumulated habits,

## CHAPTER 21 ♦ Models with Lagged Variables 951

**FIGURE 21.1** Lagged Adjustment.

experience with stable real prices, and the accumulated stocks of vehicles. Now suppose that the price of gasoline,  $P_g$ , rises permanently from  $p_t P_g$  to  $p_{t+1} P_g + 1$  in period  $s$ . The path to the new equilibrium might appear as shown in Figure 21.1. The short-run effect is the one that occurs in the same period as the change in  $x$ . This effect is  $\beta_0$  in the figure.

**DEFINITION 21.1 Impact Multiplier**

$\beta_0 = \text{impact multiplier} = \text{short-run multiplier}$ .

**DEFINITION 21.2 Cumulated Effect**

The accumulated effect  $\tau$  periods later of an impulse at time  $t$  is  $\beta_\tau = \sum_{i=0}^{\tau} \beta_i$ .

In Figure 21.1, we see that the total effect of a price change in period  $t$  after three periods have elapsed will be  $\beta_0 + \beta_1 + \beta_2 + \beta_3$ .

The difference between the old equilibrium  $D_0$  and the new one  $D_1$  is the sum of the individual period effects. The **long-run multiplier** is this total effect.

**952 PART V ♦ Time Series and Macroeconometrics**
**DEFINITION 21.3 Equilibrium Multiplier**

$$\beta = \sum_{i=0}^{\infty} \beta_i = \text{equilibrium multiplier} = \text{long-run multiplier}.$$

Because the lag coefficients are regression coefficients, their scale is determined by the scales of the variables in the model. As such, it is often useful to define the

$$\text{lag weights: } w_i = \frac{\beta_i}{\sum_{j=0}^{\infty} \beta_j}, \quad (21-5)$$

so that  $\sum_{i=0}^{\infty} w_i = 1$ , and to rewrite the model as

$$y_t = \alpha + \beta \sum_{i=0}^{\infty} w_i x_{t-i} + \varepsilon_t. \quad (21-6)$$

(Note the equation for the expected price in Example 21.1.) Two useful statistics, based on the lag weights, that characterize the period of adjustment to a new equilibrium are the **median lag** = smallest  $q^*$  such that  $\sum_{i=0}^{q^*} w_i \geq 0.5$  and the **mean lag** =  $\sum_{i=0}^{\infty} i w_i$ .<sup>2</sup>

### 21.2.2 THE LAG AND DIFFERENCE OPERATORS

A convenient device for manipulating lagged variables is the **lag operator**,

$$Lx_t = x_{t-1}.$$

Some basic results are  $La = a$  if  $a$  is a constant and  $L(Lx_t) = L^2x_t = x_{t-2}$ . Thus,  $L^p x_t = x_{t-p}$ ,  $L^q(L^p x_t) = L^{p+q}x_t = x_{t-p-q}$ , and  $(L^p + L^q)x_t = x_{t-p} + x_{t-q}$ . By convention,  $L^0 x_t = 1x_t = x_t$ . A related operation is the first difference,

$$\Delta x_t = x_t - x_{t-1}.$$

Obviously,  $\Delta x_t = (1 - L)x_t$  and  $x_t = x_{t-1} + \Delta x_t$ . These two operations can be usefully combined, for example, as in

$$\Delta^2 x_t = (1 - L)^2 x_t = (1 - 2L + L^2)x_t = x_t - 2x_{t-1} + x_{t-2}.$$

Note that

$$(1 - L)^2 x_t = (1 - L)(1 - L)x_t = (1 - L)(x_t - x_{t-1}) = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}).$$

The dynamic regression model can be written

$$y_t = \alpha + \sum_{i=0}^{\infty} \beta_i L^i x_t + \varepsilon_t = \alpha + B(L)x_t + \varepsilon_t,$$

---

<sup>2</sup>If the lag coefficients do not all have the same sign, then these results may not be meaningful. In some contexts, lag coefficients with different signs may be taken as an indication that there is a flaw in the specification of the model.

## CHAPTER 21 ♦ Models with Lagged Variables 953

where  $B(L)$  is a polynomial in  $L$ ,  $B(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \dots$ . A **polynomial in the lag operator** that reappears in many contexts is

$$A(L) = 1 + aL + (aL)^2 + (aL)^3 + \dots = \sum_{i=0}^{\infty} (aL)^i.$$

If  $|a| < 1$ , then

$$A(L) = \frac{1}{1 - aL}.$$

A **distributed lag** model in the form

$$y_t = \alpha + \beta \sum_{i=0}^{\infty} \gamma^i L^i x_t + \varepsilon_t$$

can be written

$$y_t = \alpha + \beta(1 - \gamma L)^{-1} x_t + \varepsilon_t,$$

if  $|\gamma| < 1$ . This form is called the **moving-average form** or **distributed lag form**. If we multiply through by  $(1 - \gamma L)$  and collect terms, then we obtain the **autoregressive form**,

$$y_t = \alpha(1 - \gamma) + \beta x_t + \gamma y_{t-1} + (1 - \gamma L)\varepsilon_t.$$

In more general terms, consider the  $p$ th order **autoregressive model**,

$$y_t = \alpha + \beta x_t + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \dots + \gamma_p y_{t-p} + \varepsilon_t,$$

which may be written

$$C(L)y_t = \alpha + \beta x_t + \varepsilon_t,$$

where

$$C(L) = (1 - \gamma_1 L - \gamma_2 L^2 - \dots - \gamma_p L^p).$$

Can this equation be “inverted” so that  $y_t$  is written as a function only of current and past values of  $x_t$  and  $\varepsilon_t$ ? By successively substituting the corresponding autoregressive equation for  $y_{t-1}$  in that for  $y_t$ , then likewise for  $y_{t-2}$  and so on, it would appear so. However, it is also clear that the resulting distributed lag form will have an infinite number of coefficients. Formally, the operation just described amounts to writing

$$y_t = [C(L)]^{-1}(\alpha + \beta x_t + \varepsilon_t) = A(L)(\alpha + \beta x_t + \varepsilon_t).$$

It will be of interest to be able to solve for the elements of  $A(L)$  (see, for example, Section 21.6.6). By this arrangement, it follows that  $C(L)A(L) = 1$  where

$$A(L) = (\alpha_0 L^0 + \alpha_1 L + \alpha_2 L^2 + \dots).$$

By collecting like powers of  $L$  in

$$(1 - \gamma_1 L - \gamma_2 L^2 - \dots - \gamma_p L^p)(\alpha_0 L^0 + \alpha_1 L + \alpha_2 L^2 + \dots) = 1,$$

## 954 PART V ♦ Time Series and Macroeconometrics

we find that a recursive solution for the  $\alpha$  coefficients is

$$\begin{aligned}
 L^0: \quad \alpha_0 &= 1 \\
 L^1: \quad \alpha_1 - \gamma_1\alpha_0 &= 0 \\
 L^2: \quad \alpha_2 - \gamma_1\alpha_1 - \gamma_2\alpha_0 &= 0 \\
 L^3: \quad \alpha_3 - \gamma_1\alpha_2 - \gamma_2\alpha_1 - \gamma_3\alpha_0 &= 0 \\
 L^4: \quad \alpha_4 - \gamma_1\alpha_3 - \gamma_2\alpha_2 - \gamma_3\alpha_1 - \gamma_4\alpha_0 &= 0 \\
 \dots \\
 L^p: \quad \alpha_p - \gamma_1\alpha_{p-1} - \gamma_2\alpha_{p-2} - \dots - \gamma_p\alpha_0 &= 0
 \end{aligned} \tag{21-7}$$

and, thereafter,

$$L^q: \quad \alpha_q - \gamma_1\alpha_{q-1} - \gamma_2\alpha_{q-2} - \dots - \gamma_p\alpha_{q-p} = 0.$$

After a set of  $p - 1$  starting values, the  $\alpha$  coefficients obey the same difference equation as  $y_t$  does in the dynamic equation. One problem remains. For the given set of values, the preceding gives no assurance that the solution for  $\alpha_q$  does not ultimately explode. The preceding equation system is not necessarily stable for all values of  $\gamma_j$  (although it certainly is for some). If the system is stable in this sense, then the polynomial  $C(L)$  is said to be **invertible**. The necessary conditions are precisely those discussed in Section 21.4.3, so we will defer completion of this discussion until then.

Finally, two useful results are

$$B(1) = \beta_0 1^0 + \beta_1 1^1 + \beta_2 1^2 + \dots = \beta = \text{long-run multiplier},$$

and

$$B'(1) = [dB(L)/dL]_{|L=1} = \sum_{i=0}^{\infty} i\beta_i.$$

It follows that  $B'(1)/B(1) = \text{mean lag}$ .

### 21.2.3 SPECIFICATION SEARCH FOR THE LAG LENGTH

Various procedures have been suggested for determining the appropriate lag length in a dynamic model such as

$$y_t = \alpha + \sum_{i=0}^p \beta_i x_{t-i} + \varepsilon_t. \tag{21-8}$$

One must be careful about a purely significance based specification search. Let us suppose that there is an appropriate, “true” value of  $p > 0$  that we seek. A **simple-to-general approach** to finding the right lag length would depart from a model with only the current value of the independent variable in the regression and add deeper lags until a simple  $t$  test suggested that the last one added is statistically insignificant. The problem with such an approach is that at any level at which the number of included lagged variables is less than  $p$ , the estimator of the coefficient vector is biased and inconsistent. [See the omitted variable formula (4-10).] The asymptotic covariance matrix is biased as well, so statistical inference on this basis is unlikely to be successful. A **general-to-simple approach** would begin from a model that contains more than  $p$  lagged values—it

## CHAPTER 21 ♦ Models with Lagged Variables 955

is assumed that although the precise value of  $p$  is unknown, the analyst can posit a maintained value that should be larger than  $p$ . Least squares or instrumental variables regression of  $y$  on a constant and  $(p + d)$  lagged values of  $x$  consistently estimates  $\theta = [\alpha, \beta_0, \beta_1, \dots, \beta_p, 0, 0, \dots]$ .

Because models with lagged values are often used for forecasting, researchers have tended to look for measures that have produced better results for assessing “out of sample” prediction properties. The adjusted  $R^2$  [see Section 3.5.1] is one possibility. Others include the Akaike (1973) information criterion,  $AIC(p)$ ,

$$AIC(p) = \ln \frac{\mathbf{e}'\mathbf{e}}{T} + \frac{2p}{T}, \quad (21-9)$$

and Schwarz's criterion,  $SC(p)$ :

$$SC(p) = AIC(p) + \left( \frac{p}{T} \right) (\ln T - 2). \quad (21-10)$$

(See Section 5.10.1.) If some maximum  $P$  is known, then  $p < P$  can be chosen to minimize  $AIC(p)$  or  $SC(p)$ .<sup>3</sup> An alternative approach, also based on a known  $P$ , is to do sequential  $F$  tests on the last  $P > p$  coefficients, stopping when the test rejects the hypothesis that the coefficients are jointly zero. Each of these approaches has its flaws and virtues. The Akaike information criterion retains a positive probability of leading to overfitting even as  $T \rightarrow \infty$ . In contrast,  $SC(p)$  has been seen to lead to underfitting in some finite-sample cases. They do avoid, however, the inference problems of sequential estimators. The sequential  $F$  tests require successive revision of the significance level to be appropriate, but they do have a statistical underpinning.<sup>4</sup>

### 21.3 SIMPLE DISTRIBUTED LAG MODELS

Before examining some very general specifications of the dynamic regression, we briefly consider an **infinite lag model**, which emerges from a simple model of expectations.

There are cases in which the distributed lag models the accumulation of information. The formation of expectations is an example. In these instances, intuition suggests that the most recent past will receive the greatest weight and that the influence of past observations will fade uniformly with the passage of time. The geometric lag model is often used for these settings. The general form of the model is

$$y_t = \alpha + \beta \sum_{i=0}^{\infty} (1 - \lambda) \lambda^i x_{t-i} + \varepsilon_t, \quad 0 < \lambda < 1, \quad (21-11)$$

$$= \alpha + \beta B(L)x_t + \varepsilon_t,$$

where

$$B(L) = (1 - \lambda)(1 + \lambda L + \lambda^2 L^2 + \lambda^3 L^3 + \dots) = \frac{1 - \lambda}{1 - \lambda L}.$$

<sup>3</sup>For further discussion and some alternative measures, see Geweke and Meese (1981), Amemiya (1985, pp. 146–147), Diebold (1998, pp. 85–91), and Judge et al. (1985, pp. 353–355).

<sup>4</sup>See Pagano and Hartley (1981) and Trivedi and Pagan (1979).

## 956 PART V ♦ Time Series and Macroeconometrics

The lag coefficients are  $\beta_i = \beta(1 - \lambda)\lambda^i$ . The model incorporates **infinite lags**, but it assigns arbitrarily small weights to the distant past. The lag weights decline geometrically;

$$w_i = (1 - \lambda)\lambda^i, \quad 0 \leq w_i < 1.$$

The **mean lag** is

$$\bar{w} = \frac{B'(1)}{B(1)} = \frac{\lambda}{1 - \lambda}.$$

The **median lag** is  $p^*$  such that  $\sum_{i=0}^{p^*-1} w_i = 0.5$ . We can solve for  $p^*$  by using the result

$$\sum_{i=0}^p \lambda^i = \frac{1 - \lambda^{p+1}}{1 - \lambda}.$$

Thus,

$$p^* = \frac{\ln 0.5}{\ln \lambda} - 1.$$

The impact multiplier is  $\beta(1 - \lambda)$ . The long-run multiplier is  $\beta \sum_{i=0}^{\infty} (1 - \lambda)\lambda^i = \beta$ . The equilibrium value of  $y_t$  would be found by fixing  $x_t$  at  $\bar{x}$  and  $\varepsilon_t$  at zero in (21-11), which produces  $\bar{y} = \alpha + \beta\bar{x}$ .

The geometric lag model can be motivated with an economic model of expectations. We begin with a regression in an expectations variable such as an expected future price based on information available at time  $t$ ,  $x_{t+1|t}^*$ , and perhaps a second regressor,  $w_t$ ,

$$y_t = \alpha + \beta x_{t+1|t}^* + \delta w_t + \varepsilon_t,$$

and a mechanism for the formation of the expectation,

$$x_{t+1|t}^* = \lambda x_{t|t-1}^* + (1 - \lambda)x_t = \lambda L x_{t+1|t}^* + (1 - \lambda)x_t. \quad (21-12)$$

The currently formed expectation is a weighted average of the expectation in the previous period and the most recent observation. The parameter  $\lambda$  is the adjustment coefficient. If  $\lambda$  equals 1, then the current datum is ignored and expectations are never revised. A value of zero characterizes a strict pragmatist who forgets the past immediately. The expectation variable can be written as

$$x_{t+1|t}^* = \frac{1 - \lambda}{1 - \lambda L} x_t = (1 - \lambda)[x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \dots]. \quad (21-13)$$

Inserting (21-13) into (21-12) produces the geometric distributed lag model,

$$y_t = \alpha + \beta(1 - \lambda)[x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \dots] + \delta w_t + \varepsilon_t.$$

The geometric lag model can be estimated by nonlinear least squares. Rewrite it as

$$y_t = \alpha + \gamma z_t(\lambda) + \delta w_t + \varepsilon_t, \quad \gamma = \beta(1 - \lambda). \quad (21-14)$$

The constructed variable  $z_t(\lambda)$  obeys the recursion  $z_t(\lambda) = x_t + \lambda z_{t-1}(\lambda)$ . For the first observation, we use  $z_1(\lambda) = x_{1|0}^* = x_1/(1 - \lambda)$ . If the sample is moderately long, then assuming that  $x_t$  was in long-run equilibrium, although it is an approximation, will not unduly affect the results. One can then scan over the range of  $\lambda$  from zero to one to locate the value that minimizes the sum of squares. Once the minimum is located, an estimate of the asymptotic covariance matrix of the estimators of  $(\alpha, \gamma, \delta, \lambda)$  can be

## CHAPTER 21 ♦ Models with Lagged Variables 957

found using (7-15) and Theorem 7.2. For the regression function  $h_t(\text{data} | \alpha, \gamma, \delta, \lambda)$ ,  $x_{t1}^0 = 1$ ,  $x_{t2}^0 = z_t(\lambda)$ , and  $x_{t3}^0 = w_t$ . The derivative with respect to  $\lambda$  can be computed by using the recursion  $d_t(\lambda) = \partial z_t(\lambda)/\partial \lambda = z_{t-1}(\lambda) + \lambda \partial z_{t-1}(\lambda)/\partial \lambda$ . If  $z_1 = x_1/(1 - \lambda)$ , then  $d_1(\lambda) = z_1/(1 - \lambda)$ . Then,  $x_{t4}^0 = d_t(\lambda)$ . Finally, we estimate  $\beta$  from the relationship  $\beta = \gamma/(1 - \lambda)$  and use the delta method to estimate the asymptotic standard error.

For purposes of estimating long- and short-run elasticities, researchers often use a different form of the geometric lag model. The **partial adjustment** model describes the *desired* level of  $y_t$ ,

$$y_t^* = \alpha + \beta x_t + \delta w_t + \varepsilon_t,$$

and an *adjustment equation*,

$$y_t - y_{t-1} = (1 - \lambda)(y_t^* - y_{t-1}).$$

If we solve the second equation for  $y_t$  and insert the first expression for  $y_t^*$ , then we obtain

$$\begin{aligned} y_t &= \alpha(1 - \lambda) + \beta(1 - \lambda)x_t + \delta(1 - \lambda)w_t + \lambda y_{t-1} + (1 - \lambda)\varepsilon_t \\ &= \alpha' + \beta'x_t + \delta'w_t + \lambda y_{t-1} + \varepsilon'_t. \end{aligned}$$

This formulation offers a number of significant practical advantages. It is intrinsically linear in the parameters (unrestricted), and its disturbance is nonautocorrelated if  $\varepsilon_t$  was to begin with. As such, the parameters of this model can be estimated consistently and efficiently by ordinary least squares. In this revised formulation, the short-run multipliers for  $x_t$  and  $w_t$  are  $\beta'$  and  $\delta'$ . The long-run effects are  $\beta = \beta'/(1 - \lambda)$  and  $\delta = \delta'/(1 - \lambda)$ . With the variables in logs, these effects are the short- and long-run elasticities.

### Example 21.2 Expectations-Augmented Phillips Curve

In Example 20.3, we estimated an expectations-augmented Phillips curve of the form

$$\Delta p_t - E[\Delta p_t | \Psi_{t-1}] = \beta[u_t - u^*] + \varepsilon_t.$$

Our model assumed a particularly simple model of expectations,  $E[\Delta p_t | \Psi_{t-1}] = \Delta p_{t-1}$ . The least squares results for this equation were

$$\begin{aligned} \Delta p_t - \Delta p_{t-1} &= 0.49189 - 0.090136 u_t + e_t \\ (0.7405) \quad (0.1257) \quad R^2 &= 0.002561, T = 202. \end{aligned}$$

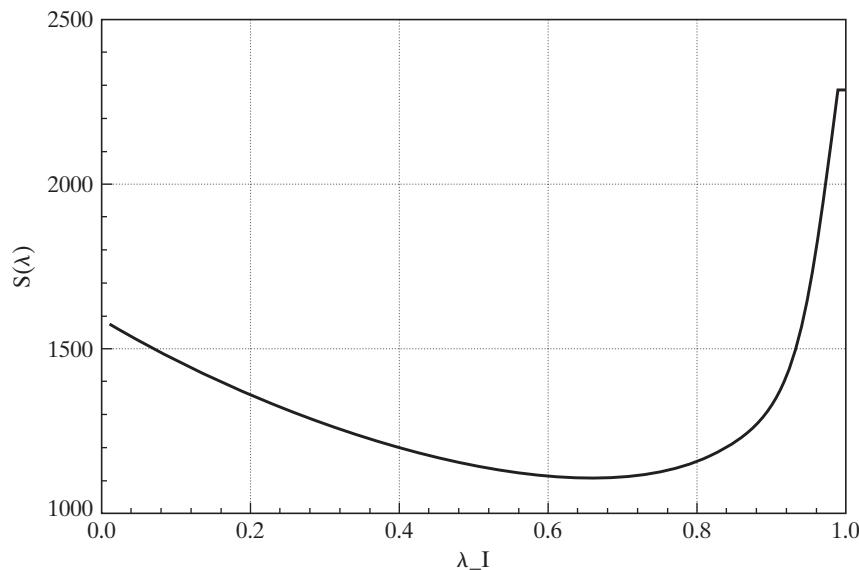
The implied estimate of the natural rate of unemployment is  $-(0.49189 / -0.090136)$  or about 5.46 percent. Suppose we allow expectations to be formulated less pragmatically with the expectations model in (21-12). For this setting, this would be

$$E[\Delta p_t | \Psi_{t-1}] = \lambda E[\Delta p_{t-1} | \Psi_{t-2}] + (1 - \lambda) \Delta p_{t-1}.$$

The strict pragmatist has  $\lambda = 0.0$ . Using the method set out earlier, we would compute this for different values of  $\lambda$ , recompute the dependent variable in the regression, and locate the value of  $\lambda$  which produces the lowest sum of squares. Figure 21.2 shows the sum of squares for the values of  $\lambda$  ranging from 0.0 to 1.0.

The minimum value of the sum of squares occurs at  $\lambda = 0.66$ . The least squares regression results are

$$\begin{aligned} \Delta p_t - \widehat{\Delta p_{t-1}} &= 1.69453 - 0.30427 u_t + e_t \\ (0.6617) \quad (0.11125) \quad T &= 202. \end{aligned}$$

**958 PART V ♦ Time Series and Macroeconometrics**


**FIGURE 21.2** Residuals Sums of Squares for Phillips Curve Estimates.

The estimated standard errors are computed using the method described earlier for the nonlinear regression. The extra variable described in the paragraph after (21-14) accounts for the estimated  $\lambda$ . The estimated asymptotic covariance matrix is then computed using  $(\mathbf{e}'\mathbf{e}/201)[\mathbf{W}'\mathbf{W}]^{-1}$  where  $w_1 = 1$ ,  $w_2 = u_t$  and  $w_3 = \partial \widehat{\Delta p_{t-1}}/\partial \lambda$ . The estimated standard error for  $\lambda$  is 0.04610. Because this is highly statistically significantly different from zero ( $t = 14.315$ ), we would reject the simple model. Finally, the implied estimate of the natural rate of unemployment is  $-(-1.69453 / 0.30427)$  or about 5.57 percent. The estimated asymptotic covariance of the slope and constant term is  $-0.0720293$ , so, using this value and the estimated standard errors given earlier and the delta method, we obtain an estimated standard error for this estimate of 0.5467. Thus, a confidence interval for the natural rate of unemployment based on these results would be (4.49 percent, 6.64 percent), which is in line with our prior expectations. There are two things to note about these results. First, because the dependent variables are different, we cannot compare the  $R^2$ 's of the models with  $\lambda = 0.00$  and  $\lambda = 0.66$ . But, the sum of squares for the two models can be compared (they are 1592.32 and 1112.89), so the second model fits far better. One of the payoffs is the much narrower confidence interval for the natural rate. The counterpart to the one given earlier when  $\lambda = 0.00$  is (1.13%, 9.79%). No doubt the model could be improved still further by expanding the equation. (This is considered in the exercises.)

**Example 21.3 Price and Income Elasticities of Demand for Gasoline**

We have extended the gasoline demand equation estimated in Examples 20.2 and 20.6 to allow for dynamic effects. Table 21.1 presents estimates of three distributed lag models for gasoline consumption. The unrestricted model allows five years of adjustment in the price and income effects. The expectations model includes the same distributed lag ( $\lambda$ ) on price and income but different long-run multipliers ( $\beta_{Pg}$  and  $\beta_I$ ). [Note, for this formulation, that the extra regressor used in computing the asymptotic covariance matrix is  $d(\lambda) = \beta_{Pg}d_{\text{price}}(\lambda) + \beta_I d_{\text{income}}(\lambda)$ .] Finally, the partial adjustment model implies lagged effects for all the variables in the model. To facilitate comparison, the constant and the first four slope coefficients in the partial adjustment model have been divided by the estimate of  $(1 - \lambda)$ . The implied long- and short-run price and income **elasticities** are shown in Table 21.2.

CHAPTER 21 ♦ Models with Lagged Variables **959****TABLE 21.1** Estimated Distributed Lag Models

<i>Coefficient</i>	<i>Unrestricted</i>	<i>Expectations</i>		<i>Partial Adjustment</i>	
		<i>Estimated</i>	<i>Derived</i>	<i>Estimated</i>	<i>Derived</i>
Constant	-28.5512	-16.1867		-4.9489	
ln <i>Pnc</i>	0.01738	-0.1050		-0.1429	
ln <i>Puc</i>	0.07602	0.02815		0.09435	
ln <i>Ppt</i>	0.04770	0.2550		0.03243	
Trend	-0.02297	0.02064		-0.004029	
ln <i>Pg</i>	-0.08282	-0.06702*	-0.06702*	-0.07627	-0.07627
ln <i>Pg</i> [−1]	-0.07152		-0.06233		-0.06116
ln <i>Pg</i> [−2]	0.03669		-0.05797		-0.04904
ln <i>Pg</i> [−3]	-0.04814		-0.05391		-0.03933
ln <i>Pg</i> [−4]	0.02958		-0.05013		-0.03153
ln <i>Pg</i> [−5]	-0.1481		-0.04663		-0.02529
ln Income	1.1074	0.04372*	0.04372*	0.3135	0.3135
ln Income[−1]	0.3776		0.04066		0.2514
ln Income[−2]	-0.01255		0.03781		0.2016
ln Income[−3]	-0.03919		0.03517		0.1616
ln Income[−4]	0.2737		0.03270		0.1296
ln Income[−5]	0.09350		0.03042		0.1039
Zt(Price)	—	-0.06702			
Zt(Income)	—	0.04372			
ln (G/Pop)[−1]	—			0.80188	
β	—	-0.9574			
γ	—	0.6245			
λ	—	0.9300		0.80188	
e'e	0.01565356		0.03911383		.01151860
T	47		52		51

\*Estimated directly

**TABLE 21.2** Estimated Elasticities

	<i>Short-Run</i>		<i>Long-Run</i>	
	<i>Price</i>	<i>Income</i>	<i>Price</i>	<i>Income</i>
Unrestricted model	-0.08282	1.1074	-0.2843	1.8004
Expectations model	-0.06702	0.04372	-0.9574	0.6246
Partial adjustment model	-0.07628	0.3135	-0.3850	1.5823

## 21.4 AUTOREGRESSIVE DISTRIBUTED LAG MODELS

Both the finite lag models and the geometric lag model impose strong, possibly incorrect restrictions on the lagged response of the dependent variable to changes in an independent variable. A very general compromise that also provides a useful platform for studying a number of interesting methodological issues is the **autoregressive distributed lag (ARDL) model**,

$$y_t = \mu + \sum_{i=1}^p \gamma_i y_{t-i} + \sum_{j=0}^r \beta_j x_{t-j} + \delta w_t + \varepsilon_t, \quad (21-15)$$

## 960 PART V ♦ Time Series and Macroeconometrics

in which  $\varepsilon_t$  is assumed to be serially uncorrelated and homoscedastic (we will relax both these assumptions in Chapter 22). We can write this more compactly as

$$C(L)y_t = \mu + B(L)x_t + \delta w_t + \varepsilon_t$$

by defining polynomials in the lag operator,

$$C(L) = 1 - \gamma_1 L - \gamma_2 L^2 - \cdots - \gamma_p L^p,$$

and

$$B(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_r L^r.$$

The model in this form is denoted ARDL( $p, r$ ) to indicate the orders of the two polynomials in  $L$ . The partial adjustment model estimated in the previous section is the special case in which  $p$  equals 1 and  $r$  equals 0. A number of other special cases are also interesting, including the familiar model of **autocorrelation** ( $p = 1, r = 1, \beta_1 = -\gamma_1 \beta_0$ ), the classical regression model ( $p = 0, r = 0$ ), and so on.

### 21.4.1 ESTIMATION OF THE ARDL MODEL

Save for the presence of the stochastic right-hand-side variables, the ARDL is a linear model with a classical disturbance. As such, ordinary least squares is the efficient estimator. The lagged dependent variable does present a complication, but we considered this in Chapter 20. Absent any obvious violations of the assumptions there, least squares continues to be the estimator of choice. Conventional testing procedures are, as before, asymptotically valid as well. Thus, for testing linear restrictions, the Wald statistic can be used, although the  $F$  statistic is generally preferable in finite samples because of its more conservative critical values.

One subtle complication in the model has attracted a large amount of attention in the recent literature. If  $C(1) = 0$ , then the model is actually inestimable. This fact is evident in the distributed lag form, which includes a term  $\mu/C(1)$ . If the equivalent condition  $\sum_i \gamma_i = 1$  holds, then the stochastic difference equation is unstable and a host of other problems arise as well. This implication suggests that one might be interested in testing this specification as a hypothesis in the context of the model. This restriction might seem to be a simple linear constraint on the alternative (unrestricted) model in (21-15). Under the null hypothesis, however, the conventional test statistics do not have the familiar distributions. The formal derivation is complicated [in the extreme, see Dickey and Fuller (1979) for an example], but intuition should suggest the reason. Under the null hypothesis, the difference equation is explosive, so our assumptions about well behaved data cannot be met. Consider a simple ARDL(1, 0) example and simplify it even further with  $B(L) = 0$ . Then,

$$y_t = \mu + \gamma y_{t-1} + \varepsilon_t.$$

If  $\gamma$  equals 1, then

$$y_t = \mu + y_{t-1} + \varepsilon_t.$$

Assuming we start the time series at time  $t = 1$ ,

$$y_t = t\mu + \sum_s \varepsilon_s = t\mu + v_t.$$

CHAPTER 21 ♦ Models with Lagged Variables **961**

The conditional mean in this **random walk with drift** model is increasing without limit, so the unconditional mean does not exist. The conditional mean of the disturbance,  $v_t$ , is zero, but its conditional variance is  $t\sigma^2$ , which shows a peculiar type of heteroscedasticity. Consider least squares estimation of  $\mu$  with  $m = (\mathbf{t}'\mathbf{y})/(\mathbf{t}'\mathbf{t})$ , where  $\mathbf{t} = [1, 2, 3, \dots, T]$ . Then  $E[m] = \mu + E[(\mathbf{t}'\mathbf{t})^{-1}(\mathbf{t}'\mathbf{v})] = \mu$ , but

$$\text{Var}[m] = \frac{\sigma^2 \sum_{t=1}^T t^3}{\left(\sum_{t=1}^T t^2\right)^2} = \frac{O(T^4)}{[O(T^3)]^2} = O\left(\frac{1}{T^2}\right).$$

So, the variance of this estimator is an order of magnitude smaller than we are used to seeing in regression models. Not only is  $m$  mean square consistent, it is also **superconsistent**. As such, without doing a formal derivation, we conclude that there is something “unusual” about this estimator and that the “usual” testing procedures whose distributions build on the distribution of  $\sqrt{T}(m - \mu)$  will not be appropriate; the variance of this normalized statistic converges to zero.

This result does not mean that the hypothesis  $\gamma = 1$  is not testable in this model. In fact, the appropriate test statistic is the conventional one that we have computed for comparable tests before. But the appropriate critical values against which to measure those statistics are quite different. We will return to this issue in our discussion of the Dickey–Fuller test in Section 23.2.4.

#### 21.4.2 COMPUTATION OF THE LAG WEIGHTS IN THE ARDL MODEL

The distributed lag form of the ARDL model is

$$\begin{aligned} y_t &= \frac{\mu}{C(L)} + \frac{B(L)}{C(L)}x_t + \frac{1}{C(L)}\delta w_t + \frac{1}{C(L)}\varepsilon_t \\ &= \frac{\mu}{1 - \gamma_1 - \dots - \gamma_p} + \sum_{j=0}^{\infty} \alpha_j x_{t-j} + \delta \sum_{l=0}^{\infty} \theta_l w_{t-l} + \sum_{l=0}^{\infty} \theta_l \varepsilon_{t-l}. \end{aligned}$$

This model provides a method of approximating a very general lag structure. In Jorgenson’s (1966) study, in which he labeled this model a **rational lag** model, he demonstrated that essentially any desired shape for the lag distribution could be produced with relatively few parameters.<sup>5</sup>

The lag coefficients on  $x_t, x_{t-1}, \dots$ , in the ARDL model are the individual terms in the ratio of polynomials that appear in the distributed lag form. We denote these as coefficients

$$\alpha_0, \alpha_1, \alpha_2, \dots = \text{the coefficient on } 1, L, L^2, \dots \text{ in } \frac{B(L)}{C(L)}. \quad (21-16)$$

A convenient way to compute these coefficients is to write (21-16) as  $A(L)C(L) = B(L)$ . Then we can just equate coefficients on the powers of  $L$ . Example 21.4 demonstrates the procedure.

---

<sup>5</sup>A long literature, highlighted by Griliches (1967), Dhrymes (1971), Nerlove (1972), Maddala (1977a), and Harvey (1990), describes estimation of models of this sort.

## 962 PART V ♦ Time Series and Macroeconometrics

The long-run effect in a rational lag model is  $\sum_{i=0}^{\infty} \alpha_i$ . This result is easy to compute because it is simply

$$\sum_{i=0}^{\infty} \alpha_i = \frac{B(1)}{C(1)}.$$

A standard error for the long-run effect can be computed using the delta method.

### 21.4.3 STABILITY OF A DYNAMIC EQUATION

In the geometric lag model, we found that a **stability** condition  $|\lambda| < 1$  was necessary for the model to be well behaved. Similarly, in the AR(1) model, the autocorrelation parameter  $\rho$  must be restricted to  $|\rho| < 1$  for the same reason. The dynamic model in (21-15) must also be restricted, but in ways that are less obvious. Consider once again the question of whether there exists an equilibrium value of  $y_t$ .

In (21-15), suppose that  $x_t$  is fixed at some value  $\bar{x}$ ,  $w_t$  is fixed at zero, and the disturbances  $\varepsilon_t$  are fixed at their expectation of zero. Would  $y_t$  converge to an equilibrium? The relevant dynamic equation is

$$y_t = \bar{\alpha} + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p},$$

where  $\bar{\alpha} = \mu + B(1)\bar{x}$ . If  $y_t$  converges to an equilibrium, then, that equilibrium is

$$\bar{y} = \frac{\mu + B(1)\bar{x}}{C(1)} = \frac{\bar{\alpha}}{C(1)}.$$

Stability of a dynamic equation hinges on the **characteristic equation** for the autoregressive part of the model. The roots of the characteristic equation,

$$C(z) = 1 - \gamma_1 z - \gamma_2 z^2 - \cdots - \gamma_p z^p = 0, \quad (21-17)$$

must be greater than one in absolute value for the model to be stable. To take a simple example, the characteristic equation for the first-order models we have examined thus far is

$$C(z) = 1 - \lambda z = 0.$$

The single root of this equation is  $z = 1/\lambda$ , which is greater than one in absolute value if  $|\lambda|$  is less than one. The roots of a more general characteristic equation are the reciprocals of the characteristic roots of the matrix

$$\mathbf{C} = \begin{bmatrix} \gamma_1 & \gamma_2 & \gamma_3 & \cdots & \gamma_{p-1} & \gamma_p \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ & & & \ddots & & \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}. \quad (21-18)$$

Because the matrix is asymmetric, its roots may include complex pairs. The reciprocal of the complex number  $a + bi$  is  $a/M - (b/M)i$ , where  $M = a^2 + b^2$  and  $i^2 = -1$ . We thus require that  $M$  be less than 1.

The case of  $z = 1$ , the unit root case, is often of special interest. If one of the roots of  $C(z) = 0$  is 1, then it follows that  $\sum_{i=1}^p \gamma_i = 1$ . This assumption would appear

CHAPTER 21 ♦ Models with Lagged Variables **963**

to be a simple hypothesis to test in the framework of the ARDL model. Instead, we find the explosive case that we examined in Section 21.4.1, so the hypothesis is more complicated than it first appears. To reiterate, under the null hypothesis that  $C(1) = 0$ , it is not possible for the standard  $F$  statistic to have a central  $F$  distribution because of the behavior of the variables in the model. We will return to this case shortly.

The **univariate autoregression**,

$$y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + \varepsilon_t,$$

can be augmented with the  $p - 1$  equations

$$y_{t-1} = y_{t-1},$$

$$y_{t-2} = y_{t-2},$$

and so on to give a **vector autoregression, VAR** (to be considered in the next section):

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{C}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t,$$

where  $\mathbf{y}_t$  has  $p$  elements,  $\boldsymbol{\varepsilon}_t = (\varepsilon_t, 0, \dots)'$ , and  $\boldsymbol{\mu} = (\mu, 0, 0, \dots)'$ . It will ultimately not be relevant to the solution, so we will let  $\varepsilon_t$  equal its expected value of zero. Now, by successive substitution, we obtain

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{C}\boldsymbol{\mu} + \mathbf{C}^2\boldsymbol{\mu} + \cdots,$$

which may or may not converge. Write  $\mathbf{C}$  in the spectral form  $\mathbf{C} = \mathbf{P}\Lambda\mathbf{Q}$ , where  $\mathbf{Q}\mathbf{P} = \mathbf{I}$  and  $\Lambda$  is a diagonal matrix of the characteristic roots. (Note that the characteristic roots in  $\Lambda$  and vectors in  $\mathbf{P}$  and  $\mathbf{Q}$  may be complex.) We then obtain

$$\mathbf{y}_t = \left[ \sum_{i=0}^{\infty} \mathbf{P}\Lambda^i \mathbf{Q} \right] \boldsymbol{\mu}. \quad (21-19)$$

If all the roots of  $\mathbf{C}$  are less than one in absolute value, then this vector will converge to the equilibrium

$$\mathbf{y}_\infty = (\mathbf{I} - \mathbf{C})^{-1} \boldsymbol{\mu}.$$

Nonexplosion of the powers of the roots of  $\mathbf{C}$  is equivalent to  $|\lambda_p| < 1$ , or  $|1/\lambda_p| > 1$ , which was our original requirement. Note finally that because  $\boldsymbol{\mu}$  is a multiple of the first column of  $\mathbf{I}_p$ , it must be the case that each element in the first column of  $(\mathbf{I} - \mathbf{C})^{-1}$  is the same. At equilibrium, therefore, we must have  $y_t = y_{t-1} = \cdots = y_\infty$ .

**Example 21.4 A Rational Lag Model**

Appendix Table F5.2 lists quarterly data on a number of macroeconomic variables including consumption and real GDP for the U.S. economy for the years 1950 to 2000, a total of 204 quarters. The model

$$c_t = \delta + \beta_0 y_t + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \gamma_1 c_{t-1} + \gamma_2 c_{t-2} + \gamma_3 c_{t-3} + \varepsilon_t$$

is estimated using the logarithms of real consumption and real GDP, denoted  $c_t$  and  $y_t$ . Ordinary least squares estimates of the parameters of the ARDL(3,3) model are

$$\begin{aligned} c_t = & 0.7233c_{t-1} + 0.3914c_{t-2} - 0.2337c_{t-3} \\ & + 0.5651y_t - 0.3909y_{t-1} - 0.2379y_{t-2} + 0.1902y_{t-3} + e_t. \end{aligned}$$

**964 PART V ♦ Time Series and Macroeconometrics**
**TABLE 21.3** Lag Coefficients in a Rational Lag Model

Lag	0	1	2	3	4	5	6	7
ARDL	0.565	0.018	-0.004	0.062	0.039	0.054	0.039	0.041
Unrestricted	0.954	-0.090	-0.063	0.100	-0.024	0.057	-0.112	0.236

(A full set of quarterly dummy variables is omitted.) The Durbin–Watson statistic is 1.78597, so remaining autocorrelation seems unlikely to be a consideration. The lag coefficients are given by the equality

$$(\alpha_0 + \alpha_1 L + \alpha_2 L^2 + \dots)(1 - \gamma_1 L - \gamma_2 L^2 - \gamma_3 L^3) = (\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3).$$

Note that  $A(L)$  is an infinite polynomial. The lag coefficients are

$$\begin{aligned} 1: \quad \alpha_0 &= \beta_0 \text{ (which will always be the case),} \\ L^1: \quad -\alpha_0\gamma_1 + \alpha_1 &= \beta_1 \text{ or } \alpha_1 = \beta_1 + \alpha_0\gamma_1, \\ L^2: \quad -\alpha_0\gamma_2 - \alpha_1\gamma_1 + \alpha_2 &= \beta_2 \text{ or } \alpha_2 = \beta_2 + \alpha_0\gamma_2 + \alpha_1\gamma_1, \\ L^3: \quad -\alpha_0\gamma_3 - \alpha_1\gamma_2 - \alpha_2\gamma_1 + \alpha_3 &= \beta_3 \text{ or } \alpha_3 = \beta_3 + \alpha_0\gamma_3 + \alpha_1\gamma_2 + \alpha_2\gamma_1, \\ L^4: \quad -\alpha_1\gamma_3 - \alpha_2\gamma_2 - \alpha_3\gamma_1 + \alpha_4 &= 0 \text{ or } \alpha_4 = \gamma_1\alpha_3 + \gamma_2\alpha_2 + \gamma_3\alpha_1, \\ L^j: \quad -\alpha_{j-3}\gamma_3 - \alpha_{j-2}\gamma_2 - \alpha_{j-1}\gamma_1 + \alpha_j &= 0 \text{ or } \alpha_j = \gamma_1\alpha_{j-1} + \gamma_2\alpha_{j-2} + \gamma_3\alpha_{j-3}, \quad j = 5, 6, \dots, \end{aligned}$$

and so on. From the fourth term onward, the series of lag coefficients follows the recursion  $\alpha_j = \gamma_1\alpha_{j-1} + \gamma_2\alpha_{j-2} + \gamma_3\alpha_{j-3}$ , which is the same as the autoregressive part of the ARDL model. The series of lag weights follows the same difference equation as the current and lagged values of  $y_t$  after  $r$  initial values, where  $r$  is the order of the DL part of the ARDL model. The three characteristic roots of the  $C$  matrix are 0.8631, -0.5949, and 0.4551. Because all are less than one, we conclude that the stochastic difference equation is stable.

The first seven lag coefficients of the estimated ARDL model are listed in Table 21.3 with the first seven coefficients in an unrestricted lag model. The coefficients from the ARDL model only vaguely resemble those from the unrestricted model, but the erratic swings of the latter are prevented by the smooth equation from the distributed lag model. The estimated long-term effects (with standard errors in parentheses) from the two models are 1.0634 (0.00791) from the ARDL model and 1.0570 (0.002135) from the unrestricted model. Surprisingly, in view of the large and highly significant estimated coefficients, the lagged effects fall off essentially to zero after the initial impact.

#### 21.4.4 FORECASTING

Consider, first, a **one-period-ahead forecast** of  $y_t$  in the  $ARDL(p, r)$  model. It will be convenient to collect the terms in  $\mu$ ,  $x_t$ ,  $w_t$ , and so on in a single term,

$$\mu_t = \mu + \sum_{j=0}^r \beta_j x_{t-j} + \delta w_t.$$

Now, the ARDL model is just

$$y_t = \mu_t + \gamma_1 y_{t-1} + \dots + \gamma_p y_{t-p} + \varepsilon_t.$$

Conditioned on the full set of information available up to time  $T$  and on forecasts of the exogenous variables, the one-period-ahead forecast of  $y_t$  would be

$$\hat{y}_{T+1|T} = \hat{\mu}_{T+1|T} + \gamma_1 y_T + \dots + \gamma_p y_{T-p+1} + \hat{\varepsilon}_{T+1|T}.$$

CHAPTER 21 ♦ Models with Lagged Variables **965**

To form a prediction interval, we will be interested in the variance of the forecast error,

$$e_{T+1|T} = \hat{y}_{T+1|T} - y_{T+1}.$$

This error will arise from three sources. First, in forecasting  $\mu_t$ , there will be two sources of error. The parameters,  $\mu$ ,  $\delta$ , and  $\beta_0, \dots, \beta_r$  will have been estimated, so  $\hat{\mu}_{T+1|T}$  will differ from  $\mu_{T+1}$  because of the sampling variation in these estimators. Second, if the exogenous variables,  $x_{T+1}$  and  $w_{T+1}$  have been forecasted, then to the extent that these forecasts are themselves imperfect, yet another source of error to the forecast will result. Finally, although we will forecast  $\varepsilon_{T+1}$  with its expectation of zero, we would not assume that the actual realization will be zero, so this step will be a third source of error. In principle, an estimate of the forecast variance,  $\text{Var}[e_{T+1|T}]$ , would account for all three sources of error. In practice, handling the second of these errors is largely intractable, while the first is merely extremely difficult. [See Harvey (1990) and Hamilton (1994, especially Section 11.7) for useful discussion. McCullough (1996) presents results that suggest that “intractable” may be too pessimistic.] For the moment, we will concentrate on the third source and return to the other issues briefly at the end of the section.

Ignoring for the moment the variation in  $\hat{\mu}_{T+1|T}$ —that is, assuming that the parameters are known and the exogenous variables are forecasted perfectly—the variance of the forecast error will be simply

$$\text{Var}[e_{T+1|T} | x_{T+1}, w_{T+1}, \mu, \beta, \delta, y_T, \dots] = \text{Var}[\varepsilon_{T+1}] = \sigma^2,$$

so at least within these assumptions, forming the forecast and computing the forecast variance are straightforward. Also, at this first step, given the data used for the forecast, the first part of the variance is also tractable. Let  $\mathbf{z}_{T+1} = [1, x_{T+1}, x_T, \dots, x_{T-p+1}, w_T, y_T, y_{T-1}, \dots, y_{T-p+1}]$ , and let  $\hat{\theta}$  denote the full estimated parameter vector. Then we would use

$$\text{Est. Var}[e_{T+1|T} | z_{T+1}] = s^2 + \mathbf{z}'_{T+1} \{\text{Est. Asy. Var}[\hat{\theta}]\} \mathbf{z}_{T+1}.$$

Now, consider forecasting further out beyond the sample period:

$$\hat{y}_{T+2|T} = \hat{\mu}_{T+2|T} + \gamma_1 \hat{y}_{T+1|T} + \dots + \gamma_p y_{T-p+2} + \hat{\varepsilon}_{T+2|T}.$$

Note that for period  $T + 1$ , the forecasted  $y_{T+1}$  is used. Making the substitution for  $\hat{y}_{T+1|T}$ , we have

$$\hat{y}_{T+2|T} = \hat{\mu}_{T+2|T} + \gamma_1 (\hat{\mu}_{T+1|T} + \gamma_1 y_T + \dots + \gamma_p y_{T-p+1} + \hat{\varepsilon}_{T+1|T}) + \dots + \gamma_p y_{T-p+2} + \hat{\varepsilon}_{T+2|T},$$

and, likewise, for subsequent periods. Our method will be simplified considerably if we use the device we constructed in the previous section. For the first forecast period, write the forecast with the previous  $p$  lagged values as

$$\begin{bmatrix} \hat{y}_{T+1|T} \\ y_T \\ y_{T-1} \\ \vdots \end{bmatrix} = \begin{bmatrix} \hat{\mu}_{T+1|T} \\ 0 \\ 0 \\ \vdots \end{bmatrix} + \begin{bmatrix} \gamma_1 & \gamma_2 & \dots & \gamma_p \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} y_T \\ y_{T-1} \\ y_{T-2} \\ \vdots \end{bmatrix} + \begin{bmatrix} \hat{\varepsilon}_{T+1|T} \\ 0 \\ 0 \\ \vdots \end{bmatrix}.$$

The coefficient matrix on the right-hand side is  $\mathbf{C}$ , which we defined in (21-18). To maintain the thread of the discussion, we will continue to use the notation  $\hat{\mu}_{T+1|T}$  for the forecast of the deterministic part of the model, although for the present, we are

## 966 PART V ♦ Time Series and Macroeconometrics

assuming that this value, as well as  $\mathbf{C}$ , is known with certainty. With this modification, then, our forecast is the top element of the vector of forecasts,

$$\hat{\mathbf{y}}_{T+1|T} = \hat{\boldsymbol{\mu}}_{T+1|T} + \mathbf{C}\mathbf{y}_T + \hat{\boldsymbol{\epsilon}}_{T+1|T}.$$

We are assuming that everything on the right-hand side is known except the period  $T + 1$  disturbance, so the covariance matrix for this  $p + 1$  vector is

$$E[(\hat{\mathbf{y}}_{T+1|T} - \mathbf{y}_{T+1})(\hat{\mathbf{y}}_{T+1|T} - \mathbf{y}_{T+1})'] = \begin{bmatrix} \sigma^2 & 0 & \cdots \\ 0 & 0 & \vdots \\ \vdots & \cdots & \ddots \end{bmatrix},$$

and the forecast variance for  $\hat{\mathbf{y}}_{T+1|T}$  is just the upper left element,  $\sigma^2$ .

Now, extend this notation to forecasting out to periods  $T + 2$ ,  $T + 3$ , and so on:

$$\begin{aligned} \hat{\mathbf{y}}_{T+2|T} &= \hat{\boldsymbol{\mu}}_{T+2|T} + \mathbf{C}\hat{\mathbf{y}}_{T+1|T} + \hat{\boldsymbol{\epsilon}}_{T+2|T} \\ &= \hat{\boldsymbol{\mu}}_{T+2|T} + \mathbf{C}\hat{\boldsymbol{\mu}}_{T+1|T} + \mathbf{C}^2\mathbf{y}_T + \hat{\boldsymbol{\epsilon}}_{T+2|T} + \mathbf{C}\hat{\boldsymbol{\epsilon}}_{T+1|T}. \end{aligned}$$

Once again, the only unknowns are the disturbances, so the forecast variance for this two-period-ahead forecasted vector is

$$\text{Var}[\hat{\boldsymbol{\epsilon}}_{T+2|T} + \mathbf{C}\hat{\boldsymbol{\epsilon}}_{T+1|T}] = \begin{bmatrix} \sigma^2 & 0 & \cdots \\ 0 & 0 & \vdots \\ \vdots & \cdots & \ddots \end{bmatrix} + \mathbf{C} \begin{bmatrix} \sigma^2 & 0 & \cdots \\ 0 & 0 & \vdots \\ \vdots & \cdots & \ddots \end{bmatrix} \mathbf{C}'.$$

Thus, the forecast variance for the two-step-ahead forecast is  $\sigma^2[1 + \Psi(1)_{11}]$ , where  $\Psi(1)_{11}$  is the  $(1, 1)$  element of  $\Psi(1) = \mathbf{C}\mathbf{j}'\mathbf{C}'$ , where  $\mathbf{j}' = [\sigma, 0, \dots, 0]$ . By extending this device to a forecast  $F$  periods beyond the sample period, we obtain

$$\hat{\mathbf{y}}_{T+F|T} = \sum_{f=1}^F \mathbf{C}^{f-1} \hat{\boldsymbol{\mu}}_{T+F-(f-1)|T} + \mathbf{C}^F \mathbf{y}_T + \sum_{f=1}^F \mathbf{C}^{f-1} \hat{\boldsymbol{\epsilon}}_{T+F-(f-1)|T}. \quad (21-20)$$

This equation shows how to compute the forecasts, which is reasonably simple. We also obtain our expression for the conditional forecast variance,

$$\text{Conditional Var}[\hat{\mathbf{y}}_{T+F|T}] = \sigma^2[1 + \Psi(1)_{11} + \Psi(2)_{11} + \dots + \Psi(F-1)_{11}], \quad (21-21)$$

where  $\Psi(i) = \mathbf{C}^i \mathbf{j}' \mathbf{C}^i$ .

The general form of the  $F$ -period-ahead forecast shows how the forecasts will behave as the forecast period extends further out beyond the sample period. If the equation is stable—that is, if all roots of the matrix  $\mathbf{C}$  are less than one in absolute value—then  $\mathbf{C}^F$  will converge to zero, and because the forecasted disturbances are zero, the forecast will be dominated by the sum in the first term. If we suppose, in addition, that the forecasts of the exogenous variables are just the period  $T + 1$  forecasted values and not revised, then, as we found at the end of the previous section, the forecast will ultimately converge to

$$\lim_{F \rightarrow \infty} \hat{\mathbf{y}}_{T+F|T} | \hat{\boldsymbol{\mu}}_{T+1|T} = [\mathbf{I} - \mathbf{C}]^{-1} \hat{\boldsymbol{\mu}}_{T+1|T}.$$

## CHAPTER 21 ♦ Models with Lagged Variables 967

To account fully for all sources of variation in the forecasts, we would have to revise the forecast variance to include the variation in the forecasts of the exogenous variables and the variation in the parameter estimates. As noted, the first of these is likely to be intractable. For the second, this revision will be extremely difficult, the more so when we also account for the matrix  $\mathbf{C}$ , as well as the vector  $\boldsymbol{\mu}$ , being built up from the estimated parameters. The level of difficulty in this case falls from impossible to merely extremely difficult. In principle, what is required is

$$\begin{aligned} \text{Est. Conditional Var}[\hat{y}_{T+F|T}] &= \sigma^2[1 + \Psi(1)_{11} + \Psi(2)_{11} + \cdots + \Psi(F-1)_{11}] \\ &\quad + \mathbf{g}' \text{Est. Asy. Var}[\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}] \mathbf{g}, \end{aligned}$$

where

$$\mathbf{g} = \frac{\partial \hat{y}_{T+F}}{\partial [\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}]}.$$

[See Hamilton (1994, Appendix to Chapter 11) for formal derivation.]

One possibility is to use the bootstrap method. For this application, bootstrapping would involve sampling new sets of disturbances from the estimated distribution of  $\varepsilon_t$ , and then repeatedly rebuilding the within-sample time series of observations on  $y_t$  by using

$$\hat{y}_t = \hat{\mu}_t + \gamma_1 y_{t-1} + \cdots + \gamma_p y_{t-p} + e_{bt}(m),$$

where  $e_{bt}(m)$  is the estimated “bootstrapped” disturbance in period  $t$  during replication  $m$ . The process is repeated  $M$  times, with new parameter estimates and a new forecast generated in each replication. The variance of these forecasts produces the estimated forecast variance.<sup>6</sup>

## 21.5 METHODOLOGICAL ISSUES IN THE ANALYSIS OF DYNAMIC MODELS

### 21.5.1 AN ERROR CORRECTION MODEL

Consider the ARDL(1, 1) model, which has become a workhorse of the modern literature on time-series analysis. By defining the first differences  $\Delta y_t = y_t - y_{t-1}$  and  $\Delta x_t = x_t - x_{t-1}$  we can rearrange

$$y_t = \mu + \gamma_1 y_{t-1} + \boldsymbol{\beta}_0 \mathbf{x}_t + \boldsymbol{\beta}_1 x_{t-1} + \varepsilon_t$$

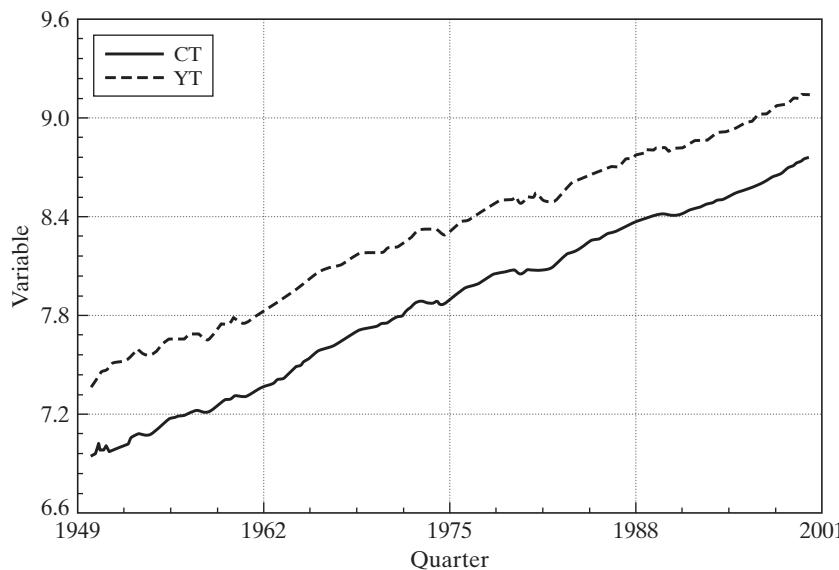
to obtain

$$\Delta y_t = \mu + \beta_0 \Delta x_t + (\gamma_1 - 1)(y_{t-1} - \theta x_{t-1}) + \varepsilon_t, \quad (21-22)$$

where  $\theta = -(\beta_0 + \beta_1)/(\gamma_1 - 1)$ . This form of the model is in the **error correction** form. In this form, we have an **equilibrium relationship**,  $\Delta y_t = \mu + \beta_0 \Delta x_t + \varepsilon_t$ , and the **equilibrium error**,  $(\gamma_1 - 1)(y_{t-1} - \theta x_{t-1})$ , which account for the deviation of the pair of variables from that equilibrium. The model states that the change in  $y_t$  from the previous period consists of the change associated with movement with  $x_t$  along the

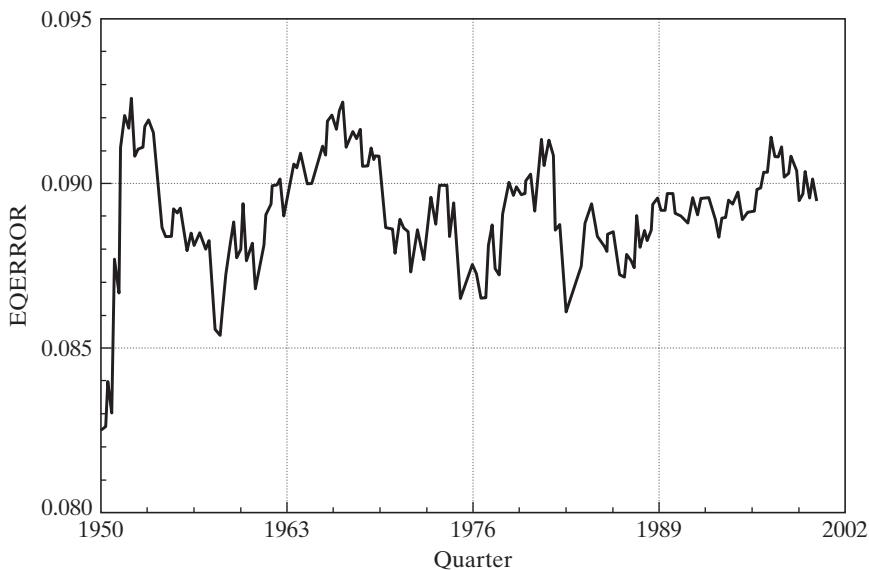
---

<sup>6</sup>Bernard and Veall (1987) give an application of this technique. See, also, McCullough (1996).

**968 PART V ♦ Time Series and Macroeconometrics**

**FIGURE 21.3** Consumption and Income Data.

long-run equilibrium path plus a part ( $\gamma_1 - 1$ ) of the deviation ( $y_{t-1} - \theta x_{t-1}$ ) from the equilibrium. With a model in logs, this relationship would be in proportional terms.

It is useful at this juncture to jump ahead a bit—we will return to this topic in some detail in Chapter 23—and explore why the error correction form might be such a useful formulation of this simple model. Consider the logged consumption and income data plotted in Figure 21.3. It is obvious on inspection of the figure that a simple regression of the log of consumption on the log of income would suggest a highly significant relationship; in fact, the simple linear regression produces a slope of 1.0567 with a  $t$  ratio of 440.5 (!) and an  $R^2$  of 0.99896. The disturbing result of a line of literature in econometrics that begins with Granger and Newbold (1974) and continues to the present is that this seemingly obvious and powerful relationship might be entirely spurious. Equally obvious from the figure is that both  $c_t$  and  $y_t$  are trending variables. If, in fact, both variables unconditionally were random walks with drift of the sort that we met at the end of Section 21.4.1—that is,  $c_t = t\mu_c + v_t$  and likewise for  $y_t$ —then we would almost certainly observe a figure such as 21.3 and compelling regression results such as those, *even if there were no relationship at all*. In addition, there is ample evidence in the recent literature that low-frequency (infrequently observed, aggregated over long periods) flow variables such as consumption and output are, indeed, often well described as random walks. In such data, the ARDL(1, 1) model might appear to be entirely appropriate even if it is not. So, how is one to distinguish between the spurious regression and a genuine relationship as shown in the ARDL(1, 1)? The first difference of consumption produces  $\Delta c_t = \mu_c + v_t - v_{t-1}$ . If the random walk proposition is indeed correct, then the spurious appearance of regression will not survive the first differencing, whereas if there is a relationship between  $c_t$  and  $y_t$ , then it will be preserved in the error correction model. We will return to this issue in Chapter 23, when we examine the issue of integration and cointegration of economic variables.



**FIGURE 21.4** Consumption–Income Equilibrium Errors.

**Example 21.5 An Error Correction Model for Consumption**

The error correction model is a nonlinear regression model, although in fact it is intrinsically linear and can be deduced simply from the unrestricted form directly above it. Because the parameter  $\theta$  is actually of some interest, it might be more convenient to use nonlinear least squares and fit the second form directly. (The model is intrinsically linear, so the nonlinear least squares estimates will be identical to the derived linear least squares estimates.) The logs of consumption and income data in Appendix Table F5.2 are plotted in Figure 21.3. Not surprisingly, the two variables are drifting upward together.

The estimated error correction model, with estimated standard errors in parentheses, is

$$c_t - c_{t-1} = -0.08533 + (0.90458 - 1)[c_{t-1} - 1.06034y_{t-1}] + 0.58421(y_t - y_{t-1}).$$

(0.02899)	(0.03029)	(0.01052)	(0.05090)
-----------	-----------	-----------	-----------

The estimated equilibrium errors are shown in Figure 21.4. Note that they are all positive, but that in each period, the adjustment is in the opposite direction. Thus (according to this model), when consumption is below its equilibrium value, the adjustment is upward, as might be expected.

### 21.5.2 AUTOCORRELATION

The disturbance in the error correction model is assumed to be nonautocorrelated. As we saw in Chapter 20, autocorrelation in a model can be induced by misspecification. An orthodox view of the modeling process might state, in fact, that this misspecification is the *only* source of autocorrelation. Although admittedly a bit optimistic in its implication, this misspecification does raise an interesting methodological question. Consider once again the simplest model of autocorrelation from Chapter 20 (with a small change in notation to make it consistent with the present discussion),

$$y_t = \beta x_t + v_t, \quad v_t = \rho v_{t-1} + \varepsilon_t, \tag{21-23}$$

## 970 PART V ♦ Time Series and Macroeconometrics

where  $\varepsilon_t$  is nonautocorrelated. As we found earlier, this model can be written as

$$y_t - \rho y_{t-1} = \beta(x_t - \rho x_{t-1}) + \varepsilon_t, \quad (21-24)$$

or

$$y_t = \rho y_{t-1} + \beta x_t - \beta \rho x_{t-1} + \varepsilon_t. \quad (21-25)$$

This model is an ARDL(1, 1) model in which  $\beta_1 = -\gamma_1 \beta_0$ . Thus, we can view (21-25) as a restricted version of

$$y_t = \gamma_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \varepsilon_t. \quad (21-26)$$

The crucial point here is that the (nonlinear) restriction on (21-26) is testable, so there is no compelling reason to proceed to (21-23) first without establishing that the restriction is in fact consistent with the data. The upshot is that the AR(1) disturbance model, as a general proposition, is a testable restriction on a simpler, linear model, not necessarily a structure unto itself.

Now, let us take this argument to its logical conclusion. The AR( $p$ ) disturbance model,

$$v_t = \rho_1 v_{t-1} + \cdots + \rho_p v_{t-p} + \varepsilon_t,$$

or  $R(L)v_t = \varepsilon_t$ , can be written in its moving average form as

$$v_t = \frac{\varepsilon_t}{R(L)}.$$

[Recall, in the AR(1) model, that  $\varepsilon_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \cdots$ .] The regression model with this AR( $p$ ) disturbance is, therefore,

$$y_t = \beta x_t + \frac{\varepsilon_t}{R(L)}.$$

But consider instead the ARDL( $p, p$ ) model

$$C(L)y_t = \beta B(L)x_t + \varepsilon_t.$$

These coefficients are the same model if  $B(L) = C(L)$ . The implication is that *any model with an AR( $p$ ) disturbance can be interpreted as a nonlinearly restricted version of an ARDL( $p, p$ ) model*.

The preceding discussion is a rather orthodox view of autocorrelation. It is predicated on the AR( $p$ ) model. Researchers have found that a more involved model for the process generating  $\varepsilon_t$  is sometimes called for. If the time-series structure of  $\varepsilon_t$  is not autoregressive, much of the preceding analysis will become intractable. As such, there remains room for disagreement with the strong conclusions. We will turn to models whose disturbances are mixtures of autoregressive and moving-average terms, which would be beyond the reach of this apparatus, in Chapter 22.

### 21.5.3 SPECIFICATION ANALYSIS

The usual explanation of autocorrelation is serial correlation in omitted variables. The preceding discussion and our results in Chapter 20 suggest another candidate: misspecification of what would otherwise be an unrestricted ARDL model. Thus, upon finding

CHAPTER 21 ♦ Models with Lagged Variables **971**

evidence of autocorrelation on the basis of a Durbin–Watson statistic or an LM statistic, we might find that relaxing the nonlinear restrictions on the ARDL model is a preferable next step to “correcting” for the autocorrelation by imposing the restrictions and refitting the model by FGLS. Because an ARDL( $p, r$ ) model with AR disturbances, even with  $p = 0$ , is implicitly an ARDL( $p + d, r + d$ ) model, where  $d$  is usually one, the approach suggested is just to add additional lags of the dependent variable to the model. Thus, one might even ask why we would ever use the familiar FGLS procedures. [See, e.g., Mizon (1995).] The payoff is that the restrictions imposed by the FGLS procedure produce a more efficient estimator than other methods. If the restrictions are in fact appropriate, then not imposing them amounts to not using information.

A related question now arises, apart from the issue of autocorrelation. In the context of the ARDL model, how should one do the specification search? (This question is not specific to the ARDL or even to the time-series setting.) Is it better to start with a small model and expand it until conventional fit measures indicate that additional variables are no longer improving the model, or is it better to start with a large model and pare away variables that conventional statistics suggest are superfluous? The first strategy, going from a *simple model to a general model*, is likely to be problematic, because the statistics computed for the narrower model are biased and inconsistent if the hypothesis is incorrect. Consider, for example, an LM test for autocorrelation in a model from which important variables have been omitted. The results are biased in favor of a finding of autocorrelation. The alternative approach is to proceed from a *general model to a simple one*. Thus, one might overfit the model and then subject it to whatever battery of tests are appropriate to produce the correct specification at the end of the procedure. In this instance, the estimates and test statistics computed from the overfit model, although inefficient, are not generally systematically biased. (We have encountered this issue at several points.)

The latter approach is common in modern analysis, but some words of caution are needed. The procedure routinely leads to overfitting the model. A typical time-series analysis might involve specifying a model with deep lags on all the variables and then paring away the model as conventional statistics indicate. The danger is that the resulting model might have an autoregressive structure with peculiar holes in it that would be hard to justify with any theory. Thus, a model for quarterly data that includes lags of 2, 3, 6, and 9 on the dependent variable would look suspiciously like the end result of a computer-driven fishing trip and, moreover, might not survive even moderate changes in the estimation sample. [As Hendry (1995) notes, a model in which the largest and most significant lag coefficient occurs at the last lag is surely misspecified.]

## 21.6 VECTOR AUTOREGRESSIONS

The preceding discussions can be extended to sets of variables. The resulting autoregressive model is

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Gamma}_1 \mathbf{y}_{t-1} + \cdots + \boldsymbol{\Gamma}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t, \quad (21-27)$$

where  $\boldsymbol{\varepsilon}_t$  is a vector of nonautocorrelated disturbances (innovations) with zero means and contemporaneous covariance matrix  $E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t'] = \boldsymbol{\Omega}$ . This equation system is a vector

## 972 PART V ♦ Time Series and Macroeconometrics

**autoregression**, or VAR. Equation (21-27) may also be written as

$$\boldsymbol{\Gamma}(L)\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t$$

where  $\boldsymbol{\Gamma}(L)$  is a matrix of polynomials in the lag operator. The individual equations are

$$y_{mt} = \mu_m + \sum_{j=1}^p (\boldsymbol{\Gamma}_j)_{m1} y_{1,t-j} + \sum_{j=1}^p (\boldsymbol{\Gamma}_j)_{m2} y_{2,t-j} + \cdots + \sum_{j=1}^p (\boldsymbol{\Gamma}_j)_{mM} y_{M,t-j} + \varepsilon_{mt},$$

where  $(\boldsymbol{\Gamma}_j)_{ml}$  indicates the  $(m, l)$  element of  $\boldsymbol{\Gamma}_j$ .

VARs have been used primarily in macroeconomics. Early in their development, it was argued by some authors [e.g., Sims (1980), Litterman (1979, 1986)] that VARs could forecast better than the sort of structural equation models discussed in Chapter 8. One could argue that as long as  $\boldsymbol{\mu}$  includes the current observations on the (truly) relevant exogenous variables, the VAR is simply an overfit reduced form of some simultaneous equations model. [See Hamilton (1994, pp. 326–327).] The overfitting results from the possible inclusion of more lags than would be appropriate in the original model. (See Example 21.7 for a detailed discussion of one such model.) On the other hand, one of the virtues of the VAR is that it obviates a decision as to what contemporaneous variables are exogenous; it has only lagged (predetermined) variables on the right-hand side, and all variables are endogenous.

The motivation behind VARs in macroeconomics runs deeper than the statistical issues.<sup>7</sup> The large structural equations models of the 1950s and 1960s were built on a theoretical foundation that has not proved satisfactory. That the forecasting performance of VARs surpassed that of large structural models—some of the later counterparts to Klein’s Model I ran to hundreds of equations—signaled to researchers a more fundamental problem with the underlying methodology. The Keynesian style systems of equations describe a structural model of decisions (consumption, investment) that seem loosely to mimic individual behavior; see Keynes’s formulation of the consumption function in Example 1.1 that is, perhaps, the canonical example. In the end, however, these decision rules are fundamentally ad hoc, and there is little basis on which to assume that they would aggregate to the macroeconomic level anyway. On a more practical level, the high inflation and high unemployment experienced in the 1970s were very badly predicted by the Keynesian paradigm. From the point of view of the underlying paradigm, the most troubling criticism of the structural modeling approach comes in the form of “the Lucas critique” (1976), in which the author argued that the *parameters* of the “decision rules” embodied in the systems of structural equations would not remain stable when economic policies changed, even if the rules themselves were appropriate. Thus, the paradigm underlying the systems of equations approach to macroeconomic modeling is arguably fundamentally flawed. More recent research has reformulated the basic equations of macroeconomic models in terms of a microeconomic optimization foundation and has, at the same time, been much less ambitious in specifying the interrelationships among economic variables.

The preceding arguments have drawn researchers to less structured equation systems for forecasting. Thus, it is not just the form of the equations that has changed. The

---

<sup>7</sup>An extremely readable, nontechnical discussion of the paradigm shift in macroeconomic forecasting is given in Diebold (2003). See also Stock and Watson (2001).

## CHAPTER 21 ♦ Models with Lagged Variables 973

variables in the equations have changed as well; the VAR is not just the reduced form of some structural model. For purposes of analyzing and forecasting macroeconomic activity and tracing the effects of policy changes and external stimuli on the economy, researchers have found that simple, small-scale VARs without a possibly flawed theoretical foundation have proved as good as or better than large-scale structural equation systems. In addition to forecasting, VARs have been used for two primary functions: testing Granger causality and studying the effects of policy through impulse response characteristics.

### 21.6.1 MODEL FORMS

To simplify things for the present, we note that the  $p$ th order VAR can be written as a first-order VAR as follows:

$$\begin{pmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-p+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix} + \begin{bmatrix} \mathbf{\Gamma}_1 & \mathbf{\Gamma}_2 & \cdots & \mathbf{\Gamma}_p \\ \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{pmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \vdots \\ \mathbf{y}_{t-p} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}.$$

[See, e.g., (21-18).] This means that we do not lose any generality in casting the treatment in terms of a first-order model

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{\Gamma}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t.$$



In Example 15.10, we examined Dahlberg and Johansson's model for municipal finances in Sweden, in which  $\mathbf{y}_t = [\Delta S_t, \Delta R_t, \Delta G_t]'$ , where  $S_t$  is spending,  $R_t$  is receipts,  $G_t$  is grants from the central government, and  $p = 3$ . We will continue that application in Example 20.7.

In principle, the VAR model is a seemingly unrelated regressions model—indeed, a particularly simple one because each equation has the same set of regressors. This is the traditional form of the model as originally proposed, for example, by Sims (1980). The VAR may also be viewed as the reduced form of a simultaneous equations model; the corresponding structure would then be

$$\boldsymbol{\Theta}\mathbf{y}_t = \boldsymbol{\alpha} + \boldsymbol{\Psi}\mathbf{y}_{t-1} + \boldsymbol{\omega}_t,$$

where  $\boldsymbol{\Theta}$  is a nonsingular matrix and  $\text{Var}[\boldsymbol{\omega}_t] = \Sigma$ . In one of Cecchetti and Rich's (2001) formulations, for example,  $\mathbf{y}_t = [\Delta y_t, \Delta \pi_t]'$  where  $y_t$  is the log of aggregate real output,  $\pi_t$  is the inflation rate from time  $t - 1$  to time  $t$ ,  $\boldsymbol{\Theta} = \begin{bmatrix} 1 & -\theta_{12} \\ -\theta_{21} & 1 \end{bmatrix}$ , and  $p = 8$ .

(We will examine their model in Section 21.6.8.) In this form, we have a conventional simultaneous equations model, which we analyzed in detail in Chapter 13. As we saw, for such a model to be identified—that is, estimable—certain restrictions must be placed on the structural coefficients. The reason for this is that ultimately, only the original VAR form, now the reduced form, is estimated from the data; the structural parameters must be deduced from these coefficients. In this model, to deduce these structural parameters, they must be extracted from the reduced form parameters,  $\mathbf{\Gamma} = \boldsymbol{\Theta}^{-1}\boldsymbol{\Psi}$ ,  $\boldsymbol{\mu} = \boldsymbol{\Theta}^{-1}\boldsymbol{\alpha}$ , and  $\boldsymbol{\Omega} = \boldsymbol{\Theta}^{-1}\sum\boldsymbol{\Theta}^{-1}'$ . In Cecchetti and Rich's application, certain restrictions were placed on the lag coefficients in order to secure identification.

## 974 PART V ♦ Time Series and Macroeconometrics

### 21.6.2 ESTIMATION

In the form of (21-27)—that is, without autocorrelation of the disturbances—VARs are particularly simple to estimate. Although the equation system can be exceedingly large, it is, in fact, a seemingly unrelated regressions model with identical regressors. As such, the equations should be estimated separately by ordinary least squares. (See Section 10.2.2 for discussion of SUR systems with identical regressors.) The disturbance covariance matrix can then be estimated with average sums of squares or cross-products of the least squares residuals. If the disturbances are normally distributed, then these least squares estimators are also maximum likelihood. If not, then OLS remains an efficient GMM estimator. The extension to instrumental variables and GMM is a bit more complicated, as the model now contains multiple equations (see Section 13.6.3), but since the equations are all linear, the necessary extensions are at least relatively straightforward. GMM estimation of the VAR system is a special case of the model discussed in Section 13.6.3. (We will examine an application in Example 21.7.)

The proliferation of parameters in VARs has been cited as a major disadvantage of their use. Consider, for example, a VAR involving five variables and three lags. Each  $\Gamma$  has 25 unconstrained elements, and there are three of them, for a total of 75 free parameters, plus any others in  $\mu$ , plus  $5(6)/2 = 15$  free parameters in  $\Omega$ . On the other hand, each single equation has only 25 parameters, and at least given sufficient degrees of freedom—there's the rub—a linear regression with 25 parameters is simple work. Moreover, applications rarely involve even as many as four variables, so the model-size issue may well be exaggerated.

### 21.6.3 TESTING PROCEDURES

Formal testing in the VAR setting usually centers either on determining the appropriate lag length (a specification search) or on whether certain blocks of zeros in the coefficient matrices are zero (a simple linear restriction on the collection of slope parameters). Both types of hypotheses may be treated as sets of linear restrictions on the elements in  $\gamma = \text{vec}[\mu, \Gamma_1, \Gamma_2, \dots, \Gamma_p]$ .

We begin by assuming that the disturbances have a joint normal distribution. Let  $\mathbf{W}$  be the  $M \times M$  residual covariance matrix based on a restricted model, and let  $\mathbf{W}^*$  be its counterpart when the model is unrestricted. Then the likelihood ratio statistic,

$$\lambda = T(\ln|\mathbf{W}| - \ln|\mathbf{W}^*|),$$

can be used to test the hypothesis. The statistic would have a limiting chi-squared distribution with degrees of freedom equal to the number of restrictions. In principle, one might base a specification search for the right lag length on this calculation. The procedure would be to test down from, say, lag  $q$  to lag  $p$ . The *general-to-simple* principle discussed in Section 21.5.3 would be to set the maximum lag length and test down from it until deletion of the last set of lags leads to a significant loss of fit. At each step at which the alternative lag model has excess terms, the estimators of the superfluous coefficient matrices would have probability limits of zero and the likelihood function would (again, asymptotically) resemble that of the model with the correct number of lags. Formally, suppose the appropriate lag length is  $p$  but the model is fit with  $q \geq p + 1$  lagged terms.

## CHAPTER 21 ♦ Models with Lagged Variables 975

Then, under the null hypothesis,

$$\lambda_q = T[\ln|\mathbf{W}(\boldsymbol{\mu}, \boldsymbol{\Gamma}_1, \dots, \boldsymbol{\Gamma}_{q-1})| - \ln|\mathbf{W}^*(\boldsymbol{\mu}, \boldsymbol{\Gamma}_1, \dots, \boldsymbol{\Gamma}_q)|] \xrightarrow{d} \chi^2[M^2].$$

The same approach would be used to test other restrictions. Thus, the Granger causality test noted in Section 21.6.5 would fit the model with and without certain blocks of zeros in the coefficient matrices, then refer the value of  $\lambda$  once again to the chi-squared distribution.

For specification searches for the right lag, the suggested procedure may be less effective than one based on the information criteria suggested for other linear models (see Section 5.10.1). Lutkepohl (2005, pp. 128–135) suggests an alternative approach based on the minimizing functions of the information criteria we have considered earlier;

$$\lambda^* = \ln(|\mathbf{W}|) + (pM^2 + M)\mathbf{IC}(T)/T,$$

where  $T$  is the sample size,  $p$  is the number of lags,  $M$  is the number of equations, and  $\mathbf{IC}(T) = 2$  for the Akaike information criterion and  $\ln T$  for the Schwarz (Bayesian) information criterion. We should note that this is not a test statistic; it is a diagnostic tool that we are using to conduct a specification search. Also, as in all such cases, the testing procedure should be from a larger model to a smaller one to avoid the misspecification problems induced by a lag length that is smaller than the appropriate one.

The preceding has relied heavily on the normality assumption. Because most recent applications of these techniques have either treated the least squares estimators as robust (distribution-free) estimators, or used GMM (as we did in Chapter 13), it is necessary to consider a different approach that does not depend on normality. An alternative approach that should be robust to variations in the underlying distributions is the Wald statistic. [See Lutkepohl (2005, pp. 93–95).] The full set of coefficients in the model may be arrayed in a single coefficient vector,  $\boldsymbol{\gamma}$ . Let  $\mathbf{c}$  be the sample estimator of  $\boldsymbol{\gamma}$  and let  $\mathbf{V}$  denote the estimated asymptotic covariance matrix. Then, the hypothesis in question (lag length, or other linear restriction) can be cast in the form  $\mathbf{R}\boldsymbol{\gamma} - \mathbf{q} = \mathbf{0}$ . The Wald statistic for testing the null hypothesis is

$$W = (\mathbf{R}\mathbf{c} - \mathbf{q})'[\mathbf{R}\mathbf{V}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{c} - \mathbf{q}).$$

Under the null hypothesis, this statistic has a limiting chi-squared distribution with degrees of freedom equal to  $J$ , the number of restrictions (rows in  $\mathbf{R}$ ). For the specification search for the appropriate lag length (or the Granger causality test discussed in the next section), the null hypothesis will be that a certain subvector of  $\boldsymbol{\gamma}$ , say  $\boldsymbol{\gamma}_0$ , equals zero. In this case, the statistic will be

$$W_0 = \mathbf{c}'_0 \mathbf{V}_{00}^{-1} \mathbf{c}_0,$$

where  $\mathbf{V}_{00}$  denotes the corresponding submatrix of  $\mathbf{V}$ .

Because time-series data sets are often only moderately long, use of the limiting distribution for the test statistic may be a bit optimistic. Also, the Wald statistic does not account for the fact that the asymptotic covariance matrix is estimated using a finite sample. In our analysis of the classical linear regression model, we accommodated these considerations by using the  $F$  distribution instead of the limiting chi-squared. (See Section 5.5.) The adjustment made was to refer  $W/J$  to the  $F[J, T - K]$  distribution. This produces a more conservative test—the corresponding critical values of  $JF$  converge

## 976 PART V ♦ Time Series and Macroeconometrics

to those of the chi-squared *from above*. A remaining complication is to decide what degrees of freedom to use for the denominator. It might seem natural to use  $MT$  minus the number of parameters, which would be correct if the restrictions are imposed on all equations simultaneously, because there are that many “observations.” In testing for causality, as in Section 21.6.5 below, Lutkepohl (2005, p. 95) argues that  $MT$  is excessive, because the restrictions are not imposed on all equations. When the causality test involves testing for zero restrictions within a single equation, the appropriate degrees of freedom would be  $T - Mp - 1$  for that one equation.

### 21.6.4 EXOGENEITY

In the classical regression model with nonstochastic regressors, there is no ambiguity about which is the independent or conditioning or “exogenous” variable in the model

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t. \quad (21-28)$$

This is the kind of characterization that might apply in an experimental situation in which the analyst is choosing the values of  $x_t$ . But, the case of nonstochastic regressors has little to do with the sort of modeling that will be of interest in this and the next chapter. There is no basis for the narrow assumption of nonstochastic regressors, and, in fact, in most of the analysis that we have done to this point, we have left this assumption far behind. With stochastic regressor(s), the regression relationship such as the preceding one becomes a conditional mean in a bivariate distribution. In this more realistic setting, what constitutes an “exogenous” variable becomes ambiguous. Assuming that the regression relationship is linear, (21-28) can be written (trivially) as

$$y_t = E[y_t | x_t] + (y_t - E[y_t | x_t]),$$

where the familiar moment condition  $E[x_t \varepsilon_t] = 0$  follows by construction. But, this form of the model is no more the “correct” equation than would be

$$x_t = \delta_1 + \delta_2 y_t + \omega_t,$$

which is (we assume)

$$x_t = E[x_t | y_t] + (x_t - E[x_t | y_t]),$$

and now,  $E[y_t \omega_t] = 0$ . Both equations are correctly specified in the context of the bivariate distribution, so there is nothing to define one variable or the other as “exogenous.” This might seem puzzling, but it is, in fact, at the heart of the matter when one considers modeling in a world in which variables are jointly determined. The definition of exogeneity depends on the analyst’s understanding of the world they are modeling, and, in the final analysis, on the purpose to which the model is to be put.

The methodological platform on which this discussion rests is the classic paper by Engle, Hendry, and Richard (1983), where they point out that exogeneity is not an absolute concept at all; it is defined in the context of the model. The central idea, which will be very useful to us here, is that we define a variable (set of variables) as exogenous *in the context of our model* if the joint density may be written

$$f(y_t, x_t) = f(y_t | \boldsymbol{\beta}, x_t) \times f(x_t | \boldsymbol{\theta})$$

where the parameters in the conditional distribution do not appear in and are functionally unrelated to those in the marginal distribution of  $x_t$ . By this arrangement, we

## CHAPTER 21 ♦ Models with Lagged Variables 977

can think of “autonomous variation” of the parameters of interest,  $\beta$ . The parameters in the conditional model for  $y_t | x_t$  can be analyzed as if they could vary independently of those in the marginal distribution of  $x_t$ . If this condition does not hold, then we cannot think of variation of those parameters without linking that variation to some effect in the marginal distribution of  $x_t$ . In this case, it makes little sense to think of  $x_t$  as somehow being determined “outside” the (conditional) model. (We considered this issue in Section 13.8 in the context of a simultaneous equations model.)

A second form of exogeneity we will consider is **strong exogeneity**, which is sometimes called **Granger noncausality**. Granger noncausality can be superficially defined by the assumption

$$E[y_t | y_{t-1}, x_{t-1}, x_{t-2}, \dots] = E[y_t | y_{t-1}].$$

That is, lagged values of  $x_t$  do not provide information about the conditional mean of  $y_t$  once lagged values of  $y_t$ , itself, are accounted for. We will consider this issue at the end of this chapter. For the present, we note that most of the models we will examine will explicitly fail this assumption.

To put this back in the context of our model, we will be assuming that in the model

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 x_{t-1} + \gamma y_{t-1} + \varepsilon_t,$$

and the extensions that we will consider,  $x_t$  is weakly exogenous—we can meaningfully estimate the parameters of the regression equation independently of the marginal distribution of  $x_t$ , but we will allow for Granger causality between  $x_t$  and  $y_t$ , thus generally not assuming strong exogeneity.

#### 21.6.5 TESTING FOR GRANGER CAUSALITY

Causality in the sense defined by Granger (1969) and Sims (1972) is inferred when lagged values of a variable, say,  $x_t$ , have explanatory power in a regression of a variable  $y_t$  on lagged values of  $y_t$  and  $x_t$ . The VAR can be used to test the hypothesis.<sup>8</sup> Tests of the restrictions can be based on simple  $F$  tests in the single equations of the VAR model. That the unrestricted equations have identical regressors means that these tests can be based on the results of simple OLS estimates. The notion can be extended in a system of equations to attempt to ascertain if a given variable is weakly exogenous to the system. If lagged values of a variable  $x_t$  have no explanatory power for *any* of the variables in a system, then we would view  $x_t$  as weakly exogenous to the system. Once again, this specification can be tested with a likelihood ratio test as described later—the restriction will be to put “holes” in one or more  $\Gamma$  matrices—or with a form of  $F$  test constructed by stacking the equations.

##### **Example 21.6 Granger Causality<sup>9</sup>**

All but one of the major recessions in the U.S. economy since World War II have been preceded by large increases in the price of crude oil. Does movement of the price of oil cause movements in U.S. GDP in the Granger sense? Let  $\mathbf{y}_t = [\text{GDP, crude oil price}]_t'$ . Then,

<sup>8</sup>See Geweke, Meese, and Dent (1983), Sims (1980), and Stock and Watson (2001).

<sup>9</sup>This example is adapted from Hamilton (1994, pp. 307–308).

## 978 PART V ♦ Time Series and Macroeconometrics

a simple VAR would be

$$\mathbf{y}_t = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{bmatrix} \mathbf{y}_{t-1} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}.$$

To assert a causal relationship between oil prices and GDP, we must find that  $\alpha_2$  is not zero; previous movements in oil prices do help explain movements in GDP even in the presence of the lagged value of GDP. Consistent with our earlier discussion, this fact, in itself, is not sufficient to assert a causal relationship. We would also have to demonstrate that there were no other intervening explanations that would explain movements in oil prices and GDP. (We will examine a more extensive application in Example 21.7.)

To establish the general result, it will prove useful to write the VAR in the multivariate regression format we used in Section 14.9.3.b. Partition the two data vectors  $\mathbf{y}_t$  and  $\mathbf{x}_t$  into  $[\mathbf{y}_{1t}, \mathbf{y}_{2t}]$  and  $[\mathbf{x}_{1t}, \mathbf{x}_{2t}]$ . Consistent with our earlier discussion,  $\mathbf{x}_1$  is lagged values of  $\mathbf{y}_1$  and  $\mathbf{x}_2$  is lagged values of  $\mathbf{y}_2$ . The VAR with this partitioning would be

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}, \quad \text{Var} \begin{bmatrix} \boldsymbol{\varepsilon}_{1t} \\ \boldsymbol{\varepsilon}_{2t} \end{bmatrix} = \begin{bmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{bmatrix}.$$

We would still obtain the unrestricted maximum likelihood estimates by least squares regressions. For testing Granger causality, the hypothesis  $\boldsymbol{\Gamma}_{12} = \mathbf{0}$  is of interest. (See Example 21.6.) For testing the hypothesis of interest,  $\boldsymbol{\Gamma}_{12} = \mathbf{0}$ , the second set of equations is irrelevant. For testing for Granger causality in the VAR model, only the restricted equations are relevant. The hypothesis can be tested using the likelihood ratio statistic. For the present application, testing means computing

$\mathbf{S}_{11}$  = residual covariance matrix when current values of  $\mathbf{y}_1$  are regressed on values of both  $\mathbf{x}_1$  and  $\mathbf{x}_2$ ,

$\mathbf{S}_{11}(0)$  = residual covariance matrix when current values of  $\mathbf{y}_1$  are regressed only on values of  $\mathbf{x}_1$ .

The likelihood ratio statistic is then

$$\lambda = T(\ln|\mathbf{S}_{11}(0)| - \ln|\mathbf{S}_{11}|).$$

The number of degrees of freedom is the number of zero restrictions.

The fact that this test is wedded to the normal distribution limits its generality. The Wald test or its transformation to an approximate  $F$  statistic as described in Section 21.6.3 is an alternative that should be more generally applicable. When the equation system is fit by GMM, as in Example 20.7, the simplicity of the likelihood ratio test is lost. The Wald statistic remains usable, however. Another possibility is to use the GMM counterpart to the likelihood ratio statistic (see Section 13.5.2) based on the GMM criterion functions. This is just the difference in the GMM criteria. Fitting both restricted and unrestricted models in this framework may be burdensome, but having set up the GMM estimator for the (larger) unrestricted model, imposing the zero restrictions of the smaller model should require only a minor modification.

There is a complication in these causality tests. The VAR can be motivated by the Wold representation theorem (see Section 22.2.5, Theorem 22.1), although with assumed nonautocorrelated disturbances, the motivation is incomplete. On the other hand, there is no formal theory behind the formulation. As such, the causality tests

## CHAPTER 21 ♦ Models with Lagged Variables 979

are predicated on a model that may, in fact, be missing either intervening variables or additional lagged effects that should be present but are not. For the first of these, the problem is that a finding of causal effects might equally well result from the omission of a variable that is correlated with both (or all) of the left-hand-side variables.

## 21.6.6 IMPULSE RESPONSE FUNCTIONS

Any VAR can be written as a first-order model by augmenting it, if necessary, with additional identity equations. For example, the model

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Gamma}_1 \mathbf{y}_{t-1} + \boldsymbol{\Gamma}_2 \mathbf{y}_{t-2} + \mathbf{v}_t$$

can be written

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Gamma}_1 & \boldsymbol{\Gamma}_2 \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \end{bmatrix} + \begin{bmatrix} \mathbf{v}_t \\ \mathbf{0} \end{bmatrix},$$

which is a first-order model. We can study the dynamic characteristics of the model in either form, but the second is more convenient, as will soon be apparent.

As we analyzed earlier, in the model

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Gamma} \mathbf{y}_{t-1} + \mathbf{v}_t,$$

dynamic stability is achieved if the characteristic roots of  $\boldsymbol{\Gamma}$  have modulus less than one. (The roots may be complex, because  $\boldsymbol{\Gamma}$  need not be symmetric. See Section 21.4.3.)

Assuming that the equation system is stable, the equilibrium is found by obtaining the final form of the system. We can do this step by repeated substitution, or more simply by using the lag operator to write

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Gamma}(L) \mathbf{y}_t + \mathbf{v}_t,$$

or

$$[\mathbf{I} - \boldsymbol{\Gamma}(L)] \mathbf{y}_t = \boldsymbol{\mu} + \mathbf{v}_t.$$

With the stability condition, we have

$$\begin{aligned} \mathbf{y}_t &= [\mathbf{I} - \boldsymbol{\Gamma}(L)]^{-1} (\boldsymbol{\mu} + \mathbf{v}_t) \\ &= (\mathbf{I} - \boldsymbol{\Gamma})^{-1} \boldsymbol{\mu} + \sum_{i=0}^{\infty} \boldsymbol{\Gamma}^i \mathbf{v}_{t-i} \\ &= \bar{\mathbf{y}} + \sum_{i=0}^{\infty} \boldsymbol{\Gamma}^i \mathbf{v}_{t-i} \\ &= \bar{\mathbf{y}} + \mathbf{v}_t + \boldsymbol{\Gamma} \mathbf{v}_{t-1} + \boldsymbol{\Gamma}^2 \mathbf{v}_{t-2} + \dots. \end{aligned} \tag{21-29}$$

The coefficients in the powers of  $\boldsymbol{\Gamma}$  are the multipliers in the system. We consider the conceptual experiment of disturbing a system in equilibrium. Suppose that  $\mathbf{v}$  has equaled  $\mathbf{0}$  for long enough that  $\mathbf{y}$  has reached equilibrium,  $\bar{\mathbf{y}}$ . Now we consider injecting a shock to the system by changing one of the  $v$ 's, for one period, and then returning it to zero thereafter. As we saw earlier,  $y_{mt}$  will move away from, then return to, its

## 980 PART V ♦ Time Series and Macroeconometrics

equilibrium. The path whereby the variables return to the equilibrium is called the **impulse response** of the VAR.<sup>10</sup>

In the autoregressive form of the model, we can identify each **innovation**,  $v_{mt}$ , with a particular variable in  $\mathbf{y}_t$ , say,  $y_{mt}$ . Consider then the effect of a one-time shock to the system,  $d v_{mt}$ . As compared with the equilibrium, we will have, in the current period,

$$y_{mt} - \bar{y}_m = d v_{mt} = \phi_{mm}(0) d v_t.$$

One period later, we will have

$$y_{m,t+1} - \bar{y}_m = (\boldsymbol{\Gamma})_{mm} d v_{mt} = \phi_{mm}(1) d v_t.$$

Two periods later,

$$y_{m,t+2} - \bar{y}_m = (\boldsymbol{\Gamma}^2)_{mm} d v_{mt} = \phi_{mm}(2) d v_t,$$

and so on. The function,  $\phi_{mm}(i)$  gives the impulse response characteristics of variable  $y_m$  to innovations in  $v_m$ . A useful way to characterize the system is to plot the impulse response functions. The preceding traces through the effect on variable  $m$  of a one-time innovation in  $v_m$ . We could also examine the effect of a one-time innovation of  $v_l$  on variable  $m$ . The impulse response function would be

$$\phi_{ml}(i) = \text{element } (m, l) \text{ in } \boldsymbol{\Gamma}^i.$$

Point estimation of  $\phi_{ml}(i)$  using the estimated model parameters is straightforward. Confidence intervals present a more difficult problem because the estimated functions  $\hat{\phi}_{ml}(i, \hat{\beta})$  are so highly nonlinear in the original parameter estimates. The delta method has thus proved unsatisfactory. Killian (1998) presents results that suggest that bootstrapping may be the more productive approach to statistical inference regarding impulse response functions.

### 21.6.7 STRUCTURAL VARs

The VAR approach to modeling dynamic behavior of economic variables has provided some interesting insights and appears [see Litterman (1986)] to bring some real benefits for forecasting. The method has received some strident criticism for its atheoretical approach, however. The “unrestricted” nature of the lag structure in (21-27) could be synonymous with “unstructured.” With no theoretical input to the model, it is difficult to claim that its output provides much of a theoretically justified result. For example, how are we to interpret the impulse response functions derived in the previous section? What lies behind much of this discussion is the idea that there is, in fact, a structure underlying the model, and the VAR that we have specified is a mere hodgepodge of all its components. Of course, that is exactly what reduced forms are. As such, to respond to this sort of criticism, analysts have begun to cast VARs formally as reduced forms and thereby attempt to deduce the structure that they had in mind all along.

A VAR model  $\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Gamma} \mathbf{y}_{t-1} + \mathbf{v}_t$  could, in principle, be viewed as the reduced form of the dynamic **structural model**

$$\boldsymbol{\Theta} \mathbf{y}_t = \boldsymbol{\alpha} + \boldsymbol{\Phi} \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t,$$

<sup>10</sup>See Hamilton (1994, pp. 318–323 and 336–350) for discussion and a number of related results.

CHAPTER 21 ♦ Models with Lagged Variables **981**

where we have embedded any exogenous variables  $x_t$  in the vector of constants  $\alpha$ . Thus,  $\Gamma = \Theta^{-1}\Phi$ ,  $\mu = \Theta^{-1}\alpha$ ,  $v = \Theta^{-1}\epsilon$ , and  $\Omega = \Theta^{-1}\Sigma(\Theta^{-1})'$ . Perhaps it is the structure, specified by an underlying theory, that is of interest. For example, we can discuss the impulse response characteristics of this system. For particular configurations of  $\Theta$ , such as a triangular matrix, we can meaningfully interpret innovations,  $\epsilon$ . As we explored at great length in the previous chapter, however, as this model stands, there is not sufficient information contained in the reduced form as just stated to deduce the structural parameters. A possibly large number of restrictions must be imposed on  $\Theta$ ,  $\Phi$ , and  $\Sigma$  to enable us to deduce structural forms from reduced-form estimates, which are always obtainable. The recent work on **structural VARs** centers on the types of restrictions and forms of the theory that can be brought to bear to allow this analysis to proceed. See, for example, the survey in Hamilton (1994, Chapter 11). At this point, the literature on this subject has come full circle because the contemporary development of “unstructured VARs” becomes very much the analysis of quite conventional dynamic structural simultaneous equations models. Indeed, current research [e.g., Diebold (1998)] brings the literature back into line with the structural modeling tradition by demonstrating how VARs can be derived formally as the reduced forms of dynamic structural models. That is, the most recent applications have begun with structures and derived the reduced forms as VARs, rather than departing from the VAR as a reduced form and attempting to deduce a structure from it by layering on restrictions.

**21.6.8 APPLICATION: POLICY ANALYSIS WITH A VAR**

Cecchetti and Rich (2001) used a structural VAR to analyze the effect of recent disinflationary policies of the Fed on aggregate output in the U.S. economy. The Fed's policy of the last two decades has leaned more toward controlling inflation and less toward stimulation of the economy. The authors argue that the long-run benefits of this policy include economic stability and increased long-term trend output growth. But, there is a short-term cost in lost output. Their study seeks to estimate the “sacrifice ratio,” which is a measure of the cumulative cost of this policy. The specific indicator they study measures the cumulative output loss after  $\tau$  periods of a policy shock at time  $t$ , where the (persistent) shock is measured as the change in the level of inflation.

**21.6.8.a A VAR Model for the Macroeconomic Variables**

The model proposed for estimating the ratio is a structural VAR,

$$\begin{aligned}\Delta y_t &= \sum_{i=1}^p b_{11}^i \Delta y_{t-i} + b_{12}^0 \Delta \pi_t + \sum_{i=1}^p b_{12}^i \Delta \pi_{t-i} + \varepsilon_t^y, \\ \Delta \pi_t &= b_{21}^0 \Delta y_t + \sum_{i=1}^p b_{21}^i \Delta y_{t-i} + \sum_{i=1}^p b_{22}^i \Delta \pi_{t-i} + \varepsilon_t^\pi,\end{aligned}$$

where  $y_t$  is aggregate real output in period  $t$  and  $\pi_t$  is the rate of inflation from period  $t-1$  to  $t$  and the model is cast in terms of rates of changes of these two variables. (Note, therefore, that sums of  $\Delta \pi_t$  measure accumulated changes in the rate of inflation, not changes in the CPI.) The vector of innovations,  $\epsilon_t = (\varepsilon_t^y, \varepsilon_t^\pi)'$  is assumed to have mean  $\mathbf{0}$ , contemporaneous covariance matrix  $E[\epsilon_t \epsilon_t'] = \Omega$  and to be strictly nonautocorrelated. (We have retained Cecchetti and Rich's notation for most of this discussion, save for

## 982 PART V ♦ Time Series and Macroeconometrics

the number of lags, which is denoted  $n$  in their paper and  $p$  here, and some other minor changes which will be noted in passing where necessary.)<sup>11</sup> The equation system may also be written

$$\mathbf{B}(L) \begin{bmatrix} \Delta y_t \\ \Delta \pi_t \end{bmatrix} = \begin{bmatrix} \varepsilon_t^y \\ \varepsilon_t^\pi \end{bmatrix},$$

where  $\mathbf{B}(L)$  is a  $2 \times 2$  matrix of polynomials in the lag operator. The components of the disturbance (innovation) vector  $\varepsilon_t$  are identified as shocks to aggregate supply and aggregate demand, respectively.

### 21.6.8.b The Sacrifice Ratio

Interest in the study centers on the impact over time of structural shocks to output and the rate of inflation. To calculate these, the authors use the **vector moving average** (VMA) form of the model, which would be

$$\begin{aligned} \begin{bmatrix} \Delta y_t \\ \Delta \pi_t \end{bmatrix} &= [\mathbf{B}(L)]^{-1} \begin{bmatrix} \varepsilon_t^y \\ \varepsilon_t^\pi \end{bmatrix} = \mathbf{A}(L) \begin{bmatrix} \varepsilon_t^y \\ \varepsilon_t^\pi \end{bmatrix} = \begin{bmatrix} A_{11}(L) & A_{12}(L) \\ A_{21}(L) & A_{22}(L) \end{bmatrix} \begin{bmatrix} \varepsilon_t^y \\ \varepsilon_t^\pi \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=0}^{\infty} a_{11}^i \varepsilon_{t-i}^y & \sum_{i=0}^{\infty} a_{12}^i \varepsilon_{t-i}^\pi \\ \sum_{i=0}^{\infty} a_{21}^i \varepsilon_{t-i}^y & \sum_{i=0}^{\infty} a_{22}^i \varepsilon_{t-i}^\pi \end{bmatrix}. \end{aligned}$$

(Note that the superscript “ $i$ ” in the last form of the preceding model is not an exponent; it is the index of the sequence of coefficients.) The impulse response functions for the model corresponding to (21-27) are precisely the coefficients in  $\mathbf{A}(L)$ . In particular, the effect on the change in inflation  $\tau$  periods later of a change in  $\varepsilon_t^\pi$  in period  $t$  is  $a_{22}^\tau$ . The total effect from time  $t + 0$  to time  $t + \tau$  would be the sum of these,  $\sum_{i=0}^{\tau} a_{22}^i$ . The counterparts for the rate of output would be  $\sum_{i=0}^{\tau} a_{12}^i$ . However, what is needed is not the effect only on period  $\tau$ ’s output, but the cumulative effect on output from the time of the shock up to period  $\tau$ . That would be obtained by summing these period-specific effects, to obtain  $\sum_{i=0}^{\tau} \sum_{j=0}^i a_{12}^j$ . Combining terms, the sacrifice ratio is

$$S_{\varepsilon^\pi}(\tau) = \frac{\sum_{j=0}^{\tau} \frac{\partial y_{t+j}}{\partial \varepsilon_t^\pi}}{\sum_{i=0}^{\tau} a_{22}^i} = \frac{\sum_{i=0}^0 a_{12}^j + \sum_{i=0}^1 a_{12}^j + \cdots + \sum_{i=0}^{\tau} a_{12}^j}{\sum_{i=0}^{\tau} a_{22}^i} = \frac{\sum_{i=0}^{\tau} \sum_{j=0}^i a_{12}^j}{\sum_{i=0}^{\tau} a_{22}^i}.$$

The function  $S(\tau)$  is then examined over long periods to study the long-term effects of monetary policy.

<sup>11</sup>The authors examine two other VAR models, a three-equation model of Shapiro and Watson (1988), which adds an equation in real interest rates ( $i_t - \pi_t$ ) and a four-equation model by Gali (1992), which models  $\Delta y_t$ ,  $\Delta i_t$ ,  $(i_t - \pi_t)$ , and the real money stock,  $(\Delta m_t - \pi_t)$ . Among the foci of Cecchetti and Rich’s paper was the surprisingly large variation in estimates of the sacrifice ratio produced by the three models. In the interest of brevity, we will restrict our analysis to Cecchetti’s (1994) two-equation model.

## CHAPTER 21 ♦ Models with Lagged Variables 983

## 21.6.8.c Identification and Estimation of a Structural VAR Model

Estimation of this model requires some manipulation. The **structural model** is a conventional linear simultaneous equations model of the form,

$$\mathbf{B}_0 \mathbf{y}_t = \mathbf{B} \mathbf{x}_t + \boldsymbol{\varepsilon}_t,$$

where  $\mathbf{y}_t$  is  $(\Delta y_t, \Delta \pi_t)'$  and  $\mathbf{x}_t$  is the lagged values on the right-hand side. As we saw in Chapter 10, without further restrictions, a model such as this is not identified (estimable). A total of  $M^2$  restrictions— $M$  is the number of equations, here two—are needed to identify the model. In the familiar cases of simultaneous equations models that we examined in Chapter 10, identification is usually secured through exclusion restrictions (i.e., zero restrictions), either in  $\mathbf{B}_0$  or  $\mathbf{B}$ . This type of exclusion restriction would be unnatural in a model such as this one—there would be no basis for poking specific holes in the coefficient matrices. The authors take a different approach, which requires us to look more closely at the different forms the time-series model can take.

Write the structural form as

$$\mathbf{B}_0 \mathbf{y}_t = \mathbf{B}_1 \mathbf{y}_{t-1} + \mathbf{B}_2 \mathbf{y}_{t-2} + \cdots + \mathbf{B}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

where

$$\mathbf{B}_0 = \begin{bmatrix} 1 & -b_{12}^0 \\ -b_{21}^0 & 1 \end{bmatrix}.$$

As noted, this is in the form of a conventional simultaneous equations model. Assuming that  $\mathbf{B}_0$  is nonsingular, which for this two-equation system requires only that  $1 - b_{12}^0 b_{21}^0$  not equal zero, we can obtain the reduced form of the model as

$$\begin{aligned} \mathbf{y}_t &= \mathbf{B}_0^{-1} \mathbf{B}_1 \mathbf{y}_{t-1} + \mathbf{B}_0^{-1} \mathbf{B}_2 \mathbf{y}_{t-2} + \cdots + \mathbf{B}_0^{-1} \mathbf{B}_p \mathbf{y}_{t-p} + \mathbf{B}_0^{-1} \boldsymbol{\varepsilon}_t \\ &= \mathbf{D}_1 \mathbf{y}_{t-1} + \mathbf{D}_2 \mathbf{y}_{t-2} + \cdots + \mathbf{D}_p \mathbf{y}_{t-p} + \boldsymbol{\mu}_t, \end{aligned} \quad (21-30)$$

where  $\boldsymbol{\mu}_t$  is the vector of reduced form innovations. Now, collect the terms in the equivalent form

$$[\mathbf{I} - \mathbf{D}_1 L - \mathbf{D}_2 L^2 - \cdots] \mathbf{y}_t = \boldsymbol{\mu}_t.$$

The moving-average form that we obtained earlier is

$$\mathbf{y}_t = [\mathbf{I} - \mathbf{D}_1 L - \mathbf{D}_2 L^2 - \cdots]^{-1} \boldsymbol{\mu}_t.$$

Assuming stability of the system, we can also write this as

$$\begin{aligned} \mathbf{y}_t &= [\mathbf{I} - \mathbf{D}_1 L - \mathbf{D}_2 L^2 - \cdots]^{-1} \boldsymbol{\mu}_t \\ &= [\mathbf{I} - \mathbf{D}_1 L - \mathbf{D}_2 L^2 - \cdots]^{-1} \mathbf{B}_0^{-1} \boldsymbol{\varepsilon}_t \\ &= [\mathbf{I} + \mathbf{C}_1 L + \mathbf{C}_2 L^2 + \cdots] \boldsymbol{\varepsilon}_t \\ &= \boldsymbol{\varepsilon}_t + \mathbf{C}_1 \boldsymbol{\varepsilon}_{t-1} + \mathbf{C}_2 \boldsymbol{\varepsilon}_{t-2} \dots \\ &= \mathbf{B}_0^{-1} \boldsymbol{\varepsilon}_t + \mathbf{C}_1 \boldsymbol{\varepsilon}_{t-1} + \mathbf{C}_2 \boldsymbol{\varepsilon}_{t-2} \dots \end{aligned}$$

So, the  $\mathbf{C}_j$  matrices correspond to our  $\mathbf{A}_j$  matrices in the original formulation. But this manipulation has added something. We can see that  $\mathbf{A}_0 = \mathbf{B}_0^{-1}$ . Looking ahead, the reduced form equations can be estimated by least squares. Whether the structural

## 984 PART V ♦ Time Series and Macroeconometrics

parameters, and thereafter, the VMA parameters can as well depends entirely on whether  $\mathbf{B}_0$  can be estimated. From (21-30) we can see that if  $\mathbf{B}_0$  can be estimated, then  $\mathbf{B}_1 \dots \mathbf{B}_p$  can also just by premultiplying the reduced form coefficient matrices by this estimated  $\mathbf{B}_0$ . So, we must now consider this issue.

Recall the initial assumption that  $E[\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}'_t] = \boldsymbol{\Omega}$ . In the reduced form, we assume  $E[\boldsymbol{\mu}_t \boldsymbol{\mu}'_t] = \boldsymbol{\Sigma}$ . As we know, reduced forms are always estimable (indeed, by least squares if the assumptions of the model are correct). That means that  $\boldsymbol{\Sigma}$  is estimable by the least squares residual variances and covariance. From the earlier derivation, we have that  $\boldsymbol{\Sigma} = \mathbf{B}_0^{-1} \boldsymbol{\Omega} (\mathbf{B}_0^{-1})' = \mathbf{A}_0 \boldsymbol{\Omega} \mathbf{A}_0'$ . (Again, see the beginning of Section 13.3.) The authors have secured identification of the model through this relationship. In particular, they assume first that  $\boldsymbol{\Omega} = \mathbf{I}$ . Assuming that  $\boldsymbol{\Omega} = \mathbf{I}$ , we now have that  $\mathbf{A}_0 \mathbf{A}_0' = \boldsymbol{\Sigma}$ , where  $\boldsymbol{\Sigma}$  is an estimable matrix with three free parameters. Because  $\mathbf{A}_0$  is  $2 \times 2$ , one more restriction is needed to secure identification. At this point, the authors, invoking Blanchard and Quah (1989), assume that “demand shocks have no permanent effect on the level of output. This is equivalent to  $A_{12}(1) = \sum_{i=0}^{\infty} a_{12}^i = 0$ .” This might seem like a cumbersome restriction to impose. But, the matrix  $\mathbf{A}(1)$  is  $[\mathbf{I} - \mathbf{D}_1 - \mathbf{D}_2 - \dots - \mathbf{D}_p]^{-1} \mathbf{A}_0 = \mathbf{F} \mathbf{A}_0$  and the components,  $\mathbf{D}_j$ , have been estimated as the reduced form coefficient matrices, so  $\mathbf{A}_{12}(1) = 0$  assumes only that the upper right element of this matrix is zero. We now obtain the equations needed to solve for  $\mathbf{A}_0$ . First,

$$\mathbf{A}_0 \mathbf{A}_0' = \begin{bmatrix} (a_{11}^0)^2 + (a_{12}^0)^2 & a_{11}^0 a_{21}^0 + a_{12}^0 a_{22}^0 \\ a_{11}^0 a_{21}^0 + a_{12}^0 a_{22}^0 & (a_{21}^0)^2 + (a_{22}^0)^2 \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{11} \end{bmatrix}, \quad (21-31)$$

which provides three equations. Second, the theoretical restriction is

$$\mathbf{F} \mathbf{A}_0 = \begin{bmatrix} * & f_{11} a_{12}^0 + f_{12} a_{22}^0 \\ * & * \end{bmatrix} = \begin{bmatrix} * & 0 \\ * & * \end{bmatrix}.$$

This provides the four equations needed to identify the four elements in  $\mathbf{A}_0$ .<sup>12</sup>

Collecting results, the estimation strategy is first to estimate  $\mathbf{D}_1, \dots, \mathbf{D}_p$  and  $\boldsymbol{\Sigma}$  in the reduced form, by least squares. (They set  $p = 8$ .) Then use the restrictions and (21-31) to obtain the elements of  $\mathbf{A}_0 = \mathbf{B}_0^{-1}$  and, finally,  $\mathbf{B}_j = \mathbf{A}_0^{-1} \mathbf{D}_j$ .

The last step is estimation of the matrices of impulse responses, which can be done as follows: We return to the reduced form which, using our augmentation trick, we

<sup>12</sup>At this point, an intriguing loose end arises. We have carried this discussion in the form of the original papers by Blanchard and Quah (1989) and Cecchetti and Rich (2001). Returning to the original structure, we see that because  $\mathbf{A}_0 = \mathbf{B}_0^{-1}$ , if  $\mathbf{B}_0$  has ones on the diagonal, then  $\mathbf{A}_0$  actually does not have four unrestricted and unknown elements, it has two. The model is thus overidentified. We could have predicted this at the outset. In our conventional simultaneous equations model, the normalizations in  $\mathbf{B}_0$  (ones on the diagonal) provide two restrictions of the  $M^2 = 4$  required for identification. Assuming that  $\boldsymbol{\Omega} = \mathbf{I}$  provides three more, and the theoretical restriction provides a sixth. Therefore, the four unknown elements in an unrestricted  $\mathbf{B}_0$  are overidentified. It might seem convenient at this point to forego the theoretical restriction on long-term impacts, but it seems more natural to omit the restrictions on the scaling of  $\boldsymbol{\Omega}$ . With the two normalizations already in place, assuming that the innovations are uncorrelated ( $\boldsymbol{\Omega}$  is diagonal) and “demand shocks have no permanent effect on the level of output” together suffice to identify the model. Blanchard and Quah appear to reach the same conclusion (page 656), but then they also assume the unit variances [page 657, equation (1)]. They argue that the assumption of unit variances is just a convenient normalization, which for their model is actually the case, because they do not assume that  $\mathbf{B}_0$  is diagonal. Cecchetti and Rich, however, do appear to normalize  $\mathbf{B}_0$  in their equation (1). They then (evidently) drop the assumption after (10), however, “[B]ecause  $\mathbf{A}_0$  has  $(n \times n)$  unique elements . . . .” This would imply that the normalization they impose on their (1) has not, in fact, been carried through the later manipulations, so, once again, the model is exactly identified.

## CHAPTER 21 ♦ Models with Lagged Variables 985

write as

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-p+1} \end{bmatrix} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{D}_2 & \cdots & \mathbf{D}_p \\ \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \vdots \\ \mathbf{y}_{t-p} \end{bmatrix} + \begin{bmatrix} \mathbf{A}_0 \boldsymbol{\varepsilon}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}. \quad (21-32)$$

For convenience, arrange this result as

$$\mathbf{Y}_t = (\mathbf{D}L)\mathbf{Y}_t + \mathbf{w}_t.$$

Now, solve this for  $\mathbf{Y}_t$  to obtain the final form

$$\mathbf{Y}_t = [\mathbf{I} - \mathbf{D}L]^{-1}\mathbf{w}_t.$$

Write this in the spectral form and expand as we did earlier, to obtain

$$\mathbf{Y}_t = \sum_{i=0}^{\infty} \mathbf{P} \mathbf{\Lambda}^i \mathbf{Q} \mathbf{w}_{t-i}. \quad (21-33)$$

We will be interested in the uppermost subvector of  $\mathbf{Y}_t$ , so we expand (21-33) to yield

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-p+1} \end{bmatrix} = \left[ \sum_{i=0}^{\infty} \mathbf{P} \mathbf{\Lambda}^i \mathbf{Q} \begin{bmatrix} \mathbf{A}_0 \boldsymbol{\varepsilon}_{t-i} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \right].$$

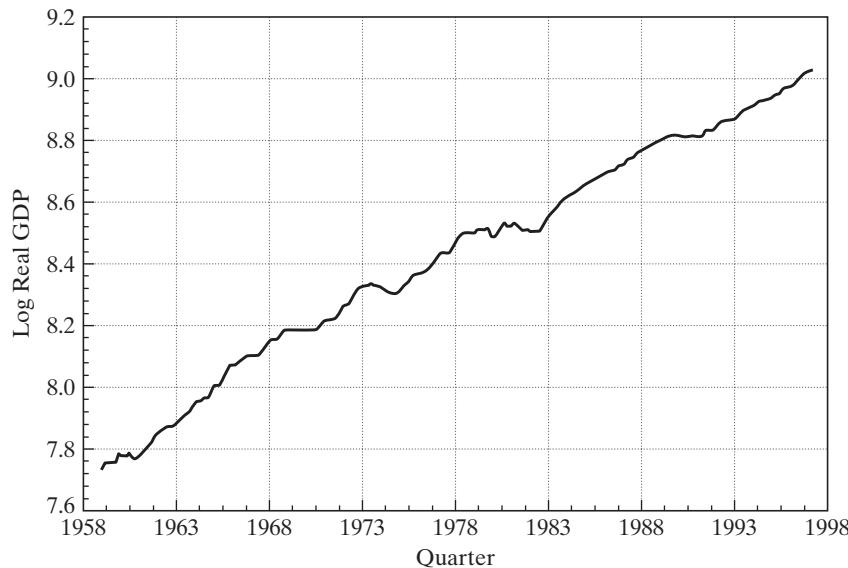
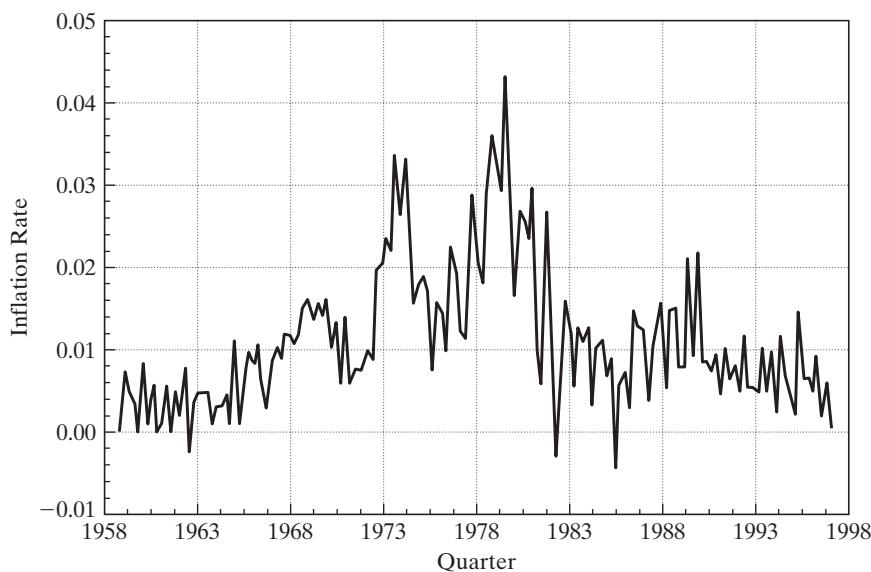
The matrix in the summation is  $Mp \times Mp$ . The impact matrices we seek are the  $M \times M$  matrices in the upper left corner of the spectral form, multiplied by  $\mathbf{A}_0$ .

#### 21.6.8.d Inference

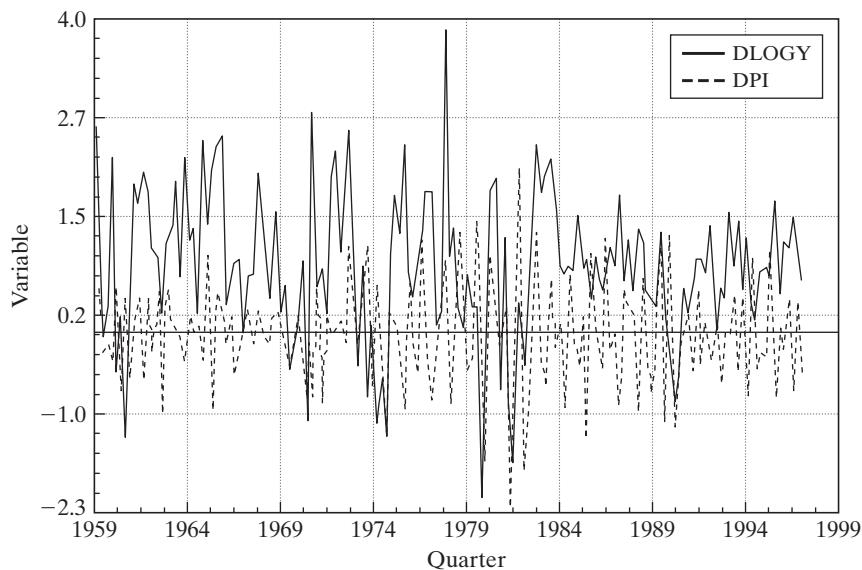
As noted at the end of Section 21.6.6, obtaining usable standard errors for estimates of impulse responses is a difficult (as yet unresolved) problem. Killian (1998) has suggested that bootstrapping is a preferable approach to using the delta method. Cecchetti and Rich reach the same conclusion and likewise resort to a bootstrapping procedure. Their bootstrap procedure is carried out as follows: Let  $\hat{\boldsymbol{\delta}}$  and  $\hat{\Sigma}$  denote the full set of estimated coefficients and estimated reduced form covariance matrix based on direct estimation. As suggested by Doan (2007), they construct a sequence of  $N$  draws for the reduced form parameters, then recompute the entire set of impulse responses. The narrowest interval, which contains 90 percent of these draws, is taken to be a confidence interval for an estimated impulse function.

#### 21.6.8.e Empirical Results

Cecchetti and Rich used quarterly observations on real aggregate output and the consumer price index. Their data set spanned 1959.1 to 1997.4. This is a subset of the data described in the Appendix Table F5.2. Before beginning their analysis, they subjected the data to the standard tests for stationarity. Figures 21.5–21.7 show the log of real output, the rate of inflation, and the changes in these two variables. The first two figures do suggest that neither variable is stationary. On the basis of the Dickey–Fuller (1981) test (see Section 23.2.4), they found (as might be expected) that the  $y_t$  and  $\pi_t$  series both contain unit roots. They conclude that because output has a unit root, the identification restriction that the long-run effect of aggregate demand shocks on output is well

**986 PART V ♦ Time Series and Macroeconometrics**

**FIGURE 21.5** Log GDP.

**FIGURE 21.6** The Quarterly Rate of Inflation.

defined and meaningful. The unit root in inflation allows for permanent shifts in its level. The lag length for the model is set at  $p = 8$ . Long-run impulse response functions are truncated at 20 years (80 quarters). Analysis is based on the rate of change data shown in Figure 21.7.



**FIGURE 21.7** Rates of Change, Log Real GDP, and the Rate of Inflation.

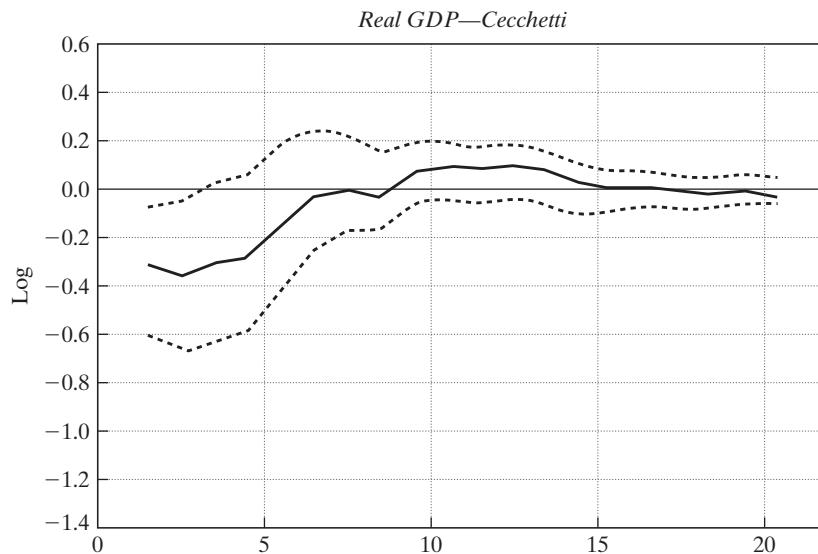
As a final check on the model, the authors examined the data for the possibility of a structural shift using the tests described in Andrews (1993) and Andrews and Ploberger (1994). None of the Andrews/Quandt supremum LM test, Andrews/Ploberger exponential LM test, or the Andrews/Ploberger average LM test suggested that the underlying structure had changed (in spite of what seems likely to have been a major shift in Fed policy in the 1970s). On this basis, they concluded that the VAR is stable over the sample period.

Figure 21.8 (Figures 3A and 3B taken from the article) shows their two separate estimated impulse response functions. The dotted lines in the figures show the bootstrap-generated confidence bounds. Estimates of the sacrifice ratio for Cecchetti's model are 1.3219 for  $\tau = 4$ , 1.3204 for  $\tau = 8$ , 1.5700 for  $\tau = 12$ , 1.5219 for  $\tau = 16$ , and 1.3763 for  $\tau = 20$ .

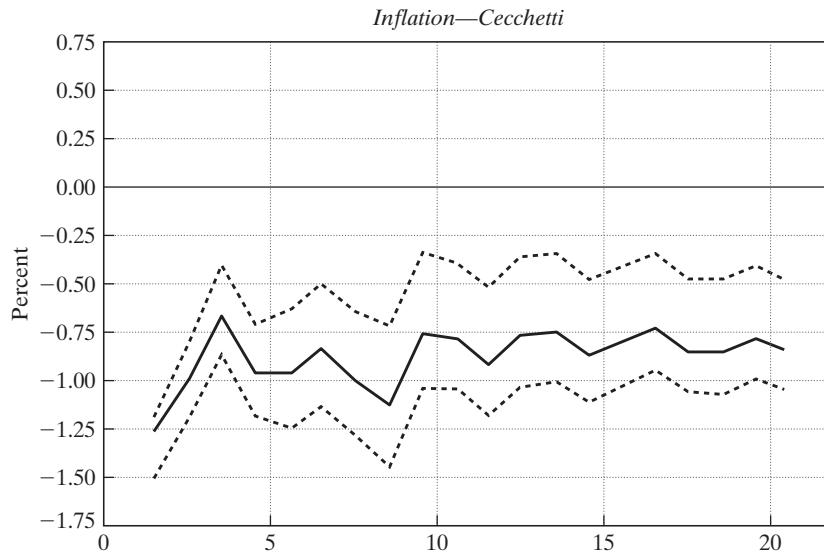
The authors also examined the forecasting performance of their model compared to Shapiro and Watson's and Gali's. The device used was to produce one step ahead, period  $T + 1 | T$  forecasts for the model estimated using periods  $1, \dots, T$ . The first reduced form of the model is fit using 1959.1 to 1975.1 and used to forecast 1975.2. Then, it is reestimated using 1959.1 to 1975.2 and used to forecast 1975.3, and so on. Finally, the root mean squared error of these out of sample forecasts is compared for three models. In each case, the level, rather than the rate of change of the inflation rate is forecasted. Overall, the results suggest that the smaller model does a better job of estimating the impulse responses (has smaller confidence bounds and conforms more nearly with theoretical predictions) but performs worst of the three (slightly) in terms of the mean squared error of the out-of-sample forecasts. Because the unrestricted reduced form model is being used for the latter, this comes as no surprise. The end result follows essentially from the result that adding variables to a regression model improves its fit.

**988 PART V ♦ Time Series and Macroeconometrics**

A: Dynamic Response to a Monetary Policy Shock



B: Dynamic Response to a Monetary Policy Shock

**FIGURE 21.8** Estimated Impulse Response Functions.
**21.6.9 VARs IN MICROECONOMICS**

VARs have appeared in the microeconomics literature as well. Chamberlain (1980) suggested that a useful approach to the analysis of panel data would be to treat each period's observation as a separate equation. For the case of  $T = 2$ , we would have

$$y_{i1} = \alpha_i + \mathbf{x}'_{i1}\boldsymbol{\beta} + \varepsilon_{i1},$$

$$y_{i2} = \alpha_i + \mathbf{x}'_{i2}\boldsymbol{\beta} + \varepsilon_{i2},$$

CHAPTER 21 ♦ Models with Lagged Variables **989**

where  $i$  indexes individuals and  $\alpha_i$  are unobserved individual effects. This specification produces a multivariate regression, to which Chamberlain added restrictions related to the individual effects. Holtz-Eakin, Newey, and Rosen's (1988) approach is to specify the equation as

$$y_{it} = \alpha_{0t} + \sum_{l=1}^m \alpha_{lt} y_{i,t-l} + \sum_{l=1}^m \delta_{lt} x_{i,t-l} + \Psi_t f_i + \mu_{it}.$$

In their study,  $y_{it}$  is hours worked by individual  $i$  in period  $t$  and  $x_{it}$  is the individual's wage in that period. A second equation for earnings is specified with lagged values of hours and earnings on the right-hand side. The individual, unobserved effects are  $f_i$ . This model is similar to the VAR in (21–27), but it differs in several ways as well. The number of periods is quite small (14 yearly observations for each individual), but there are nearly 1,000 individuals. The dynamic equation is specified for a specific period, however, so the relevant sample size in each case is  $n$ , not  $T$ . Also, the number of lags in the model used is relatively small; the authors fixed it at three. They thus have a two-equation VAR containing 12 unknown parameters, six in each equation. The authors used the model to analyze causality, measurement error, and parameter stability—that is, constancy of  $\alpha_{lt}$  and  $\delta_{lt}$  across time.

**Example 21.7 VAR for Municipal Expenditures**

In Example 13.10, we examined a model of municipal expenditures proposed by Dahlberg and Johansson (2000). Their equation of interest is

$$\Delta S_{i,t} = \mu_t + \sum_{j=1}^m \beta_j \Delta S_{i,t-j} + \sum_{j=1}^m \gamma_j \Delta R_{i,t-j} + \sum_{j=1}^m \delta_j \Delta G_{i,t-j} + u_{i,t}^S$$

for  $i = 1, \dots, N = 265$  and  $t = m+1, \dots, 9$ .  $S_{i,t}$ ,  $R_{i,t}$ , and  $G_{i,t}$  are municipal spending, receipts (taxes and fees), and central government grants, respectively. Analogous equations are specified for the current values of  $R_{i,t}$  and  $G_{i,t}$ . This produces a vector autoregression for each municipality,

$$\begin{bmatrix} \Delta S_{i,t} \\ \Delta R_{i,t} \\ \Delta G_{i,t} \end{bmatrix} = \begin{pmatrix} \mu_{S,t} \\ \mu_{R,t} \\ \mu_{G,t} \end{pmatrix} + \begin{pmatrix} \beta_{S,1} & \gamma_{S,1} & \delta_{S,1} \\ \beta_{R,1} & \gamma_{R,1} & \delta_{R,1} \\ \beta_{G,1} & \gamma_{G,1} & \delta_{G,1} \end{pmatrix} \begin{bmatrix} \Delta S_{i,t-1} \\ \Delta R_{i,t-1} \\ \Delta G_{i,t-1} \end{bmatrix} + \dots \\ + \begin{pmatrix} \beta_{S,m} & \gamma_{S,m} & \delta_{S,m} \\ \beta_{R,m} & \gamma_{R,m} & \delta_{R,m} \\ \beta_{G,m} & \gamma_{G,m} & \delta_{G,m} \end{pmatrix} \begin{bmatrix} \Delta S_{i,t-m} \\ \Delta R_{i,t-m} \\ \Delta G_{i,t-m} \end{bmatrix} + \begin{bmatrix} u_{i,t}^S \\ u_{i,t}^R \\ u_{i,t}^G \end{bmatrix}.$$

The model was estimated by GMM, so the discussion at the end of the preceding section applies here. We will be interested in testing whether changes in municipal spending,  $\Delta S_{i,t}$ , are Granger-caused by changes in revenues,  $\Delta R_{i,t}$ , and grants,  $\Delta G_{i,t}$ . The hypothesis to be tested is  $\gamma_{S,j} = \delta_{S,j} = 0$  for all  $j$ . This hypothesis can be tested in the context of only the first equation. Parameter estimates and diagnostic statistics are given in Example 13.10. We can carry out the test in two ways. In the unrestricted equation with all three lagged values of all three variables, the minimized GMM criterion is  $q = 22.8287$ . If the lagged values of  $\Delta R$  and  $\Delta G$  are omitted from the  $\Delta S$  equation, the criterion rises to 42.9182.<sup>13</sup> There are six restrictions. The difference is 20.090 so the  $F$  statistic is  $20.09/6 = 3.348$ . We have more

<sup>13</sup>Once again, these results differ from those given by Dahlberg and Johansson. As before, the difference results from our use of the same weighting matrix for all GMM computations in contrast to their recomputation of the matrix for each new coefficient vector estimated.

**990 PART V ♦ Time Series and Macroeconometrics**

than 1,000 degrees of freedom for the denominator, with 265 municipalities and five years, so we can use the limiting value for the critical value. This is 2.10, so we may reject the hypothesis of noncausality and conclude that changes in revenues and grants do Granger cause changes in spending. (This hardly seems surprising.) The alternative approach is to use a Wald statistic to test the six restrictions. Using the full GMM results for the  $\Delta S$  equation with 14 coefficients we obtain a Wald statistic of 15.3030. The critical chi-squared would be  $6 \times 2.1 = 12.6$ , so once again, the hypothesis is rejected.

Dahlberg and Johansson approach the causality test somewhat differently by using a sequential testing procedure. (See their page 413 for discussion.) They suggest that the intervening variables be dropped in turn. By dropping first  $G$ , then  $R$  and  $G$ , and then first  $R$  then  $G$  and  $R$ , they conclude that grants do not Granger-cause changes in spending ( $\Delta q = \text{only } 0.07$ ) but in the absence of grants, revenues do ( $\Delta q|\text{grants excluded} = 24.6$ ). The reverse order produces test statistics of 12.2 and 12.4, respectively. Our own calculations of the four values of  $q$  yields 22.829 for the full model, 23.1302 with only grants excluded, 23.0894 with only  $R$  excluded, and 42.9182 with both excluded, which disagrees with their results but is consistent with our earlier ones.

**Instability of a VAR Model**

The coefficients for the three-variable VAR model in Example 21.7 appear in Table 13.5. The characteristic roots of the  $9 \times 9$  coefficient matrix are  $-0.6025, 0.2529, 0.0840, (1.4586 \pm 0.6584)i, (-0.6992 \pm 0.2019i)$ , and  $(0.0611 \pm 0.6291i)$ . The first pair of complex roots has modulus greater than one, so the estimated VAR is unstable. The data do not appear to be consistent with this result, though with only five usable years of data, that conclusion is a bit fragile. One might suspect that the model is overfit. Because the disturbances are assumed to be uncorrelated across equations, the three equations have been estimated separately. The GMM criterion for the system is then the sum of those for the three equations. For  $p = 3, 2$ , and 1, respectively, these are  $(22.8287 + 30.5398 + 17.5810) = 70.9495$ ,  $(30.4526 + 34.2590 + 20.5416) = 85.2532$ , and  $(34.4986 + 53.2506 + 27.5927) = 115.6119$ . The difference statistic for testing down from three lags to two is 14.3037. The critical chi-squared for nine degrees of freedom is 19.62, so it would appear that  $m = 3$  may be too large. The results clearly reject the hypothesis that  $m = 1$ , however. The coefficients for a model with two lags instead of one appear in Table 15.5. If we construct  $\Gamma$  from these results instead, we obtain a  $6 \times 6$  matrix whose characteristic roots are  $1.5817, -0.2196, -0.3509 \pm 0.4362i$ , and  $0.0968 \pm 0.2791i$ . The system remains unstable.

**21.7 SUMMARY AND CONCLUSIONS**

This chapter has surveyed a particular type of regression model, the dynamic regression. The signature feature of the dynamic model is effects that are delayed or that persist through time. In a static regression setting, effects embodied in coefficients are assumed to take place all at once. In the dynamic model, the response to an innovation is distributed through several periods. The first three sections of this chapter examined several different forms of single-equation models that contained lagged effects. The progression, which mirrors the current literature, is from tightly structured lag "models" (which were sometimes formulated to respond to a shortage of data rather than to correspond to an underlying theory) to unrestricted models with multiple period lag structures. We also examined several hybrids of these two forms, models that allow long lags but build some regular structure into the lag weights. Thus, our model of the formation of expectations of inflation is reasonably flexible but does assume a specific behavioral mechanism. We then examined several methodological issues. In this context as elsewhere, there is a preference in the methods toward forming broad unrestricted

## CHAPTER 21 ♦ Models with Lagged Variables 991

models and using familiar inference tools to reduce them to the final appropriate specification. The second half of the chapter was devoted to a type of seemingly unrelated regressions model. The vector autoregression, or VAR, has been a major tool in recent research. After developing the econometric framework, we examined two applications, one in macroeconomics centered on monetary policy and one from microeconomics.

### **Key Terms and Concepts**

- Autocorrelation
- Autoregression
- Autoregressive distributed lag (ARDL)
- Autoregressive form
- Autoregressive model
- Characteristic equation
- Distributed lag form
- Dynamic regression model
- Elasticity
- Equilibrium
- Equilibrium error
- Equilibrium multiplier
- Equilibrium relationship
- Error correction
- Exogeneity
- Expectation
- Finite lags
- General-to-simple method
- Granger causality
- Impact multiplier
- Impulse response
- Infinite lag model
- Infinite lags
- Innovation
- Invertible
- Lagged variables
- Lag operator
- Lag weight
- Long-run multiplier
- Mean lag
- Median lag
- Moving-average form
- One-period-ahead forecast
- Partial adjustment
- Phillips curve
- Polynomial in lag operator
- Random walk with drift
- Rational lag
- Simple-to-general approach
- Specification
- Stability
- Stationary
- Strong exogeneity
- Structural model
- Structural VAR
- Superconsistent
- Univariate autoregression
- Vector autoregression (VAR)
- Vector moving average (VMA)

### **Exercises**

1. Obtain the mean lag and the long- and short-run multipliers for the following distributed lag models:
  - a.  $y_t = 0.55(0.02x_t + 0.15x_{t-1} + 0.43x_{t-2} + 0.23x_{t-3} + 0.17x_{t-4}) + e_t$ .
  - b. The model in Exercise 3.
  - c. The model in Exercise 4. (Do for either  $x$  or  $z$ .)
2. Expand the rational lag model  $y_t = [(0.6 + 2L)/(1 - 0.6L + 0.5L^2)]x_t + e_t$ . What are the coefficients on  $x_t, x_{t-1}, x_{t-2}, x_{t-3}$ , and  $x_{t-4}$ ?
3. Suppose that the model of Exercise 2 were specified as

$$y_t = \alpha + \frac{\beta + \gamma L}{1 - \delta_1 L - \delta_2 L^2} x_t + e_t.$$

Describe a method of estimating the parameters. Is ordinary least squares consistent?

4. Describe how to estimate the parameters of the model

$$y_t = \alpha + \beta \frac{x_t}{1 - \gamma L} + \delta \frac{z_t}{1 - \phi L} + \varepsilon_t,$$

where  $\varepsilon_t$  is a serially uncorrelated, homoscedastic, classical disturbance.

**992 PART V ♦ Time Series and Macroeconometrics****Applications**

1. We are interested in the long-run multiplier in the model

$$y_t = \beta_0 + \sum_{j=0}^6 \beta_j x_{t-j} + \varepsilon_t.$$

Assume that  $x_t$  is an autoregressive series,  $x_t = rx_{t-1} + v_t$  where  $|r| < 1$ .

- a. What is the long run multiplier in this model?
  - b. How would you estimate the long-run multiplier in this model?
  - c. Suppose you knew that the preceding is the true model, but you linearly regress  $y_t$  only on a constant and the first five lags of  $x_t$ . How does this affect your estimate of the long run multiplier?
  - d. Same as part c. for four lags instead of five.
  - e. Using the macroeconomic data in Appendix Table F5.2, let  $y_t$  be the log of real investment and  $x_t$  be the log of real output. Carry out the computations suggested and report your findings. Specifically, how does the omission of a lagged value affect estimates of the short-run and long-run multipliers in the unrestricted lag model?
2. Explain how to estimate the parameters of the following model:

$$y_t = \alpha + \beta x_t + \gamma y_{t-1} + \delta y_{t-2} + e_t,$$

$$e_t = \rho e_{t-1} + u_t.$$

Is there any problem with ordinary least squares? Let  $y_t$  be consumption and let  $x_t$  be disposable income. Using the method you have described, fit the previous model to the data in Appendix Table F5.2. Report your results.



## 22

## TIME-SERIES MODELS



## 22.1 INTRODUCTION

For forecasting purposes, a simple model that *describes* the behavior of a variable (or a set of variables) in terms of past values, without the benefit of a well-developed theory, may well prove quite satisfactory. Researchers have observed that the large simultaneous equations macroeconomic models constructed in the 1960s frequently have poorer forecasting performance than fairly simple, univariate time-series models based on just a few parameters and compact specifications. It is just this observation that has raised to prominence the univariate time-series forecasting models pioneered by Box and Jenkins (1984).

In this chapter, we introduce some of the tools employed in the analysis of time-series data.<sup>1</sup> Section 22.2 describes stationary stochastic processes. We encountered this body of theory in Chapters 20 and 21, where we discovered that certain assumptions were required to ascribe familiar properties to a time series of data. We continue that discussion by defining several characteristics of a stationary time series. The recent literature in macroeconomics has seen an explosion of studies of nonstationary time series. Nonstationarity mandates a revision of the standard inference tools we have used thus far. Chapter 23 introduces some extensions of the results of this chapter to nonstationary time series.

Some of the concepts to be discussed here were introduced in Section 20.2. Section 20.2 also contains a cursory introduction to the nature of time-series processes. It will be useful to review that material before proceeding with the rest of this chapter. Finally, Sections 13.6 on estimation and 13.9.2 and 21.4.3 on stability of dynamic models will be especially useful for the latter sections of this chapter.

---

<sup>1</sup>Each topic discussed here is the subject of a vast literature with articles and book-length treatments at all levels. For example, two survey papers on the subject of unit roots in economic time-series data, Diebold and Nerlove (1990) and Campbell and Perron (1991), cite between them more than 200 basic sources on the subject. The literature on unit roots and cointegration is almost surely the most rapidly moving target in econometrics. Stock's (1994) survey adds hundreds of references to those in the aforementioned surveys and brings the literature up to date as of then. Useful basic references on the subjects of this chapter are Box and Jenkins (1984); Judge et al. (1985); Mills (1990); Granger and Newbold (1996); Granger and Watson (1984); Hendry, Pagan, and Sargan (1984); Geweke (1984); and especially Harvey (1989, 1990); Enders (2004); Tsay (2005); Hamilton (1994); and Patterson (2000). There are also many survey style and pedagogical articles on these subjects. The aforementioned paper by Diebold and Nerlove is a useful tour guide through some of the literature. We recommend Dickey, Bell, and Miller (1986) and Dickey, Jansen, and Thornton (1991) as well. The latter is an especially clear introduction at a very basic level of the fundamental tools for empirical researchers.

## 994 PART V ♦ Time Series and Macroeconometrics

### 22.2 STATIONARY STOCHASTIC PROCESSES

The essential building block for the models to be discussed in this chapter is the **white noise** time-series process,

$$\{\varepsilon_t\}, t = -\infty, +\infty,$$

where each element in the sequence has  $E[\varepsilon_t] = 0$ ,  $E[\varepsilon_t^2] = \sigma_\varepsilon^2$ , and  $\text{Cov}[\varepsilon_t, \varepsilon_s] = 0$  for all  $s \neq t$ . Each element in the series is a random draw from a population with zero mean and constant variance. It is occasionally assumed that the draws are independent or normally distributed, although for most of our analysis, neither assumption will be essential.

A **univariate time-series model** describes the behavior of a variable in terms of its own past values. Consider, for example, the autoregressive disturbance models introduced in Chapter 20,

$$u_t = \rho u_{t-1} + \varepsilon_t. \quad (22-1)$$

Autoregressive disturbances are generally the residual variation in a regression model built up from what may be an elaborate underlying theory,  $y_t = \mathbf{x}'\boldsymbol{\beta} + u_t$ . The theory usually stops short of stating what enters the disturbance. But the presumption that some time-series process generates  $\mathbf{x}_t$  should extend equally to  $u_t$ . There are two ways to interpret this simple series. As stated,  $u_t$  equals the previous value of  $u_t$  plus an “innovation,”  $\varepsilon_t$ . Alternatively, by manipulating the series, we showed that  $u_t$  could be interpreted as an aggregation of the entire history of the  $\varepsilon_t$ ’s.

Occasionally, statistical evidence is convincing that a more intricate process is at work in the disturbance. Perhaps a second-order **autoregression**,

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t, \quad (22-2)$$

better explains the movement of the disturbances in the regression. The model may not arise naturally from an underlying behavioral theory. But in the face of certain kinds of statistical evidence, one might conclude that the more elaborate model would be preferable.<sup>2</sup> This section will describe several alternatives to the AR(1) model that we have relied on in most of the preceding applications.

#### 22.2.1 AUTOREGRESSIVE MOVING-AVERAGE PROCESSES

The variable  $y_t$  in the model

$$y_t = \mu + \gamma y_{t-1} + \varepsilon_t \quad (22-3)$$

is said to be **autoregressive** (or **self-regressive**) because under certain assumptions,

$$E[y_t | y_{t-1}] = \mu + \gamma y_{t-1}.$$

A more general  $p$ th-order autoregression or AR( $p$ ) process would be written

$$y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + \varepsilon_t. \quad (22-4)$$

---

<sup>2</sup>For example, the estimates  computed after a correction for first-order autocorrelation may fail tests of randomness such as the LM (Section 20.7.1) test.

## CHAPTER 22 ♦ Time-Series Models 995

The analogy to the classical regression is clear. Now consider the first-order moving average, or MA(1) specification<sup>3</sup>

$$y_t = \mu + \varepsilon_t - \theta \varepsilon_{t-1}. \quad (22-5)$$

By writing

$$y_t = \mu + (1 - \theta L)\varepsilon_t,$$

or

$$\frac{y_t}{1 - \theta L} = \frac{\mu}{1 - \theta} + \varepsilon_t,$$

we find that

$$y_t = \frac{\mu}{1 - \theta} - \theta y_{t-1} - \theta^2 y_{t-2} - \cdots + \varepsilon_t.$$

Once again, the effect is to represent  $y_t$  as a function of its own past values.

An extremely general model that encompasses (22-4) and (22-5) is the **autoregressive moving average**, or ARMA( $p, q$ ), model:

$$y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}. \quad (22-6)$$

Note the convention that the ARMA( $p, q$ ) process has  $p$  autoregressive (lagged dependent-variable) terms and  $q$  lagged **moving-average** terms. Researchers have found that models of this sort with relatively small values of  $p$  and  $q$  have proved quite effective as forecasting models.

The disturbances  $\varepsilon_t$  are labeled the **innovations** in the model. The term is fitting because the only new information that enters the processes in period  $t$  is this innovation. Consider, then, the AR(1) process

$$y_t = \mu + \gamma y_{t-1} + \varepsilon_t. \quad (22-7)$$

Either by successive substitution or by using the lag operator, we obtain

$$(1 - \gamma L)y_t = \mu + \varepsilon_t,$$

or

$$y_t = \frac{\mu}{1 - \gamma} + \sum_{i=0}^{\infty} \gamma^i \varepsilon_{t-i}.^4 \quad (22-8)$$

The observed series is a particular type of aggregation of the history of the innovations. The moving average, MA( $q$ ) model,

$$y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} = \mu + D(L)\varepsilon_t, \quad (22-9)$$

is yet another, particularly simple form of aggregation in that only information from the  $q$  most recent periods is retained. The general result is that many time-series processes can be viewed either as regressions on lagged values with additive disturbances or as aggregations of a history of innovations. They differ from one to the next in the form of that aggregation.

<sup>3</sup>The lag operator is discussed in Section 21.2.2. Because  $\mu$  is a constant,  $(1 - \theta L)^{-1}\mu = \mu + \theta\mu + \theta^2\mu + \cdots = \mu/(1 - \theta)$ . The lag operator may be set equal to one when it operates on a constant.

<sup>4</sup>See Section 21.3 for discussion of models with infinite lag structures.

## 996 PART V ♦ Time Series and Macroeconometrics

More involved processes can be similarly represented in either an autoregressive or moving-average form. (We will turn to the mathematical requirements later.) Consider, for example, the ARMA(2, 1) process,

$$y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \varepsilon_t - \theta \varepsilon_{t-1},$$

which we can write as

$$(1 - \theta L)\varepsilon_t = y_t - \mu - \gamma_1 y_{t-1} - \gamma_2 y_{t-2}.$$

If  $|\theta| < 1$ , then we can divide both sides of the equation by  $(1 - \theta L)$  and obtain

$$\varepsilon_t = \sum_{i=0}^{\infty} \theta^i (y_{t-i} - \mu - \gamma_1 y_{t-i-1} - \gamma_2 y_{t-i-2}).$$

After some tedious manipulation, this equation produces the autoregressive form,

$$y_t = \frac{\mu}{1 - \theta} + \sum_{i=1}^{\infty} \pi_i y_{t-i} + \varepsilon_t,$$

where

$$\pi_1 = \gamma_1 - \theta \quad \text{and} \quad \pi_j = -(\theta^j - \gamma_1 \theta^{j-1} - \gamma_2 \theta^{j-2}), \quad j = 2, 3, \dots \quad (22-10)$$

Alternatively, by similar (yet more tedious) manipulation, we can write

$$y_t = \frac{\mu}{1 - \gamma_1 - \gamma_2} + \left[ \frac{1 - \theta L}{1 - \gamma_1 L - \gamma_2 L^2} \right] \varepsilon_t = \frac{\mu}{1 - \gamma_1 - \gamma_2} + \sum_{i=0}^{\infty} \delta_i \varepsilon_{t-i}. \quad (22-11)$$

In each case, the weights,  $\pi_i$  in the **autoregressive form** and  $\delta_i$  in the **moving-average form**, are complicated functions of the original parameters. But nonetheless, each is just an alternative representation of the same time-series process that produces the current value of  $y_t$ . This result is a fundamental property of certain time series. We will return to the issue after we formally define the assumption that we have used at the preceding several steps that allows these transformations.

### 22.2.2 STATIONARITY AND INVERTIBILITY

At several points in the preceding, we have alluded to the notion of **stationarity**, either directly or indirectly by making certain assumptions about the parameters in the model. In Section 20.3.2, we characterized an AR(1) disturbance process

$$u_t = \rho u_{t-1} + \varepsilon_t,$$

as stationary if  $|\rho| < 1$  and  $\varepsilon_t$  is **white noise**. Then

$$\begin{aligned} E[u_t] &= 0 \quad \text{for all } t, \\ \text{Var}[u_t] &= \frac{\sigma_\varepsilon^2}{1 - \rho^2}, \\ \text{Cov}[u_t, u_s] &= \frac{\rho^{|t-s|} \sigma_\varepsilon^2}{1 - \rho^2}. \end{aligned} \quad (22-12)$$

If  $|\rho| \geq 1$ , then the variance and covariances are undefined.

In the following, we use  $\varepsilon_t$  to denote the white noise innovations in the process. The ARMA( $p, q$ ) process will be denoted as in (22-6).

### DEFINITION 22.1 Covariance Stationarity

A stochastic process  $y_t$  is **weakly stationary** or **covariance stationary** if it satisfies the following requirements:<sup>5</sup>

1.  $E[y_t]$  is independent of  $t$ .
2.  $\text{Var}[y_t]$  is a finite, positive constant, independent of  $t$ .
3.  $\text{Cov}[y_t, y_s]$  is a finite function of  $|t - s|$ , but not of  $t$  or  $s$ .

The third requirement is that the covariance between observations in the series is a function only of how far apart they are in time, not the time at which they occur. These properties clearly hold for the AR(1) process shown earlier. Whether they apply for the other models we have examined remains to be seen.

We define the **autocovariance** at lag  $k$  as

$$\lambda_k = \text{Cov}[y_t, y_{t-k}].$$

Note that

$$\lambda_{-k} = \text{Cov}[y_t, y_{t+k}] = \lambda_k.$$

Stationarity implies that autocovariances are a function of  $k$ , but not of  $t$ . For example, in (22-12), we see that the autocovariances of the AR(1) process  $y_t = \mu + \gamma y_{t-1} + \varepsilon_t$  are

$$\text{Cov}[y_t, y_{t-k}] = \frac{\gamma^k \sigma_\varepsilon^2}{1 - \gamma^2}, \quad k = 0, 1, \dots \quad (22-13)$$

If  $|\gamma| < 1$ , then this process is stationary. For any MA( $q$ ) series,

$$\begin{aligned} y_t &= \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}, \\ E[y_t] &= \mu + E[\varepsilon_t] - \theta_1 E[\varepsilon_{t-1}] - \cdots - \theta_q E[\varepsilon_{t-q}] = \mu, \\ \text{Var}[y_t] &= (1 + \theta_1^2 + \cdots + \theta_q^2) \sigma_\varepsilon^2, \\ \text{Cov}[y_t, y_{t-1}] &= (-\theta_1 + \theta_1 \theta_2 + \theta_2 \theta_3 + \cdots + \theta_{q-1} \theta_q) \sigma_\varepsilon^2, \end{aligned} \quad (22-14)$$

and so on until

$$\begin{aligned} \text{Cov}[y_t, y_{t-(q-1)}] &= [-\theta_{q-1} + \theta_1 \theta_2 \cdots \theta_{q-1}] \sigma_\varepsilon^2, \\ \text{Cov}[y_t, y_{t-q}] &= -\theta_q \sigma_\varepsilon^2, \end{aligned}$$

and, for lags greater than  $q$ , the autocovariances are zero. It follows, therefore, that finite moving-average processes are stationary regardless of the values of the parameters. The MA(1) process  $y_t = \varepsilon_t - \theta \varepsilon_{t-1}$  is an important special case that has  $\text{Var}[y_t] = (1 + \theta^2) \sigma_\varepsilon^2$ ,  $\lambda_1 = -\theta \sigma_\varepsilon^2$ , and  $\lambda_k = 0$  for  $|k| > 1$ .

<sup>5</sup>Strong stationarity requires that the joint distribution of all sets of observations  $(y_t, y_{t-1}, \dots)$  be invariant to when the observations are made. For practical purposes in econometrics, this statement is a theoretical fine point. Although weak stationary suffices for our applications, we would not normally analyze weakly stationary time series that were not **strongly stationary** as well. Indeed, we often go even beyond this step and assume joint normality.

## 998 PART V ♦ Time Series and Macroeconometrics

For the AR(1) process, the stationarity requirement is that  $|\gamma| < 1$ , which in turn, implies that the variance of the moving average representation in (22-8) is finite. Consider the AR(2) process

$$y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \varepsilon_t.$$

Write this equation as

$$C(L)y_t = \mu + \varepsilon_t,$$

where

$$C(L) = 1 - \gamma_1 L - \gamma_2 L^2.$$

Then, if it is possible, we invert this result to produce

$$y_t = [C(L)]^{-1}(\mu + \varepsilon_t).$$

Whether the inversion of the polynomial in the lag operator leads to a convergent series depends on the values of  $\gamma_1$  and  $\gamma_2$ . If so, then the moving-average representation will be

$$y_t = \sum_{i=0}^{\infty} \delta_i (\mu + \varepsilon_{t-i}),$$

so that

$$\text{Var}[y_t] = \sum_{i=0}^{\infty} \delta_i^2 \sigma_{\varepsilon}^2.$$

Whether this result is finite or not depends on whether the series of  $\delta_i$ s is exploding or converging. For the AR(2) case, the series converges if  $|\gamma_2| < 1$ ,  $\gamma_1 + \gamma_2 < 1$ , and  $\gamma_2 - \gamma_1 < 1$ .<sup>6</sup>

For the more general case, the autoregressive process is stationary if the roots of the **characteristic equation**,

$$C(z) = 1 - \gamma_1 z - \gamma_2 z^2 - \cdots - \gamma_p z^p = 0,$$

have modulus greater than one, or “lie outside the **unit circle**.<sup>7</sup> It follows that if a stochastic process is stationary, it has an infinite moving-average representation (and, if not, it does not). The AR(1) process is the simplest case. The characteristic equation is

$$C(z) = 1 - \gamma z = 0,$$

and its single root is  $1/\gamma$ . This root lies outside the unit circle if  $|\gamma| < 1$ , which we saw earlier.

Finally, consider the inversion of the moving-average process in (22-9). Whether this inversion is possible depends on the coefficients in  $D(L)$  in the same fashion that stationarity hinges on the coefficients in  $C(L)$ . This counterpart to stationarity of an autoregressive process is called **invertibility**. For it to be possible to invert a moving-average process to produce an autoregressive representation, the roots of  $D(L) = 0$

<sup>6</sup>This requirement restricts  $(\gamma_1, \gamma_2)$  to within a triangle with points at  $(2, -1)$ ,  $(-2, -1)$ , and  $(0, 1)$ .

<sup>7</sup>The roots may be complex. (See Sections 21.4.3.) They are of the form  $a \pm bi$ , where  $i = \sqrt{-1}$ . The unit circle refers to the two-dimensional set of values of  $a$  and  $b$  defined by  $a^2 + b^2 = 1$ , which defines a circle centered at the origin with radius 1.

## CHAPTER 22 ♦ Time-Series Models 999

must be outside the unit circle. Notice, for example, that in (22-5), the inversion of the moving-average process is possible only if  $|\theta| < 1$ . Because the characteristic equation for the MA(1) process is  $1 - \theta L = 0$ , the root is  $1/\theta$ , which must be larger than one.

If the roots of the characteristic equation of a moving-average process all lie outside the unit circle, then the series is said to be invertible. Note that invertibility has no bearing on the stationarity of a process. All moving-average processes with finite coefficients are stationary. Whether an ARMA process is stationary or not depends only on the AR part of the model.

### 22.2.3 AUTOCORRELATIONS OF A STATIONARY STOCHASTIC PROCESS

The function

$$\lambda_k = \text{Cov}[y_t, y_{t-k}]$$

is called the **autocovariance function** of the process  $y_t$ . The **autocorrelation function**, or **ACF**, is obtained by dividing by the variance,  $\lambda_0$ , to obtain

$$\rho_k = \frac{\lambda_k}{\lambda_0}, \quad -1 \leq \rho_k \leq 1.$$

For a stationary process, the ACF will be a function of  $k$  and the parameters of the process. The ACF is a useful device for describing a time-series process in much the same way that the moments are used to describe the distribution of a random variable. One of the characteristics of a stationary stochastic process is an autocorrelation function that either abruptly drops to zero at some finite lag or eventually tapers off to zero. The AR(1) process provides the simplest example, because

$$\rho_k = \gamma^k,$$

which is a geometric series that either declines monotonically from  $\rho_0 = 1$  if  $\gamma$  is positive or with a damped sawtooth pattern if  $\gamma$  is negative. Note as well that for the process  $y_t = \gamma y_{t-1} + \varepsilon_t$ ,

$$\rho_k = \gamma \rho_{k-1}, \quad k \geq 1,$$

which bears a noteworthy resemblance to the process itself.

For higher-order autoregressive series, the autocorrelations may decline monotonically or may progress in the fashion of a damped sine wave.<sup>8</sup> Consider, for example, the second-order autoregression, where we assume without loss of generality that  $\mu = 0$  (because we are examining second moments in deviations from the mean):

$$y_t = \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \varepsilon_t.$$

If the process is stationary, then  $\text{Var}[y_t] = \text{Var}[y_{t-s}]$  for all  $s$ . Also,  $\text{Var}[y_t] = \text{Cov}[y_t, y_t]$ , and  $\text{Cov}[\varepsilon_t, y_{t-s}] = 0$  if  $s > 0$ . These relationships imply that

$$\lambda_0 = \gamma_1 \lambda_1 + \gamma_2 \lambda_2 + \sigma_\varepsilon^2.$$

---

<sup>8</sup>The behavior is a function of the roots of the characteristic equation. This aspect is discussed in Section 21.4.3.

**1000 PART V ♦ Time Series and Macroeconometrics**

Now, using additional lags, we find that

$$\begin{aligned} \lambda_1 &= \gamma_1\lambda_0 + \gamma_2\lambda_1, \\ \text{and} \\ \lambda_2 &= \gamma_1\lambda_1 + \gamma_2\lambda_0. \end{aligned} \tag{22-15}$$

These three equations provide the solution:

$$\lambda_0 = \sigma_\varepsilon^2 \frac{[(1 - \gamma_2)/(1 + \gamma_2)]}{(1 - \gamma_2)^2 - \gamma_1^2}.$$

The variance is unchanging, so we can divide throughout by  $\lambda_0$  to obtain the relationships for the autocorrelations,

$$\rho_1 = \gamma_1\rho_0 + \gamma_2\rho_1.$$

Because  $\rho_0 = 1$ ,  $\rho_1 = \gamma_1/(1 - \gamma_2)$ . Using the same procedure for additional lags, we find that

$$\rho_2 = \gamma_1\rho_1 + \gamma_2,$$

so  $\rho_2 = \gamma_1^2/(1 - \gamma_2) + \gamma_2$ . Generally, then, for lags of two or more,

$$\rho_k = \gamma_1\rho_{k-1} + \gamma_2\rho_{k-2}.$$

Once again, the autocorrelations follow the same difference equation as the series itself. The behavior of this function depends on  $\gamma_1$ ,  $\gamma_2$ , and  $k$ , although not in an obvious way. The inherent behavior of the autocorrelation function can be deduced from the characteristic equation.<sup>9</sup> For the second-order process we are examining, the autocorrelations are of the form

$$\rho_k = \phi_1(1/z_1)^k + \phi_2(1/z_2)^k,$$

where the two roots are<sup>10</sup>

$$1/z = \frac{1}{2}[\gamma_1 \pm \sqrt{\gamma_1^2 + 4\gamma_2}].$$

If the two roots are real, then we know that their reciprocals will be less than one in absolute value, so that  $\rho_k$  will be the sum of two terms that are decaying to zero. If the two roots are complex, then  $\rho_k$  will be the sum of two terms that are oscillating in the form of a damped sine wave.

Applications that involve autoregressions of order greater than two are relatively unusual. Nonetheless, higher-order models can be handled in the same fashion. For the AR( $p$ ) process

$$y_t = \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + \varepsilon_t,$$

the autocovariances will obey the **Yule–Walker equations**

$$\begin{aligned} \lambda_0 &= \gamma_1\lambda_1 + \gamma_2\lambda_2 + \cdots + \gamma_p\lambda_p + \sigma_\varepsilon^2, \\ \lambda_1 &= \gamma_1\lambda_0 + \gamma_2\lambda_1 + \cdots + \gamma_p\lambda_{p-1}, \end{aligned}$$

<sup>9</sup>The set of results that we would use to derive this result are exactly those we used in Section 21.4.3 to analyze the stability of a dynamic equation, which makes sense, of course, because the equation linking the autocorrelations is a simple difference equation.

<sup>10</sup>We used the device in Section 21.4.3 to find the characteristic roots. For a second-order equation, the quadratic is easy to manipulate.

## CHAPTER 22 ♦ Time-Series Models 1001

and so on. The autocorrelations will once again follow the same difference equation as the original series,

$$\rho_k = \gamma_1 \rho_{k-1} + \gamma_2 \rho_{k-2} + \cdots + \gamma_p \rho_{k-p}.$$

The ACF for a moving-average process is very simple to obtain. For the first-order process,

$$\begin{aligned} y_t &= \varepsilon_t - \theta \varepsilon_{t-1}, \\ \lambda_0 &= (1 + \theta^2)\sigma_\varepsilon^2, \\ \lambda_1 &= -\theta\sigma_\varepsilon^2, \end{aligned}$$

then  $\lambda_k = 0$  for  $k > 1$ . Higher-order processes appear similarly. For the MA(2) process, by multiplying out the terms and taking expectations, we find that

$$\begin{aligned} \lambda_0 &= (1 + \theta_1^2 + \theta_2^2)\sigma_\varepsilon^2, \\ \lambda_1 &= (-\theta_1 + \theta_1\theta_2)\sigma_\varepsilon^2, \\ \lambda_2 &= -\theta_2\sigma_\varepsilon^2, \\ \lambda_k &= 0, \quad k > 2. \end{aligned}$$

The pattern for the general MA( $q$ ) process  $y_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q}$  is analogous. The signature of a moving-average process is an autocorrelation function that abruptly drops to zero at one lag past the order of the process. As we will explore later, this sharp distinction provides a statistical tool that will help us distinguish between these two types of processes empirically.

The mixed process, ARMA( $p, q$ ), is more complicated because it is a mixture of the two forms. For the ARMA(1, 1) process

$$y_t = \gamma y_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1},$$

the Yule–Walker equations are

$$\begin{aligned} \lambda_0 &= E[y_t(\gamma y_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1})] = \gamma\lambda_1 + \sigma_\varepsilon^2(\theta\gamma - \theta^2), \\ \lambda_1 &= \gamma\lambda_0 - \theta\sigma_\varepsilon^2, \end{aligned}$$

and

$$\lambda_k = \gamma\lambda_{k-1}, \quad k > 1.$$

The general characteristic of ARMA processes is that when the moving-average component is of order  $q$ , then in the series of autocorrelations there will be an initial  $q$  terms that are complicated functions of both the AR and MA parameters, but after  $q$  periods,

$$\rho_k = \gamma_1 \rho_{k-1} + \gamma_2 \rho_{k-2} + \cdots + \gamma_p \rho_{k-p}, \quad k > q.$$

### 22.2.4 PARTIAL AUTOCORRELATIONS OF A STATIONARY STOCHASTIC PROCESS

The autocorrelation function ACF( $k$ ) gives the gross correlation between  $y_t$  and  $y_{t-k}$ . But as we saw in our analysis of the classical regression model in Section 3.4, a gross correlation such as this one can mask a completely different underlying relationship. In

## 1002 PART V ♦ Time Series and Macroeconometrics

in this setting, we observe, for example, that a correlation between  $y_t$  and  $y_{t-2}$  could arise primarily because both variables are correlated with  $y_{t-1}$ . Consider the AR(1) process  $y_t = \gamma y_{t-1} + \varepsilon_t$ , where  $E[\varepsilon_t] = 0$  so  $E[y_t] = E[y_{t-1}]/(1 - \gamma) = 0$ . The second gross autocorrelation is  $\rho_2 = \gamma^2$ . But in the same spirit, we might ask what is the correlation between  $y_t$  and  $y_{t-2}$  net of the intervening effect of  $y_{t-1}$ ? In this model, if we remove the effect of  $y_{t-1}$  from  $y_t$ , then only  $\varepsilon_t$  remains, and this disturbance is uncorrelated with  $y_{t-2}$ . We would conclude that the **partial autocorrelation** between  $y_t$  and  $y_{t-2}$  in this model is zero.

### DEFINITION 22.2 Partial Autocorrelation Coefficient

The partial correlation between  $y_t$  and  $y_{t-k}$  is the simple correlation between  $y_{t-k}$  and  $y_t$  minus that part explained linearly by the intervening lags. That is,

$$\rho_k^* = \text{Corr}[y_t - E^*(y_t | y_{t-1}, \dots, y_{t-k+1}), y_{t-k}],$$

where  $E^*(y_t | y_{t-1}, \dots, y_{t-k+1})$  is the minimum mean-squared error predictor of  $y_t$  by  $y_{t-1}, \dots, y_{t-k+1}$ .

The function  $E^*(.)$  might be the linear regression if the conditional mean happened to be linear, but it might not. The optimal *linear* predictor is the linear regression, however, so what we have is

$$\rho_k^* = \text{Corr}[y_t - \beta_1 y_{t-1} - \beta_2 y_{t-2} - \dots - \beta_{k-1} y_{t-k+1}, y_{t-k}],$$

where  $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_{k-1}] = \{\text{Var}[y_{t-1}, y_{t-2}, \dots, y_{t-k+1}]\}^{-1} \times \text{Cov}[y_t, (y_{t-1}, y_{t-2}, \dots, y_{t-k+1})']'$ . This equation will be recognized as a vector of regression coefficients. As such, what we are computing here (of course) is the correlation between a vector of residuals and  $y_{t-k}$ . There are various ways to formalize this computation [see, e.g., Enders (2004)]. One intuitively appealing approach is suggested by the equivalent definition (which is also a prescription for computing it), as follows.

### DEFINITION 22.3 Partial Autocorrelation Coefficient

The partial correlation between  $y_t$  and  $y_{t-k}$  is the last coefficient in the linear projection of  $y_t$  on  $[y_{t-1}, y_{t-2}, \dots, y_{t-k}]$ ,

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k-1} \\ \rho_k^* \end{bmatrix} = \begin{bmatrix} \lambda_0 & \lambda_1 & \dots & \lambda_{k-2} & \lambda_{k-1} \\ \lambda_1 & \lambda_0 & \dots & \lambda_{k-3} & \lambda_{k-2} \\ \dots & \dots & \ddots & \dots & \dots \\ \lambda_{k-1} & \lambda_{k-2} & \dots & \lambda_1 & \lambda_0 \end{bmatrix}^{-1} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_k \end{bmatrix}.$$

## CHAPTER 22 ♦ Time-Series Models 1003

As before, there are some distinctive patterns for particular time-series processes. Consider first the autoregressive processes,

$$y_t = \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + \varepsilon_t.$$

We are interested in the last coefficient in the projection of  $y_t$  on  $y_{t-1}$ , then on  $[y_{t-1}, y_{t-2}]$ , and so on. The first of these is the simple regression coefficient of  $y_t$  on  $y_{t-1}$ , so

$$\rho_1^* = \frac{\text{Cov}[y_t, y_{t-1}]}{\text{Var}[y_{t-1}]} = \frac{\lambda_1}{\lambda_0} = \rho_1.$$

The first partial autocorrelation coefficient for any process equals the first autocorrelation coefficient.

Without doing the messy algebra, we also observe that for the AR( $p$ ) process,  $\rho_1^*$  is a mixture of all the  $\gamma$  coefficients. Of course, if  $p$  equals 1, then  $\rho_1^* = \rho_1 = \gamma$ . For the higher-order processes, the autocorrelations are likewise mixtures of the autoregressive coefficients until we reach  $\rho_p^*$ . In view of the form of the AR( $p$ ) model, the last coefficient in the linear projection on  $p$  lagged values is  $\gamma_p$ . Also, we can see the signature pattern of the AR( $p$ ) process, any additional partial autocorrelations must be zero, because they will be simply  $\rho_k^* = \text{Corr}[\varepsilon_t, y_{t-k}] = 0$  if  $k > p$ .

Combining results thus far, we have the characteristic pattern for an autoregressive process. The ACF,  $\rho_k$ , will gradually decay to zero, either monotonically if the characteristic roots are real or in a sinusoidal pattern if they are complex. The PACF,  $\rho_k^*$ , will be irregular out to lag  $p$ , when they abruptly drop to zero and remain there.

The moving-average process has the mirror image of this pattern. We have already examined the ACF for the MA( $q$ ) process; it has  $q$  irregular spikes, then it falls to zero and stays there. For the PACF, write the model as

$$y_t = (1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_q L^q) \varepsilon_t.$$

If the series is invertible, which we will assume throughout, then we have

$$\frac{y_t}{1 - \theta_1 L - \cdots - \theta_q L^q} = \varepsilon_t,$$

or

$$\begin{aligned} y_t &= \pi_1 y_{t-1} + \pi_2 y_{t-2} + \cdots + \varepsilon_t \\ &= \sum_{i=1}^{\infty} \pi_i y_{t-i} + \varepsilon_t. \end{aligned}$$

The autoregressive form of the MA( $q$ ) process has an infinite number of terms, which means that the PACF will not fall off to zero the way that the PACF of the AR process does. Rather, the PACF of an MA process will resemble the ACF of an AR process. For example, for the MA(1) process  $y_t = \varepsilon_t - \theta \varepsilon_{t-1}$ , the AR representation is

$$y_t = \theta y_{t-1} + \theta^2 y_{t-2} + \cdots + \varepsilon_t,$$

which is the familiar form of an AR(1) process. Thus, the PACF of an MA(1) process is identical to the ACF of an AR(1) process,  $\rho_1^* = \theta^k$ .

The ARMA( $p, q$ ) is a mixture of the two types of processes, so its ACF and PACF are likewise mixtures of the two forms discussed above. Generalities are difficult to draw, but normally, the ACF of an ARMA process will have a few distinctive spikes in

## 1004 PART V ♦ Time Series and Macroeconometrics

the early lags corresponding to the number of MA terms, followed by the characteristic smooth pattern of the AR part of the model. High-order MA processes are relatively uncommon in general, and high-order AR processes (greater than two) seem primarily to arise in the form of the nonstationary processes described in the next section. For a stationary process, the workhorses of the applied literature are the (2, 0) and (1, 1) processes. For the ARMA(1, 1) process, both the ACF and the PACF will display a distinctive spike at lag 1 followed by an exponentially decaying pattern thereafter.

### 22.2.5 MODELING UNIVARIATE TIME SERIES

The preceding discussion is largely descriptive. There is no underlying economic theory that states *why* a compact ARMA( $p, q$ ) representation should adequately describe the movement of a given economic time series. Nonetheless, as a methodology for building forecasting models, this set of tools and its empirical counterpart have proved as good as and even superior to much more elaborate specifications (perhaps to the consternation of the builders of large macroeconomic models).<sup>11</sup> Box and Jenkins (1984) pioneered a forecasting framework based on the preceding that has been used in a great many fields and that has, certainly in terms of numbers of applications, largely supplanted the use of large integrated econometric models.

Box and Jenkins's approach to modeling a stochastic process can be motivated by the following.

#### THEOREM 22.1 Wold's Decomposition Theorem

*Every zero-mean covariance stationary stochastic process can be represented in the form*

$$y_t = E^*[y_t | y_{t-1}, y_{t-2}, \dots, y_{t-p}] + \sum_{i=0}^{\infty} \pi_i \varepsilon_{t-i},$$

*where  $\varepsilon_t$  is white noise,  $\pi_0 = 1$ , and the weights are square summable—that is,*

$$\sum_{i=1}^{\infty} \pi_i^2 < \infty$$

*— $E^*[y_t | y_{t-1}, y_{t-2}, \dots, y_{t-p}]$  is the optimal linear predictor of  $y_t$  based on its lagged values, and the predictor  $E^*$  is uncorrelated with  $\varepsilon_{t-i}$ .*

Thus, the theorem decomposes the process generating  $y_t$  into

$E_t^* = E^*[y_t | y_{t-1}, y_{t-2}, \dots, y_{t-p}]$  = the **linearly deterministic component**

and

$\sum_{i=0}^{\infty} \pi_i \varepsilon_{t-i}$  = the **linearly indeterministic component**.

<sup>11</sup>This observation can be overstated. Even the most committed advocate of the Box–Jenkins methods would concede that an ARMA model of, for example, housing starts will do little to reveal the link between the interest rate policies of the Federal Reserve and their variable of interest. That is, the *covariation* of economic variables remains as interesting as ever.

**CHAPTER 22 ♦ Time-Series Models 1005**

The theorem states that for any stationary stochastic process, for a given choice of  $p$ , there is a Wold representation of the stationary series

$$y_t = \sum_{i=1}^p \gamma_i y_{t-i} + \sum_{i=0}^{\infty} \pi_i \varepsilon_{t-i}.$$

Note that for a specific ARMA( $P, Q$ ) process, if  $p \geq P$ , then  $\pi_i = 0$  for  $i > Q$ . For practical purposes, the problem with the Wold representation is that we cannot estimate the infinite number of parameters needed to produce the full right-hand side, and, of course,  $P$  and  $Q$  are unknown. The compromise, then, is to base an estimate of the representation on a model with a finite number of moving-average terms. We can seek the one that best fits the data in hand.

It is important to note that neither the ARMA representation of a process nor the Wold representation is unique. In general terms, suppose that the process generating  $y_t$  is

$$\Gamma(L)y_t = \Theta(L)\varepsilon_t.$$

We assume that  $\Gamma(L)$  is finite but  $\Theta(L)$  need not be. Let  $\Phi(L)$  be some other polynomial in the lag operator with roots that are outside the unit circle. Then

$$\left[ \frac{\Phi(L)}{\Gamma(L)} \right] \Gamma(L)y_t = \left[ \frac{\Phi(L)}{\Gamma(L)} \right] \Theta(L)\varepsilon_t,$$

or

$$\Phi(L)y_t = \Pi(L)\varepsilon_t.$$

The new representation is fully equivalent to the old one, but it might have a different number of autoregressive parameters, which is exactly the point of the Wold decomposition. The implication is that part of the model-building process will be to determine the lag structures. Further discussion on the methodology is given by Box and Jenkins (1984).

The Box-Jenkins approach to modeling stochastic processes consists of the following steps:

1. Satisfactorily transform the data so as to obtain a stationary series. This step will usually mean taking first differences, logs, or both to obtain a series whose autocorrelation function eventually displays the characteristic exponential decay of a stationary series.
2. Estimate the parameters of the resulting ARMA model, generally by nonlinear least squares.
3. Generate the set of residuals from the estimated model and verify that they satisfactorily resemble a white noise series. If not, respecify the model and return to step 2.
4. The model can now be used for forecasting purposes.

Space limitations prevent us from giving a full presentation of the set of techniques. Because this methodology has spawned a mini-industry of its own, however, there is no shortage of book length analyses and prescriptions to which the reader may refer. Five to consider are the canonical source, Box and Jenkins (1984), Granger and Newbold (1996), Mills (1993), Enders (2004), and Patterson (2000). Some of the aspects of the

## 1006 PART V ♦ Time Series and Macroeconometrics

estimation and analysis steps do have broader relevance for our work here, so we will continue to examine them in some detail.

### 22.2.6 ESTIMATION OF THE PARAMETERS OF A UNIVARIATE TIME SERIES

The broad problem of regression estimation with time-series data, which carries through to all the discussions of this chapter, is that the consistency and asymptotic normality results that we derived based on random sampling will no longer apply. For example, for a stationary series, we have assumed that  $\text{Var}[y_t] = \lambda_0$  regardless of  $t$ . But we have yet to establish that an estimated variance,

$$c_0 = \frac{1}{T-1} \sum_{t=1}^T (y_t - \bar{y})^2,$$

will converge to  $\lambda_0$ , or anything else for that matter. It is necessary to assume that the process is **ergodic**. (We first encountered this assumption in Section 20.4.1—see Definition 20.3.) Ergodicity is a crucial element of our theory of estimation. When a time series has this property (with stationarity), then we can consider estimation of parameters in a meaningful sense. If the process is stationary and ergodic, then, by the Ergodic theorem (Theorems 20.1 and 20.2), moments such as  $\bar{y}$  and  $c_0$  converge to their population counterparts  $\mu$  and  $\lambda_0$ .<sup>12</sup> The essential component of the condition is one that we have met at many points in this discussion, that autocovariances must decline sufficiently rapidly as the separation in time increases. It is possible to construct theoretical examples of processes that are stationary but not ergodic, but for practical purposes, a stationarity assumption will be sufficient for us to proceed with estimation. For example, in our models of stationary processes, if we assume that  $\varepsilon_t \sim N[0, \sigma^2]$ , which is common, then the stationary processes are ergodic as well.

Estimation of the parameters of a time-series process must begin with a determination of the type of process that we have in hand. (Box and Jenkins label this the **identification** step. But identification is a term of art in econometrics, so we will steer around that admittedly standard name.) For this purpose, the empirical estimates of the autocorrelation and partial autocorrelation functions are useful tools.

The sample counterpart to the ACF is the **correlogram**,

$$r_k = \frac{\sum_{t=k+1}^T (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^T (y_t - \bar{y})^2}.$$

A plot of  $r_k$  against  $k$  provides a description of a process and can be used to help discern what type of process is generating the data. The sample PACF is the counterpart to the ACF, but net of the intervening lags; that is,

$$r_k^* = \frac{\sum_{t=k+1}^T y_t^* y_{t-k}^*}{\sum_{t=k+1}^T (y_{t-k}^*)^2},$$

---

<sup>12</sup>The formal conditions for ergodicity are quite involved; see Davidson and MacKinnon (1993) or Hamilton (1994, Chapter 7).

**CHAPTER 22 ♦ Time-Series Models 1007**

where  $y_t^*$  and  $y_{t-k}^*$  are residuals from the regressions of  $y_t$  and  $y_{t-k}$  on  $[1, y_{t-1}, y_{t-2}, \dots, y_{t-k+1}]$ . We have seen this at many points before;  $r_k^*$  is simply the last linear least squares regression coefficient in the regression of  $y_t$  on  $[1, y_{t-1}, y_{t-2}, \dots, y_{t-k+1}, y_{t-k}]$ . Plots of the ACF and PACF of a series are usually presented together. Because the sample estimates of the autocorrelations and partial autocorrelations are not likely to be identically zero even when the population values are, we use diagnostic tests to discern whether a time series appears to be nonautocorrelated.<sup>13</sup> Individual sample autocorrelations will be approximately distributed with mean zero and variance  $1/T$  under the hypothesis that the series is white noise. The Box–Pierce (1970) statistic

$$Q = T \sum_{k=1}^p r_k^2$$

is commonly used to test whether a series is white noise. Under the null hypothesis that the series is white noise,  $Q$  has a limiting chi-squared distribution with  $p$  degrees of freedom. A refinement that appears to have better finite-sample properties is the Ljung–Box (1979) statistic,

$$Q' = T(T+2) \sum_{k=1}^p \frac{r_k^2}{T-k}.$$

The limiting distribution of  $Q'$  is the same as that of  $Q$ .

The process of finding the appropriate specification is essentially trial and error. An initial specification based on the sample ACF and PACF can be found. The parameters of the model can then be estimated by least squares. For pure AR( $p$ ) processes, the estimation step is simple. The parameters can be estimated by linear least squares. If there are moving-average terms, then linear least squares is inconsistent, but the parameters of the model can be fit by nonlinear least squares. Once the model has been estimated, a set of residuals is computed to assess the adequacy of the specification. In an AR model, the residuals are just the deviations from the regression line.

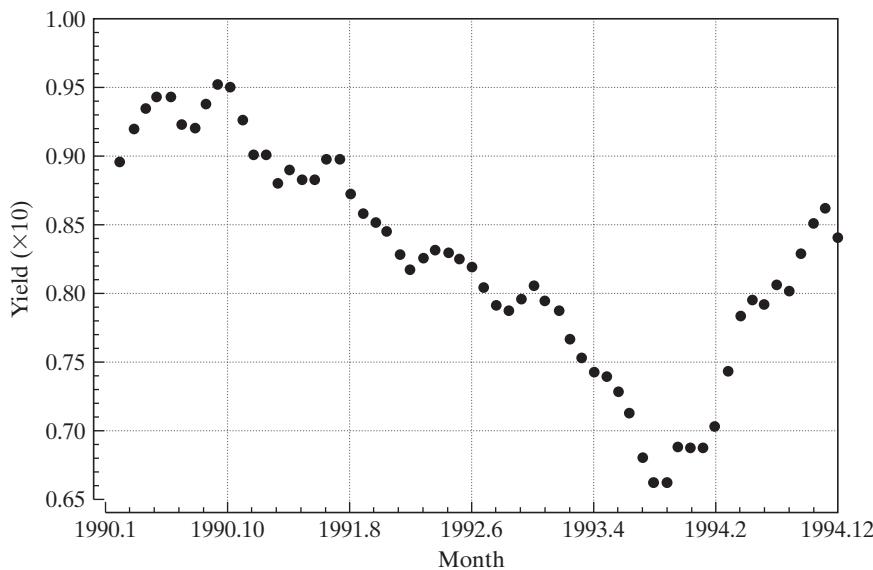
The adequacy of the specification can be examined by applying the foregoing techniques to the estimated residuals. If they appear satisfactorily to mimic a white noise process, then analysis can proceed to the forecasting step. If not, a new specification should be considered.

**Example 22.1 ACF and PACF for a Series of Bond Yields**

Appendix Table F22.1 lists five years of monthly averages of the yield on a Moody's Aaa-rated corporate bond. (Note: In previous editions of this text, the second observation in the data file was incorrectly recorded as 9.72. The correct value is 9.22. Computations to follow are based on the corrected value.) The series is plotted in Figure 22.1. From the figure, it would appear that stationarity may not be a reasonable assumption. We will return to this question below. The ACF and PACF for the original series are shown in Table 22.1, with the diagnostic statistics discussed earlier.

Based on the two spikes in the PACF, the results appear to be consistent with an AR(2) process, although the ACF at longer lags seems a bit more persistent than might have been expected. Once again, this condition may indicate that the series is not stationary. Maintaining that assumption for the present, we computed the residuals from an AR(2) model and subjected them to the same test as the original observations. To compute the regression,

<sup>13</sup>The LM test discussed in Section 20.7.1 is one of these.

**1008 PART V ♦ Time Series and Macroeometrics**

**FIGURE 22.1** Monthly Data on Bond Yields.

**TABLE 22.1** ACF and PACF for Bond Yields

Time-series identification for YIELD

Box-Pierce statistic = 326.0507

Degrees of freedom = 14

Significance level = 0.0000

 $* \Rightarrow |\text{coefficient}| > 2/\sqrt{N}$  or  $> 95\%$  significant.

PACF is computed using Yule-Walker equations.

Box-Ljung statistic = 364.6475

Degrees of freedom = 14

Significance level = 0.0000

<b>Lag</b>	<b>Autocorrelation Function</b>			<b>Box/Prc</b>	<b>Partial Autocorrelations X</b>		
	1	2	3		4	5	6
1	0.973*		*****	56.81*	0.973*		*****
2	0.922*		*****	107.76*	-0.477*	*****	
3	0.863*		*****	152.47*	0.057		*
4	0.806*		*****	191.43*	0.021		*
5	0.745*		*****	224.71*	-0.186	***	
6	0.679*		*****	252.39*	-0.046	*	
7	0.606*		*****	274.44*	-0.174	***	
8	0.529*		****	291.22*	-0.039	*	
9	0.450*		****	303.37*	-0.049	*	
10	0.379*		***	311.98*	0.146		**
11	0.316*		***	317.95*	-0.023	*	
12	0.259*		***	321.97*	-0.001	*	
13	0.205		**	324.49*	-0.018	*	
14	0.161		**	326.05*	0.185		***

Note: \*s in first column and bars in the right-hand panel have changed from earlier edition.

we first subtracted the overall mean from all 60 observations. We then fit the AR(2) without the first two observations. The coefficients of the AR(2) process are 1.47701 and -.51553, which also satisfy the restrictions for stationarity given in Section 22.2. Despite the earlier suggestions, the residuals do appear to resemble a stationary process. (See Table 22.2.)

**TABLE 22.2** ACF and PACF for Residuals

Time series identification for U

Box-Pierce Statistic = 10.6480

Degrees of freedom = 14

Significance level = 0.7134

\* =>  $|coefficient| > 2/\sqrt{N}$  or > 95% significant.

PACF is computed using Yule-Walker equations.

Box-Ljung Statistic = 12.3380

Degrees of freedom = 14

Significance level = 0.5792



Lag	Autocorrelation Function		Box/Prc	Partial Autocorrelation Function		X
1	0.063	*	0.23	0.063	*	X
2	-0.119	*	1.06	-0.133	*	X
3	-0.235	***	4.27	-0.241	***	X
4	0.108	*	4.95	0.142	**	X
5	0.142	**	6.11	0.113	*	X
6	0.117	*	6.91	0.108	*	X
7	-0.091	*	7.39	-0.047	*	X
8	0.058	*	7.58	0.189	**	X
9	-0.167	**	9.19	-0.321*	****	X
10	0.034	*	9.25	0.021	*	X
11	-0.004	*	9.25	0.043	*	X
12	0.013	*	9.26	-0.072	*	X
13	-0.134	*	10.31	-0.179	**	X
14	-0.076	*	10.65	-0.114	*	X

Note: \* in third column and bars in both panels have changed from earlier edition.

## 22.3 SUMMARY AND CONCLUSIONS

This chapter has developed the standard tools for analyzing a stationary time series. The analysis takes place in one of two frameworks, the time domain or the frequency domain. The analysis in the time domain focuses on the different representations of the series in terms of autoregressive and moving-average components. This interpretation of the time series is closely related to the concept of regression—though in this case it is “auto-” or self-regression, that is, on past values of the random variable itself. The autocorrelations and partial autocorrelations are the central tools for characterizing a time series in this framework. Constructing a time series in this fashion allows the analyst to construct forecasting equations that exploit the internal structure of the time series. (We have left for additional courses and the many references on the subject the embellishments of these models in terms of seasonal patterns, differences, and so on, that are the staples of Box-Jenkins model building.)

The analysis in this chapter, of modern economic time-series analysis, is a prelude to the analysis of nonstationary series in the next chapter. Nonstationarity is, in large measure, the norm in recent time-series modeling.

### Key Terms and Concepts

- Autocorrelation function(ACF)
- Autocovariance
- Autocovariance function
- Autoregression
- Autoregressive
- Autoregressive form
- Autoregressive moving average
- Characteristic equation
- Covariance stationarity
- Ergodic
- Identification
- Innovations
- Invertibility

**1010 PART V ♦ Time Series and Macroeconometrics**

- Linearly deterministic component
- Linearly indeterministic component
- Moving average
- Moving-average form

- Nonstationarity
- Partial autocorrelation
- Self-regressive
- Square summable
- Stationarity
- Strong stationarity
- Unit circle

- Univariate time-series model
- Weak stationarity
- White noise
- **Wold's decomposition theorem**
- Yule–Walker equations

8/1

## 23

## NONSTATIONARY DATA



## 23.1 INTRODUCTION

Most economic variables that exhibit strong trends, such as GDP, consumption, or the price level, are not stationary and are thus not amenable to the analysis of the previous three chapters. In many cases, stationarity can be achieved by simple differencing or some other simple transformation. But, new statistical issues arise in analyzing nonstationary series that are understated by this superficial observation. This chapter will survey a few of the major issues in the analysis of nonstationary data.<sup>1</sup> We begin in Section 23.2 with results on analysis of a single nonstationary time series. Section 23.3 examines the implications of nonstationarity for analyzing regression relationship. Finally, Section 23.4 turns to the extension of the time-series results to panel data.

## 23.2 NONSTATIONARY PROCESSES AND UNIT ROOTS

This section will begin the analysis of nonstationary time series with some basic results for univariate time series. The fundamental results concern the characteristics of nonstationary series and statistical tests for identification of nonstationarity in observed data.

## 23.2.1 INTEGRATED PROCESSES AND DIFFERENCING

A process that figures prominently in recent work is the **random walk with drift**,

$$y_t = \mu + y_{t-1} + \varepsilon_t.$$

By direct substitution,

$$y_t = \sum_{i=0}^{\infty} (\mu + \varepsilon_{t-i}).$$

That is,  $y_t$  is the simple sum of what will eventually be an infinite number of random variables, possibly with nonzero mean. If the innovations are being generated by the same zero-mean, constant-variance distribution, then the variance of  $y_t$  would obviously be infinite. As such, the random walk is clearly a **nonstationary process**, even if  $\mu$  equals

---

<sup>1</sup>With panel data, this is one of the rapidly growing areas in econometrics, and the literature advances rapidly. We can only scratch the surface. Several recent surveys and books provide useful extensions. Two that will be very helpful are Enders (2004) and Tsay (2005).

## 1012 PART V ♦ Time Series and Macroeconometrics

zero. On the other hand, the first difference of  $y_t$ ,

$$z_t = y_t - y_{t-1} = \mu + \varepsilon_t,$$

is simply the innovation plus the mean of  $z_t$ , which we have already assumed is stationary.

The series  $y_t$  is said to be **integrated of order one**, denoted  $I(1)$ , because taking a first difference produces a stationary process. A nonstationary series is integrated of order  $d$ , denoted  $I(d)$ , if it becomes stationary after being first differenced  $d$  times. A further generalization of the ARMA model discussed in Section 22.2.1 would be the series

$$z_t = (1 - L)^d y_t = \Delta^d y_t.$$

The resulting model is denoted an **autoregressive integrated moving-average** model, or **ARIMA**  $(p, d, q)$ .<sup>2</sup> In full, the model would be

$$\Delta^d y_t = \mu + \gamma_1 \Delta^d y_{t-1} + \gamma_2 \Delta^d y_{t-2} + \cdots + \gamma_p \Delta^d y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q},$$

where

$$\Delta y_t = y_t - y_{t-1} = (1 - L)y_t.$$

This result may be written compactly as

$$C(L)[(1 - L)^d y_t] = \mu + D(L)\varepsilon_t,$$

where  $C(L)$  and  $D(L)$  are the polynomials in the lag operator and  $(1 - L)^d y_t = \Delta^d y_t$  is the  $d$ th difference of  $y_t$ .

An  $I(1)$  series in its raw (undifferenced) form will typically be constantly growing, or wandering about with no tendency to revert to a fixed mean. Most macroeconomic flows and stocks that relate to population size, such as output or employment, are  $I(1)$ . An  $I(2)$  series is growing at an ever-increasing rate. The price-level data in Appendix Table F23.1 and shown later appear to be  $I(2)$ . Series that are  $I(3)$  or greater are extremely unusual, but they do exist. Among the few manifestly  $I(3)$  series that could be listed, one would find, for example, the money stocks or price levels in hyperinflationary economies such as interwar Germany or Hungary after World War II.

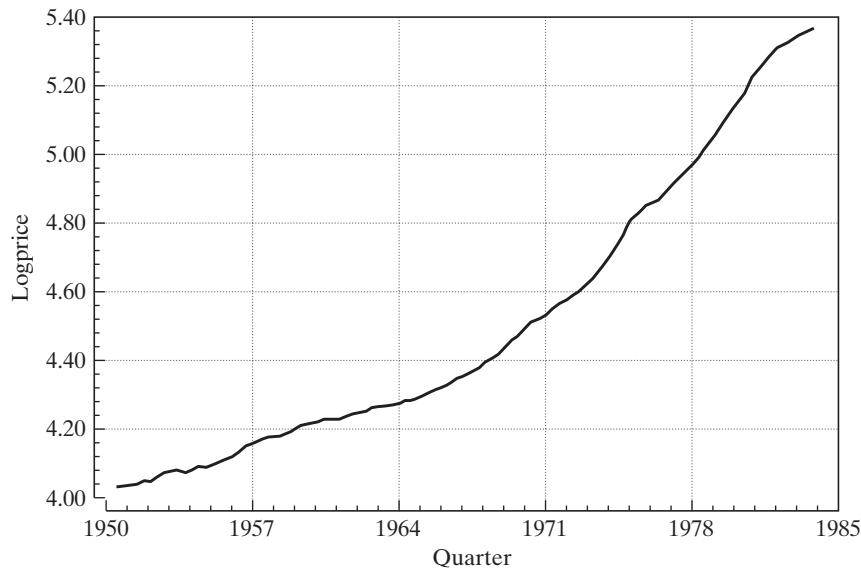
### **Example 23.1 A Nonstationary Series**

The nominal GNP and price deflator variables in Appendix Table F23.1 are strongly trended, so the mean is changing over time. Figures 23.1 through 23.3 plot the log of the GNP deflator series in Table F23.1 and its first and second differences. The original series and first differences are obviously nonstationary, but the second differencing appears to have rendered the series stationary.

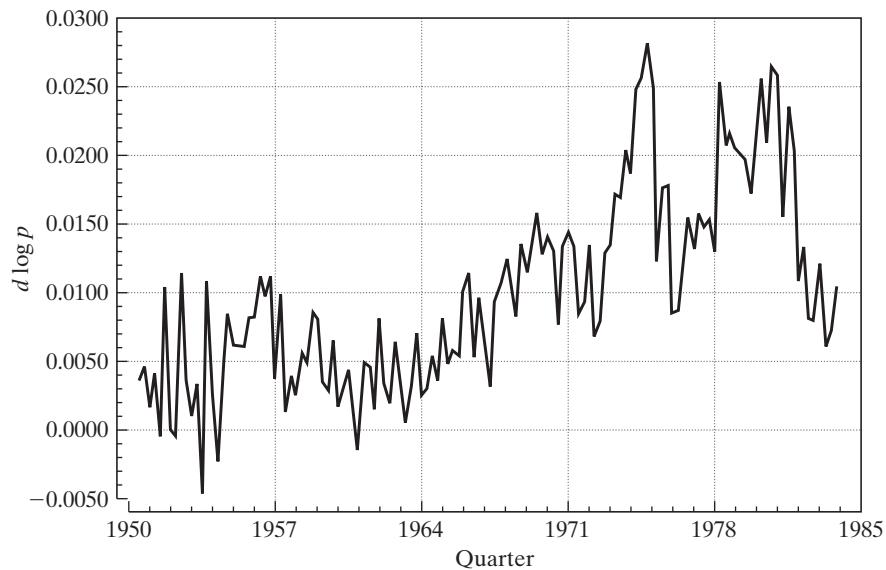
The first 10 autocorrelations of the log of the GNP deflator series are shown in Table 23.1. The autocorrelations of the original series show the signature of a strongly trended, nonstationary series. The first difference also exhibits nonstationarity, because the autocorrelations are still very large after a lag of 10 periods. The second difference appears to be stationary, with mild negative autocorrelation at the first lag, but essentially none after that. Intuition might suggest that further differencing would reduce the autocorrelation further, but that would be incorrect. We leave as an exercise to show that, in fact, for values of  $\gamma$  less than about 0.5, first differencing of an AR(1) process actually increases autocorrelation.

---

<sup>2</sup>There are yet further refinements one might consider, such as removing seasonal effects from  $z_t$  by differencing by quarter or month. See Harvey (1990) and Davidson and MacKinnon (1993). Some recent work has relaxed the assumption that  $d$  is an integer. The **fractionally integrated** series, or ARFIMA has been used to model series in which the very long-run multipliers decay more slowly than would be predicted otherwise.



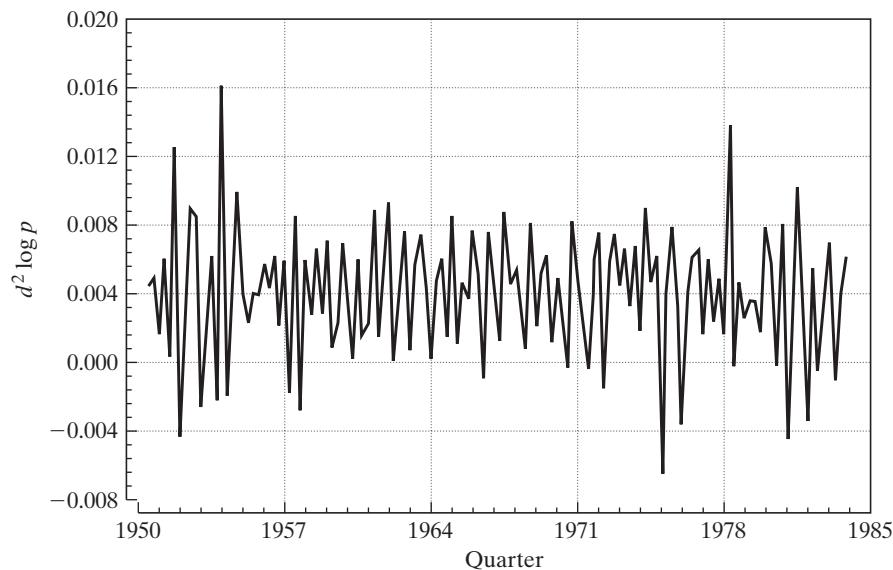
**FIGURE 23.1** Quarterly Data on log GNP Deflator.



**FIGURE 23.2** First Difference of log GNP Deflator.

### 23.2.2 RANDOM WALKS, TRENDS, AND SPURIOUS REGRESSIONS

In a seminal paper, Granger and Newbold (1974) argued that researchers had not paid sufficient attention to the warning of very high autocorrelation in the residuals from conventional regression models. Among their conclusions were that macroeconomic

**1014 PART V ♦ Time Series and Macroeometrics**

**FIGURE 23.3** Second Difference of log GNP Deflator.

**TABLE 23.1** Autocorrelations for ln GNP Deflator

Lag	Autocorrelation Function Original Series, log Price	Autocorrelation Function First Difference of log Price	Autocorrelation Function Second Difference of log Price
1	1.000	0.812	-0.395
2	1.000	0.765	-0.112
3	0.999	0.776	0.258
4	0.999	0.682	-0.101
5	0.999	0.631	-0.022
6	0.998	0.592	0.076
7	0.998	0.523	-0.163
8	0.997	0.513	0.052
9	0.997	0.488	-0.054
10	0.997	0.491	0.062

data, as a rule, were integrated and that in regressions involving the levels of such data, the standard significance tests were usually misleading. The conventional  $t$  and  $F$  tests would tend to reject the hypothesis of no relationship when, in fact, there might be none. The general result at the center of these findings is that conventional linear regression, ignoring serial correlation, of one random walk on another is virtually certain to suggest a significant relationship, even if the two are, in fact, independent. Among their extreme conclusions, Granger and Newbold suggested that researchers use a critical  $t$  value of 11.2 rather than the standard normal value of 1.96 to assess the significance of a coefficient estimate. Phillips (1986) took strong issue with this conclusion. Based on a more general model and on an analytical rather than a Monte Carlo approach, he suggested that the normalized statistic  $t_\beta/\sqrt{T}$  be used for testing purposes rather than  $t_\beta$  itself. For the 50 observations used by Granger and Newbold, the appropriate

## CHAPTER 23 ♦ Nonstationary Data 1015

critical value would be close to 15! If anything, Granger and Newbold were too optimistic.

The random walk with drift,

$$z_t = \mu + z_{t-1} + \varepsilon_t, \quad (23-1)$$

and the **trend stationary process**,

$$z_t = \mu + \beta t + \varepsilon_t, \quad (23-2)$$

where, in both cases,  $\varepsilon_t$  is a white noise process, appear to be reasonable characterizations of many macroeconomic time series.<sup>3</sup> Clearly both of these will produce strongly trended, nonstationary series,<sup>4</sup> so it is not surprising that regressions involving such variables almost always produce significant relationships. The strong correlation would seem to be a consequence of the underlying trend, whether or not there really is any regression at work. But Granger and Newbold went a step further. The intuition is less clear if there is a pure **random walk** at work,

$$z_t = z_{t-1} + \varepsilon_t, \quad (23-3)$$

but even here, they found that regression “relationships” appear to persist even in unrelated series.

Each of these three series is characterized by a **unit root**. In each case, the **data-generating process (DGP)** can be written

$$(1 - L)z_t = \alpha + v_t, \quad (23-4)$$

where  $\alpha = \mu, \beta$ , and 0, respectively, and  $v_t$  is a stationary process. Thus, the characteristic equation has a single root equal to one, hence the name. The upshot of Granger and Newbold’s and Phillips’s findings is that the use of data characterized by unit roots has the potential to lead to serious errors in inferences.

In all three settings, differencing or detrending would seem to be a natural first step. On the other hand, it is not going to be immediately obvious which is the correct way to proceed—the data are strongly trended in all three cases—and taking the incorrect approach will not necessarily improve matters. For example, first differencing in (23-1) or (23-3) produces a white noise series, but first differencing in (23-2) trades the trend for autocorrelation in the form of an MA(1) process. On the other hand, detrending—that is, computing the residuals from a regression on time—is obviously counterproductive in (23-1) and (23-3), even though the regression of  $z_t$  on a trend will appear to be significant for the reasons we have been discussing, whereas detrending in (23-2) appears to be the right approach.<sup>5</sup> Because none of these approaches is likely to be obviously preferable

<sup>3</sup>The analysis to follow has been extended to more general disturbance processes, but that complicates matters substantially. In this case, in fact, our assumption does cost considerable generality, but the extension is beyond the scope of our work. Some references on the subject are Phillips and Perron (1988) and Davidson and MacKinnon (1993).

<sup>4</sup>The constant term  $\mu$  produces the deterministic trend in the random walk with drift. For convenience, suppose that the process starts at time zero. Then  $z_t = \sum_{s=0}^t (\mu + \varepsilon_s) = \mu t + \sum_{s=0}^t \varepsilon_s$ . Thus,  $z_t$  consists of a deterministic trend plus a stochastic trend consisting of the sum of the innovations. The result is a variable with increasing variance around a linear trend.

<sup>5</sup>See Nelson and Kang (1984).

## 1016 PART V ♦ Time Series and Macroeconometrics

at the outset, some means of choosing is necessary. Consider nesting all three models in a single equation,

$$z_t = \mu + \beta t + z_{t-1} + \varepsilon_t.$$

Now subtract  $z_{t-1}$  from both sides of the equation and introduce the artificial parameter  $\gamma$ .

$$\begin{aligned} z_t - z_{t-1} &= \mu\gamma + \beta\gamma t + (\gamma - 1)z_{t-1} + \varepsilon_t \\ &= \alpha_0 + \alpha_1 t + (\gamma - 1)z_{t-1} + \varepsilon_t, \end{aligned} \tag{23-5}$$

where, by hypothesis,  $\gamma = 1$ . Equation (23-5) provides the basis for a variety of tests for unit roots in economic data. In principle, a test of the hypothesis that  $\gamma - 1$  equals zero gives confirmation of the random walk with drift, because if  $\gamma$  equals 1 (and  $\alpha_1$  equals zero), then (23-1) results. If  $\gamma - 1$  is less than zero, then the evidence favors the trend stationary (or some other) model, and detrending (or some alternative) is the preferable approach. The practical difficulty is that standard inference procedures based on least squares and the familiar test statistics are not valid in this setting. The issue is discussed in the next section.

### 23.2.3 TESTS FOR UNIT ROOTS IN ECONOMIC DATA

The implications of unit roots in macroeconomic data are, at least potentially, profound. If a structural variable, such as real output, is truly  $I(1)$ , then shocks to it will have permanent effects. If confirmed, then this observation would mandate some rather serious reconsideration of the analysis of macroeconomic policy. For example, the argument that a change in monetary policy could have a transitory effect on real output would vanish.<sup>6</sup> The literature is not without its skeptics, however. This result rests on a razor's edge. Although the literature is thick with tests that have failed to reject the hypothesis that  $\gamma = 1$ , many have also not rejected the hypothesis that  $\gamma \geq 0.95$ , and at 0.95 (or even at 0.99), the entire issue becomes moot.<sup>7</sup>

Consider the simple AR(1) model with zero-mean, white noise innovations,

$$y_t = \gamma y_{t-1} + \varepsilon_t.$$

The downward bias of the least squares estimator when  $\gamma$  approaches one has been widely documented.<sup>8</sup> For  $|\gamma| < 1$ , however, the least squares estimator

$$c = \frac{\sum_{t=2}^T y_t y_{t-1}}{\sum_{t=2}^T y_{t-1}^2}$$

does have

$$\text{plim } c = \gamma$$

---

<sup>6</sup>The 1980s saw the appearance of literally hundreds of studies, both theoretical and applied, of unit roots in economic data. An important example is the seminal paper by Nelson and Plosser (1982). There is little question but that this observation is an early part of the radical paradigm shift that has occurred in empirical macroeconomics.

<sup>7</sup>A large number of issues are raised in Maddala (1992, pp. 582–588).

<sup>8</sup>See, for example, Evans and Savin (1981, 1984).

CHAPTER 23 ♦ Nonstationary Data **1017**

and

$$\sqrt{T}(c - \gamma) \xrightarrow{d} N[0, 1 - \gamma^2].$$

Does the result hold up if  $\gamma = 1$ ? The case is called the unit root case, because in the ARMA representation  $C(L)y_t = \varepsilon_t$ , the characteristic equation  $1 - \gamma z = 0$  has one root equal to one. That the limiting variance appears to go to zero should raise suspicions. The literature on the question dates back to Mann and Wald (1943) and Rubin (1950). But for econometric purposes, the literature has a focal point at the celebrated papers of Dickey and Fuller (1979, 1981). They showed that if  $\gamma$  equals one, then

$$T(c - \gamma) \xrightarrow{d} v,$$

where  $v$  is a random variable with finite, positive variance, and in finite samples,  $E[c] < 1$ .<sup>9</sup>

There are two important implications in the Dickey–Fuller results. First, the estimator of  $\gamma$  is biased downward if  $\gamma$  equals one. Second, the OLS estimator of  $\gamma$  converges to its probability limit more rapidly than the estimators to which we are accustomed. That is, the variance of  $c$  under the null hypothesis is  $O(1/T^2)$ , not  $O(1/T)$ . (In a mean squared error sense, the OLS estimator is superconsistent.) It turns out that the implications of this finding for the regressions with trended data are considerable.

We have already observed that in some cases, differencing or detrending is required to achieve stationarity of a series. Suppose, though, that the preceding AR(1) model is fit to an  $I(1)$  series, despite that fact. The upshot of the preceding discussion is that the conventional measures will tend to hide the true value of  $\gamma$ ; the sample estimate is biased downward, and by dint of the very small *true* sampling variance, the conventional  $t$  test will tend, incorrectly, to reject the hypothesis that  $\gamma = 1$ . The practical solution to this problem devised by Dickey and Fuller was to derive, through Monte Carlo methods, an appropriate set of critical values for testing the hypothesis that  $\gamma$  equals one in an AR(1) regression when there truly is a unit root. One of their general results is that the test may be carried out using a conventional  $t$  statistic, but the critical values for the test must be revised: The standard  $t$  table is inappropriate. A number of variants of this form of testing procedure have been developed. We will consider several of them.

#### 23.2.4 THE DICKEY–FULLER TESTS

The simplest version of the model to be analyzed is the random walk,

$$y_t = \gamma y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N[0, \sigma^2], \quad \text{and} \quad \text{Cov}[\varepsilon_t, \varepsilon_s] = 0 \forall t \neq s.$$

Under the null hypothesis that  $\gamma = 1$ , there are two approaches to carrying out the test. The conventional  $t$  ratio

$$DF_\tau = \frac{\hat{\gamma} - 1}{\text{Est. Std. Error}(\hat{\gamma})}$$

---

<sup>9</sup>A full derivation of this result is beyond the scope of this book. For the interested reader, a fairly comprehensive treatment at an accessible level is given in Chapter 17 of Hamilton (1994, pp. 475–542).

**1018 PART V ♦ Time Series and Macroeconometrics**
**TABLE 23.2** Critical Values for the Dickey–Fuller  $DF_\tau$  Test

	<i>Sample Size</i>			
	<b>25</b>	<b>50</b>	<b>100</b>	<b><math>\infty</math></b>
<i>F</i> ratio (D–F) <sup>a</sup>	7.24	6.73	6.49	6.25
<i>F</i> ratio (standard)	3.42	3.20	3.10	3.00
AR model <sup>b</sup> (random walk)				
0.01	-2.66	-2.62	-2.60	-2.58
0.025	-2.26	-2.25	-2.24	-2.23
0.05	-1.95	-1.95	-1.95	-1.95
0.10	-1.60	-1.61	-1.61	-1.62
0.975	1.70	1.66	1.64	1.62
AR model with constant (random walk with drift)				
0.01	-3.75	-3.59	-3.50	-3.42
0.025	-3.33	-3.23	-3.17	-3.12
0.05	-2.99	-2.93	-2.90	-2.86
0.10	-2.64	-2.60	-2.58	-2.57
0.975	0.34	0.29	0.26	0.23
AR model with constant and time trend (trend stationary)				
0.01	-4.38	-4.15	-4.04	-3.96
0.025	-3.95	-3.80	-3.69	-3.66
0.05	-3.60	-3.50	-3.45	-3.41
0.10	-3.24	-3.18	-3.15	-3.13
0.975	-0.50	-0.58	-0.62	-0.66

<sup>a</sup>From Dickey and Fuller (1981, p. 1063). Degrees of freedom are 2 and  $T - p - 3$ .

<sup>b</sup>From Fuller (1976, p. 373 and 1996, Table 10.A.2).

with the revised set of critical values may be used for a one-sided test. Critical values for this test are shown in the top panel of Table 23.2. Note that in general, the critical value is considerably larger in absolute value than its counterpart from the  $t$  distribution. The second approach is based on the statistic

$$DF_\gamma = T(\hat{\gamma} - 1).$$

Critical values for this test are shown in the top panel of Table 23.2.

The simple random walk model is inadequate for many series. Consider the rate of inflation from 1950.2 to 2000.4 (plotted in Figure 23.4) and the log of GDP over the same period (plotted in Figure 23.5). The first of these may be a random walk, but it is clearly drifting. The log GDP series, in contrast, has a strong trend. For the first of these, a random walk with drift may be specified,

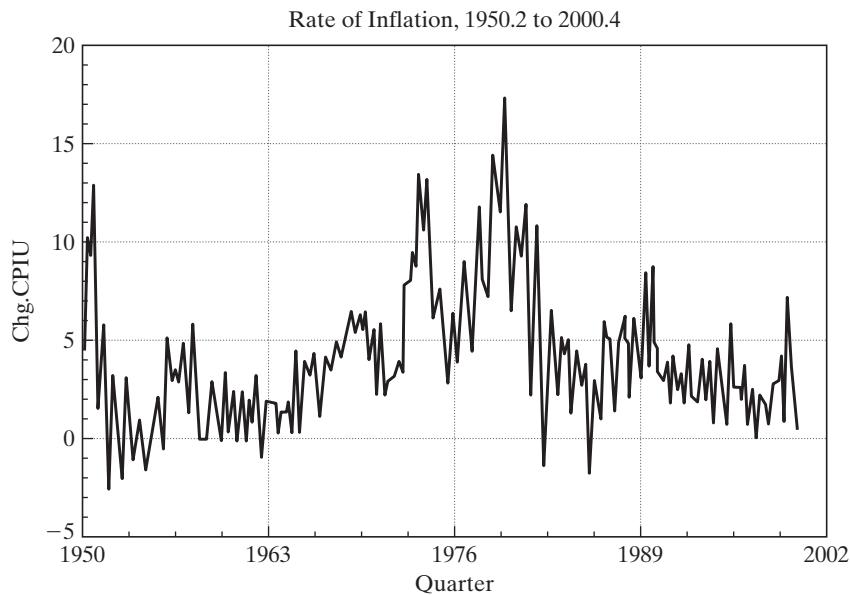
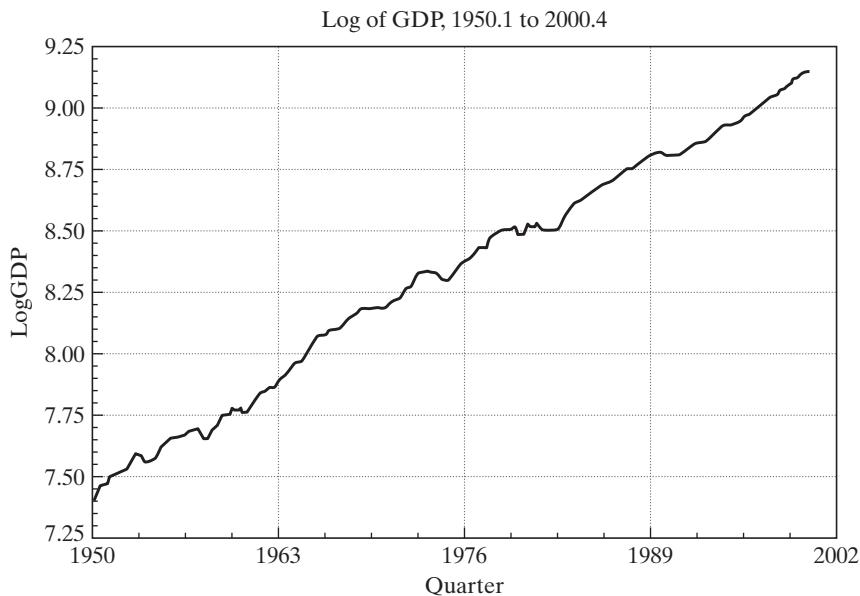
$$\begin{aligned}y_t &= \mu + z_t, \\z_t &= \gamma z_{t-1} + \varepsilon_t,\end{aligned}$$

or

$$y_t = \mu(1 - \gamma) + \gamma y_{t-1} + \varepsilon_t.$$

For the second type of series, we may specify the trend stationary form,

$$\begin{aligned}y_t &= \mu + \beta t + z_t, \\z_t &= \gamma z_{t-1} + \varepsilon_t\end{aligned}$$

CHAPTER 23 ♦ Nonstationary Data **1019****FIGURE 23.4** Rate of Inflation in the Consumer Price Index.**FIGURE 23.5** Log of Gross Domestic Product.

or

$$y_t = [\mu(1 - \gamma) + \gamma\beta] + \beta(1 - \gamma) + \gamma y_{t-1} + \varepsilon_t.$$

The tests for these forms may be carried out in the same fashion. For the model with drift only, the center panels of Tables 23.2 and 23.3 are used. When the trend is included, the lower panel of each table is used.

**1020 PART V ♦ Time Series and Macroeconometrics**
**TABLE 23.3** Critical Values for the Dickey–Fuller  $DF_\gamma$  Test

	<i>Sample Size</i>			
	<i>25</i>	<i>50</i>	<i>100</i>	$\infty$
AR model <sup>a</sup> (random walk)				
0.01	-11.8	-12.8	-13.3	-13.8
0.025	-9.3	-9.9	-10.2	-10.5
0.05	-7.3	-7.7	-7.9	-8.1
0.10	-5.3	-5.5	-5.6	-5.7
0.975	1.78	1.69	1.65	1.60
AR model with constant (random walk with drift)				
0.01	-17.2	-18.9	-19.8	-20.7
0.025	-14.6	-15.7	-16.3	-16.9
0.05	-12.5	-13.3	-13.7	-14.1
0.10	-10.2	-10.7	-11.0	-11.3
0.975	0.65	0.53	0.47	0.41
AR model with constant and time trend (trend stationary)				
0.01	-22.5	-25.8	-27.4	-29.4
0.025	-20.0	-22.4	-23.7	-24.4
0.05	-17.9	-19.7	-20.6	-21.7
0.10	-15.6	-16.8	-17.5	-18.3
0.975	-1.53	-1.667	-1.74	-1.81

<sup>a</sup>From Fuller (1976, p. 373 and 1996, Table 10.A.1).

**Example 23.2 Tests for Unit Roots**

In Section 21.6.8, we examined Cecchetti and Rich's study of the effect of recent monetary policy on the U.S. economy. The data used in their study were the following variables:

$\pi$  = one period rate of inflation = the rate of change in the CPI

$y$  = log of real GDP

$i$  = nominal interest rate = the quarterly average yield on a 90-day T-bill

$\Delta m$  = change in the log of the money stock, M1

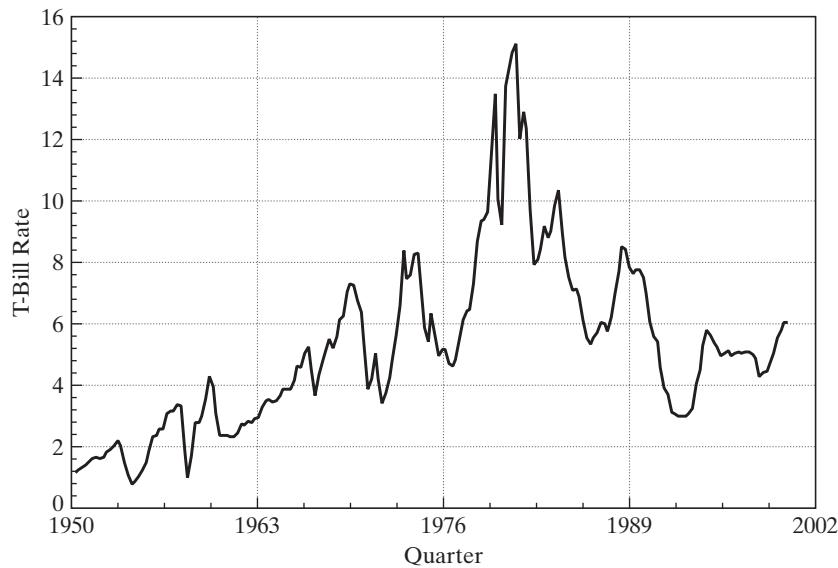
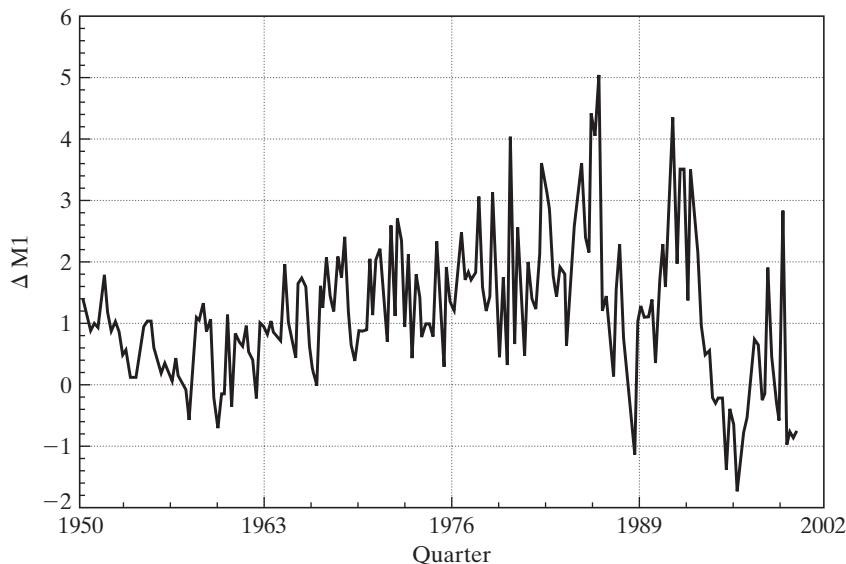
$i - \pi$  = ex post real interest rate

$\Delta m - \pi$  = real growth in the money stock

Data used in their analysis were from the period 1959.1 to 1997.4. As part of their analysis, they checked each of these series for a unit root and suggested that the hypothesis of a unit root could only be rejected for the last two variables. We will reexamine these data for the longer interval, 1950.2 to 2000.4. The data are in Appendix Table F5.2. Figures 23.6 through 23.9 show the behavior of the last four variables. The first two are shown in Figures 23.4 and 23.5. Only the real output figure shows a strong trend, so we will use the random walk with drift for all the variables except this one.

The Dickey–Fuller tests are carried out in Table 23.4. There are 203 observations used in each one. The first observation is lost when computing the rate of inflation and the change in the money stock, and one more is lost for the difference term in the regression. The critical values from interpolating to the second row, last column in each panel for 95 percent significance and a one-tailed test are -3.68 and -24.2, respectively, for  $DF_\tau$  and  $DF_\gamma$  for the output equation, which contains the time trend, and -3.14 and -16.8 for the other equations, which contain a constant but no trend. For the output equation ( $y$ ), the test statistics are

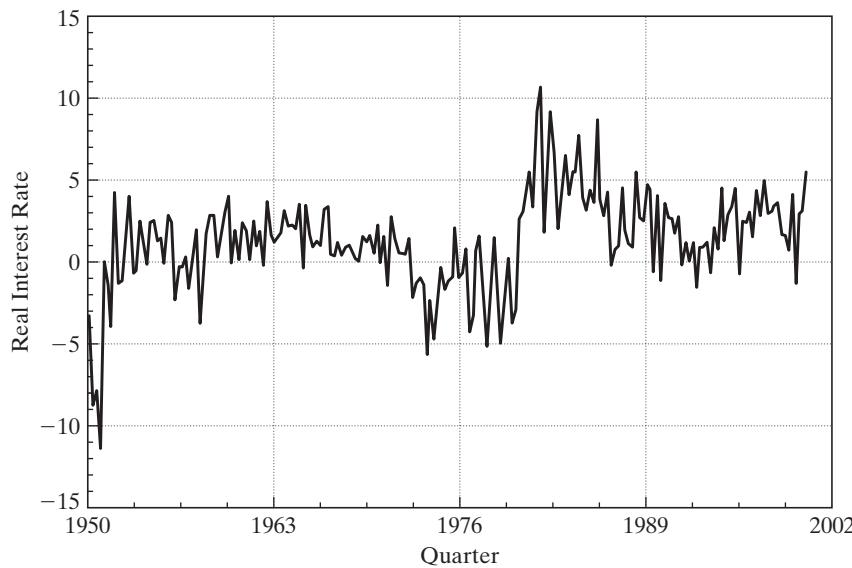
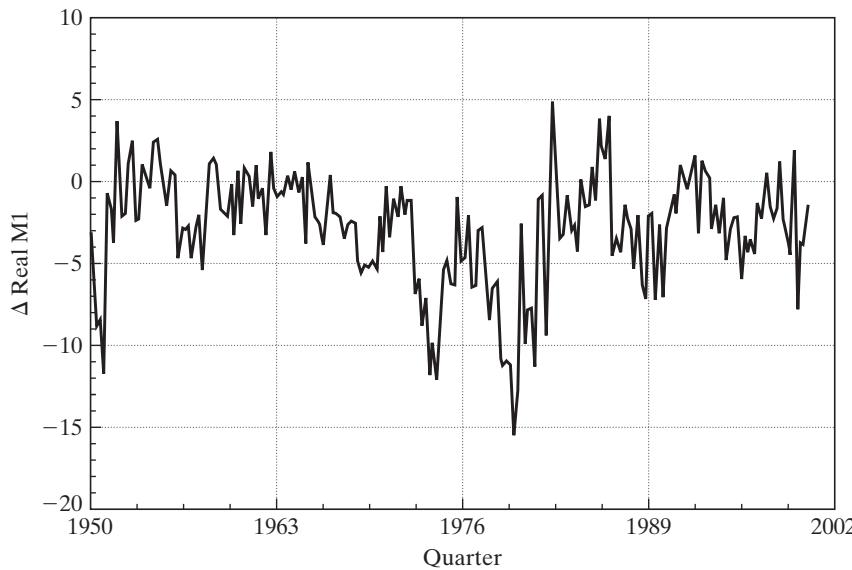
$$DF_\tau = \frac{0.9584940384 - 1}{.017880922} = -2.32 > -3.44,$$

**FIGURE 23.6** T-Bill Rate.**FIGURE 23.7** Change in the Money Stock.

and

$$DF_\gamma = 202(0.9584940384 - 1) = -8.38 > -21.2.$$

Neither is less than the critical value, so we conclude (as have others) that there is a unit root in the log GDP process. The results of the other tests are shown in Table 23.4. Surprisingly,

**1022 PART V ♦ Time Series and Macroeometrics**

**FIGURE 23.8** Ex Post Real T-Bill Rate.

**FIGURE 23.9** Change in the Real Money Stock.

these results do differ sharply from those obtained by Cecchetti and Rich (2001) for  $\pi$  and  $\Delta m$ . The sample period appears to matter; if we repeat the computation using Cecchetti and Rich's interval, 1959.4 to 1997.4, then  $DF_\tau$  equals  $-3.51$ . This is borderline, but less contradictory. For  $\Delta m$  we obtain a value of  $-4.204$  for  $DF_\tau$  when the sample is restricted to the shorter interval.

## CHAPTER 23 ♦ Nonstationary Data 1023

**TABLE 23.4** Unit Root Tests (Standard errors of estimates in parentheses)

	$\mu$	$\beta$	$\gamma$	$DF_\tau$	$DF_\gamma$	Conclusion
$\pi$	0.332 (0.0696)		0.659 (0.0532)	-6.40 $R^2 = 0.432, s = 0.643$	-68.88	Reject $H_0$
$y$	0.320 (0.134)	0.00033 (0.00015)	0.958 (0.0179)	-2.35 $R^2 = 0.999, s = 0.001$	-8.48	Do not reject $H_0$
$i$	0.228 (0.109)		0.961 (0.0182)	-2.14 $R^2 = 0.933, s = 0.743$	-7.88	Do not reject $H_0$
$\Delta m$	0.448 (0.0923)		0.596 (0.0573)	-7.05 $R^2 = 0.351, s = 0.929$	-81.61	Reject $H_0$
$i - \pi$	0.615 (0.185)		0.557 (0.0585)	-7.57 $R^2 = 0.311, s = 2.395$	-89.49	Reject $H_0$
$\Delta m - \pi$	0.0700 (0.0833)		0.490 (0.0618)	-8.25 $R^2 = 0.239, s = 1.176$	-103.02	Reject $H_0$

The Dickey–Fuller tests described in this section assume that the disturbances in the model as stated are white noise. An extension which will accommodate some forms of serial correlation is the **augmented Dickey–Fuller test**. The augmented Dickey–Fuller test is the same one as described earlier, carried out in the context of the model

$$y_t = \mu + \beta t + \gamma y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_p \Delta y_{t-p} + \varepsilon_t.$$

The random walk form is obtained by imposing  $\mu = 0$  and  $\beta = 0$ ; the random walk with drift has  $\beta = 0$ ; and the trend stationary model leaves both parameters free. The two test statistics are

$$DF_\tau = \frac{\hat{\gamma} - 1}{\text{Est. Std. Error}(\hat{\gamma})},$$

exactly as constructed before, and

$$DF_\gamma = \frac{T(\hat{\gamma} - 1)}{1 - \hat{\gamma}_1 - \cdots - \hat{\gamma}_p}.$$

The advantage of this formulation is that it can accommodate higher-order autoregressive processes in  $\varepsilon_t$ .

An alternative formulation may prove convenient. By subtracting  $y_{t-1}$  from both sides of the equation, we obtain

$$\Delta y_t = \mu + \beta t + \gamma^* y_{t-1} + \sum_{j=1}^p \phi_j \Delta y_{t-j} + \varepsilon_t,$$

where

$$\phi_j = - \sum_{k=j+1}^p \gamma_k \quad \text{and} \quad \gamma^* = \left( \sum_{i=1}^p \gamma_i \right) - 1.$$

## 1024 PART V ♦ Time Series and Macroeconometrics

The unit root test is carried out as before by testing the null hypothesis  $\gamma^* = 0$  against  $\gamma^* < 0$ .<sup>10</sup> The  $t$  test,  $DF_\tau$ , may be used. If the failure to reject the unit root is taken as evidence that a unit root is present, that is,  $\gamma^* = 0$ , then the model specializes to the AR( $p - 1$ ) model in the first differences which is an ARIMA( $p - 1, 1, 0$ ) model for  $y_t$ . For a model with a time trend,

$$\Delta y_t = \mu + \beta t + \gamma^* y_{t-1} + \sum_{j=1}^{p-1} \phi_j \Delta y_{t-j} + \varepsilon_t,$$

the test is carried out by testing the joint hypothesis that  $\beta = \gamma^* = 0$ . Dickey and Fuller (1981) present counterparts to the critical  $F$  statistics for testing the hypothesis. Some of their values are reproduced in the first row of Table 23.2. (Authors frequently focus on  $\gamma^*$  and ignore the time trend, maintaining it only as part of the appropriate formulation. In this case, one may use the simple test of  $\gamma^* = 0$  as before, with the  $DF_\tau$  critical values.)

The lag length,  $p$ , remains to be determined. As usual, we are well advised to test down to the right value instead of up. One can take the familiar approach and sequentially examine the  $t$  statistic on the last coefficient—the usual  $t$  test is appropriate. An alternative is to combine a measure of model fit, such as the regression  $s^2$  with one of the information criteria. The Akaike and Schwarz (Bayesian) information criteria would produce the two information measures

$$IC(p) = \ln \left( \frac{\mathbf{e}'\mathbf{e}}{T - p_{\max} - K^*} \right) + (p + K^*) \left( \frac{A^*}{T - p_{\max} - K^*} \right),$$

$K^* = 1$  for random walk, 2 for random walk with drift, 3 for trend stationary,

$A^* = 2$  for Akaike criterion,  $\ln(T - p_{\max} - K^*)$  for Bayesian criterion,

$p_{\max}$  = the largest lag length being considered.

The remaining detail is to decide upon  $p_{\max}$ . The theory provides little guidance here. On the basis of a large number of simulations, Schwert (1989) found that

$$p_{\max} = \text{integer part of } [12 \times (T/100)^{.25}]$$

gave good results.

Many alternatives to the Dickey–Fuller tests have been suggested, in some cases to improve on the finite sample properties and in others to accommodate more general modeling frameworks. The Phillips (1987) and Phillips and Perron (1988) statistic may be computed for the same three functional forms,

$$y_t = \delta_t + \gamma y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_p \Delta y_{t-p} + \varepsilon_t, \quad (23-6)$$

where  $\delta_t$  may be 0,  $\mu$ , or  $\mu + \beta t$ . The procedure modifies the two Dickey–Fuller statistics we previously examined:

$$Z_\tau = \sqrt{\frac{c_0}{a}} \left( \frac{\hat{\gamma} - 1}{v} \right) - \frac{1}{2}(a - c_0) \frac{T v}{\sqrt{as^2}},$$

$$\mathbf{Z}_\gamma = \frac{T(\hat{\gamma} - 1)}{1 - \hat{\gamma}_1 - \cdots - \hat{\gamma}_p} - \frac{1}{2} \left( \frac{T^2 v^2}{s^2} \right)(a - c_0),$$

---

<sup>10</sup>It is easily verified that one of the roots of the characteristic polynomial is  $1/(\gamma_1 + \gamma_2 + \cdots + \gamma_p)$ .

## CHAPTER 23 ♦ Nonstationary Data 1025

where

$$s^2 = \frac{\sum_{t=1}^T e_t^2}{T - K},$$

$v^2$  = estimated asymptotic variance of  $\hat{\gamma}$ ,

$$c_j = \frac{1}{T} \sum_{s=j+1}^T e_s e_{t-s}, \quad j = 0, \dots, L = j\text{th autocovariance of residuals},$$

$$c_0 = [(T - K)/T]s^2,$$

$$a = c_0 + 2 \sum_{j=1}^L \left(1 - \frac{j}{L+1}\right) c_j.$$

[Note the Newey–West (Bartlett) weights in the computation of  $a$ . As before, the analyst must choose  $L$ .] The test statistics are referred to the same Dickey–Fuller tables we have used before.

Elliot, Rothenberg, and Stock (1996) have proposed a method they denote the ADF-GLS procedure, which is designed to accommodate more general formulations of  $\varepsilon$ ; the process generating  $\varepsilon_t$  is assumed to be an  $I(0)$  stationary process, possibly an ARMA( $r, s$ ). The null hypothesis, as before, is  $\gamma = 1$  in (23-6) where  $\delta_t = \mu$  or  $\mu + \beta t$ . The method proceeds as follows:

**Step 1.** Linearly regress

$$\mathbf{y}^* = \begin{bmatrix} y_1 \\ y_2 - \bar{r}y_1 \\ \dots \\ y_T - \bar{r}y_{T-1} \end{bmatrix} \quad \text{on} \quad \mathbf{X}^* = \begin{bmatrix} 1 \\ 1 - \bar{r} \\ \dots \\ 1 - \bar{r} \end{bmatrix} \quad \text{or} \quad \mathbf{X}^* = \begin{bmatrix} 1 & 1 \\ 1 - \bar{r} & 2 - \bar{r} \\ \dots & \dots \\ 1 - \bar{r} & T - \bar{r}(T-1) \end{bmatrix}$$

for the random walk with drift and trend stationary cases, respectively. (Note that the second column of the matrix is simply  $\bar{r} + (1 - \bar{r})t$ .) Compute the residuals from this regression,  $\tilde{y}_t = y_t - \hat{\delta}_t$ .  $\bar{r} = 1 - 7/T$  for the random walk model and  $1 - 13.5/T$  for the model with a trend.

**Step 2.** The Dickey–Fuller DF<sub>t</sub> test can now be carried out using the model

$$\tilde{y}_t = \gamma \tilde{y}_{t-1} + \gamma_1 \Delta \tilde{y}_{t-1} + \dots + \gamma_p \Delta \tilde{y}_{t-p} + \eta_t.$$

If the model does not contain the time trend, then the  $t$  statistic for  $(\gamma - 1)$  may be referred to the critical values in the center panel of Table 23.2. For the trend stationary model, the critical values are given in a table presented in Elliot et al. The 97.5 percent critical values for a one-tailed test from their table is  $-3.15$ .

As in many such cases of a new technique, as researchers develop large and small modifications of these tests, the practitioner is likely to have some difficulty deciding how to proceed. The Dickey–Fuller procedures have stood the test of time as robust tools that appear to give good results over a wide range of applications. The **Phillips–Perron**

## 1026 PART V ♦ Time Series and Macroeconometrics

**tests** are very general but appear to have less than optimal small sample properties. Researchers continue to examine it and the others such as Elliot et al. method. Other tests are catalogued in Maddala and Kim (1998).

### Example 23.3 Augmented Dickey–Fuller Test for a Unit Root in GDP

The Dickey–Fuller 1981 JASA paper is a classic in the econometrics literature—it is probably the single most frequently cited paper in the field. It seems appropriate, therefore, to revisit at least some of their work. Dickey and Fuller apply their methodology to a model for the log of a quarterly series on output, the Federal Reserve Board Production Index. The model used is

$$y_t = \mu + \beta t + \gamma y_{t-1} + \phi(y_{t-1} - y_{t-2}) + \varepsilon_t. \quad (23-7)$$

The test is carried out by testing the joint hypothesis that both  $\beta$  and  $\gamma^*$  are zero in the model

$$y_t - y_{t-1} = \mu^* + \beta t + \gamma^* y_{t-1} + \phi(y_{t-1} - y_{t-2}) + \varepsilon_t.$$

(If  $\gamma = 0$ , then  $\mu^*$  will also by construction.) We will repeat the study with our data on real GDP from Appendix Table F5.1 using observations 1950.1 to 2000.4.

We will use the augmented Dickey–Fuller test first. Thus, the first step is to determine the appropriate lag length for the augmented regression. Using Schwert's suggestion, we find that the maximum lag length should be allowed to reach  $p_{\max} = \{\text{the integer part of } 12[204/100]\}^{25} = 14$ . The specification search uses observations 18 to 204, because as many as 17 coefficients will be estimated in the equation

$$y_t = \mu + \beta t + \gamma y_{t-1} + \sum_{j=1}^p \gamma_j \Delta y_{t-j} + \varepsilon_t.$$

In the sequence of 14 regressions with  $j = 14, 13, \dots$ , the only statistically significant lagged difference is the first one, in the last regression, so it would appear that the model used by Dickey and Fuller would be chosen on this basis. The two information criteria produce a similar conclusion. Both of them decline monotonically from  $j = 14$  all the way down to  $j = 1$ , so on this basis, we end the search with  $j = 1$ , and proceed to analyze Dickey and Fuller's model.

The linear regression results for the equation in (23-7) are

$$\begin{aligned} y_t &= 0.368 + 0.000391t + 0.952y_{t-1} + 0.36025\Delta y_{t-1} + \varepsilon_t, & s &= 0.00912 \\ (0.125) & (0.000138) & (0.0167) & (0.0647) & R^2 &= 0.999647. \end{aligned}$$

The two test statistics are

$$DF_\tau = \frac{0.95166 - 1}{0.016716} = -2.892$$

and

$$DF_\gamma = \frac{201(0.95166 - 1)}{1 - 0.36025} = -15.263.$$

Neither statistic is less than the respective critical values, which are  $-3.70$  and  $-24.5$ . On this basis, we conclude, as have many others, that there is a unit root in log GDP.

For the Phillips and Perron statistic, we need several additional intermediate statistics. Following Hamilton (1994, p. 512), we choose  $L = 4$  for the long-run variance calculation. Other values we need are  $T = 202$ ,  $\dot{y} = 0.9516613$ ,  $s^2 = 0.00008311488$ ,  $v^2 = 0.00027942647$ , and the first five autocovariances,  $c_0 = 0.000081469$ ,  $c_1 = -0.00000351162$ ,  $c_2 = 0.00000688053$ ,  $c_3 = 0.000000597305$ , and  $c_4 = -0.00000128163$ . Applying these to the weighted sum produces  $a = 0.0000840722$ , which is only a minor correction to  $c_0$ . Collecting the results, we obtain the Phillips–Perron statistics,  $Z_\tau = -2.89921$  and  $Z_\gamma = -15.44133$ . Because these are applied to the same critical values in the Dickey–Fuller tables, we reach the same conclusion as before—we do not reject the hypothesis of a unit root in log GDP.

## CHAPTER 23 ♦ Nonstationary Data 1027

## 23.2.5 THE KPSS TEST OF STATIONARITY

Kwiatkowski et al. (1992) (KPSS) have devised an alternative to the Dickey–Fuller test for stationarity of a time series. The procedure is a test of nonstationarity against the null hypothesis of stationarity in the model

$$\begin{aligned} y_t &= \alpha + \beta t + \gamma \sum_{i=1}^t z_i + \varepsilon_t, \quad t = 1, \dots, T \\ &= \alpha + \beta t + \gamma \mathbf{Z}_t + \varepsilon_t, \end{aligned}$$

where  $\varepsilon_t$  is a stationary series and  $z_t$  is an i.i.d. stationary series with mean zero and variance one. (These are merely convenient normalizations because a nonzero mean would move to  $\alpha$  and a nonunit variance is absorbed in  $\gamma$ .) If  $\gamma$  equals zero, then the process is stationary if  $\beta = 0$  and trend stationary if  $\beta \neq 0$ . Because  $\mathbf{Z}_t$  is  $I(1)$ ,  $y_t$  is nonstationary if  $\gamma$  is nonzero.

The KPSS test of the null hypothesis,  $H_0: \gamma = 0$ , against the alternative that  $\gamma$  is nonzero reverses the strategy of the Dickey–Fuller statistic (which tests the null hypothesis  $\gamma < 1$  against the alternative  $\gamma = 1$ ). Under the null hypothesis,  $\alpha$  and  $\beta$  can be estimated by OLS. Let  $e_t$  denote the  $t$ th OLS residual,

$$e_t = y_t - \hat{\alpha} - \hat{\beta}t,$$

and let the sequence of partial sums be

$$E_t = \sum_{i=1}^t e_i, \quad t = 1, \dots, T.$$

(Note  $E_T = 0$ .) The KPSS statistic is

$$\text{KPSS} = \frac{\sum_{t=1}^T E_t^2}{T^2 \hat{\sigma}^2},$$

where

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^T e_t^2}{T} + 2 \sum_{j=1}^L \left( 1 - \frac{j}{L+1} \right) r_j \quad \text{and} \quad r_j = \frac{\sum_{s=j+1}^T e_s e_{s-j}}{T},$$

and  $L$  is chosen by the analyst. [See (20-17).] Under normality of the disturbances,  $\varepsilon_t$ , the KPSS statistic is an LM statistic. The authors derive the statistic under more general conditions. Critical values for the test statistic are estimated by simulation. Table 23.5 gives the values reported by the authors (in their Table 1, p. 166).

**Example 23.4 Is There a Unit Root in GDP?**

Using the data used for the Dickey–Fuller tests in Example 23.3, we repeated the procedure using the KPSS test with  $L = 10$ . The two statistics are 1.953 without the trend and 0.312

TABLE 23.5 Critical Values for the KPSS Test

<i>Critical Value</i>	<i>Upper Tail Percentiles</i>			
	<b>0.100</b>	<b>0.050</b>	<b>0.025</b>	<b>0.010</b>
$\beta = 0$	0.347	0.463	0.573	0.739
$\beta \neq 0$	0.119	0.146	0.176	0.216

## 1028 PART V ♦ Time Series and Macroeometrics

with it. Comparing these results to the values in Table 23.4 we conclude (again) that there is, indeed, a unit root in  $\ln \text{GDP}$ . Or, more precisely, we conclude that  $\ln \text{GDP}$  is not a stationary series, nor even a trend stationary series.

### 23.3 COINTEGRATION

Studies in empirical macroeconomics almost always involve nonstationary and trending variables, such as income, consumption, money demand, the price level, trade flows, and exchange rates. Accumulated wisdom and the results of the previous sections suggest that the appropriate way to manipulate such series is to use differencing and other transformations (such as seasonal adjustment) to reduce them to stationarity and then to analyze the resulting series as VARs or with the methods of Box and Jenkins. But recent research and a growing literature has shown that there are more interesting, appropriate ways to analyze trending variables.

In the *fully specified* regression model

$$y_t = \beta x_t + \varepsilon_t,$$

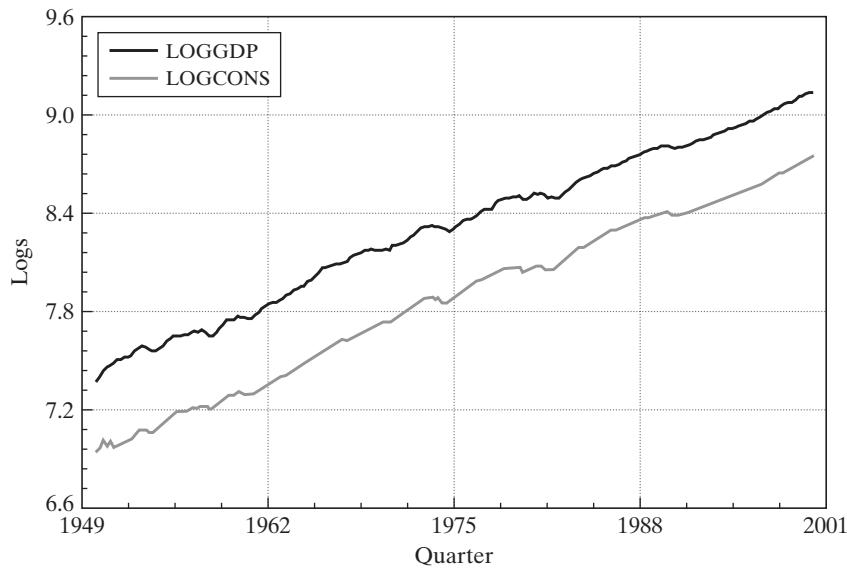
there is a presumption that the disturbances  $\varepsilon_t$  are a stationary, white noise series.<sup>11</sup> But this presumption is unlikely to be true if  $y_t$  and  $x_t$  are integrated series. Generally, if two series are integrated to different orders, then linear combinations of them will be integrated to the higher of the two orders. Thus, if  $y_t$  and  $x_t$  are  $I(1)$ —that is, if both are trending variables—then we would normally expect  $y_t - \beta x_t$  to be  $I(1)$  regardless of the value of  $\beta$ , not  $I(0)$  (i.e., not stationary). If  $y_t$  and  $x_t$  are each drifting upward with their own trend, then unless there is some relationship between those trends, the difference between them should also be growing, with yet another trend. There must be some kind of inconsistency in the model. On the other hand, if the two series are both  $I(1)$ , then there *may* be a  $\beta$  such that

$$\varepsilon_t = y_t - \beta x_t$$

is  $I(0)$ . Intuitively, if the two series are both  $I(1)$ , then this partial difference between them might be stable around a fixed mean. The implication would be that the series are drifting together at roughly the same rate. Two series that satisfy this requirement are said to be **cointegrated**, and the vector  $[1, -\beta]$  (or any multiple of it) is a **cointegrating vector**. In such a case, we can distinguish between a long-run relationship between  $y_t$  and  $x_t$ , that is, the manner in which the two variables drift upward together, and the short-run dynamics, that is, the relationship between deviations of  $y_t$  from its long-run trend and deviations of  $x_t$  from its long-run trend. If this is the case, then differencing of the data would be counterproductive, since it would obscure the long-run relationship between  $y_t$  and  $x_t$ . Studies of cointegration and a related technique, **error correction**, are concerned with methods of estimation that preserve the information about both forms of covariation.<sup>12</sup>

<sup>11</sup>If there is autocorrelation in the model, then it has been removed through an appropriate transformation.

<sup>12</sup>See, for example, Engle and Granger (1987) and the lengthy literature cited in Hamilton (1994). A survey paper on VARs and cointegration is Watson (1994).



**FIGURE 23.10** Cointegrated Variables: Logs of Consumption and GDP.

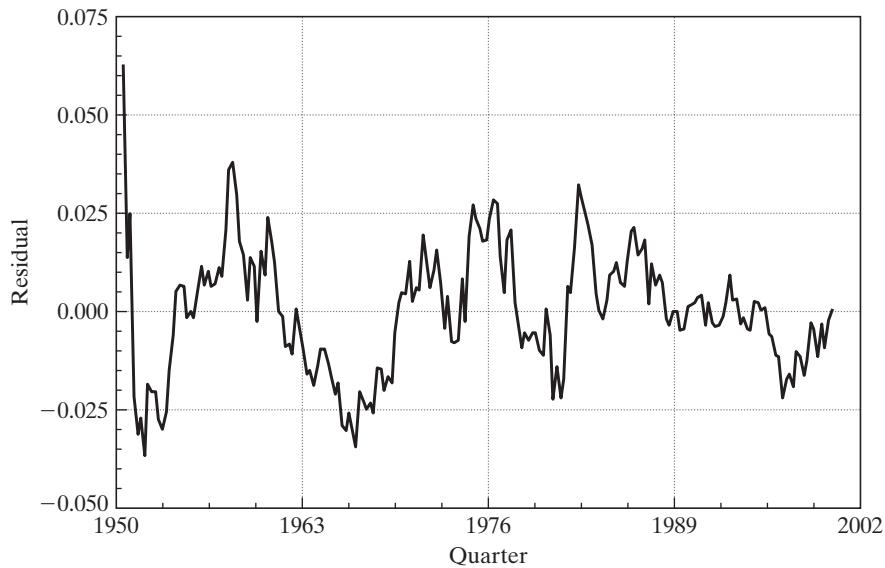
#### Example 23.5 Cointegration in Consumption and Output

Consumption and income provide one of the more familiar examples of the phenomenon described previously. The logs of GDP and consumption for 1950.1 to 2000.4 are plotted in Figure 23.10. Both variables are obviously nonstationary. We have already verified that there is a unit root in the income data. We leave as an exercise for the reader to verify that the consumption variable is likewise  $I(1)$ . Nonetheless, there is a clear relationship between consumption and output. To see where this discussion of relationships among variables is going, consider a simple regression of the log of consumption on the log of income, where both variables are manipulated in mean deviation form (so, the regression includes a constant). The slope in that regression is 1.056765. The residuals from the regression,  $u_t = [\ln \text{InCons}^*, \ln \text{GDP}^*][1, -1.056765]'$  (where the “\*” indicates mean deviations) are plotted in Figure 23.11. The trend is clearly absent from the residuals. But, it remains to verify whether the series of residuals is stationary. In the ADF regression of the least squares residuals on a constant (random walk with drift), the lagged value and the lagged first difference, the coefficient on  $u_{t-1}$  is 0.838488 (0.0370205) and that on  $u_{t-1} - u_{t-2}$  is -0.098522. (The constant differs trivially from zero because two observations are lost in computing the ADF regression.) With 202 observations, we find  $DF_\tau = -4.63$  and  $DF_\gamma = -29.55$ . Both are well below the critical values, which suggests that the residual series does not contain a unit root. We conclude (at least it appears so) that even after accounting for the trend, although neither of the original variables is stationary, there is a linear combination of them that is. If this conclusion holds up after a more formal treatment of the testing procedure, we will state that logGDP and log consumption are cointegrated.

#### Example 23.6 Several Cointegrated Series

The theory of purchasing power parity specifies that in long-run equilibrium, exchange rates will adjust to erase differences in purchasing power across different economies. Thus, if  $p_1$  and  $p_0$  are the price levels in two countries and  $E$  is the exchange rate between the two currencies, then in equilibrium,

$$v_t = E_t \frac{p_1}{p_0} = \mu, \quad \text{a constant.}$$

**1030 PART V ♦ Time Series and Macroeconometrics**


**FIGURE 23.11** Residuals from Consumption—Income Regression.

The price levels in any two countries are likely to be strongly trended. But allowing for short-term deviations from equilibrium, the theory suggests that for a particular  $\beta = (\ln \mu, -1, 1)$ , in the model

$$\ln E_t = \beta_1 + \beta_2 \ln p_{1t} + \beta_3 \ln p_{0t} + \varepsilon_t,$$

$\varepsilon_t = \ln v_t$  would be a stationary series, which would imply that the logs of the three variables in the model are cointegrated.

We suppose that the model involves  $M$  variables,  $\mathbf{y}_t = [y_{1t}, \dots, y_{Mt}]'$ , which individually may be  $I(0)$  or  $I(1)$ , and a long-run equilibrium relationship,

$$\mathbf{y}'_t \boldsymbol{\gamma} - \mathbf{x}'_t \boldsymbol{\beta} = 0.$$

The “regressors” may include a constant, exogenous variables assumed to be  $I(0)$ , and/or a time trend. The vector of parameters  $\boldsymbol{\gamma}$  is the cointegrating vector. In the short run, the system may deviate from its equilibrium, so the relationship is rewritten as

$$\mathbf{y}'_t \boldsymbol{\gamma} - \mathbf{x}'_t \boldsymbol{\beta} = \varepsilon_t,$$

where the **equilibrium error**  $\varepsilon_t$  must be a stationary series. In fact, because there are  $M$  variables in the system, at least in principle, there could be more than one cointegrating vector. In a system of  $M$  variables, there can only be up to  $M - 1$  linearly independent cointegrating vectors. A proof of this proposition is very simple, but useful at this point.

**Proof:** Suppose that  $\boldsymbol{\gamma}_i$  is a cointegrating vector and that there are  $M$  linearly independent cointegrating vectors. Then, neglecting  $\mathbf{x}'_t \boldsymbol{\beta}$  for the moment, for every  $\boldsymbol{\gamma}_i$ ,  $\mathbf{y}'_t \boldsymbol{\gamma}_i$  is a stationary series  $v_{it}$ . Any linear combination of a set of stationary series is stationary, so it follows that every linear combination of the cointegrating vectors is also a cointegrating vector. If there are  $M$  such  $M \times 1$

## CHAPTER 23 ♦ Nonstationary Data 1031

linearly independent vectors, then they form a basis for the  $M$ -dimensional space, so any  $M \times 1$  vector can be formed from these cointegrating vectors, including the columns of an  $M \times M$  identity matrix. Thus, the first column of an identity matrix would be a cointegrating vector, or  $y_{t1}$  is  $I(0)$ . This result is a contradiction, because we are allowing  $y_{t1}$  to be  $I(1)$ . It follows that there can be at most  $M - 1$  cointegrating vectors.

The number of linearly independent cointegrating vectors that exist in the equilibrium system is called its **cointegrating rank**. The cointegrating rank may range from 1 to  $M - 1$ . If it exceeds one, then we will encounter an interesting identification problem. As a consequence of the observation in the preceding proof, we have the unfortunate result that, in general, *if the cointegrating rank of a system exceeds one*, then without out-of-sample, *exact* information, it is not possible to estimate behavioral relationships as cointegrating vectors. Enders (1995) provides a useful example.

**Example 23.7 Multiple Cointegrating Vectors**

We consider the logs of four variables, money demand  $m$ , the price level  $p$ , real income  $y$ , and an interest rate  $r$ . The basic relationship is

$$m = \gamma_0 + \gamma_1 p + \gamma_2 y + \gamma_3 r + \varepsilon.$$

The price level and real income are assumed to be  $I(1)$ . The existence of long-run equilibrium in the money market implies a cointegrating vector  $\alpha_1$ . If the Fed follows a certain feedback rule, increasing the money stock when *nominal* income ( $y + p$ ) is low and decreasing it when nominal income is high—which might make more sense in terms of rates of growth—then there is a second cointegrating vector in which  $\gamma_1 = \gamma_2$  and  $\gamma_3 = 0$ . Suppose that we label this vector  $\alpha_2$ . The parameters in the money demand equation, notably the interest elasticity, are interesting quantities, and we might seek to estimate  $\alpha_1$  to learn the value of this quantity. But since every linear combination of  $\alpha_1$  and  $\alpha_2$  is a cointegrating vector, to this point we are only able to estimate a hash of the two cointegrating vectors.

In fact, the parameters of this model are identifiable from sample information (in principle). We have specified two cointegrating vectors,

$$\alpha_1 = [1, -\gamma_{10}, -\gamma_{11}, -\gamma_{12}, -\gamma_{13}]$$

and

$$\alpha_2 = [1, -\gamma_{20}, \gamma_{21}, \gamma_{21}, 0]'$$

Although it is true that every linear combination of  $\alpha_1$  and  $\alpha_2$  is a cointegrating vector, only the original two vectors, as they are, have a 1 in the first position of both and a 0 in the last position of the second. (The equality restriction actually overidentifies the parameter matrix.) This result is, of course, exactly the sort of analysis that we used in establishing the identifiability of a simultaneous equations system.

### 23.3.1 COMMON TRENDS

If two  $I(1)$  variables are cointegrated, then some linear combination of them is  $I(0)$ . Intuition should suggest that the linear combination does not mysteriously create a well-behaved new variable; rather, something present in the original variables must be missing from the aggregated one. Consider an example. Suppose that two  $I(1)$  variables have a linear trend,

$$y_{1t} = \alpha + \beta t + u_t,$$

$$y_{2t} = \gamma + \delta t + v_t,$$

## 1032 PART V ♦ Time Series and Macroeconometrics

where  $u_t$  and  $v_t$  are white noise. A linear combination of  $y_{1t}$  and  $y_{2t}$  with vector  $(1, \theta)$  produces the new variable,

$$z_t = (\alpha + \theta\gamma) + (\beta + \theta\delta)t + u_t + \theta v_t,$$

which, in general, is still  $I(1)$ . In fact, the only way the  $z_t$  series can be made stationary is if  $\theta = -\beta/\delta$ . If so, then the effect of combining the two variables linearly is *to remove the common linear trend*, which is the basis of Stock and Watson's (1988) analysis of the problem. But their observation goes an important step beyond this one. *The only way that  $y_{1t}$  and  $y_{2t}$  can be cointegrated to begin with is if they have a common trend of some sort.* To continue, suppose that instead of the linear trend  $t$ , the terms on the right-hand side,  $y_1$  and  $y_2$ , are functions of a random walk,  $w_t = w_{t-1} + \eta_t$ , where  $\eta_t$  is white noise. The analysis is identical. But now suppose that each variable  $y_{it}$  has its own random walk component  $w_{it}$ ,  $i = 1, 2$ . Any linear combination of  $y_{1t}$  and  $y_{2t}$  must involve *both* random walks. It is clear that they cannot be cointegrated unless, in fact,  $w_{1t} = w_{2t}$ . That is, once again, they must have a **common trend**. Finally, suppose that  $y_{1t}$  and  $y_{2t}$  share two common trends,

$$\begin{aligned} y_{1t} &= \alpha + \beta t + \lambda w_t + u_t, \\ y_{2t} &= \gamma + \delta t + \pi w_t + v_t. \end{aligned}$$

We place no restriction on  $\lambda$  and  $\pi$ . Then, a bit of manipulation will show that it is not possible to find a linear combination of  $y_{1t}$  and  $y_{2t}$  that is cointegrated, even though they share common trends. The end result for this example is that if  $y_{1t}$  and  $y_{2t}$  are cointegrated, then they must share exactly one common trend.

As Stock and Watson determined, the preceding is the crux of the cointegration of economic variables. A set of  $M$  variables that are cointegrated can be written as a stationary component plus linear combinations of a smaller set of common trends. If the cointegrating rank of the system is  $r$ , then there can be up to  $M - r$  linear trends and  $M - r$  common random walks. [See Hamilton (1994, p. 578).] (The two-variable case is special. In a two-variable system, there can be only one common trend in total.) The effect of the cointegration is to purge these common trends from the resultant variables.

### 23.3.2 ERROR CORRECTION AND VAR REPRESENTATIONS

Suppose that the two  $I(1)$  variables  $y_t$  and  $z_t$  are cointegrated and that the cointegrating vector is  $[1, -\theta]$ . Then all three variables,  $\Delta y_t = y_t - y_{t-1}$ ,  $\Delta z_t$ , and  $(y_t - \theta z_t)$  are  $I(0)$ . The **error correction model**

$$\Delta y_t = \mathbf{x}'_t \boldsymbol{\beta} + \gamma(\Delta z_t) + \lambda(y_{t-1} - \theta z_{t-1}) + \varepsilon_t$$

describes the variation in  $y_t$  around its long-run trend in terms of a set of  $I(0)$  exogenous factors  $\mathbf{x}_t$ , the variation of  $z_t$  around its long-run trend, and the error correction  $(y_t - \theta z_t)$ , which is the equilibrium error in the model of cointegration. There is a tight connection between models of cointegration and models of error correction. The model in this form is reasonable as it stands, but in fact, it is only internally consistent if the two variables are cointegrated. If not, then the third term, and hence the right-hand side, cannot be  $I(0)$ , even though the left-hand side must be. The upshot is that the same assumption

## CHAPTER 23 ♦ Nonstationary Data 1033

that we make to produce the cointegration implies (and is implied by) the existence of an error correction model.<sup>13</sup> As we will examine in the next section, the utility of this representation is that it suggests a way to build an elaborate model of the long-run variation in  $y_t$  as well as a test for cointegration. Looking ahead, the preceding suggests that residuals from an estimated cointegration model—that is, estimated equilibrium errors—can be included in an elaborate model of the long-run covariation of  $y_t$  and  $z_t$ . Once again, we have the foundation of Engel and Granger's approach to analyzing cointegration.

Pesaran, Shin, and Smith (2001) suggest a method of testing for a relationship in levels between a  $y_t$  and  $\mathbf{x}_t$  when there exits significant lags in the error correction form. Their **bounds test** accommodates the possibility that the regressors may be trend or difference stationary. The critical values they provide give a band that covers the polar cases in which all regressors are  $I(0)$ , or are  $I(1)$ , or are mutually cointegrated. The statistic is able to test for the existence of a levels equation regardless of whether the variables are  $I(0)$ ,  $I(1)$ , or are cointegrated. In their application,  $y_t$  is real earnings in the UK while  $\mathbf{x}_t$  includes a measure of productivity, the unemployment rate, unionization of the workforce, a “replacement ratio” that measures the difference between unemployment benefits and real wages, and a “wedge” between the real product wage and the real consumption wage. It is found that wages and productivity have unit roots. The issue then is to discern whether unionization, the wedge, and the unemployment rate, which might be  $I(0)$ , have level effects in the model.

Consider the VAR representation of the model

$$\mathbf{y}_t = \boldsymbol{\Gamma} \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t,$$

where the vector  $\mathbf{y}_t$  is  $[y_t, z_t]'$ . Now take first differences to obtain

$$\mathbf{y}_t - \mathbf{y}_{t-1} = (\boldsymbol{\Gamma} - \mathbf{I}) \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t,$$

or

$$\Delta \mathbf{y}_t = \boldsymbol{\Pi} \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t.$$

If all variables are  $I(1)$ , then all  $M$  variables on the left-hand side are  $I(0)$ . Whether those on the right-hand side are  $I(0)$  remains to be seen. The matrix  $\boldsymbol{\Pi}$  produces linear combinations of the variables in  $\mathbf{y}_t$ . But as we have seen, not all linear combinations can be cointegrated. The number of such independent linear combinations is  $r < M$ . Therefore, although there must be a VAR representation of the model, cointegration implies a restriction on the rank of  $\boldsymbol{\Pi}$ . It cannot have full rank; its rank is  $r$ . From another viewpoint, a different approach to discerning cointegration is suggested. Suppose that we estimate this model as an unrestricted VAR. The resultant coefficient matrix should be short-ranked. The implication is that if we fit the VAR model and impose short rank on the coefficient matrix as a restriction—how we could do that remains to be seen—then if the variables really are cointegrated, this restriction should not lead to a loss of fit. This implication is the basis of Johansen's (1988) and Stock and Watson's (1988) analysis of cointegration.

<sup>13</sup>The result in its general form is known as the Granger representation theorem. See Hamilton (1994, p. 582).

## 1034 PART V ♦ Time Series and Macroeconometrics

### 23.3.3 TESTING FOR COINTEGRATION

A natural first step in the analysis of cointegration is to establish that it is indeed a characteristic of the data. Two broad approaches for testing for cointegration have been developed. The Engle and Granger (1987) method is based on assessing whether single-equation estimates of the equilibrium errors appear to be stationary. The second approach, due to Johansen (1988, 1991) and Stock and Watson (1988), is based on the VAR approach. As noted earlier, if a set of variables is truly cointegrated, then we should be able to detect the implied restrictions in an otherwise unrestricted VAR. We will examine these two methods in turn.

Let  $\mathbf{y}_t$  denote the set of  $M$  variables that are believed to be cointegrated. Step one of either analysis is to establish that the variables are indeed integrated to the same order. The Dickey–Fuller tests discussed in Section 23.2.4 can be used for this purpose. If the evidence suggests that the variables are integrated to different orders or not at all, then the specification of the model should be reconsidered.

If the cointegration rank of the system is  $r$ , then there are  $r$  independent vectors,  $\mathbf{y}_i = [1, -\boldsymbol{\theta}_i]$ , where each vector is distinguished by being normalized on a different variable. If we suppose that there are also a set of  $I(0)$  exogenous variables, including a constant, in the model, then each cointegrating vector produces the equilibrium relationship

$$\mathbf{y}'_i \mathbf{y}_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_{it},$$

which we may rewrite as

$$y_{it} = \mathbf{Y}'_{it} \boldsymbol{\theta}_i + \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_{it}.$$

We can obtain estimates of  $\boldsymbol{\theta}_i$  by least squares regression. If the theory is correct *and* if this OLS estimator is consistent, then residuals from this regression should estimate the equilibrium errors. There are two obstacles to consistency. First, because both sides of the equation contain  $I(1)$  variables, the problem of spurious regressions appears. Second, a moment's thought should suggest that what we have done is extract an equation from an otherwise ordinary simultaneous equations model and ~~try to~~ estimate its parameters by ordinary least squares. As we examined in Chapter 8, consistency is unlikely in that case. It is one of the extraordinary results of this body of theory that in this setting, neither of these considerations is a problem. In fact, as shown by a number of authors [see, e.g., Davidson and MacKinnon (1993)], not only is  $\mathbf{c}_i$ , the OLS estimator of  $\boldsymbol{\theta}_i$ , consistent, it is **superconsistent** in that its asymptotic variance is  $O(1/T^2)$  rather than  $O(1/T)$  as in the usual case. Consequently, the problem of spurious regressions disappears as well. Therefore, the next step is to estimate the cointegrating vector(s), by OLS. Under all the assumptions thus far, the residuals from these regressions,  $e_{it}$ , are estimates of the equilibrium errors,  $\varepsilon_{it}$ . As such, they should be  $I(0)$ . The natural approach would be to apply the familiar Dickey–Fuller tests to these residuals. The logic is sound, but the Dickey–Fuller tables are inappropriate for these estimated errors. Estimates of the appropriate critical values for the tests are given by Engle and Granger (1987), Engle and Yoo (1987), Phillips and Ouliaris (1990), and Davidson and MacKinnon (1993). If autocorrelation in the equilibrium errors is suspected, then an augmented Engle and Granger test can be based on the template

$$\Delta e_{it} = \delta e_{i,t-1} + \phi_1(\Delta e_{i,t-1}) + \cdots + u_t.$$

## CHAPTER 23 ♦ Nonstationary Data 1035

If the null hypothesis that  $\delta = 0$  cannot be rejected (against the alternative  $\delta < 0$ ), then we conclude that the variables are not cointegrated. (Cointegration can be rejected by this method. Failing to reject does not confirm it, of course. But having failed to reject the presence of cointegration, we will proceed as if our finding had been affirmative.)

**Example 23.8 (Continued) Cointegration in Consumption and Output**

In the example presented at the beginning of this discussion, we proposed precisely the sort of test suggested by Phillips and Ouliaris (1990) to determine if (log) consumption and (log) GDP are cointegrated. As noted, the logic of our approach is sound, but a few considerations remain. The Dickey–Fuller critical values suggested for the test are appropriate only in a few cases, and not when several trending variables appear in the equation. For the case of only a pair of trended variables, as we have here, one may use infinite sample values in the Dickey–Fuller tables for the trend stationary form of the equation. (The drift and trend would have been removed from the residuals by the original regression, which would have these terms either embedded in the variables or explicitly in the equation.) Finally, there remains an issue of how many lagged differences to include in the ADF regression. We have specified one, although further analysis might be called for. [A lengthy discussion of this set of issues appears in Hayashi (2000, pp. 645–648).] Thus, but for the possibility of this specification issue, the ADF approach suggested in the introduction does pass muster. The sample value found earlier was  $-4.63$ . The critical values from the table are  $-3.45$  for 5 percent and  $-3.67$  for 2.5 percent. Thus, we conclude (as have many other analysts) that log consumption and log GDP are cointegrated.

The Johansen (1988, 1992) and Stock and Watson (1988) methods are similar, so we will describe only the first one. The theory is beyond the scope of this text, although the operational details are suggestive. To carry out the Johansen test, we first formulate the VAR:

$$\mathbf{y}_t = \boldsymbol{\Gamma}_1 \mathbf{y}_{t-1} + \boldsymbol{\Gamma}_2 \mathbf{y}_{t-2} + \cdots + \boldsymbol{\Gamma}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t.$$

The order of the model,  $p$ , must be determined in advance. Now, let  $\mathbf{z}_t$  denote the vector of  $M(p - 1)$  variables,

$$\mathbf{z}_t = [\Delta \mathbf{y}_{t-1}, \Delta \mathbf{y}_{t-2}, \dots, \Delta \mathbf{y}_{t-p+1}].$$

That is,  $\mathbf{z}_t$  contains the lags 1 to  $p - 1$  of the first differences of all  $M$  variables. Now, using the  $T$  available observations, we obtain two  $T \times M$  matrices of least squares residuals:

$\mathbf{D}$  = the residuals in the regressions of  $\Delta \mathbf{y}_t$  on  $\mathbf{z}_t$ ,

$\mathbf{E}$  = the residuals in the regressions of  $\mathbf{y}_{t-p}$  on  $\mathbf{z}_t$ .

We now require the  $M^2$  **canonical correlations** between the columns in  $\mathbf{D}$  and those in  $\mathbf{E}$ . To continue, we will digress briefly to define the canonical correlations. Let  $\mathbf{d}_1^*$  denote a linear combination of the columns of  $\mathbf{D}$ , and let  $\mathbf{e}_1^*$  denote the same from  $\mathbf{E}$ . We wish to choose these two linear combinations so as to maximize the correlation between them. This pair of variables are the first canonical variates, and their correlation  $r_1^*$  is the first canonical correlation. In the setting of cointegration, this computation has some intuitive appeal. Now, with  $\mathbf{d}_1^*$  and  $\mathbf{e}_1^*$  in hand, we seek a second pair of variables  $\mathbf{d}_2^*$  and  $\mathbf{e}_2^*$  to maximize *their* correlation, subject to the constraint that this second variable in each pair be orthogonal to the first. This procedure continues for all  $M$  pairs of variables. It turns out that the computation of all these is quite simple. We will not need to compute the coefficient vectors for the linear combinations. The squared canonical

## 1036 PART V ♦ Time Series and Macroeometrics

correlations are simply the ordered characteristic roots of the matrix

$$\mathbf{R}^* = \mathbf{R}_{DD}^{-1/2} \mathbf{R}_{DE} \mathbf{R}_{EE}^{-1} \mathbf{R}_{ED} \mathbf{R}_{DD}^{-1/2},$$

where  $\mathbf{R}_{ij}$  is the (cross-) correlation matrix between variables in set  $i$  and set  $j$ , for  $i, j = D, E$ .

Finally, the null hypothesis that there are  $r$  or fewer cointegrating vectors is tested using the test statistic

$$\text{TRACE TEST} = -T \sum_{i=r+1}^M \ln[1 - (r_i^*)^2].$$

If the correlations based on actual disturbances had been observed instead of estimated, then we would refer this statistic to the chi-squared distribution with  $M - r$  degrees of freedom. Alternative sets of appropriate tables are given by Johansen and Juselius (1990) and Osterwald-Lenum (1992). Large values give evidence against the hypothesis of  $r$  or fewer cointegrating vectors.

### 23.3.4 ESTIMATING COINTEGRATION RELATIONSHIPS

Both of the testing procedures discussed earlier involve actually estimating the cointegrating vectors, so this additional section is actually superfluous. In the Engle and Granger framework, at a second step after the cointegration test, we can use the residuals from the static regression as an error correction term in a dynamic, first-difference regression, as shown in Section 23.3.2. One can then “test down” to find a satisfactory structure. In the Johansen test shown earlier, the characteristic vectors corresponding to the canonical correlations are the sample estimates of the cointegrating vectors. Once again, computation of an error correction model based on these first step results is a natural next step. We will explore these in an application.

### 23.3.5 APPLICATION: GERMAN MONEY DEMAND

The demand for money has provided a convenient and well targeted illustration of methods of cointegration analysis. The central equation of the model is

$$m_t - p_t = \mu + \beta y_t + \gamma i_t + \varepsilon_t, \quad (23-8)$$

where  $m_t$ ,  $p_t$ , and  $y_t$  are the logs of nominal money demand, the price level, and output, and  $i$  is the nominal interest rate (not the log of). The equation involves trending variables ( $m_t$ ,  $p_t$ ,  $y_t$ ), and one that we found earlier appears to be a random walk with drift ( $i_t$ ). As such, the usual form of statistical inference for estimation of the income elasticity and interest semielasticity based on stationary data is likely to be misleading.

Beyer (1998) analyzed the demand for money in Germany over the period 1975 to 1994. A central focus of the study was whether the 1990 reunification produced a structural break in the long-run demand function. (The analysis extended an earlier study by the same author that was based on data that predated the reunification.) One of the interesting questions pursued in this literature concerns the stability of the long-term demand equation,

$$(m - p)_t - y_t = \mu + \gamma i_t + \varepsilon_t. \quad (23-9)$$

## CHAPTER 23 ♦ Nonstationary Data 1037

**TABLE 23.6** Augmented Dickey–Fuller Tests for Variables in the Beyer Model

<b>Variable</b>	<b>m</b>	<b><math>\Delta m</math></b>	<b><math>\Delta^2 m</math></b>	<b>p</b>	<b><math>\Delta p</math></b>	<b><math>\Delta^2 p</math></b>	<b><math>\Delta_4 p</math></b>	<b><math>\Delta \Delta_4 p</math></b>
Spec.	TS	RW	RW	TS	RW/D	RW	RW/D	RW
lag	0	4	3	4	3	2	2	2
DF <sub>r</sub>	-1.82	-1.61	-6.87	-2.09	-2.14	-10.6	-2.66	-5.48
Crit. Value	-3.47	-1.95	-1.95	-3.47	-2.90	-1.95	-2.90	-1.95
<b>Variable</b>	<b>y</b>	<b><math>\Delta y</math></b>	<b>RS</b>	<b><math>\Delta RS</math></b>	<b>RL</b>	<b><math>\Delta RL</math></b>	<b>(m – p)</b>	<b><math>\Delta(m – p)</math></b>
Spec.	TS	RW/D	TS	RW	TS	RW	RW/D	RW/D
lag	4	3	1	0	1	0	0	0
DF <sub>r</sub>	-1.83	-2.91	-2.33	-5.26	-2.40	-6.01	-1.65	-8.50
Crit. Value	-3.47	-2.90	-2.90	-1.95	-2.90	-1.95	-3.47	-2.90

The left-hand side is the log of the inverse of the velocity of money, as suggested by Lucas (1988). An issue to be confronted in this specification is the exogeneity of the interest variable—exogeneity [in the Engle, Hendry, and Richard (1993) sense] of income is moot in the long-run equation as its coefficient is assumed (per Lucas) to equal one. Beyer  explored this latter issue in the framework developed by Engle et al. (see Section 3.5) and through the Granger causality testing methods discussed in Section 20..

The analytical platform of Beyer's study is a long-run function for the real money stock M3 (we adopt the author's notation)

$$(m - p)^* = \delta_0 + \delta_1 y + \delta_2 RS + \delta_3 RL + \delta_4 \Delta_4 p, \quad (23-10)$$

where  $RS$  is a short-term interest rate,  $RL$  is a long-term interest rate, and  $\Delta_4 p$  is the annual inflation rate—the data are quarterly. The first step is an examination of the data. Augmented Dickey–Fuller tests suggest that for these German data in this period,  $m_t$  and  $p_t$  are  $I(2)$ , while  $(m_t - p_t)$ ,  $y_t$ ,  $\Delta_4 p_t$ ,  $RS_t$ , and  $RL_t$  are all  $I(1)$ . Some of Beyer's results which produced these conclusions are shown in Table 23.6. Note that although both  $m_t$  and  $p_t$  appear to be  $I(2)$ , their simple difference (linear combination) is  $I(1)$ , that is, integrated to a lower order. That produces the long-run specification given by (23-10). The Lucas specification is layered onto this to produce the model for the long-run velocity

$$(m - p - y)^* = \delta_0^* + \delta_2^* RS + \delta_3^* RL + \delta_4^* \Delta_4 p. \quad (23-11)$$

### 23.3.5.a Cointegration Analysis and a Long-Run Theoretical Model

For (23-10) to be a valid model, there must be at least one cointegrating vector that transforms  $\mathbf{z}_t = [(m_t - p_t), y_t, RS_t, RL_t, \Delta_4 p_t]$  to stationarity. The Johansen trace test described in Section 22.3.3 was applied to the VAR consisting of these five  $I(1)$  variables. A lag length of two was chosen for the analysis. The results of the trace test are a bit ambiguous; the hypothesis that  $r = 0$  is rejected, albeit not strongly (sample value = 90.17 against a 95 percent critical value = 87.31) while the hypothesis that  $r \leq 1$  is not rejected (sample value = 60.15 against a 95 percent critical value of 62.99). (These borderline results follow from the result that Beyer's first three eigenvalues—canonical correlations in the trace test statistic—are nearly equal. Variation in the test statistic results from variation in the correlations.) On this basis, it is concluded that the

**1038 PART V ♦ Time Series and Macroeconometrics**

cointegrating rank equals one. The unrestricted cointegrating vector for the equation, with a time trend added is found to be

$$(m - p) = 0.936y - 1.780\Delta_4 p + 1.601RS - 3.279RL + 0.002t. \quad (23-12)$$

(These are the coefficients from the first characteristic vector of the canonical correlation analysis in the Johansen computations detailed in Section 23.3.3.) An exogeneity test—we have not developed this in detail; see Beyer (1998, p. 59), Hendry and Ericsson (1991), and Engle and Hendry (1993)—confirms weak exogeneity of all four right-hand-side variables in this specification. The final specification test is for the Lucas formulation and elimination of the time trend, both of which are found to pass, producing the cointegration vector

$$(m - p - y) = -1.832\Delta_4 p + 4.352RS - 10.89RL.$$

The conclusion drawn from the cointegration analysis is that a single-equation model for the long-run money demand is appropriate and a valid way to proceed. A last step before this analysis is a series of Granger causality tests for feedback between changes in the money stock and the four right-hand-side variables in (23-12) (not including the trend). (See Section 21.6.5.) The test results are generally favorable, with some mixed results for exogeneity of GDP.

**23.3.5.b Testing for Model Instability**

Let  $\mathbf{z}_t = [(m_t - p_t), y_t, \Delta_4 p_t, RS_t, RL_t]$  and let  $\mathbf{z}_{t-1}^0$  denote the entire history of  $\mathbf{z}_t$  up to the previous period. The joint distribution for  $\mathbf{z}_t$ , conditioned on  $\mathbf{z}_{t-1}^0$  and a set of parameters  $\Psi$  factors one level further into

$$\begin{aligned} f(\mathbf{z}_t | \mathbf{z}_{t-1}^0, \Psi) &= f[(m - p)_t | y_t, \Delta_4 p_t, RS_t, RL_t, \mathbf{z}_{t-1}^0, \Psi_1] \\ &\quad \times g(y_t, \Delta_4 p_t, RS_t, RL_t | \mathbf{z}_{t-1}^0, \Psi_2). \end{aligned}$$

The result of the exogeneity tests carried out earlier implies that the conditional distribution may be analyzed apart from the marginal distribution—that is, the implication of the Engle, Hendry, and Richard results noted earlier. Note the partitioning of the parameter vector. Thus, the conditional model is represented by an error correction form that explains  $\Delta(m - p)_t$  in terms of its own lags, the error correction term and contemporaneous and lagged changes in the (now established) weakly exogenous variables as well as other terms such as a constant term, trend, and certain dummy variables which pick up particular events. The error correction model specified is

$$\begin{aligned} \Delta(m - p)_t &= \sum_{i=1}^4 c_i \Delta(m - p)_{t-i} + \sum_{i=0}^4 d_{1,i} \Delta(\Delta_4 p_{t-i}) + \sum_{i=0}^4 d_{2,i} \Delta y_{t-i} \\ &\quad + \sum_{i=0}^4 d_{3,i} \Delta RS_{t-i} + \sum_{i=0}^4 d_{4,i} \Delta RL_{t-i} + \lambda(m - p - y)_{t-1} \quad (23-13) \\ &\quad + \gamma_1 RS_{t-1} + \gamma_2 RL_{t-1} + \mathbf{d}'_t \boldsymbol{\phi} + \omega_t, \end{aligned}$$

where  $\mathbf{d}_t$  is the set of additional variables, including the constant and five one-period dummy variables that single out specific events such as a currency crisis in September, 1992 [Beyer (1998, p. 62, fn. 4)]. The model is estimated by least squares, “stepwise

## CHAPTER 23 ♦ Nonstationary Data 1039

simplified and reparameterized.” (The number of parameters in the equation is reduced from 32 to 15.<sup>14</sup>)

The estimated form of (23-13) is an autoregressive distributed lag model. We proceed to use the model to solve for the long-run, steady-state growth path of the real money stock, (23-10). The annual growth rates  $\Delta_4 m = g_m$ ,  $\Delta_4 p = g_p$ ,  $\Delta_4 y = g_y$  and (assumed)  $\Delta_4 RS = g_{RS} = \Delta_4 RL = g_{RL} = 0$  are used for the solution<sup>15</sup>

$$\frac{1}{4}(g_m - g_p) = \frac{c_4}{4}(g_m - g_p) - d_{1,1}g_p + \frac{d_{2,2}}{2}g_y + \gamma_1 RS + \gamma_2 RL + \lambda(m - p - y).$$

This equation is solved for  $(m - p)^*$  under the assumption that  $g_m = (g_y + g_p)$ ,

$$(m - p)^* = \hat{\delta}_0 + \hat{\delta}_1 g_y + y + \hat{\delta}_2 \Delta_4 p + \hat{\delta}_3 RS + \hat{\delta}_4 RL.$$

Analysis then proceeds based on this estimated long-run relationship.

The primary interest of the study is the stability of the demand equation pre- and postunification. A comparison of the parameter estimates from the same set of procedures using the period 1976–1989 shows them to be surprisingly similar, [(1.22 – 3.67 $g_y$ ), 1, –3.67, 3.67, –6.44] for the earlier period and [(1.25 – 2.09 $g_y$ ), 1, –3.625, 3.5, –7.25] for the later one. This suggests, albeit informally, that the function has not changed (at least by much). A variety of testing procedures for structural break led to the conclusion that in spite of the dramatic changes of 1990, the long-run money demand function had not materially changed in the sample period.

## 23.4 NONSTATIONARY PANEL DATA

 In Section 11.12, we began to examine panel data settings in which  $T$ , the number of observations in each group (e.g., country), became large as well as  $n$ . Applications include cross-country studies of growth using the Penn World Tables [Im, Pesaran, and Shin (2003) and Sala-i-Martin (1996)], studies of purchasing power parity [Pedroni (2001)], and analyses of health care expenditures [McCoskey and Selden (1998)]. In the small  $T$  cases of longitudinal, microeconomic data sets, the time-series properties of the data are a side issue that is usually of little interest. But when  $T$  is growing at essentially the same rate as  $n$ , for example, in the cross-country studies, these properties become a central focus of the analysis.

The large  $T$ , large  $n$  case presents several complications for the analyst. In the longitudinal analysis, pooling of the data is usually a given, although we developed several extensions of the models to accommodate parameter heterogeneity (see Section 11.11). In a long-term cross-country model, any type of pooling would be especially suspect. The time series are long, so this would seem to suggest that the appropriate modeling strategy would be simply to analyze each country separately. But this would neglect the hypothesized commonalities across countries such as a (proposed) common growth rate. Thus, the recent “time-series panel data” literature seeks to reconcile these opposing features of the data.

<sup>14</sup>The equation ultimately used is  $\Delta(m_t - p_t) = h[\Delta(m - p)_{t-4}, \Delta\Delta_4 p_t, \Delta^2 y_{t-2}, \Delta RS_{t-1} + \Delta RS_{t-3}, \Delta^2 RL_t, RS_{t-1}, RL_{t-1}, \Delta_4 p_{t-1}, (m - p - y)_{t-1}, \mathbf{d}_t]$ .

<sup>15</sup>The division of the coefficients is done because the intervening lags do not appear in the estimated equation.

## 1040 PART V ♦ Time Series and Macroeconometrics

As in the single time-series cases examined earlier in this chapter, long-term aggregate series  usually nonstationary, which calls conventional methods (such as those in Section 17.12) into question. A focus of the recent literature, for example, is on testing for unit roots in an analog to the platform for the augmented Dickey–Fuller tests (Section 23.2),

$$\Delta y_{it} = \rho_i y_{i,t-1} + \sum_{m=1}^{L_i} \gamma_{im} \Delta y_{i,t-m} + \alpha_i + \beta_i t + \varepsilon_{it}.$$

Different formulations of this model have been analyzed, for example, by Levin, Lin, and Chu (2002), who assume  $\rho_i = \rho$ ; Im, Pesaran, and Shin (2003), who relax that restriction; and Breitung (2000), who considers various mixtures of the cases. An extension of the KPSS test in Section 23.2.5 that is particularly simple to compute is Hadri's (2000) LM statistic,

$$LM = \frac{1}{n} \sum_{i=1}^n \left( \frac{\sum_{t=1}^T E_{it}^2}{T^2 \hat{\sigma}_\varepsilon^2} \right) = \frac{\sum_{i=1}^n KPSS_i}{n}.$$

This is the sample average of the KPSS statistics for the  $n$  countries. Note that it includes two assumptions: that the countries are independent and that there is a common  $\sigma_\varepsilon^2$  for all countries. An alternative is suggested that allows  $\sigma_\varepsilon^2$  to vary across countries.

As it stands, the preceding model would suggest that separate analyses for each country would be appropriate. An issue to consider, then, would be how to combine, if possible, the separate results in some optimal fashion. Maddala and Wu (1999), for example, suggested a “Fisher-type” chi-squared test based on  $P = -2 \sum_i \ln p_i$ , where  $p_i$  is the  $p$ -value from the individual tests. Under the null hypothesis that  $\rho_i$  equals zero, the limiting distribution is chi-squared with  $2n$  degrees of freedom.

Analysis of cointegration, and models of cointegrated series in the panel data setting, parallel the single time-series case, but also differ in a crucial respect. [See, e.g., Kao (1999), McCoskey and Kao (1999), and Pedroni (2000, 2004)]. Whereas in the single time-series case, the analysis of cointegration focuses on the long-run relationships between, say,  $x_t$  and  $z_t$  for two variables for the same country, in the panel data setting, say, in the analysis of exchange rates, inflation, purchasing power parity or international R & D spillovers, interest may focus on a long-run relationship between  $x_{it}$  and  $x_{mt}$  for two different countries (or  $n$  countries). This substantially complicates the analyses. It is also well beyond the scope of this text. Extensive surveys of these issues may be found in Baltagi (2005, Chapter 12) and Smith (2000).

### 23.5 SUMMARY AND CONCLUSIONS

This chapter has completed our survey of techniques for the analysis of time-series data. While Chapters 21 and 22 were about extensions of regression modeling to the time-series setting, most of the results in this chapter focus on the internal structure of the individual time series, themselves. Chapter 22 presented the standard models for time-series processes. While the empirical distinction between, say, AR( $p$ ) and MA( $q$ ) series may seem ad hoc, the Wold decomposition assures that with enough care, a variety of models can be used to analyze a time series. This chapter described what is arguably the fundamental tool of modern macroeconomics: the tests for nonstationarity. Contemporary econometric analysis of macroeconomic data has added considerable structure and

**CHAPTER 23 ♦ Nonstationary Data 1041**

formality to trending variables, which are more common than not in that setting. The variants of the Dickey–Fuller and KPSS tests for unit roots are an indispensable tool for the analyst of time-series data. Section 23.4 then considered the subject of cointegration. This modeling framework is a distinct extension of the regression modeling where this discussion began. Cointegrated relationships and equilibrium relationships form the basis of the time-series counterpart to regression relationships. But, in this case, it is not the conditional mean as such that is of interest. Here, both the long-run equilibrium and short-run relationships around trends are of interest and are studied in the data.

**Key Terms and Concepts**

- Autoregressive integrated moving-average (ARIMA) process
- Augmented Dickey–Fuller test
- Canonical correlation
- Cointegration
- Cointegration rank
- Cointegration relationship
- Cointegrating vector
- Common trend
- Data generating process (DGP)
- Dickey–Fuller test
- Equilibrium error
- Error correction model
- Fractional integration
- Integrated of order one
- KPSS test
- Phillips–Perron test
- Random walk
- Random walk with drift
- Spurious regression
- Superconsistent
- Trend stationary process
- Unit root

**Exercise**

1. Find the autocorrelations and partial autocorrelations for the MA(2) process

$$\varepsilon_t = v_t - \theta_1 v_{t-1} - \theta_2 v_{t-2}.$$

**Applications**

1. Carry out the ADF test for a unit root in the bond yield data of Example 22.1.
2. Using the macroeconomic data in Appendix Table F5.2, estimate by least squares the parameters of the model

$$c_t = \beta_0 + \beta_1 y_t + \beta_2 c_{t-1} + \beta_3 c_{t-2} + \varepsilon_t,$$

where  $c_t$  is the log of real consumption and  $y_t$  is the log of real disposable income.

- a. Use the Breusch and Pagan test to examine the residuals for autocorrelation.
- b. Is the estimated equation stable? What is the characteristic equation for the autoregressive part of this model? What are the roots of the characteristic equation, using your estimated parameters?
- c. What is your implied estimate of the short-run (impact) multiplier for change in  $y_t$  on  $c_t$ ? Compute the estimated long-run multiplier.
3. Carry out an ADF test for a unit root in the rate of inflation using the subset of the data in Appendix Table F5.2 since 1974.1. (This is the first quarter after the oil shock of 1973.)
4. Estimate the parameters of the model in Example 16, using two-stage least squares. Obtain the residuals from the two equations. Do these residuals appear to be white noise series? Based on your findings, what do you conclude about the specification of the model?



## APPENDIX A

---



# MATRIX ALGEBRA

### A.1 TERMINOLOGY

**A matrix** is a rectangular array of numbers, denoted

$$\mathbf{A} = [a_{ik}] = [\mathbf{A}]_{ik} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nK} \end{bmatrix}. \quad (\mathbf{A}-1)$$

The typical element is used to denote the matrix. A subscripted element of a matrix is always read as  $a_{\text{row},\text{column}}$ . An example is given in Table A.1. In these data, the rows are identified with years and the columns with particular variables.

**A vector** is an ordered set of numbers arranged either in a row or a column. In view of the preceding, a **row vector** is also a matrix with one row, whereas a **column vector** is a matrix with one column. Thus, in Table A.1, the five variables observed for 1972 (including the date) constitute a row vector, whereas the time series of nine values for consumption is a column vector.

A matrix can also be viewed as a set of column vectors or as a set of row vectors.<sup>1</sup> The **dimensions** of a matrix are the numbers of rows and columns it contains. “**A** is an  $n \times K$  matrix” (read “ $n$  by  $K$ ”) will always mean that **A** has  $n$  rows and  $K$  columns. If  $n$  equals  $K$ , then **A** is a **square matrix**. Several particular types of square matrices occur frequently in econometrics.

- A **symmetric matrix** is one in which  $a_{ik} = a_{ki}$  for all  $i$  and  $k$ .
- A **diagonal matrix** is a square matrix whose only nonzero elements appear on the **main diagonal**, that is, moving from upper left to lower right.
- A **scalar matrix** is a diagonal matrix with the same value in all diagonal elements.
- An **identity matrix** is a scalar matrix with ones on the diagonal. This matrix is always denoted **I**. A subscript is sometimes included to indicate its size, or **order**. For example, **I**<sub>4</sub> indicates a  $4 \times 4$  identity matrix.
- A **triangular matrix** is one that has only zeros either above or below the main diagonal. If the zeros are above the diagonal, the matrix is **lower triangular**.

### A.2 ALGEBRAIC MANIPULATION OF MATRICES

#### A.2.1 EQUALITY OF MATRICES

Matrices (or vectors) **A** and **B** are equal if and only if they have the same dimensions and each element of **A** equals the corresponding element of **B**. That is,

$$\mathbf{A} = \mathbf{B} \quad \text{if and only if } a_{ik} = b_{ik} \quad \text{for all } i \text{ and } k. \quad (\mathbf{A}-2)$$

---

<sup>1</sup>Henceforth, we shall denote a matrix by a boldfaced capital letter, as is **A** in (A-1), and a vector as a boldfaced lowercase letter, as in **a**. Unless otherwise noted, a vector will always be assumed to be a *column vector*.

## APPENDIX A ♦ Matrix Algebra 1043

**TABLE A.1** Matrix of Macroeconomic Data

Row	1 Year	Column			
		2 <i>Consumption</i> (billions of dollars)	3 <i>GNP</i> (billions of dollars)	4 <i>GNP Deflator</i>	5 <i>Discount Rate</i> (N.Y Fed., avg.)
1	1972	737.1	1185.9	1.0000	4.50
2	1973	812.0	1326.4	1.0575	6.44
3	1974	808.1	1434.2	1.1508	7.83
4	1975	976.4	1549.2	1.2579	6.25
5	1976	1084.3	1718.0	1.3234	5.50
6	1977	1204.4	1918.3	1.4005	5.46
7	1978	1346.5	2163.9	1.5042	7.46
8	1979	1507.2	2417.8	1.6342	10.28
9	1980	1667.2	2633.1	1.7864	11.77

Source: Data from the *Economic Report of the President* (Washington, D.C.: U.S. Government Printing Office, 1983).

**A.2.2 TRANSPOSITION**

The **transpose** of a matrix  $\mathbf{A}$ , denoted  $\mathbf{A}'$ , is obtained by creating the matrix whose  $k$ th row is the  $k$ th column of the original matrix. Thus, if  $\mathbf{B} = \mathbf{A}'$ , then each column of  $\mathbf{A}$  will appear as the corresponding row of  $\mathbf{B}$ . If  $\mathbf{A}$  is  $n \times K$ , then  $\mathbf{A}'$  is  $K \times n$ .

An equivalent definition of the transpose of a matrix is

$$\mathbf{B} = \mathbf{A}' \Leftrightarrow b_{ik} = a_{ki} \quad \text{for all } i \text{ and } k. \quad (\text{A-3})$$

The definition of a symmetric matrix implies that

$$\text{if (and only if) } \mathbf{A} \text{ is symmetric, then } \mathbf{A} = \mathbf{A}'. \quad (\text{A-4})$$

It also follows from the definition that for any  $\mathbf{A}$ ,

$$(\mathbf{A}')' = \mathbf{A}. \quad (\text{A-5})$$

Finally, the transpose of a column vector,  $\mathbf{a}$ , is a row vector:

$$\mathbf{a}' = [a_1 \ a_2 \ \cdots \ a_n].$$

**A.2.3 MATRIX ADDITION**

The operations of addition and subtraction are extended to matrices by defining

$$\mathbf{C} = \mathbf{A} + \mathbf{B} = [a_{ik} + b_{ik}]. \quad (\text{A-6})$$

$$\mathbf{A} - \mathbf{B} = [a_{ik} - b_{ik}]. \quad (\text{A-7})$$

Matrices cannot be added unless they have the same dimensions, in which case they are said to be **conformable for addition**. A **zero matrix** or **null matrix** is one whose elements are all zero. In the addition of matrices, the zero matrix plays the same role as the scalar 0 in scalar addition; that is,

$$\mathbf{A} + \mathbf{0} = \mathbf{A}. \quad (\text{A-8})$$

It follows from (A-6) that matrix addition is commutative,

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}. \quad (\text{A-9})$$

## 1044 PART VI ♦ Appendices

and associative,

$$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}), \quad (\text{A-10})$$

and that

$$(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'. \quad (\text{A-11})$$

### A.2.4 VECTOR MULTIPLICATION

Matrices are multiplied by using the **inner product**. The inner product, or **dot product**, of two vectors,  $\mathbf{a}$  and  $\mathbf{b}$ , is a scalar and is written

$$\mathbf{a}'\mathbf{b} = a_1b_1 + a_2b_2 + \cdots + a_nb_n. \quad (\text{A-12})$$

Note that the inner product is written as the transpose of vector  $\mathbf{a}$  times vector  $\mathbf{b}$ , a row vector times a column vector. In (A-12), each term  $a_jb_j$  equals  $b_ja_j$ ; hence

$$\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a}. \quad (\text{A-13})$$

### A.2.5 A NOTATION FOR ROWS AND COLUMNS OF A MATRIX

We need a notation for the  $i$ th row of a matrix. Throughout this book, an untransposed vector will always be a column vector. However, we will often require a notation for the column vector that is the transpose of a row of a matrix. This has the potential to create some ambiguity, but the following convention based on the subscripts will suffice for our work throughout this text:

- $\mathbf{a}_k$ , or  $\mathbf{a}_l$  or  $\mathbf{a}_m$  will denote column  $k$ ,  $l$ , or  $m$  of the matrix  $\mathbf{A}$ ,
- $\mathbf{a}'_i$ , or  $\mathbf{a}_j$  or  $\mathbf{a}_t$  or  $\mathbf{a}_s$  will denote the column vector formed by the transpose of row  $i$ ,  $j$ ,  $t$ , or  $s$  of matrix  $\mathbf{A}$ . Thus,  $\mathbf{a}'_i$  is row  $i$  of  $\mathbf{A}$ .

For example, from the data in Table A.1 it might be convenient to speak of  $\mathbf{x}_i$ , where  $i = 1972$  as the  $5 \times 1$  vector containing the five variables measured for the year 1972, that is, the transpose of the 1972 row of the matrix. In our applications, the common association of subscripts “ $i$ ” and “ $j$ ” with individual  $i$  or  $j$ , and “ $t$ ” and “ $s$ ” with time periods  $t$  and  $s$  will be natural.

### A.2.6 MATRIX MULTIPLICATION AND SCALAR MULTIPLICATION

For an  $n \times K$  matrix  $\mathbf{A}$  and a  $K \times M$  matrix  $\mathbf{B}$ , the product matrix,  $\mathbf{C} = \mathbf{AB}$ , is an  $n \times M$  matrix whose  $ik$ th element is the inner product of row  $i$  of  $\mathbf{A}$  and column  $k$  of  $\mathbf{B}$ . Thus, the product matrix  $\mathbf{C}$  is

$$\mathbf{C} = \mathbf{AB} \Rightarrow c_{ik} = \mathbf{a}'_i \mathbf{b}_k. \quad (\text{A-15})$$

[Note our use of (A-14) in (A-15).] To multiply two matrices, the number of columns in the first must be the same as the number of rows in the second, in which case they are **conformable for multiplication**.<sup>2</sup> Multiplication of matrices is generally not commutative. In some cases,  $\mathbf{AB}$  may exist, but  $\mathbf{BA}$  may be undefined or, if it does exist, may have different dimensions. In general, however, even if  $\mathbf{AB}$  and  $\mathbf{BA}$  do have the same dimensions, they will not be equal. In view of this, we define **premultiplication** and **postmultiplication** of matrices. In the product  $\mathbf{AB}$ ,  $\mathbf{B}$  is **premultiplied** by  $\mathbf{A}$ , whereas  $\mathbf{A}$  is **postmultiplied** by  $\mathbf{B}$ .

---

<sup>2</sup>A simple way to check the conformability of two matrices for multiplication is to write down the dimensions of the operation, for example,  $(n \times K)$  times  $(K \times M)$ . The inner dimensions must be equal; the result has dimensions equal to the outer values.

## APPENDIX A ♦ Matrix Algebra 1045

**Scalar multiplication** of a matrix is the operation of multiplying every element of the matrix by a given scalar. For scalar  $c$  and matrix  $\mathbf{A}$ ,

$$c\mathbf{A} = [ca_{ik}]. \quad (\text{A-16})$$

The product of a matrix and a vector is written

$$\mathbf{c} = \mathbf{Ab}.$$

The number of elements in  $\mathbf{b}$  must equal the number of columns in  $\mathbf{A}$ ; the result is a vector with number of elements equal to the number of rows in  $\mathbf{A}$ . For example,

$$\begin{bmatrix} 5 \\ 4 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix}.$$

We can interpret this in two ways. First, it is a compact way of writing the three equations

$$5 = 4a + 2b + 1c,$$

$$4 = 2a + 6b + 1c,$$

$$1 = 1a + 1b + 0c.$$

Second, by writing the set of equations as

$$\begin{bmatrix} 5 \\ 4 \\ 1 \end{bmatrix} = a \begin{bmatrix} 4 \\ 2 \\ 1 \end{bmatrix} + b \begin{bmatrix} 2 \\ 6 \\ 1 \end{bmatrix} + c \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

we see that the right-hand side is a **linear combination** of the columns of the matrix where the coefficients are the elements of the vector. For the general case,

$$\mathbf{c} = \mathbf{Ab} = b_1\mathbf{a}_1 + b_2\mathbf{a}_2 + \cdots + b_K\mathbf{a}_K. \quad (\text{A-17})$$

In the calculation of a matrix product  $\mathbf{C} = \mathbf{AB}$ , each column of  $\mathbf{C}$  is a linear combination of the columns of  $\mathbf{A}$ , where the coefficients are the elements in the corresponding column of  $\mathbf{B}$ . That is,

$$\mathbf{C} = \mathbf{AB} \Leftrightarrow \mathbf{c}_k = \mathbf{Ab}_k. \quad (\text{A-18})$$

Let  $\mathbf{e}_k$  be a column vector that has zeros everywhere except for a one in the  $k$ th position. Then  $\mathbf{A}\mathbf{e}_k$  is a linear combination of the columns of  $\mathbf{A}$  in which the coefficient on every column but the  $k$ th is zero, whereas that on the  $k$ th is one. The result is

$$\mathbf{a}_k = \mathbf{Ae}_k. \quad (\text{A-19})$$

Combining this result with (A-17) produces

$$(\mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_n) = \mathbf{A}(\mathbf{e}_1 \quad \mathbf{e}_2 \quad \cdots \quad \mathbf{e}_n) = \mathbf{AI} = \mathbf{A}. \quad (\text{A-20})$$

In matrix multiplication, the identity matrix is analogous to the scalar 1. For any matrix or vector  $\mathbf{A}$ ,  $\mathbf{AI} = \mathbf{A}$ . In addition,  $\mathbf{IA} = \mathbf{A}$ , although if  $\mathbf{A}$  is not a square matrix, the two identity matrices are of different orders.

A conformable matrix of zeros produces the expected result:  $\mathbf{A}\mathbf{0} = \mathbf{0}$ .

Some general rules for matrix multiplication are as follows:

- **Associative law:**  $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}).$  (A-21)

- **Distributive law:**  $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}.$  (A-22)

**1046 PART VI ♦ Appendices**

- **Transpose of a product:**  $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$ . (A-23)
- **Transpose of an extended product:**  $(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'$ . (A-24)

**A.2.7 SUMS OF VALUES**

Denote by  $\mathbf{i}$  a vector that contains a column of ones. Then,

$$\sum_{i=1}^n x_i = x_1 + x_2 + \cdots + x_n = \mathbf{i}'\mathbf{x}. \quad (\text{A-25})$$

If all elements in  $\mathbf{x}$  are equal to the same constant  $a$ , then  $\mathbf{x} = a\mathbf{i}$  and

$$\sum_{i=1}^n x_i = \mathbf{i}'(a\mathbf{i}) = a(\mathbf{i}'\mathbf{i}) = na. \quad (\text{A-26})$$

For any constant  $a$  and vector  $\mathbf{x}$ ,

$$\sum_{i=1}^n ax_i = a \sum_{i=1}^n x_i = a\mathbf{i}'\mathbf{x}. \quad (\text{A-27})$$

If  $a = 1/n$ , then we obtain the arithmetic mean,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n}\mathbf{i}'\mathbf{x}, \quad (\text{A-28})$$

from which it follows that

$$\sum_{i=1}^n x_i = \mathbf{i}'\mathbf{x} = n\bar{x}.$$

The sum of squares of the elements in a vector  $\mathbf{x}$  is

$$\sum_{i=1}^n x_i^2 = \mathbf{x}'\mathbf{x}; \quad (\text{A-29})$$

while the sum of the products of the  $n$  elements in vectors  $\mathbf{x}$  and  $\mathbf{y}$  is

$$\sum_{i=1}^n x_i y_i = \mathbf{x}'\mathbf{y}. \quad (\text{A-30})$$

By the definition of matrix multiplication,

$$[\mathbf{X}'\mathbf{X}]_{kl} = [\mathbf{x}_k'\mathbf{x}_l] \quad (\text{A-31})$$

is the inner product of the  $k$ th and  $l$ th columns of  $\mathbf{X}$ . For example, for the data set given in Table A.1, if we define  $\mathbf{X}$  as the  $9 \times 3$  matrix containing (year, consumption, GNP), then

$$\begin{aligned} [\mathbf{X}'\mathbf{X}]_{23} &= \sum_{t=1972}^{1980} \text{consumption}_t \text{GNP}_t = 737.1(1185.9) + \cdots + 1667.2(2633.1) \\ &= 19,743,711.34. \end{aligned}$$

If  $\mathbf{X}$  is  $n \times K$ , then [again using (A-14)]

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'.$$

## APPENDIX A ♦ Matrix Algebra 1047

This form shows that the  $K \times K$  matrix  $\mathbf{X}'\mathbf{X}$  is the sum of  $n$   $K \times K$  matrices, each formed from a single row (year) of  $\mathbf{X}$ . For the example given earlier, this sum is of nine  $3 \times 3$  matrices, each formed from one row (year) of the original data matrix.

### A.2.8 A USEFUL IDEMPOTENT MATRIX

A fundamental matrix in statistics is the “centering matrix” that is used to transform data to deviations from their mean. First,

$$\mathbf{i}\bar{x} = \frac{1}{n}\mathbf{i}'\mathbf{x} = \begin{bmatrix} \bar{x} \\ \bar{x} \\ \vdots \\ \bar{x} \end{bmatrix} = \frac{1}{n}\mathbf{i}\mathbf{i}'\mathbf{x}. \quad (\text{A-32})$$

The matrix  $(1/n)\mathbf{i}\mathbf{i}'$  is an  $n \times n$  matrix with every element equal to  $1/n$ . The set of values in deviations form is

$$\begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \dots \\ x_n - \bar{x} \end{bmatrix} = [\mathbf{x} - \mathbf{i}\bar{x}] = \left[ \mathbf{x} - \frac{1}{n}\mathbf{i}\mathbf{i}'\mathbf{x} \right]. \quad (\text{A-33})$$

Because  $\mathbf{x} = \mathbf{I}\mathbf{x}$ ,

$$\left[ \mathbf{x} - \frac{1}{n}\mathbf{i}\mathbf{i}'\mathbf{x} \right] = \left[ \mathbf{I}\mathbf{x} - \frac{1}{n}\mathbf{i}\mathbf{i}'\mathbf{x} \right] = \left[ \mathbf{I} - \frac{1}{n}\mathbf{i}\mathbf{i}' \right]\mathbf{x} = \mathbf{M}^0\mathbf{x}. \quad (\text{A-34})$$

Henceforth, the symbol  $\mathbf{M}^0$  will be used only for this matrix. Its diagonal elements are all  $(1 - 1/n)$ , and its off-diagonal elements are  $-1/n$ . The matrix  $\mathbf{M}^0$  is primarily useful in computing sums of squared deviations. Some computations are simplified by the result

$$\mathbf{M}^0\mathbf{i} = \left[ \mathbf{I} - \frac{1}{n}\mathbf{i}\mathbf{i}' \right]\mathbf{i} = \mathbf{i} - \frac{1}{n}\mathbf{i}(\mathbf{i}'\mathbf{i}) = \mathbf{0},$$

which implies that  $\mathbf{i}'\mathbf{M}^0 = \mathbf{0}'$ . The sum of deviations about the mean is then

$$\sum_{i=1}^n (x_i - \bar{x}) = \mathbf{i}'[\mathbf{M}^0\mathbf{x}] = \mathbf{0}'\mathbf{x} = 0. \quad (\text{A-35})$$

For a single variable  $\mathbf{x}$ , the sum of squared deviations about the mean is

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \left( \sum_{i=1}^n x_i^2 \right) - n\bar{x}^2. \quad (\text{A-36})$$

In matrix terms,

$$\sum_{i=1}^n (x_i - \bar{x})^2 = (\mathbf{x} - \bar{x}\mathbf{i})'(\mathbf{x} - \bar{x}\mathbf{i}) = (\mathbf{M}^0\mathbf{x})'(\mathbf{M}^0\mathbf{x}) = \mathbf{x}'\mathbf{M}^0\mathbf{x}.$$

Two properties of  $\mathbf{M}^0$  are useful at this point. First, because all off-diagonal elements of  $\mathbf{M}^0$  equal  $-1/n$ ,  $\mathbf{M}^0$  is symmetric. Second, as can easily be verified by multiplication,  $\mathbf{M}^0$  is equal to its square;  $\mathbf{M}^0\mathbf{M}^0 = \mathbf{M}^0$ .

**1048 PART VI ♦ Appendices**
**DEFINITION A.1 Idempotent Matrix**

An **idempotent** matrix,  $\mathbf{M}$ , is one that is equal to its square, that is,  $\mathbf{M}^2 = \mathbf{MM} = \mathbf{M}$ . If  $\mathbf{M}$  is a symmetric idempotent matrix (all of the idempotent matrices we shall encounter are symmetric), then  $\mathbf{M}'\mathbf{M} = \mathbf{M}$ .

Thus,  $\mathbf{M}^0$  is a symmetric idempotent matrix. Combining results, we obtain

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \mathbf{x}'\mathbf{M}^0\mathbf{x}. \quad (\text{A-37})$$

Consider constructing a matrix of sums of squares and cross products in deviations from the column means. For two vectors  $\mathbf{x}$  and  $\mathbf{y}$ ,

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = (\mathbf{M}^0\mathbf{x})'(\mathbf{M}^0\mathbf{y}), \quad (\text{A-38})$$

so

$$\begin{bmatrix} \sum_{i=1}^n (x_i - \bar{x})^2 & \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) & \sum_{i=1}^n (y_i - \bar{y})^2 \end{bmatrix} = \begin{bmatrix} \mathbf{x}'\mathbf{M}^0\mathbf{x} & \mathbf{x}'\mathbf{M}^0\mathbf{y} \\ \mathbf{y}'\mathbf{M}^0\mathbf{x} & \mathbf{y}'\mathbf{M}^0\mathbf{y} \end{bmatrix}. \quad (\text{A-39})$$

If we put the two column vectors  $\mathbf{x}$  and  $\mathbf{y}$  in an  $n \times 2$  matrix  $\mathbf{Z} = [\mathbf{x}, \mathbf{y}]$ , then  $\mathbf{M}^0\mathbf{Z}$  is the  $n \times 2$  matrix in which the two columns of data are in mean deviation form. Then

$$(\mathbf{M}^0\mathbf{Z})'(\mathbf{M}^0\mathbf{Z}) = \mathbf{Z}'\mathbf{M}^0\mathbf{M}^0\mathbf{Z} = \mathbf{Z}'\mathbf{M}^0\mathbf{Z}.$$

### A.3 GEOMETRY OF MATRICES

#### A.3.1 VECTOR SPACES

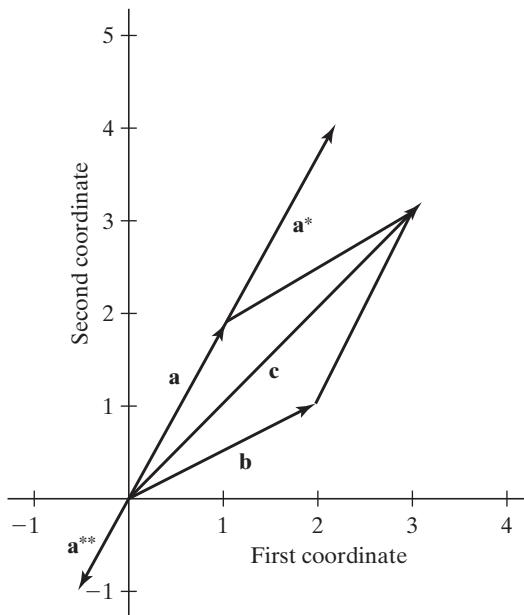
The  $K$  elements of a column vector

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_K \end{bmatrix}$$

can be viewed as the coordinates of a point in a  $K$ -dimensional space, as shown in Figure A.1 for two dimensions, or as the definition of the line segment connecting the origin and the point defined by  $\mathbf{a}$ .

Two basic arithmetic operations are defined for vectors, **scalar multiplication** and **addition**. A scalar multiple of a vector,  $\mathbf{a}$ , is another vector, say  $\mathbf{a}^*$ , whose coordinates are the scalar multiple of  $\mathbf{a}$ 's coordinates. Thus, in Figure A.1,

$$\mathbf{a} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{a}^* = 2\mathbf{a} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}, \quad \mathbf{a}^{**} = -\frac{1}{2}\mathbf{a} = \begin{bmatrix} -\frac{1}{2} \\ -1 \end{bmatrix}.$$

APPENDIX A ♦ Matrix Algebra **1049****FIGURE A.1** Vector Space.

The set of all possible scalar multiples of **a** is the line through the origin, **0** and **a**. Any scalar multiple of **a** is a segment of this line. The sum of two vectors **a** and **b** is a third vector whose coordinates are the sums of the corresponding coordinates of **a** and **b**. For example,

$$\mathbf{c} = \mathbf{a} + \mathbf{b} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}.$$

Geometrically, **c** is obtained by moving in the distance and direction defined by **b** from the tip of **a** or, because addition is commutative, from the tip of **b** in the distance and direction of **a**. Note that scalar multiplication and addition of vectors are special cases of (A-16) and (A-6) for matrices.

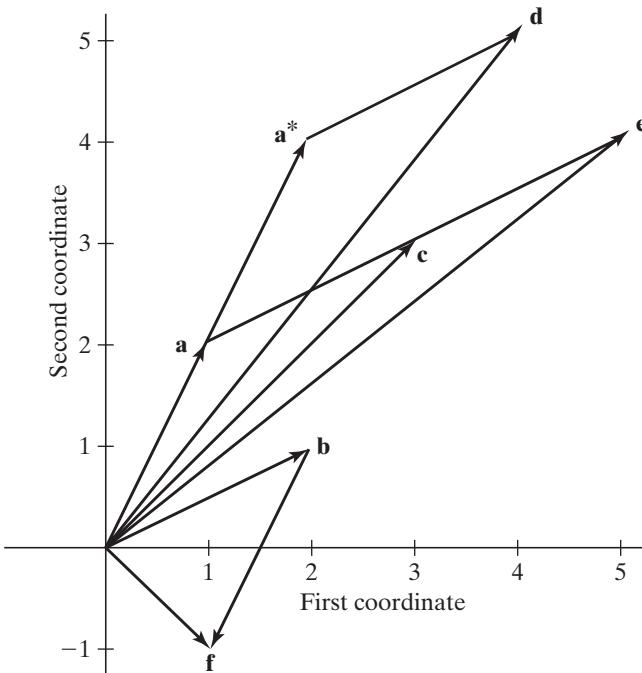
The two-dimensional plane is the set of all vectors with two real-valued coordinates. We label this set  $\mathbb{R}^2$  ("R two," not "R squared"). It has two important properties.

- $\mathbb{R}^2$  is closed under scalar multiplication; every scalar multiple of a vector in  $\mathbb{R}^2$  is also in  $\mathbb{R}^2$ .
- $\mathbb{R}^2$  is closed under addition; the sum of any two vectors in the plane is always a vector in  $\mathbb{R}^2$ .

**DEFINITION A.2** Vector Space

*A vector space is any set of vectors that is closed under scalar multiplication and addition.*

Another example is the set of all real numbers, that is,  $\mathbb{R}^1$ , that is, the set of vectors with one real element. In general, that set of  $K$ -element vectors all of whose elements are real numbers is a  $K$ -dimensional vector space, denoted  $\mathbb{R}^K$ . The preceding examples are drawn in  $\mathbb{R}^2$ .

**1050 PART VI ♦ Appendices**

**FIGURE A.2** Linear Combinations of Vectors.

**A.3.2 LINEAR COMBINATIONS OF VECTORS AND BASIS VECTORS**

In Figure A.2,  $\mathbf{c} = \mathbf{a} + \mathbf{b}$  and  $\mathbf{d} = \mathbf{a}^* + \mathbf{b}$ . But since  $\mathbf{a}^* = 2\mathbf{a}$ ,  $\mathbf{d} = 2\mathbf{a} + \mathbf{b}$ . Also,  $\mathbf{e} = \mathbf{a} + 2\mathbf{b}$  and  $\mathbf{f} = \mathbf{b} + (-\mathbf{a}) = \mathbf{b} - \mathbf{a}$ . As this exercise suggests, any vector in  $\mathbb{R}^2$  could be obtained as a **linear combination** of **a** and **b**.

**DEFINITION A.3 Basis Vectors**

*A set of vectors in a vector space is a **basis** for that vector space if they are linearly independent and any vector in the vector space can be written as a linear combination of that set of vectors.*

As is suggested by Figure A.2, any pair of two-element vectors, including **a** and **b**, that point in different directions will form a basis for  $\mathbb{R}^2$ . Consider an arbitrary set of vectors in  $\mathbb{R}^2$ , **a**, **b**, and **c**. If **a** and **b** are a basis, then we can find numbers  $\alpha_1$  and  $\alpha_2$  such that  $\mathbf{c} = \alpha_1\mathbf{a} + \alpha_2\mathbf{b}$ . Let

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}.$$

Then

$$\begin{aligned} c_1 &= \alpha_1 a_1 + \alpha_2 b_1, \\ c_2 &= \alpha_1 a_2 + \alpha_2 b_2. \end{aligned} \tag{A-40}$$

## APPENDIX A ♦ Matrix Algebra 1051

The solutions to this pair of equations are

$$\alpha_1 = \frac{b_2 c_1 - b_1 c_2}{a_1 b_2 - b_1 a_2}, \quad \alpha_2 = \frac{a_1 c_2 - a_2 c_1}{a_1 b_2 - b_1 a_2}. \quad (\text{A-41})$$

This result gives a unique solution unless  $(a_1 b_2 - b_1 a_2) = 0$ . If  $(a_1 b_2 - b_1 a_2) = 0$ , then  $a_1/a_2 = b_1/b_2$ , which means that **b** is just a multiple of **a**. This returns us to our original condition, that **a** and **b** must point in different directions. The implication is that if **a** and **b** are any pair of vectors for which the denominator in (A-41) is not zero, then any other vector **c** can be formed as a *unique* linear combination of **a** and **b**. The basis of a vector space is not unique, since any set of vectors that satisfies the definition will do. But for any particular basis, only one linear combination of them will produce another particular vector in the vector space.

### A.3.3 LINEAR DEPENDENCE

As the preceding should suggest,  $K$  vectors are required to form a basis for  $\mathbb{R}^K$ . Although the basis for a vector space is not unique, not every set of  $K$  vectors will suffice. In Figure A.2, **a** and **b** form a basis for  $\mathbb{R}^2$ , but **a** and **a\*** do not. The difference between these two pairs is that **a** and **b** are linearly *independent*, whereas **a** and **a\*** are linearly *dependent*.

#### DEFINITION A.4 Linear Dependence

*A set of  $k \geq 2$  vectors is **linearly dependent** if at least one of the vectors in the set can be written as a linear combination of the others.*

Because **a\*** is a multiple of **a**, **a** and **a\*** are linearly dependent. For another example, if

$$\mathbf{a} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \quad \text{and} \quad \mathbf{c} = \begin{bmatrix} 10 \\ 14 \end{bmatrix},$$

then

$$2\mathbf{a} + \mathbf{b} - \frac{1}{2}\mathbf{c} = \mathbf{0},$$

so **a**, **b**, and **c** are linearly dependent. Any of the three possible pairs of them, however, are linearly independent.

#### DEFINITION A.5 Linear Independence

*A set of vectors is **linearly independent** if and only if the only solution to*

$$\alpha_1 \mathbf{a}_1 + \alpha_2 \mathbf{a}_2 + \cdots + \alpha_K \mathbf{a}_K = \mathbf{0}$$

*is*

$$\alpha_1 = \alpha_2 = \cdots = \alpha_K = 0.$$

The preceding implies the following equivalent definition of a basis.

**1052 PART VI ♦ Appendices**
**DEFINITION A.6 Basis for a Vector Space**

*A basis for a vector space of  $K$  dimensions is any set of  $K$  linearly independent vectors in that vector space.*

Because any  $(K + 1)$ st vector can be written as a linear combination of the  $K$  basis vectors, it follows that any set of more than  $K$  vectors in  $\mathbb{R}^K$  must be linearly dependent.

**A.3.4 SUBSPACES**
**DEFINITION A.7 Spanning Vectors**

*The set of all linear combinations of a set of vectors is the vector space that is **spanned** by those vectors.*

For example, by definition, the space spanned by a basis for  $\mathbb{R}^K$  is  $\mathbb{R}^K$ . An implication of this is that if **a** and **b** are a basis for  $\mathbb{R}^2$  and **c** is another vector in  $\mathbb{R}^2$ , the space spanned by [**a**, **b**, **c**] is, again,  $\mathbb{R}^2$ . Of course, **c** is superfluous. Nonetheless, any vector in  $\mathbb{R}^2$  *can* be expressed as a linear combination of **a**, **b**, and **c**. (The linear combination will not be unique. Suppose, for example, that **a** and **c** are also a basis for  $\mathbb{R}^2$ .)

Consider the set of three coordinate vectors whose third element is zero. In particular,

$$\mathbf{a}' = [a_1 \ a_2 \ 0] \quad \text{and} \quad \mathbf{b}' = [b_1 \ b_2 \ 0].$$

Vectors **a** and **b** do not span the three-dimensional space  $\mathbb{R}^3$ . Every linear combination of **a** and **b** has a third coordinate equal to zero; thus, for instance,  $\mathbf{c}' = [1 \ 2 \ 3]$  could not be written as a linear combination of **a** and **b**. If  $(a_1 b_2 - a_2 b_1)$  is not equal to zero [see (A-41)]; however, then *any vector whose third element is zero can be expressed as a linear combination of **a** and **b***. So, although **a** and **b** do not span  $\mathbb{R}^3$ , they do span something, the set of vectors in  $\mathbb{R}^3$  whose third element is zero. This area is a plane (the “floor” of the box in a three-dimensional figure). This plane in  $\mathbb{R}^3$  is a **subspace**, in this instance, a two-dimensional subspace. Note that it is not  $\mathbb{R}^2$ ; it is the set of vectors in  $\mathbb{R}^3$  whose third coordinate is 0. Any plane in  $\mathbb{R}^3$  contains the origin,  $(0, 0, 0)$ , regardless of how it is oriented, forms a two-dimensional subspace. Any two independent vectors that lie in that subspace will span it. But without a third vector that points in some other direction, we cannot span any more of  $\mathbb{R}^3$  than this two-dimensional part of it. By the same logic, any line in  $\mathbb{R}^3$  that passes through the origin is a one-dimensional subspace, in this case, the set of all vectors in  $\mathbb{R}^3$  whose coordinates are multiples of those of the vector that define the line. A subspace is a vector space in all the respects in which we have defined it. We emphasize that it is *not* a vector space of lower dimension. For example,  $\mathbb{R}^2$  is not a subspace of  $\mathbb{R}^3$ . The essential difference is the number of dimensions in the vectors. The vectors in  $\mathbb{R}^3$  that form a two-dimensional subspace are still three-element vectors; they all just happen to lie in the same plane.

The space spanned by a set of vectors in  $\mathbb{R}^K$  has at most  $K$  dimensions. If this space has fewer than  $K$  dimensions, it is a subspace, or **hyperplane**. But the important point in the preceding discussion is that *every set of vectors spans some space*; it may be the entire space in which the vectors reside, or it may be some subspace of it.

## APPENDIX A ♦ Matrix Algebra 1053

### A.3.5 RANK OF A MATRIX

We view a matrix as a set of column vectors. The number of columns in the matrix equals the number of vectors in the set, and the number of rows equals the number of coordinates in each column vector.

#### **DEFINITION A.8 Column Space**

*The column space of a matrix is the vector space that is spanned by its column vectors.*

If the matrix contains  $K$  rows, its column space might have  $K$  dimensions. But, as we have seen, it might have fewer dimensions; the column vectors might be linearly dependent, or there might be fewer than  $K$  of them. Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 5 & 6 \\ 2 & 6 & 8 \\ 7 & 1 & 8 \end{bmatrix}.$$

It contains three vectors from  $\mathbb{R}^3$ , but the third is the sum of the first two, so the column space of this matrix cannot have three dimensions. Nor does it have only one, because the three columns are not all scalar multiples of one another. Hence, it has two, and the column space of this matrix is a two-dimensional subspace of  $\mathbb{R}^3$ .

#### **DEFINITION A.9 Column Rank**

*The column rank of a matrix is the dimension of the vector space that is spanned by its column vectors.*

It follows that the column rank of a matrix is equal to the largest number of linearly independent column vectors it contains. The column rank of  $\mathbf{A}$  is 2. For another specific example, consider

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 1 & 5 \\ 6 & 4 & 5 \\ 3 & 1 & 4 \end{bmatrix}.$$

It can be shown (we shall see how later) that this matrix has a column rank equal to 3. Each column of  $\mathbf{B}$  is a vector in  $\mathbb{R}^4$ , so the column space of  $\mathbf{B}$  is a three-dimensional subspace of  $\mathbb{R}^4$ .

Consider, instead, the set of vectors obtained by using the *rows* of  $\mathbf{B}$  instead of the columns. The new matrix would be

$$\mathbf{C} = \begin{bmatrix} 1 & 5 & 6 & 3 \\ 2 & 1 & 4 & 1 \\ 3 & 5 & 5 & 4 \end{bmatrix}.$$

This matrix is composed of four column vectors from  $\mathbb{R}^3$ . (Note that  $\mathbf{C}$  is  $\mathbf{B}'$ .) The column space of  $\mathbf{C}$  is at most  $\mathbb{R}^3$ , since four vectors in  $\mathbb{R}^3$  must be linearly dependent. In fact, the column space of

## 1054 PART VI ♦ Appendices

$\mathbf{C}$  is  $\mathbb{R}^3$ . Although this is not the same as the column space of  $\mathbf{B}$ , it does have the same dimension. Thus, the column rank of  $\mathbf{C}$  and the column rank of  $\mathbf{B}$  are the same. But the columns of  $\mathbf{C}$  are the rows of  $\mathbf{B}$ . Thus, the column rank of  $\mathbf{C}$  equals the **row rank** of  $\mathbf{B}$ . That the column and row ranks of  $\mathbf{B}$  are the same is not a coincidence. The general results (which are equivalent) are as follows.

### THEOREM A.1 Equality of Row and Column Rank

The **column rank** and **row rank** of a matrix are equal. By the definition of row rank and its counterpart for column rank, we obtain the corollary,

the **row space** and **column space** of a matrix have the same dimension. (A-42)

Theorem A.1 holds regardless of the actual row and column rank. If the column rank of a matrix happens to equal the number of columns it contains, then the matrix is said to have **full column rank**. **Full row rank** is defined likewise. Because the row and column ranks of a matrix are always equal, we can speak unambiguously of the **rank of a matrix**. For either the row rank or the column rank (and, at this point, we shall drop the distinction),

$$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}') \leq \min(\text{number of rows}, \text{number of columns}). \quad (\text{A-43})$$

In most contexts, we shall be interested in the columns of the matrices we manipulate. We shall use the term **full rank** to describe a matrix whose rank is equal to the number of columns it contains.

Of particular interest will be the distinction between **full rank** and **short rank matrices**. The distinction turns on the solutions to  $\mathbf{Ax} = \mathbf{0}$ . If a nonzero  $\mathbf{x}$  for which  $\mathbf{Ax} = \mathbf{0}$  exists, then  $\mathbf{A}$  does not have full rank. Equivalently, if the nonzero  $\mathbf{x}$  exists, then the columns of  $\mathbf{A}$  are linearly dependent and at least one of them can be expressed as a linear combination of the others. For example, a nonzero set of solutions to

$$\begin{bmatrix} 1 & 3 & 10 \\ 2 & 3 & 14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

is any multiple of  $\mathbf{x}' = (2, 1, -\frac{1}{2})$ .

In a product matrix  $\mathbf{C} = \mathbf{AB}$ , every column of  $\mathbf{C}$  is a linear combination of the columns of  $\mathbf{A}$ , so each column of  $\mathbf{C}$  is in the column space of  $\mathbf{A}$ . It is possible that the set of columns in  $\mathbf{C}$  could span this space, but it is not possible for them to span a higher-dimensional space. At best, they could be a full set of linearly independent vectors in  $\mathbf{A}$ 's column space. We conclude that the column rank of  $\mathbf{C}$  could not be greater than that of  $\mathbf{A}$ . Now, apply the same logic to the rows of  $\mathbf{C}$ , which are all linear combinations of the rows of  $\mathbf{B}$ . For the same reason that the column rank of  $\mathbf{C}$  cannot exceed the column rank of  $\mathbf{A}$ , the row rank of  $\mathbf{C}$  cannot exceed the row rank of  $\mathbf{B}$ . Row and column ranks are always equal, so we can conclude that

$$\text{rank}(\mathbf{AB}) \leq \min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B})). \quad (\text{A-44})$$

A useful corollary to (A-44) is

$$\text{If } \mathbf{A} \text{ is } M \times n \text{ and } \mathbf{B} \text{ is a square matrix of rank } n, \text{ then } \text{rank}(\mathbf{AB}) = \text{rank}(\mathbf{A}). \quad (\text{A-45})$$

## APPENDIX A ♦ Matrix Algebra 1055

Another application that plays a central role in the development of regression analysis is, for any matrix  $\mathbf{A}$ ,

$$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}'\mathbf{A}) = \text{rank}(\mathbf{A}\mathbf{A}'). \quad (\mathbf{A}-46)$$

### A.3.6 DETERMINANT OF A MATRIX

The determinant of a square matrix—determinants are not defined for nonsquare matrices—is a function of the elements of the matrix. There are various definitions, most of which are not useful for our work. Determinants figure into our results in several ways, however, that we can enumerate before we need formally to define the computations.

#### **PROPOSITION**

*The determinant of a matrix is nonzero if and only if it has full rank.*

Full rank and short rank matrices can be distinguished by whether or not their determinants are nonzero. There are some settings in which the value of the determinant is also of interest, so we now consider some algebraic results.

It is most convenient to begin with a diagonal matrix

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & 0 & \cdots & 0 \\ 0 & d_2 & 0 & \cdots & 0 \\ & & \ddots & & \\ 0 & 0 & 0 & \cdots & d_K \end{bmatrix}.$$

The column vectors of  $\mathbf{D}$  define a “box” in  $\mathbb{R}^K$  whose sides are all at right angles to one another.<sup>3</sup> Its “volume,” or determinant, is simply the product of the lengths of the sides, which we denote

$$|\mathbf{D}| = d_1 d_2 \dots d_K = \prod_{k=1}^K d_k. \quad (\mathbf{A}-47)$$

A special case is the identity matrix, which has, regardless of  $K$ ,  $|\mathbf{I}_K| = 1$ . Multiplying  $\mathbf{D}$  by a scalar  $c$  is equivalent to multiplying the length of each side of the box by  $c$ , which would multiply its volume by  $c^K$ . Thus,

$$|c\mathbf{D}| = c^K |\mathbf{D}|. \quad (\mathbf{A}-48)$$

Continuing with this admittedly special case, we suppose that only one column of  $\mathbf{D}$  is multiplied by  $c$ . In two dimensions, this would make the box wider but not higher, or vice versa. Hence, the “volume” (area) would also be multiplied by  $c$ . Now, suppose that each side of the box were multiplied by a different  $c$ , the first by  $c_1$ , the second by  $c_2$ , and so on. The volume would, by an obvious extension, now be  $c_1 c_2 \dots c_K |\mathbf{D}|$ . The matrix with columns defined by  $[c_1 \mathbf{d}_1 \ c_2 \mathbf{d}_2 \dots]$  is just  $\mathbf{DC}$ , where  $\mathbf{C}$  is a diagonal matrix with  $c_i$  as its  $i$ th diagonal element. The computation just described is, therefore,

$$|\mathbf{DC}| = |\mathbf{D}| \cdot |\mathbf{C}|. \quad (\mathbf{A}-49)$$

(The determinant of  $\mathbf{C}$  is the product of the  $c_i$ 's since  $\mathbf{C}$ , like  $\mathbf{D}$ , is a diagonal matrix.) In particular, note what happens to the whole thing if one of the  $c_i$ 's is zero.

---

<sup>3</sup>Each column vector defines a segment on one of the axes.

## 1056 PART VI ♦ Appendices

For  $2 \times 2$  matrices, the computation of the determinant is

$$\begin{vmatrix} a & c \\ b & d \end{vmatrix} = ad - bc. \quad (\text{A-50})$$

Notice that it is a function of all the elements of the matrix. This statement will be true, in general. For more than two dimensions, the determinant can be obtained by using an **expansion by cofactors**. Using *any* row, say,  $i$ , we obtain

$$|\mathbf{A}| = \sum_{k=1}^K a_{ik}(-1)^{i+k} |\mathbf{A}_{ik}|, \quad k = 1, \dots, K, \quad (\text{A-51})$$

where  $\mathbf{A}_{ik}$  is the matrix obtained from  $\mathbf{A}$  by deleting row  $i$  and column  $k$ . The determinant of  $\mathbf{A}_{ik}$  is called a **minor** of  $\mathbf{A}$ .<sup>4</sup> When the correct sign,  $(-1)^{i+k}$ , is added, it becomes a **cofactor**. This operation can be done using any column as well. For example, a  $4 \times 4$  determinant becomes a sum of four  $3 \times 3$ s, whereas a  $5 \times 5$  is a sum of five  $4 \times 4$ s, each of which is a sum of four  $3 \times 3$ s, and so on. Obviously, it is a good idea to base (A-51) on a row or column with many zeros in it, if possible. In practice, this rapidly becomes a heavy burden. It is unlikely, though, that you will ever calculate any determinants over  $3 \times 3$  without a computer. A  $3 \times 3$ , however, might be computed on occasion; if so, the following shortcut due to P. Sarrus will prove useful:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{31}a_{22}a_{13} - a_{21}a_{12}a_{33} - a_{11}a_{23}a_{32}.$$

Although (A-48) and (A-49) were given for diagonal matrices, they hold for general matrices **C** and **D**. One special case of (A-48) to note is that of  $c = -1$ . Multiplying a matrix by  $-1$  does not necessarily change the sign of its determinant. It does so only if the order of the matrix is odd. By using the expansion by cofactors formula, an additional result can be shown:

$$|\mathbf{A}| = |\mathbf{A}'| \quad (\text{A-52})$$

### A.3.7 A LEAST SQUARES PROBLEM

Given a vector  $\mathbf{y}$  and a matrix  $\mathbf{X}$ , we are interested in expressing  $\mathbf{y}$  as a linear combination of the columns of  $\mathbf{X}$ . There are two possibilities. If  $\mathbf{y}$  lies in the column space of  $\mathbf{X}$ , then we shall be able to find a vector  $\mathbf{b}$  such that

$$\mathbf{y} = \mathbf{X}\mathbf{b}. \quad (\text{A-53})$$

Figure A.3 illustrates such a case for three dimensions in which the two columns of  $\mathbf{X}$  both have a third coordinate equal to zero. Only  $\mathbf{y}$ 's whose third coordinate is zero, such as  $\mathbf{y}^0$  in the figure, can be expressed as  $\mathbf{X}\mathbf{b}$  for some  $\mathbf{b}$ . For the general case, assuming that  $\mathbf{y}$  is, indeed, in the column space of  $\mathbf{X}$ , we can find the coefficients  $\mathbf{b}$  by solving the set of equations in (A-53). The solution is discussed in the next section.

Suppose, however, that  $\mathbf{y}$  is not in the column space of  $\mathbf{X}$ . In the context of this example, suppose that  $\mathbf{y}$ 's third component is not zero. Then there is no  $\mathbf{b}$  such that (A-53) holds. We can, however, write

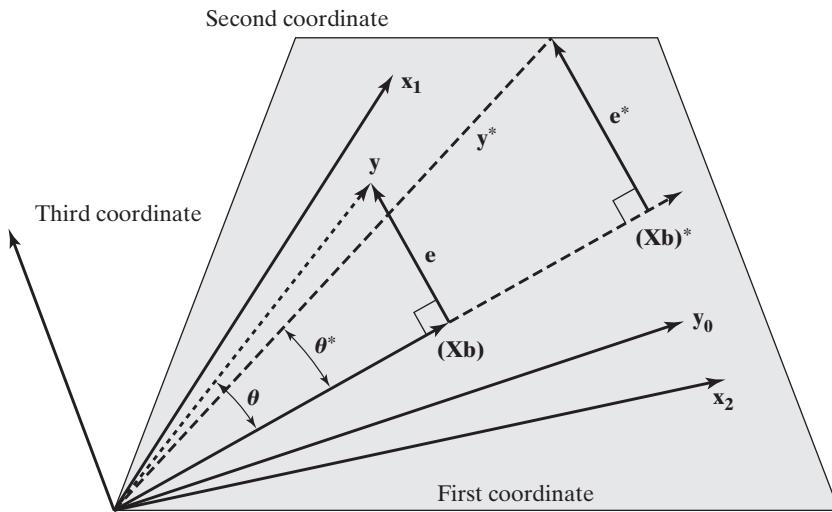
$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \quad (\text{A-54})$$

where  $\mathbf{e}$  is the difference between  $\mathbf{y}$  and  $\mathbf{X}\mathbf{b}$ . By this construction, we find an  $\mathbf{X}\mathbf{b}$  that is in the column space of  $\mathbf{X}$ , and  $\mathbf{e}$  is the difference, or “residual.” Figure A.3 shows two examples,  $\mathbf{y}$  and  $\mathbf{y}^*$ .

---

<sup>4</sup>If  $i$  equals  $k$ , then the determinant is a **principal minor**.

## APPENDIX A ♦ Matrix Algebra 1057

**FIGURE A.3** Least Squares Projections.

For the present, we consider only  $\mathbf{y}$ . We are interested in finding the  $\mathbf{b}$  such that  $\mathbf{y}$  is as close as possible to  $\mathbf{X}\mathbf{b}$  in the sense that  $\mathbf{e}$  is as short as possible.

**DEFINITION A.10 Length of a Vector**

*The length, or norm, of a vector  $\mathbf{e}$  is given by the Pythagorean theorem:*

$$\|\mathbf{e}\| = \sqrt{\mathbf{e}'\mathbf{e}}. \quad (\text{A-55})$$

The problem is to find the  $\mathbf{b}$  for which

$$\|\mathbf{e}\| = \|\mathbf{y} - \mathbf{X}\mathbf{b}\|$$

is as small as possible. The solution is that  $\mathbf{b}$  that makes  $\mathbf{e}$  perpendicular, or *orthogonal*, to  $\mathbf{X}\mathbf{b}$ .

**DEFINITION A.11 Orthogonal Vectors**

*Two nonzero vectors  $\mathbf{a}$  and  $\mathbf{b}$  are orthogonal, written  $\mathbf{a} \perp \mathbf{b}$ , if and only if*

$$\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a} = 0.$$

Returning once again to our fitting problem, we find that the  $\mathbf{b}$  we seek is that for which

$$\mathbf{e} \perp \mathbf{X}\mathbf{b}.$$

Expanding this set of equations gives the requirement

$$\begin{aligned} (\mathbf{X}\mathbf{b})'\mathbf{e} &= 0 \\ &= \mathbf{b}'\mathbf{X}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \\ &= \mathbf{b}'[\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\mathbf{b}], \end{aligned}$$

## 1058 PART VI ♦ Appendices

or, assuming  $\mathbf{b}$  is not  $\mathbf{0}$ , the set of equations

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}.$$

The means of solving such a set of equations is the subject of Section A.5.

In Figure A.3, the linear combination  $\mathbf{X}\mathbf{b}$  is called the **projection** of  $\mathbf{y}$  into the column space of  $\mathbf{X}$ . The figure is drawn so that, although  $\mathbf{y}$  and  $\mathbf{y}^*$  are different, they are similar in that the projection of  $\mathbf{y}$  lies on top of that of  $\mathbf{y}^*$ . The question we wish to pursue here is, Which vector,  $\mathbf{y}$  or  $\mathbf{y}^*$ , is closer to its projection in the column space of  $\mathbf{X}$ ? Superficially, it would appear that  $\mathbf{y}$  is closer, because  $\mathbf{e}$  is shorter than  $\mathbf{e}^*$ . Yet  $\mathbf{y}^*$  is much more nearly parallel to its projection than  $\mathbf{y}$ . A measure of comparison that would be unaffected by the length of the vectors is the angle between the vector and its projection (assuming that angle is not zero). By this measure,  $\theta^*$  is smaller than  $\theta$ , which would reverse the earlier conclusion.

### THEOREM A.2 The Cosine Law

*The angle  $\theta$  between two vectors  $\mathbf{a}$  and  $\mathbf{b}$  satisfies*

$$\cos \theta = \frac{\mathbf{a}'\mathbf{b}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|}.$$

The two vectors in the calculation would be  $\mathbf{y}$  or  $\mathbf{y}^*$  and  $\mathbf{X}\mathbf{b}$  or  $(\mathbf{X}\mathbf{b})^*$ . A zero cosine implies that the vectors are orthogonal. If the cosine is one, then the angle is zero, which means that the vectors are the same. (They would be if  $\mathbf{y}$  were in the column space of  $\mathbf{X}$ .) By dividing by the lengths, we automatically compensate for the length of  $\mathbf{y}$ . By this measure, we find in Figure A.3 that  $\mathbf{y}^*$  is closer to its projection,  $(\mathbf{X}\mathbf{b})^*$  than  $\mathbf{y}$  is to its projection,  $\mathbf{X}\mathbf{b}$ .

## A.4 SOLUTION OF A SYSTEM OF LINEAR EQUATIONS

Consider the set of  $n$  linear equations

$$\mathbf{Ax} = \mathbf{b}, \tag{A-56}$$

in which the  $K$  elements of  $\mathbf{x}$  constitute the unknowns.  $\mathbf{A}$  is a known matrix of coefficients, and  $\mathbf{b}$  is a specified vector of values. We are interested in knowing whether a solution exists; if so, then how to obtain it; and finally, if it does exist, then whether it is unique.

### A.4.1 SYSTEMS OF LINEAR EQUATIONS

For most of our applications, we shall consider only square systems of equations, that is, those in which  $\mathbf{A}$  is a square matrix. In what follows, therefore, we take  $n$  to equal  $K$ . Because the number of rows in  $\mathbf{A}$  is the number of equations, whereas the number of columns in  $\mathbf{A}$  is the number of variables, this case is the familiar one of “ $n$  equations in  $n$  unknowns.”

There are two types of systems of equations.

**DEFINITION A.12 Homogeneous Equation System**

*A homogeneous system is of the form  $\mathbf{Ax} = \mathbf{0}$ .*

By definition, a nonzero solution to such a system will exist if and only if  $\mathbf{A}$  does not have **full rank**. If so, then for at least one column of  $\mathbf{A}$ , we can write the preceding as

$$\mathbf{a}_k = - \sum_{m \neq k} \frac{x_m}{x_k} \mathbf{a}_m.$$

This means, as we know, that the columns of  $\mathbf{A}$  are linearly dependent and that  $|\mathbf{A}| = 0$ .

**DEFINITION A.13 Nonhomogeneous Equation System**

*A nonhomogeneous system of equations is of the form  $\mathbf{Ax} = \mathbf{b}$ , where  $\mathbf{b}$  is a nonzero vector.*

The vector  $\mathbf{b}$  is chosen arbitrarily and is to be expressed as a linear combination of the columns of  $\mathbf{A}$ . Because  $\mathbf{b}$  has  $K$  elements, this solution will exist only if the columns of  $\mathbf{A}$  span the entire  $K$ -dimensional space,  $\mathbb{R}^K$ .<sup>5</sup> Equivalently, we shall require that the columns of  $\mathbf{A}$  be linearly independent or that  $|\mathbf{A}|$  not be equal to zero.

**A.4.2 INVERSE MATRICES**

To solve the system  $\mathbf{Ax} = \mathbf{b}$  for  $\mathbf{x}$ , something akin to division by a matrix is needed. Suppose that we could find a square matrix  $\mathbf{B}$  such that  $\mathbf{BA} = \mathbf{I}$ . If the equation system is premultiplied by this  $\mathbf{B}$ , then the following would be obtained:

$$\mathbf{B}\mathbf{Ax} = \mathbf{Ix} = \mathbf{x} = \mathbf{Bb}. \quad (\text{A-57})$$

If the matrix  $\mathbf{B}$  exists, then it is the **inverse** of  $\mathbf{A}$ , denoted

$$\mathbf{B} = \mathbf{A}^{-1}.$$

From the definition,

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}.$$

In addition, by premultiplying by  $\mathbf{A}$ , postmultiplying by  $\mathbf{A}^{-1}$ , and then canceling terms, we find

$$\mathbf{AA}^{-1} = \mathbf{I}$$

as well.

If the inverse exists, then it must be unique. Suppose that it is not and that  $\mathbf{C}$  is a different inverse of  $\mathbf{A}$ . Then  $\mathbf{CAB} = \mathbf{CAB}$ , but  $(\mathbf{CA})\mathbf{B} = \mathbf{IB} = \mathbf{B}$  and  $\mathbf{C}(\mathbf{AB}) = \mathbf{C}$ , which would be a

---

<sup>5</sup>If  $\mathbf{A}$  does not have full rank, then the nonhomogeneous system will have solutions for *some* vectors  $\mathbf{b}$ , namely, any  $\mathbf{b}$  in the column space of  $\mathbf{A}$ . But we are interested in the case in which there are solutions for *all* nonzero vectors  $\mathbf{b}$ , which requires  $\mathbf{A}$  to have full rank.

## 1060 PART VI ♦ Appendices

contradiction if  $\mathbf{C}$  did not equal  $\mathbf{B}$ . Because, by (A-57), the solution is  $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ , the solution to the equation system is unique as well.

We now consider the calculation of the inverse matrix. For a  $2 \times 2$  matrix,  $\mathbf{AB} = \mathbf{I}$  implies that

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} = 1 \\ a_{11}b_{12} + a_{12}b_{22} = 0 \\ a_{21}b_{11} + a_{22}b_{21} = 0 \\ a_{21}b_{12} + a_{22}b_{22} = 1 \end{bmatrix}.$$

The solutions are

$$\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} = \frac{1}{|\mathbf{A}|} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}. \quad (\mathbf{A-58})$$

Notice the presence of the reciprocal of  $|\mathbf{A}|$  in  $\mathbf{A}^{-1}$ . This result is not specific to the  $2 \times 2$  case. We infer from it that if the determinant is zero, then the inverse does not exist.

### DEFINITION A.14 Nonsingular Matrix

*A matrix is nonsingular if and only if its inverse exists.*

The simplest inverse matrix to compute is that of a diagonal matrix. If

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & 0 & \cdots & 0 \\ 0 & d_2 & 0 & \cdots & 0 \\ \dots & & & & \\ 0 & 0 & 0 & \cdots & d_K \end{bmatrix}, \quad \text{then} \quad \mathbf{D}^{-1} = \begin{bmatrix} 1/d_1 & 0 & 0 & \cdots & 0 \\ 0 & 1/d_2 & 0 & \cdots & 0 \\ \dots & & & & \\ 0 & 0 & 0 & \cdots & 1/d_K \end{bmatrix},$$

which shows, incidentally, that  $\mathbf{I}^{-1} = \mathbf{I}$ .

We shall use  $a^{ik}$  to indicate the  $ik$ th element of  $\mathbf{A}^{-1}$ . The general formula for computing an inverse matrix is

$$a^{ik} = \frac{|\mathbf{C}_{ki}|}{|\mathbf{A}|}, \quad (\mathbf{A-59})$$

where  $|\mathbf{C}_{ki}|$  is the  $k$ th cofactor of  $\mathbf{A}$ . [See (A-51).] It follows, therefore, that for  $\mathbf{A}$  to be non-singular,  $|\mathbf{A}|$  must be nonzero. Notice the reversal of the subscripts

Some computational results involving inverses are

$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}, \quad (\mathbf{A-60})$$

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A}, \quad (\mathbf{A-61})$$

$$(\mathbf{A}^{-1})' = (\mathbf{A}')^{-1}. \quad (\mathbf{A-62})$$

$$\text{If } \mathbf{A} \text{ is symmetric, then } \mathbf{A}^{-1} \text{ is symmetric.} \quad (\mathbf{A-63})$$

When both inverse matrices exist,

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}. \quad (\mathbf{A-64})$$

## APPENDIX A ♦ Matrix Algebra 1061

Note the condition preceding (A-64). It may be that  $\mathbf{AB}$  is a square, nonsingular matrix when neither  $\mathbf{A}$  nor  $\mathbf{B}$  is even square. (Consider, e.g.,  $\mathbf{A}'\mathbf{A}$ .) Extending (A-64), we have

$$(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}(\mathbf{AB})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}. \quad (\text{A-65})$$

Recall that for a data matrix  $\mathbf{X}$ ,  $\mathbf{X}'\mathbf{X}$  is the sum of the *outer products* of the rows  $\mathbf{X}$ . Suppose that we have already computed  $\mathbf{S} = (\mathbf{X}'\mathbf{X})^{-1}$  for a number of years of data, such as those given in Table A.1. The following result, which is called an **updating formula**, shows how to compute the new  $\mathbf{S}$  that would result when a new row is added to  $\mathbf{X}$ : For symmetric, nonsingular matrix  $\mathbf{A}$ ,

$$[\mathbf{A} \pm \mathbf{bb}']^{-1} = \mathbf{A}^{-1} \mp \left[ \frac{1}{1 \pm \mathbf{b}'\mathbf{A}^{-1}\mathbf{b}} \right] \mathbf{A}^{-1}\mathbf{bb}'\mathbf{A}^{-1}. \quad (\text{A-66})$$

Note the reversal of the sign in the inverse. Two more general forms of (A-66) that are occasionally useful are

$$[\mathbf{A} \pm \mathbf{bc}']^{-1} = \mathbf{A}^{-1} \mp \left[ \frac{1}{1 \pm \mathbf{c}'\mathbf{A}^{-1}\mathbf{b}} \right] \mathbf{A}^{-1}\mathbf{bc}'\mathbf{A}^{-1}. \quad (\text{A-66a})$$

$$[\mathbf{A} \pm \mathbf{BCB}']^{-1} = \mathbf{A}^{-1} \mp \mathbf{A}^{-1}\mathbf{B}[\mathbf{C}^{-1} \pm \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}]^{-1}\mathbf{B}'\mathbf{A}^{-1}. \quad (\text{A-66b})$$

### A.4.3 NONHOMOGENEOUS SYSTEMS OF EQUATIONS

For the nonhomogeneous system

$$\mathbf{Ax} = \mathbf{b},$$

if  $\mathbf{A}$  is nonsingular, then the unique solution is

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}.$$

### A.4.4 SOLVING THE LEAST SQUARES PROBLEM

We now have the tool needed to solve the least squares problem posed in Section A3.7. We found the solution vector,  $\mathbf{b}$  to be the solution to the nonhomogenous system  $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{b}$ . Let  $\mathbf{a}$  equal the vector  $\mathbf{X}'\mathbf{y}$  and let  $\mathbf{A}$  equal the square matrix  $\mathbf{X}'\mathbf{X}$ . The equation system is then

$$\mathbf{Ab} = \mathbf{a}.$$

By the preceding results, if  $\mathbf{A}$  is nonsingular, then

$$\mathbf{b} = \mathbf{A}^{-1}\mathbf{a} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})$$

assuming that the matrix to be inverted is nonsingular. We have reached the irreducible minimum. If the columns of  $\mathbf{X}$  are linearly independent, that is, if  $\mathbf{X}$  has full rank, then this is the solution to the least squares problem. If the columns of  $\mathbf{X}$  are linearly dependent, then this system has no unique solution.

## A.5 PARTITIONED MATRICES

In formulating the elements of a matrix, it is sometimes useful to group some of the elements in **submatrices**. Let

$$\mathbf{A} = \left[ \begin{array}{cc|c} 1 & 4 & 5 \\ 2 & 9 & 3 \\ \hline 8 & 9 & 6 \end{array} \right] = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}.$$

## 1062 PART VI ♦ Appendices

**A** is a **partitioned matrix**. The subscripts of the submatrices are defined in the same fashion as those for the elements of a matrix. A common special case is the **block-diagonal matrix**:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix},$$

where  $\mathbf{A}_{11}$  and  $\mathbf{A}_{22}$  are square matrices.

### A.5.1 ADDITION AND MULTIPLICATION OF PARTITIONED MATRICES

For conformably partitioned matrices **A** and **B**,

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} \mathbf{A}_{11} + \mathbf{B}_{11} & \mathbf{A}_{12} + \mathbf{B}_{12} \\ \mathbf{A}_{21} + \mathbf{B}_{21} & \mathbf{A}_{22} + \mathbf{B}_{22} \end{bmatrix}, \quad (\text{A-67})$$

and

$$\mathbf{AB} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{bmatrix}. \quad (\text{A-68})$$

In all these, the matrices must be conformable for the operations involved. For addition, the dimensions of  $\mathbf{A}_{ik}$  and  $\mathbf{B}_{ik}$  must be the same. For multiplication, the number of columns in  $\mathbf{A}_{ij}$  must equal the number of rows in  $\mathbf{B}_{jl}$  for all pairs  $i$  and  $j$ . That is, all the necessary matrix products of the submatrices must be defined. Two cases frequently encountered are of the form

$$\begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix}' \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix} = [\mathbf{A}'_1 \quad \mathbf{A}'_2] \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix} = [\mathbf{A}'_1\mathbf{A}_1 + \mathbf{A}'_2\mathbf{A}_2], \quad (\text{A-69})$$

and

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}' \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{A}'_{11}\mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}'_{22}\mathbf{A}_{22} \end{bmatrix}. \quad (\text{A-70})$$

### A.5.2 DETERMINANTS OF PARTITIONED MATRICES

The determinant of a block-diagonal matrix is obtained analogously to that of a diagonal matrix:

$$\begin{vmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{A}_{11}| \cdot |\mathbf{A}_{22}|. \quad (\text{A-71})$$

The determinant of a general  $2 \times 2$  partitioned matrix is

$$\begin{vmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{A}_{22}| \cdot |\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}| = |\mathbf{A}_{11}| \cdot |\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}|. \quad (\text{A-72})$$

### A.5.3 INVERSES OF PARTITIONED MATRICES

The inverse of a block-diagonal matrix is

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22}^{-1} \end{bmatrix}, \quad (\text{A-73})$$

which can be verified by direct multiplication.

## APPENDIX A ♦ Matrix Algebra 1063

For the general  $2 \times 2$  partitioned matrix, one form of the **partitioned inverse** is

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1}(\mathbf{I} + \mathbf{A}_{12}\mathbf{F}_2\mathbf{A}_{21}\mathbf{A}_{11}^{-1}) & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{F}_2 \\ -\mathbf{F}_2\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{F}_2 \end{bmatrix}, \quad (\text{A-74})$$

where

$$\mathbf{F}_2 = (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}.$$

The upper left block could also be written as

$$\mathbf{F}_1 = (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}.$$

#### A.5.4 DEVIATIONS FROM MEANS

Suppose that we begin with a column vector of  $n$  values  $\mathbf{x}$  and let

$$\mathbf{A} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} = \begin{bmatrix} \mathbf{i}'\mathbf{i} & \mathbf{i}'\mathbf{x} \\ \mathbf{x}'\mathbf{i} & \mathbf{x}'\mathbf{x} \end{bmatrix}.$$

We are interested in the lower-right-hand element of  $\mathbf{A}^{-1}$ . Upon using the definition of  $\mathbf{F}_2$  in (A-74), this is

$$\begin{aligned} \mathbf{F}_2 &= [\mathbf{x}'\mathbf{x} - (\mathbf{x}'\mathbf{i})(\mathbf{i}'\mathbf{i})^{-1}(\mathbf{i}'\mathbf{x})]^{-1} = \left\{ \mathbf{x}' \left[ \mathbf{I}\mathbf{x} - \mathbf{i} \left( \frac{1}{n} \right) \mathbf{i}'\mathbf{x} \right] \right\}^{-1} \\ &= \left\{ \mathbf{x}' \left[ \mathbf{I} - \left( \frac{1}{n} \right) \mathbf{i}\mathbf{i}' \right] \mathbf{x} \right\}^{-1} = (\mathbf{x}'\mathbf{M}^0\mathbf{x})^{-1}. \end{aligned}$$

Therefore, the lower-right-hand value in the inverse matrix is

$$(\mathbf{x}'\mathbf{M}^0\mathbf{x})^{-1} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} = a^{22}.$$

Now, suppose that we replace  $\mathbf{x}$  with  $\mathbf{X}$ , a matrix with several columns. We seek the lower-right block of  $(\mathbf{Z}'\mathbf{Z})^{-1}$ , where  $\mathbf{Z} = [\mathbf{i}, \mathbf{X}]$ . The analogous result is

$$(\mathbf{Z}'\mathbf{Z})^{22} = [\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}'\mathbf{X}]^{-1} = (\mathbf{X}'\mathbf{M}^0\mathbf{X})^{-1},$$

which implies that the  $K \times K$  matrix in the lower-right corner of  $(\mathbf{Z}'\mathbf{Z})^{-1}$  is the inverse of the  $K \times K$  matrix whose  $jk$ th element is  $\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ . Thus, when a data matrix contains a column of ones, the elements of the inverse of the matrix of sums of squares and cross products will be computed from the original data in the form of deviations from the respective column means.

#### A.5.5 KRONECKER PRODUCTS

A calculation that helps to condense the notation when dealing with sets of regression models (see Chapter 10) is the **Kronecker product**. For general matrices  $\mathbf{A}$  and  $\mathbf{B}$ ,

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1K}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2K}\mathbf{B} \\ \vdots & & & \\ a_{n1}\mathbf{B} & a_{n2}\mathbf{B} & \cdots & a_{nK}\mathbf{B} \end{bmatrix}. \quad (\text{A-75})$$

**1064 PART VI ♦ Appendices**

Notice that there is no requirement for conformability in this operation. The Kronecker product can be computed for any pair of matrices. If  $\mathbf{A}$  is  $K \times L$  and  $\mathbf{B}$  is  $m \times n$ , then  $\mathbf{A} \otimes \mathbf{B}$  is  $(Km) \times (Ln)$ .

For the Kronecker product,

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = (\mathbf{A}^{-1} \otimes \mathbf{B}^{-1}), \quad (\text{A-76})$$

If  $\mathbf{A}$  is  $M \times M$  and  $\mathbf{B}$  is  $n \times n$ , then

$$\begin{aligned} |\mathbf{A} \otimes \mathbf{B}| &= |\mathbf{A}|^n |\mathbf{B}|^M, \\ (\mathbf{A} \otimes \mathbf{B})' &= \mathbf{A}' \otimes \mathbf{B}', \\ \text{trace}(\mathbf{A} \otimes \mathbf{B}) &= \text{tr}(\mathbf{A})\text{tr}(\mathbf{B}). \end{aligned}$$

For  $\mathbf{A}$ ,  $\mathbf{B}$ ,  $\mathbf{C}$ , and  $\mathbf{D}$  such that the products are defined is

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}.$$

**A.6 CHARACTERISTIC ROOTS AND VECTORS**

A useful set of results for analyzing a square matrix  $\mathbf{A}$  arises from the solutions to the set of equations

$$\mathbf{Ac} = \lambda \mathbf{c}. \quad (\text{A-77})$$

The pairs of solutions are the **characteristic vectors**  $\mathbf{c}$  and **characteristic roots**  $\lambda$ . If  $\mathbf{c}$  is any nonzero solution vector, then  $k\mathbf{c}$  is also for any value of  $k$ . To remove the indeterminacy,  $\mathbf{c}$  is **normalized** so that  $\mathbf{c}'\mathbf{c} = 1$ .

The solution then consists of  $\lambda$  and the  $n - 1$  unknown elements in  $\mathbf{c}$ .

**A.6.1 THE CHARACTERISTIC EQUATION**

Solving (A-77) can, in principle, proceed as follows. First, (A-77) implies that

$$\mathbf{Ac} = \lambda \mathbf{I}\mathbf{c},$$

or that

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{c} = \mathbf{0}.$$

This equation is a homogeneous system that has a nonzero solution only if the matrix  $(\mathbf{A} - \lambda \mathbf{I})$  is singular or has a zero determinant. Therefore, if  $\lambda$  is a solution, then

$$|\mathbf{A} - \lambda \mathbf{I}| = 0. \quad (\text{A-78})$$

This polynomial in  $\lambda$  is the **characteristic equation** of  $\mathbf{A}$ . For example, if

$$\mathbf{A} = \begin{bmatrix} 5 & 1 \\ 2 & 4 \end{bmatrix},$$

then

$$|\mathbf{A} - \lambda \mathbf{I}| = \begin{vmatrix} 5 - \lambda & 1 \\ 2 & 4 - \lambda \end{vmatrix} = (5 - \lambda)(4 - \lambda) - 2(1) = \lambda^2 - 9\lambda + 18.$$

The two solutions are  $\lambda = 6$  and  $\lambda = 3$ .

**APPENDIX A ♦ Matrix Algebra 1065**

In solving the characteristic equation, there is no guarantee that the characteristic roots will be real. In the preceding example, if the 2 in the lower-left-hand corner of the matrix were  $-2$  instead, then the solution would be a pair of complex values. The same result can emerge in the general  $n \times n$  case. The characteristic roots of a symmetric matrix such as  $\mathbf{X}'\mathbf{X}$  are real, however.<sup>6</sup> This result will be convenient because most of our applications will involve the characteristic roots and vectors of symmetric matrices.

For an  $n \times n$  matrix, the characteristic equation is an  $n$ th-order polynomial in  $\lambda$ . Its solutions may be  $n$  distinct values, as in the preceding example, or may contain repeated values of  $\lambda$ , and may contain some zeros as well.

**A.6.2 CHARACTERISTIC VECTORS**

With  $\lambda$  in hand, the characteristic vectors are derived from the original problem,

$$\mathbf{A}\mathbf{c} = \lambda\mathbf{c},$$

or

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{c} = \mathbf{0}. \quad (\text{A-79})$$

Neither pair determines the values of  $c_1$  and  $c_2$ . But this result was to be expected; it was the reason  $\mathbf{c}'\mathbf{c} = 1$  was specified at the outset. The additional equation  $\mathbf{c}'\mathbf{c} = 1$ , however, produces complete solutions for the vectors.

**A.6.3 GENERAL RESULTS FOR CHARACTERISTIC ROOTS AND VECTORS**

A  $K \times K$  symmetric matrix has  $K$  distinct characteristic vectors,  $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K$ . The corresponding characteristic roots,  $\lambda_1, \lambda_2, \dots, \lambda_K$ , although real, need not be distinct. The characteristic vectors of a symmetric matrix are orthogonal,<sup>7</sup> which implies that for every  $i \neq j$ ,  $\mathbf{c}_i'\mathbf{c}_j = 0$ .<sup>8</sup> It is convenient to collect the  $K$ -characteristic vectors in a  $K \times K$  matrix whose  $i$ th column is the  $\mathbf{c}_i$  corresponding to  $\lambda_i$ ,

$$\mathbf{C} = [\mathbf{c}_1 \ \mathbf{c}_2 \ \cdots \ \mathbf{c}_K],$$

and the  $K$ -characteristic roots in the same order, in a diagonal matrix,

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \lambda_K \end{bmatrix}.$$

Then, the full set of equations

$$\mathbf{A}\mathbf{c}_k = \lambda_k\mathbf{c}_k$$

is contained in

$$\mathbf{AC} = \mathbf{C}\Lambda. \quad (\text{A-80})$$

<sup>6</sup>A proof may be found in Theil (1971).

<sup>7</sup>For proofs of these propositions, see Strang (1988).

<sup>8</sup>This statement is not true if the matrix is not symmetric. For instance, it does not hold for the characteristic vectors computed in the first example. For nonsymmetric matrices, there is also a distinction between “right” characteristic vectors,  $\mathbf{Ac} = \lambda\mathbf{c}$ , and “left” characteristic vectors,  $\mathbf{d}'\mathbf{A} = \lambda\mathbf{d}'$ , which may not be equal.

**1066 PART VI ♦ Appendices**

Because the vectors are orthogonal and  $\mathbf{c}_i' \mathbf{c}_i = 1$ , we have

$$\mathbf{C}' \mathbf{C} = \begin{bmatrix} \mathbf{c}_1' \mathbf{c}_1 & \mathbf{c}_1' \mathbf{c}_2 & \cdots & \mathbf{c}_1' \mathbf{c}_K \\ \mathbf{c}_2' \mathbf{c}_1 & \mathbf{c}_2' \mathbf{c}_2 & \cdots & \mathbf{c}_2' \mathbf{c}_K \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{c}_K' \mathbf{c}_1 & \mathbf{c}_K' \mathbf{c}_2 & \cdots & \mathbf{c}_K' \mathbf{c}_K \end{bmatrix} = \mathbf{I}. \quad (\text{A-81})$$

Result (A-81) implies that

$$\mathbf{C}' = \mathbf{C}^{-1}. \quad (\text{A-82})$$

Consequently,

$$\mathbf{C} \mathbf{C}' = \mathbf{C} \mathbf{C}^{-1} = \mathbf{I} \quad (\text{A-83})$$

as well, so the rows as well as the columns of  $\mathbf{C}$  are orthogonal.

#### A.6.4 DIAGONALIZATION AND SPECTRAL DECOMPOSITION OF A MATRIX

By premultiplying (A-80) by  $\mathbf{C}'$  and using (A-81), we can extract the characteristic roots of  $\mathbf{A}$ .

#### **DEFINITION A.15 Diagonalization of a Matrix**

The *diagonalization* of a matrix  $\mathbf{A}$  is

$$\mathbf{C}' \mathbf{A} \mathbf{C} = \mathbf{C}' \mathbf{C} \Lambda = \mathbf{I} \Lambda = \Lambda. \quad (\text{A-84})$$

Alternatively, by postmultiplying (A-80) by  $\mathbf{C}'$  and using (A-83), we obtain a useful representation of  $\mathbf{A}$ .

#### **DEFINITION A.16 Spectral Decomposition of a Matrix**

The *spectral decomposition* of  $\mathbf{A}$  is

$$\mathbf{A} = \mathbf{C} \Lambda \mathbf{C}' = \sum_{k=1}^K \lambda_k \mathbf{c}_k \mathbf{c}_k'. \quad (\text{A-85})$$

In this representation, the  $K \times K$  matrix  $\mathbf{A}$  is written as a sum of  $K$  rank one matrices. This sum is also called the **eigenvalue** (or, “own” value) decomposition of  $\mathbf{A}$ . In this connection, the term *signature* of the matrix is sometimes used to describe the characteristic roots and vectors. Yet another pair of terms for the parts of this decomposition are the **latent roots** and **latent vectors** of  $\mathbf{A}$ .

#### A.6.5 RANK OF A MATRIX

The diagonalization result enables us to obtain the rank of a matrix very easily. To do so, we can use the following result.

**THEOREM A.3 Rank of a Product**

*For any matrix  $\mathbf{A}$  and nonsingular matrices  $\mathbf{B}$  and  $\mathbf{C}$ , the rank of  $\mathbf{BAC}$  is equal to the rank of  $\mathbf{A}$ .*

**Proof:** By (A-45),  $\text{rank}(\mathbf{BAC}) = \text{rank}[(\mathbf{BA})\mathbf{C}] = \text{rank}(\mathbf{BA})$ . By (A-43),  $\text{rank}(\mathbf{BA}) = \text{rank}(\mathbf{A}'\mathbf{B}')$ , and applying (A-45) again,  $\text{rank}(\mathbf{A}'\mathbf{B}') = \text{rank}(\mathbf{A}')$  because  $\mathbf{B}'$  is nonsingular if  $\mathbf{B}$  is nonsingular [once again, by (A-43)]. Finally, applying (A-43) again to obtain  $\text{rank}(\mathbf{A}') = \text{rank}(\mathbf{A})$  gives the result.

Because  $\mathbf{C}$  and  $\mathbf{C}'$  are nonsingular, we can use them to apply this result to (A-84). By an obvious substitution,

$$\text{rank}(\mathbf{A}) = \text{rank}(\Lambda). \quad (\text{A-86})$$

Finding the rank of  $\Lambda$  is trivial. Because  $\Lambda$  is a diagonal matrix, its rank is just the number of nonzero values on its diagonal. By extending this result, we can prove the following theorems. (Proofs are brief and are left for the reader.)

**THEOREM A.4 Rank of a Symmetric Matrix**

*The rank of a symmetric matrix is the number of nonzero characteristic roots it contains.*

Note how this result enters the spectral decomposition given earlier. If any of the characteristic roots are zero, then the number of rank one matrices in the sum is reduced correspondingly. It would appear that this simple rule will not be useful if  $\mathbf{A}$  is not square. But recall that

$$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}'\mathbf{A}). \quad (\text{A-87})$$

Because  $\mathbf{A}'\mathbf{A}$  is always square, we can use it instead of  $\mathbf{A}$ . Indeed, we can use it even if  $\mathbf{A}$  is square, which leads to a fully general result.

**THEOREM A.5 Rank of a Matrix**

*The rank of any matrix  $\mathbf{A}$  equals the number of nonzero characteristic roots in  $\mathbf{A}'\mathbf{A}$ .*

The row rank and column rank of a matrix are equal, so we should be able to apply Theorem A.5 to  $\mathbf{AA}'$  as well. This process, however, requires an additional result.

**THEOREM A.6 Roots of an Outer Product Matrix**

*The nonzero characteristic roots of  $\mathbf{AA}'$  are the same as those of  $\mathbf{A}'\mathbf{A}$ .*

## 1068 PART VI ♦ Appendices

The proof is left as an exercise. A useful special case the reader can examine is the characteristic roots of  $\mathbf{aa}'$  and  $\mathbf{a}'\mathbf{a}$ , where  $\mathbf{a}$  is an  $n \times 1$  vector.

If a characteristic root of a matrix is zero, then we have  $\mathbf{Ac} = \mathbf{0}$ . Thus, if the matrix has a zero root, it must be singular. Otherwise, no nonzero  $\mathbf{c}$  would exist. In general, therefore, a matrix is singular; that is, it does not have full rank if and only if it has at least one zero root.

### A.6.6 CONDITION NUMBER OF A MATRIX

As the preceding might suggest, there is a discrete difference between full rank and short rank matrices. In analyzing data matrices such as the one in Section A.2, however, we shall often encounter cases in which a matrix is not quite short ranked, because it has all nonzero roots, but it is close. That is, by some measure, we can come very close to being able to write one column as a linear combination of the others. This case is important; we shall examine it at length in our discussion of multicollinearity in Section 4.7.1. Our definitions of rank and determinant will fail to indicate this possibility, but an alternative measure, the **condition number**, is designed for that purpose. Formally, the condition number for a square matrix  $\mathbf{A}$  is

$$\gamma = \left[ \frac{\text{maximum root}}{\text{minimum root}} \right]^{1/2}. \quad (\text{A-88})$$

For nonsquare matrices  $\mathbf{X}$ , such as the data matrix in the example, we use  $\mathbf{A} = \mathbf{X}'\mathbf{X}$ . As a further refinement, because the characteristic roots are affected by the scaling of the columns of  $\mathbf{X}$ , we scale the columns to have length 1 by dividing each column by its norm [see (A-55)]. For the  $\mathbf{X}$  in Section A.2, the largest characteristic root of  $\mathbf{A}$  is 4.9255 and the smallest is 0.0001543. Therefore, the condition number is 178.67, which is extremely large. (Values greater than 20 are large.) That the smallest root is close to zero compared with the largest means that this matrix is nearly singular. Matrices with large condition numbers are difficult to invert accurately.

### A.6.7 TRACE OF A MATRIX

The **trace** of a square  $K \times K$  matrix is the sum of its diagonal elements:

$$\text{tr}(\mathbf{A}) = \sum_{k=1}^K a_{kk}.$$

Some easily proven results are

$$\text{tr}(c\mathbf{A}) = c(\text{tr}(\mathbf{A})), \quad (\text{A-89})$$

$$\text{tr}(\mathbf{A}') = \text{tr}(\mathbf{A}), \quad (\text{A-90})$$

$$\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B}), \quad (\text{A-91})$$

$$\text{tr}(\mathbf{I}_K) = K. \quad (\text{A-92})$$

$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA}). \quad (\text{A-93})$$

$$\mathbf{a}'\mathbf{a} = \text{tr}(\mathbf{a}'\mathbf{a}) = \text{tr}(\mathbf{aa}')$$

$$\text{tr}(\mathbf{A}'\mathbf{A}) = \sum_{k=1}^K \mathbf{a}'_k \mathbf{a}_k = \sum_{i=1}^K \sum_{k=1}^K a_{ik}^2.$$

The permutation rule can be extended to any *cyclic* permutation in a product:

$$\text{tr}(\mathbf{ABCD}) = \text{tr}(\mathbf{BCDA}) = \text{tr}(\mathbf{CDAB}) = \text{tr}(\mathbf{DABC}). \quad (\text{A-94})$$

**APPENDIX A ♦ Matrix Algebra 1069**

By using (A-84), we obtain

$$\text{tr}(\mathbf{C}'\mathbf{AC}) = \text{tr}(\mathbf{ACC}') = \text{tr}(\mathbf{AI}) = \text{tr}(\mathbf{A}) = \text{tr}(\Lambda). \quad (\text{A-95})$$

Because  $\mathbf{A}$  is diagonal with the roots of  $\mathbf{A}$  on its diagonal, the general result is the following.

**THEOREM A.7 Trace of a Matrix**

*The trace of a matrix equals the sum of its characteristic roots.*

(A-96)

**A.6.8 DETERMINANT OF A MATRIX**

Recalling how tedious the calculation of a determinant promised to be, we find that the following is particularly useful. Because

$$\begin{aligned} \mathbf{C}'\mathbf{AC} &= \Lambda, \\ |\mathbf{C}'\mathbf{AC}| &= |\Lambda|. \end{aligned} \quad (\text{A-97})$$

Using a number of earlier results, we have, for orthogonal matrix  $\mathbf{C}$ ,

$$\begin{aligned} |\mathbf{C}'\mathbf{AC}| &= |\mathbf{C}'| \cdot |\mathbf{A}| \cdot |\mathbf{C}| = |\mathbf{C}'| \cdot |\mathbf{C}| \cdot |\mathbf{A}| = |\mathbf{C}'\mathbf{C}| \cdot |\mathbf{A}| = |\mathbf{I}| \cdot |\mathbf{A}| = 1 \cdot |\mathbf{A}| \\ &= |\mathbf{A}| \\ &= |\Lambda|. \end{aligned} \quad (\text{A-98})$$

Because  $|\Lambda|$  is just the product of its diagonal elements, the following is implied.

**THEOREM A.8 Determinant of a Matrix**

*The determinant of a matrix equals the product of its characteristic roots.*

(A-99)

Notice that we get the expected result if any of these roots is zero. The determinant is the product of the roots, so it follows that a matrix is singular if and only if its determinant is zero and, in turn, if and only if it has at least one zero characteristic root.

**A.6.9 POWERS OF A MATRIX**

We often use expressions involving powers of matrices, such as  $\mathbf{AA} = \mathbf{A}^2$ . For positive integer powers, these expressions can be computed by repeated multiplication. But this does not show how to handle a problem such as finding a  $\mathbf{B}$  such that  $\mathbf{B}^2 = \mathbf{A}$ , that is, the square root of a matrix. The characteristic roots and vectors provide a solution. Consider first

$$\begin{aligned} \mathbf{AA} &= \mathbf{A}^2 = (\mathbf{CAC}')(\mathbf{CAC}') = \mathbf{CAC}'\mathbf{CAC}' = \mathbf{C}\Lambda\mathbf{C}' = \mathbf{C}\Lambda\Lambda\mathbf{C}' \\ &= \mathbf{C}\Lambda^2\mathbf{C}'. \end{aligned} \quad (\text{A-100})$$

Two results follow. Because  $\Lambda^2$  is a diagonal matrix whose nonzero elements are the squares of those in  $\Lambda$ , the following is implied.

*For any symmetric matrix, the characteristic roots of  $\mathbf{A}^2$  are the squares of those of  $\mathbf{A}$ , and the characteristic vectors are the same.* (A-101)

## 1070 PART VI ♦ Appendices

The proof is obtained by observing that the second line in (A-100) is the spectral decomposition of the matrix  $\mathbf{B} = \mathbf{AA}$ . Because  $\mathbf{A}^3 = \mathbf{AA}^2$  and so on, (A-101) extends to any positive integer. By convention, for any  $\mathbf{A}$ ,  $\mathbf{A}^0 = \mathbf{I}$ . Thus, for any symmetric matrix  $\mathbf{A}$ ,  $\mathbf{A}^K = \mathbf{CA}^K\mathbf{C}'$ ,  $K = 0, 1, \dots$ . Hence, the characteristic roots of  $\mathbf{A}^K$  are  $\lambda^K$ , whereas the characteristic vectors are the same as those of  $\mathbf{A}$ . If  $\mathbf{A}$  is nonsingular, so that all its roots  $\lambda_i$  are nonzero, then this proof can be extended to negative powers as well.

If  $\mathbf{A}^{-1}$  exists, then

$$\mathbf{A}^{-1} = (\mathbf{CAC}')^{-1} = (\mathbf{C}')^{-1}\mathbf{A}^{-1}\mathbf{C}^{-1} = \mathbf{CA}^{-1}\mathbf{C}', \quad (\text{A-102})$$

where we have used the earlier result,  $\mathbf{C}' = \mathbf{C}^{-1}$ . This gives an important result that is useful for analyzing inverse matrices.

### THEOREM A.9 Characteristic Roots of an Inverse Matrix

*If  $\mathbf{A}^{-1}$  exists, then the characteristic roots of  $\mathbf{A}^{-1}$  are the reciprocals of those of  $\mathbf{A}$ , and the characteristic vectors are the same.*

By extending the notion of repeated multiplication, we now have a more general result.

### THEOREM A.10 Characteristic Roots of a Matrix Power

*For any nonsingular symmetric matrix  $\mathbf{A} = \mathbf{CAC}'$ ,  $\mathbf{A}^K = \mathbf{CA}^K\mathbf{C}'$ ,  $K = \dots, -2, -1, 0, 1, 2, \dots$*

We now turn to the general problem of how to compute the square root of a matrix. In the scalar case, the value would have to be nonnegative. The matrix analog to this requirement is that all the characteristic roots are nonnegative. Consider, then, the candidate

$$\mathbf{A}^{1/2} = \mathbf{CA}^{1/2}\mathbf{C}' = \mathbf{C} \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{bmatrix} \mathbf{C}'. \quad (\text{A-103})$$

This equation satisfies the requirement for a square root, because

$$\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{CA}^{1/2}\mathbf{C}'\mathbf{CA}^{1/2}\mathbf{C}' = \mathbf{CA}\mathbf{C}' = \mathbf{A}. \quad (\text{A-104})$$

If we continue in this fashion, we can define the powers of a matrix more generally, still assuming that all the characteristic roots are nonnegative. For example,  $\mathbf{A}^{1/3} = \mathbf{CA}^{1/3}\mathbf{C}'$ . If all the roots are strictly positive, we can go one step further and extend the result to any real power. For reasons that will be made clear in the next section, we say that a matrix with positive characteristic roots is **positive definite**. It is the matrix analog to a positive number.

### DEFINITION A.17 Real Powers of a Positive Definite Matrix

*For a positive definite matrix  $\mathbf{A}$ ,  $\mathbf{A}^r = \mathbf{CA}^r\mathbf{C}'$ , for any real number,  $r$ .* (A-105)

## APPENDIX A ♦ Matrix Algebra 1071

The characteristic roots of  $\mathbf{A}^r$  are the  $r$ th power of those of  $\mathbf{A}$ , and the characteristic vectors are the same.

If  $\mathbf{A}$  is only **nonnegative definite**—that is, has roots that are either zero or positive—then (A-105) holds only for nonnegative  $r$ .

### A.6.10 IDEMPOTENT MATRICES

Idempotent matrices are equal to their squares [see (A-37) to (A-39)]. In view of their importance in econometrics, we collect a few results related to idempotent matrices at this point. First, (A-101) implies that if  $\lambda$  is a characteristic root of an idempotent matrix, then  $\lambda = \lambda^K$  for all nonnegative integers  $K$ . As such, if  $\mathbf{A}$  is a symmetric idempotent matrix, then all its roots are one or zero. Assume that all the roots of  $\mathbf{A}$  are one. Then  $\mathbf{A} = \mathbf{I}$ , and  $\mathbf{A} = \mathbf{C}\Lambda\mathbf{C}' = \mathbf{C}\mathbf{I}\mathbf{C}' = \mathbf{C}\mathbf{C}' = \mathbf{I}$ . If the roots are not all one, then one or more are zero. Consequently, we have the following results for symmetric idempotent matrices:<sup>9</sup>

- *The only full rank, symmetric idempotent matrix is the identity matrix  $\mathbf{I}$ .* (A-106)
- *All symmetric idempotent matrices except the identity matrix are singular.* (A-107)

The final result on idempotent matrices is obtained by observing that the count of the nonzero roots of  $\mathbf{A}$  is also equal to their sum. By combining Theorems A.5 and A.7 with the result that for an idempotent matrix, the roots are all zero or one, we obtain this result:

- *The rank of a symmetric idempotent matrix is equal to its trace.* (A-108)

### A.6.11 FACTORING A MATRIX

In some applications, we shall require a matrix  $\mathbf{P}$  such that

$$\mathbf{P}'\mathbf{P} = \mathbf{A}^{-1}.$$

One choice is

$$\mathbf{P} = \Lambda^{-1/2}\mathbf{C},$$

so that

$$\mathbf{P}'\mathbf{P} = (\mathbf{C}')'(\Lambda^{-1/2})'\Lambda^{-1/2}\mathbf{C}' = \mathbf{C}\Lambda^{-1}\mathbf{C}',$$

as desired.<sup>10</sup> Thus, the **spectral decomposition** of  $\mathbf{A}$ ,  $\mathbf{A} = \mathbf{C}\Lambda\mathbf{C}'$  is a useful result for this kind of computation.

The **Cholesky factorization** of a symmetric positive definite matrix is an alternative representation that is useful in regression analysis. Any symmetric positive definite matrix  $\mathbf{A}$  may be written as the product of a **lower triangular matrix**  $\mathbf{L}$  and its transpose (which is an **upper triangular matrix**)  $\mathbf{L}' = \mathbf{U}$ . Thus,  $\mathbf{A} = \mathbf{L}\mathbf{U}$ . This result is the Cholesky decomposition of  $\mathbf{A}$ . The square roots of the diagonal elements of  $\mathbf{L}$ ,  $d_i$ , are the **Cholesky values** of  $\mathbf{A}$ . By arraying these in a diagonal matrix  $\mathbf{D}$ , we may also write  $\mathbf{A} = \mathbf{L}\mathbf{D}^{-1}\mathbf{D}^2\mathbf{D}^{-1}\mathbf{U} = \mathbf{L}^*\mathbf{D}^2\mathbf{U}^*$ , which is similar to the spectral decomposition in (A-85). The usefulness of this formulation arises when the inverse of  $\mathbf{A}$  is required. Once  $\mathbf{L}$  is

---

<sup>9</sup>Not all idempotent matrices are symmetric. We shall not encounter any asymmetric ones in our work, however.

<sup>10</sup>We say that this is “one” choice because if  $\mathbf{A}$  is symmetric, as it will be in all our applications, there are other candidates. The reader can easily verify that  $\mathbf{C}\Lambda^{-1/2}\mathbf{C}' = \mathbf{A}^{-1/2}$  works as well.

## 1072 PART VI ♦ Appendices

computed, finding  $\mathbf{A}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1}$  is also straightforward as well as extremely fast and accurate. Most recently developed econometric software packages use this technique for inverting positive definite matrices.

A third type of decomposition of a matrix is useful for numerical analysis when the inverse is difficult to obtain because the columns of  $\mathbf{A}$  are “nearly” collinear. Any  $n \times K$  matrix  $\mathbf{A}$  for which  $n \geq K$  can be written in the form  $\mathbf{A} = \mathbf{U}\mathbf{W}\mathbf{V}'$ , where  $\mathbf{U}$  is an orthogonal  $n \times K$  matrix—that is,  $\mathbf{U}'\mathbf{U} = \mathbf{I}_K$ — $\mathbf{W}$  is a  $K \times K$  diagonal matrix such that  $w_i \geq 0$ , and  $\mathbf{V}$  is a  $K \times K$  matrix such that  $\mathbf{V}'\mathbf{V} = \mathbf{I}_K$ . This result is called the **singular value decomposition** (SVD) of  $\mathbf{A}$ , and  $w_i$  are the singular values of  $\mathbf{A}$ .<sup>11</sup> (Note that if  $\mathbf{A}$  is square, then the spectral decomposition is a singular value decomposition.) As with the Cholesky decomposition, the usefulness of the SVD arises in inversion, in this case, of  $\mathbf{A}'\mathbf{A}$ . By multiplying it out, we obtain that  $(\mathbf{A}'\mathbf{A})^{-1}$  is simply  $\mathbf{V}\mathbf{W}^{-2}\mathbf{V}'$ . Once the SVD of  $\mathbf{A}$  is computed, the inversion is trivial. The other advantage of this format is its numerical stability, which is discussed at length in Press et al. (1986).

Press et al. (1986) recommend the SVD approach as the method of choice for solving least squares problems because of its accuracy and numerical stability. A commonly used alternative method similar to the SVD approach is the QR decomposition. Any  $n \times K$  matrix,  $\mathbf{X}$ , with  $n \geq K$  can be written in the form  $\mathbf{X} = \mathbf{Q}\mathbf{R}$  in which the columns of  $\mathbf{Q}$  are orthonormal ( $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$ ) and  $\mathbf{R}$  is an upper triangular matrix. Decomposing  $\mathbf{X}$  in this fashion allows an extremely accurate solution to the least squares problem that does not involve inversion or direct solution of the normal equations. Press et al. suggest that this method may have problems with rounding errors in problems when  $\mathbf{X}$  is nearly of short rank, but based on other published results, this concern seems relatively minor.<sup>12</sup>

### A.6.12 THE GENERALIZED INVERSE OF A MATRIX

Inverse matrices are fundamental in econometrics. Although we shall not require them much in our treatment in this book, there are more general forms of inverse matrices than we have considered thus far. A **generalized inverse** of a matrix  $\mathbf{A}$  is another matrix  $\mathbf{A}^+$  that satisfies the following requirements:

1.  $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$ .
2.  $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$ .
3.  $\mathbf{A}^+\mathbf{A}$  is symmetric.
4.  $\mathbf{A}\mathbf{A}^+$  is symmetric.

A unique  $\mathbf{A}^+$  can be found for any matrix, whether  $\mathbf{A}$  is singular or not, or even if  $\mathbf{A}$  is not square.<sup>13</sup> The unique matrix that satisfies all four requirements is called the **Moore–Penrose inverse** or **pseudoinverse** of  $\mathbf{A}$ . If  $\mathbf{A}$  happens to be square and nonsingular, then the generalized inverse will be the familiar ordinary inverse. But if  $\mathbf{A}^{-1}$  does not exist, then  $\mathbf{A}^+$  can still be computed.

An important special case is the overdetermined system of equations

$$\mathbf{Ab} = \mathbf{y},$$

<sup>11</sup>Discussion of the singular value decomposition (and listings of computer programs for the computations) may be found in Press et al. (1986).

<sup>12</sup>The National Institute of Standards and Technology (NIST) has published a suite of benchmark problems that test the accuracy of least squares computations (<http://www.nist.gov/itl/div898/strd>). Using these problems, which include some extremely difficult, ill-conditioned data sets, we found that the QR method would reproduce all the NIST certified solutions to 15 digits of accuracy, which suggests that the QR method should be satisfactory for all but the worst problems.

<sup>13</sup>A proof of uniqueness, with several other results, may be found in Theil (1983).

## APPENDIX A ♦ Matrix Algebra 1073

where  $\mathbf{A}$  has  $n$  rows,  $K < n$  columns, and column rank equal to  $R \leq K$ . Suppose that  $R$  equals  $K$ , so that  $(\mathbf{A}'\mathbf{A})^{-1}$  exists. Then the Moore–Penrose inverse of  $\mathbf{A}$  is

$$\mathbf{A}^+ = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}',$$

which can be verified by multiplication. A “solution” to the system of equations can be written

$$\mathbf{b} = \mathbf{A}^+\mathbf{y}.$$

This is the vector that minimizes the length of  $\mathbf{Ab} - \mathbf{y}$ . Recall this was the solution to the least squares problem obtained in Section A.4.4. If  $\mathbf{y}$  lies in the column space of  $\mathbf{A}$ , this vector will be zero, but otherwise, it will not.

Now suppose that  $\mathbf{A}$  does not have full rank. The previous solution cannot be computed. An alternative solution can be obtained, however. We continue to use the matrix  $\mathbf{A}'\mathbf{A}$ . In the spectral decomposition of Section A.6.4, if  $\mathbf{A}$  has rank  $R$ , then there are  $R$  terms in the summation in (A-85). In (A-102), the spectral decomposition using the reciprocals of the characteristic roots is used to compute the inverse. To compute the Moore–Penrose inverse, we apply this calculation to  $\mathbf{A}'\mathbf{A}$ , using only the nonzero roots, then postmultiply the result by  $\mathbf{A}'$ . Let  $\mathbf{C}_1$  be the  $R$  characteristic vectors corresponding to the nonzero roots, which we array in the diagonal matrix,  $\Lambda_1$ . Then the Moore–Penrose inverse is

$$\mathbf{A}^+ = \mathbf{C}_1\Lambda_1^{-1}\mathbf{C}_1'\mathbf{A}',$$

which is very similar to the previous result.

If  $\mathbf{A}$  is a symmetric matrix with rank  $R \leq K$ , the Moore–Penrose inverse is computed precisely as in the preceding equation without postmultiplying by  $\mathbf{A}'$ . Thus, for a symmetric matrix  $\mathbf{A}$ ,

$$\mathbf{A}^+ = \mathbf{C}_1\Lambda_1^{-1}\mathbf{C}_1',$$

where  $\Lambda_1^{-1}$  is a diagonal matrix containing the reciprocals of the *nonzero* roots of  $\mathbf{A}$ .

## A.7 QUADRATIC FORMS AND DEFINITE MATRICES

Many optimization problems involve double sums of the form

$$q = \sum_{i=1}^n \sum_{j=1}^n x_i x_j a_{ij}. \quad (\text{A-109})$$

This **quadratic form** can be written

$$q = \mathbf{x}'\mathbf{A}\mathbf{x},$$

where  $\mathbf{A}$  is a symmetric matrix. In general,  $q$  may be positive, negative, or zero; it depends on  $\mathbf{A}$  and  $\mathbf{x}$ . There are some matrices, however, for which  $q$  will be positive regardless of  $\mathbf{x}$ , and others for which  $q$  will always be negative (or nonnegative or nonpositive). For a given matrix  $\mathbf{A}$ ,

1. If  $\mathbf{x}'\mathbf{A}\mathbf{x} > (<) 0$  for all nonzero  $\mathbf{x}$ , then  $\mathbf{A}$  is **positive (negative) definite**.
2. If  $\mathbf{x}'\mathbf{A}\mathbf{x} \geq (\leq) 0$  for all nonzero  $\mathbf{x}$ , then  $\mathbf{A}$  is **nonnegative definite** or **positive semidefinite** (nonpositive definite).

It might seem that it would be impossible to check a matrix for definiteness, since  $\mathbf{x}$  can be chosen arbitrarily. But we have already used the set of results necessary to do so. Recall that a

## 1074 PART VI ♦ Appendices

symmetric matrix can be decomposed into

$$\mathbf{A} = \mathbf{C}\Lambda\mathbf{C}'.$$

Therefore, the quadratic form can be written as

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{C}\Lambda\mathbf{C}'\mathbf{x}.$$

Let  $\mathbf{y} = \mathbf{C}'\mathbf{x}$ . Then

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{y}'\Lambda\mathbf{y} = \sum_{i=1}^n \lambda_i y_i^2. \quad (\text{A-110})$$

If  $\lambda_i$  is positive for all  $i$ , then regardless of  $\mathbf{y}$ —that is, regardless of  $\mathbf{x}$ — $q$  will be positive. This case was identified earlier as a positive definite matrix. Continuing this line of reasoning, we obtain the following theorem.

### THEOREM A.11 Definite Matrices

*Let  $\mathbf{A}$  be a symmetric matrix. If all the characteristic roots of  $\mathbf{A}$  are positive (negative), then  $\mathbf{A}$  is positive definite (negative definite). If some of the roots are zero, then  $\mathbf{A}$  is nonnegative (nonpositive) definite if the remainder are positive (negative). If  $\mathbf{A}$  has both negative and positive roots, then  $\mathbf{A}$  is indefinite.*

The preceding statements give, in each case, the “if” parts of the theorem. To establish the “only if” parts, assume that the condition on the roots does not hold. This must lead to a contradiction. For example, if some  $\lambda$  can be negative, then  $\mathbf{y}'\Lambda\mathbf{y}$  could be negative for some  $\mathbf{y}$ , so  $\mathbf{A}$  cannot be positive definite.

#### A.7.1 NONNEGATIVE DEFINITE MATRICES

A case of particular interest is that of nonnegative definite matrices. Theorem A.11 implies a number of related results.

- If  $\mathbf{A}$  is nonnegative definite, then  $|\mathbf{A}| \geq 0$ . (A-111)

**Proof:** The determinant is the product of the roots, which are nonnegative.

The converse, however, is not true. For example, a  $2 \times 2$  matrix with two negative roots is clearly not positive definite, but it does have a positive determinant.

- If  $\mathbf{A}$  is positive definite, so is  $\mathbf{A}^{-1}$ . (A-112)

**Proof:** The roots are the reciprocals of those of  $\mathbf{A}$ , which are, therefore positive.

- The identity matrix  $\mathbf{I}$  is positive definite. (A-113)

**Proof:**  $\mathbf{x}'\mathbf{I}\mathbf{x} = \mathbf{x}'\mathbf{x} > 0$  if  $\mathbf{x} \neq \mathbf{0}$ .

A very important result for regression analysis is

- If  $\mathbf{A}$  is  $n \times K$  with full column rank and  $n > K$ , then  $\mathbf{A}'\mathbf{A}$  is positive definite and  $\mathbf{A}\mathbf{A}'$  is nonnegative definite. (A-114)

**Proof:** By assumption,  $\mathbf{A}\mathbf{x} \neq \mathbf{0}$ . So  $\mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{x} = (\mathbf{A}\mathbf{x})'(\mathbf{A}\mathbf{x}) = \mathbf{y}'\mathbf{y} = \sum_j y_j^2 > 0$ .

## APPENDIX A ♦ Matrix Algebra 1075

A similar proof establishes the nonnegative definiteness of  $\mathbf{A}\mathbf{A}'$ . The difference in the latter case is that because  $\mathbf{A}$  has more rows than columns there is an  $\mathbf{x}$  such that  $\mathbf{A}'\mathbf{x} = \mathbf{0}$ . Thus, in the proof, we only have  $\mathbf{y}'\mathbf{y} \geq 0$ . The case in which  $\mathbf{A}$  does not have full column rank is the same as that of  $\mathbf{A}\mathbf{A}'$ .

- If  $\mathbf{A}$  is positive definite and  $\mathbf{B}$  is a nonsingular matrix, then  $\mathbf{B}'\mathbf{A}\mathbf{B}$  is positive definite. **(A-115)**

*Proof:*  $\mathbf{x}'\mathbf{B}'\mathbf{A}\mathbf{B}\mathbf{x} = \mathbf{y}'\mathbf{A}\mathbf{y} > 0$ , where  $\mathbf{y} = \mathbf{B}\mathbf{x}$ . But  $\mathbf{y}$  cannot be  $\mathbf{0}$  because  $\mathbf{B}$  is nonsingular.

Finally, note that for  $\mathbf{A}$  to be negative definite, all  $\mathbf{A}$ 's characteristic roots must be negative. But, in this case,  $|\mathbf{A}|$  is positive if  $\mathbf{A}$  is of even order and negative if  $\mathbf{A}$  is of odd order.

### A.7.2 IDEMPOTENT QUADRATIC FORMS

Quadratic forms in idempotent matrices play an important role in the distributions of many test statistics. As such, we shall encounter them fairly often. Two central results are of interest.

- Every symmetric idempotent matrix is nonnegative definite. **(A-116)**

*Proof:* All roots are one or zero; hence, the matrix is nonnegative definite by definition.

Combining this with some earlier results yields a result used in determining the sampling distribution of most of the standard test statistics.

- If  $\mathbf{A}$  is symmetric and idempotent,  $n \times n$  with rank  $J$ , then every quadratic form in  $\mathbf{A}$  can be written  $\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{j=1}^J y_j^2$ . **(A-117)**

*Proof:* This result is (A-110) with  $\lambda =$  one or zero.

### A.7.3 COMPARING MATRICES

Derivations in econometrics often focus on whether one matrix is “larger” than another. We now consider how to make such a comparison. As a starting point, the two matrices must have the same dimensions. A useful comparison is based on

$$d = \mathbf{x}'\mathbf{A}\mathbf{x} - \mathbf{x}'\mathbf{B}\mathbf{x} = \mathbf{x}'(\mathbf{A} - \mathbf{B})\mathbf{x}.$$

If  $d$  is always positive for any nonzero vector,  $\mathbf{x}$ , then by this criterion, we can say that  $\mathbf{A}$  is larger than  $\mathbf{B}$ . The reverse would apply if  $d$  is always negative. It follows from the definition that

$$\text{if } d > 0 \text{ for all nonzero } \mathbf{x}, \text{ then } \mathbf{A} - \mathbf{B} \text{ is positive definite.} \quad \text{(A-118)}$$

If  $d$  is only greater than or equal to zero, then  $\mathbf{A} - \mathbf{B}$  is nonnegative definite. The ordering is not complete. For some pairs of matrices,  $d$  could have either sign, depending on  $\mathbf{x}$ . In this case, there is no simple comparison.

A particular case of the general result which we will encounter frequently is.

$$\begin{aligned} &\text{If } \mathbf{A} \text{ is positive definite and } \mathbf{B} \text{ is nonnegative definite,} \\ &\text{then } \mathbf{A} + \mathbf{B} \geq \mathbf{A}. \end{aligned} \quad \text{(A-119)}$$

Consider, for example, the “updating formula” introduced in (A-66). This uses a matrix

$$\mathbf{A} = \mathbf{B}'\mathbf{B} + \mathbf{b}\mathbf{b}' \geq \mathbf{B}'\mathbf{B}.$$

Finally, in comparing matrices, it may be more convenient to compare their inverses. The result analogous to a familiar result for scalars is:

$$\text{If } \mathbf{A} > \mathbf{B}, \text{ then } \mathbf{B}^{-1} > \mathbf{A}^{-1}. \quad \text{(A-120)}$$

## 1076 PART VI ♦ Appendices

To establish this intuitive result, we would make use of the following, which is proved in Goldberger (1964, Chapter 2):

### THEOREM A.12 Ordering for Positive Definite Matrices

*If  $\mathbf{A}$  and  $\mathbf{B}$  are two positive definite matrices with the same dimensions and if every characteristic root of  $\mathbf{A}$  is larger than (at least as large as) the corresponding characteristic root of  $\mathbf{B}$  when both sets of roots are ordered from largest to smallest, then  $\mathbf{A} - \mathbf{B}$  is positive (nonnegative) definite.*

The roots of the inverse are the reciprocals of the roots of the original matrix, so the theorem can be applied to the inverse matrices.

## A.8 CALCULUS AND MATRIX ALGEBRA<sup>14</sup>

### A.8.1 DIFFERENTIATION AND THE TAYLOR SERIES

A variable  $y$  is a function of another variable  $x$  written

$$y = f(x), \quad y = g(x), \quad y = y(x),$$

and so on, if each value of  $x$  is associated with a single value of  $y$ . In this relationship,  $y$  and  $x$  are sometimes labeled the **dependent variable** and the **independent variable**, respectively. Assuming that the function  $f(x)$  is continuous and differentiable, we obtain the following derivatives:

$$f'(x) = \frac{dy}{dx}, \quad f''(x) = \frac{d^2y}{dx^2},$$

and so on.

A frequent use of the derivatives of  $f(x)$  is in the **Taylor series approximation**. A Taylor series is a polynomial approximation to  $f(x)$ . Letting  $x^0$  be an arbitrarily chosen expansion point

$$f(x) \approx f(x^0) + \sum_{i=1}^P \frac{1}{i!} \frac{d^i f(x^0)}{dx^{(0)i}} (x - x^0)^i. \quad (\text{A-121})$$

The choice of the number of terms is arbitrary; the more that are used, the more accurate the approximation will be. The approximation used most frequently in econometrics is the **linear approximation**,

$$f(x) \approx \alpha + \beta x, \quad (\text{A-122})$$

where, by collecting terms in (A-121),  $\alpha = [f(x^0) - f'(x^0)x^0]$  and  $\beta = f'(x^0)$ . The superscript “0” indicates that the function is evaluated at  $x^0$ . The **quadratic approximation** is

$$f(x) \approx \alpha + \beta x + \gamma x^2, \quad (\text{A-123})$$

where  $\alpha = [f^0 - f'^0 x^0 + \frac{1}{2} f''^0 (x^0)^2]$ ,  $\beta = [f'^0 - f''^0 x^0]$  and  $\gamma = \frac{1}{2} f''^0$ .

<sup>14</sup>For a complete exposition, see Magnus and Neudecker (1988).

## APPENDIX A ♦ Matrix Algebra 1077

We can regard a function  $y = f(x_1, x_2, \dots, x_n)$  as a **scalar-valued function** of a vector; that is,  $y = f(\mathbf{x})$ . The vector of partial derivatives, or **gradient vector**, or simply **gradient**, is

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}. \quad (\text{A-124})$$

The vector  $\mathbf{g}(\mathbf{x})$  or  $\mathbf{g}$  is used to represent the gradient. Notice that it is a column vector. The shape of the derivative is determined by the denominator of the derivative.

A **second derivatives matrix** or **Hessian** is computed as

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 y}{\partial x_1 \partial x_1} & \frac{\partial^2 y}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 y}{\partial x_1 \partial x_n} \\ \frac{\partial^2 y}{\partial x_2 \partial x_1} & \frac{\partial^2 y}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 y}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 y}{\partial x_n \partial x_1} & \frac{\partial^2 y}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 y}{\partial x_n \partial x_n} \end{bmatrix} = [f_{ij}]. \quad (\text{A-125})$$

In general,  $\mathbf{H}$  is a square, symmetric matrix. (The symmetry is obtained for continuous and continuously differentiable functions from Young's theorem.) Each column of  $\mathbf{H}$  is the derivative of  $\mathbf{g}$  with respect to the corresponding variable in  $\mathbf{x}'$ . Therefore,

$$\mathbf{H} = \left[ \frac{\partial(\partial y / \partial \mathbf{x})}{\partial x_1} \frac{\partial(\partial y / \partial \mathbf{x})}{\partial x_2} \cdots \frac{\partial(\partial y / \partial \mathbf{x})}{\partial x_n} \right] = \frac{\partial(\partial y / \partial \mathbf{x})}{\partial(x_1 \ x_2 \ \cdots \ x_n)} = \frac{\partial(\partial y / \partial \mathbf{x})}{\partial \mathbf{x}'} = \frac{\partial^2 y}{\partial \mathbf{x} \partial \mathbf{x}'},$$

The first-order, or linear Taylor series approximation is

$$y \approx f(\mathbf{x}^0) + \sum_{i=1}^n f_i(\mathbf{x}^0)(x_i - x_i^0). \quad (\text{A-126})$$

The right-hand side is

$$f(\mathbf{x}^0) + \left[ \frac{\partial f(\mathbf{x}^0)}{\partial \mathbf{x}^0} \right]' (\mathbf{x} - \mathbf{x}^0) = [f(\mathbf{x}^0) - \mathbf{g}(\mathbf{x}^0)' \mathbf{x}^0] + \mathbf{g}(\mathbf{x}^0)' \mathbf{x} = [f^0 - \mathbf{g}^0' \mathbf{x}^0] + \mathbf{g}^0' \mathbf{x}.$$

This produces the linear approximation,

$$y \approx \alpha + \beta' \mathbf{x}.$$

The second-order, or quadratic, approximation adds the second-order terms in the expansion,

$$\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n f_{ij}^0 (x_i - x_i^0)(x_j - x_j^0) = \frac{1}{2} (\mathbf{x} - \mathbf{x}^0)' \mathbf{H}^0 (\mathbf{x} - \mathbf{x}^0),$$

to the preceding one. Collecting terms in the same manner as in (A-126), we have

$$y \approx \alpha + \beta' \mathbf{x} + \frac{1}{2} \mathbf{x}' \boldsymbol{\Gamma} \mathbf{x}, \quad (\text{A-127})$$

where

$$\alpha = f^0 - \mathbf{g}^0' \mathbf{x}^0 + \frac{1}{2} \mathbf{x}^0' \mathbf{H}^0 \mathbf{x}^0, \quad \beta = \mathbf{g}^0 - \mathbf{H}^0 \mathbf{x}^0 \quad \text{and} \quad \boldsymbol{\Gamma} = \mathbf{H}^0.$$

A linear function can be written

$$y = \mathbf{a}' \mathbf{x} = \mathbf{x}' \mathbf{a} = \sum_{i=1}^n a_i x_i,$$

**1078 PART VI ♦ Appendices**

so

$$\frac{\partial(\mathbf{a}'\mathbf{x})}{\partial \mathbf{x}} = \mathbf{a}. \quad (\text{A-128})$$

Note, in particular, that  $\partial(\mathbf{a}'\mathbf{x})/\partial \mathbf{x} = \mathbf{a}$ , not  $\mathbf{a}'$ . In a set of linear functions

$$\mathbf{y} = \mathbf{Ax},$$

each element  $y_i$  of  $\mathbf{y}$  is

$$y_i = \mathbf{a}'_i \mathbf{x},$$

where  $\mathbf{a}'_i$  is the  $i$ th row of  $\mathbf{A}$  [see (A-14)]. Therefore,

$$\frac{\partial y_i}{\partial \mathbf{x}} = \mathbf{a}_i = \text{transpose of } i\text{th row of } \mathbf{A},$$

and

$$\begin{bmatrix} \partial y_1 / \partial \mathbf{x}' \\ \partial y_2 / \partial \mathbf{x}' \\ \dots \\ \partial y_n / \partial \mathbf{x}' \end{bmatrix} = \begin{bmatrix} \mathbf{a}'_1 \\ \mathbf{a}'_2 \\ \dots \\ \mathbf{a}'_n \end{bmatrix}.$$

Collecting all terms, we find that  $\partial \mathbf{Ax} / \partial \mathbf{x}' = \mathbf{A}$ , whereas the more familiar form will be

$$\frac{\partial \mathbf{Ax}}{\partial \mathbf{x}} = \mathbf{A}'. \quad (\text{A-129})$$

A quadratic form is written

$$\mathbf{x}' \mathbf{Ax} = \sum_{i=1}^n \sum_{j=1}^n x_i x_j a_{ij}. \quad (\text{A-130})$$

For example,

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix},$$

so that

$$\mathbf{x}' \mathbf{Ax} = 1x_1^2 + 4x_2^2 + 6x_1x_2.$$

Then

$$\frac{\partial \mathbf{x}' \mathbf{Ax}}{\partial \mathbf{x}} = \begin{bmatrix} 2x_1 + 6x_2 \\ 6x_1 + 8x_2 \end{bmatrix} = \begin{bmatrix} 2 & 6 \\ 6 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2\mathbf{Ax}, \quad (\text{A-131})$$

which is the general result when  $\mathbf{A}$  is a symmetric matrix. If  $\mathbf{A}$  is not symmetric, then

$$\frac{\partial(\mathbf{x}' \mathbf{Ax})}{\partial a_{ij}} = (\mathbf{A} + \mathbf{A}')\mathbf{x}. \quad (\text{A-132})$$

Referring to the preceding double summation, we find that for each term, the coefficient on  $a_{ij}$  is  $x_i x_j$ . Therefore,

$$\frac{\partial(\mathbf{x}' \mathbf{Ax})}{\partial a_{ij}} = x_i x_j.$$

## APPENDIX A ♦ Matrix Algebra 1079

The square matrix whose  $ij$ th element is  $x_i x_j$  is  $\mathbf{xx}'$ , so

$$\frac{\partial(\mathbf{x}' \mathbf{Ax})}{\partial \mathbf{A}} = \mathbf{xx}' . \quad (\text{A-133})$$

Derivatives involving determinants appear in maximum likelihood estimation. From the cofactor expansion in (A-51),

$$\frac{\partial |\mathbf{A}|}{\partial a_{ij}} = (-1)^{i+j} |\mathbf{A}_{ij}| = c_{ij}$$

where  $|\mathbf{C}_{ji}|$  is the  $j$ th cofactor in  $\mathbf{A}$ . The inverse of  $\mathbf{A}$  can be computed using

$$\mathbf{A}_{ij}^{-1} = \frac{|\mathbf{C}_{ji}|}{|\mathbf{A}|}$$

(note the reversal of the subscripts), which implies that

$$\frac{\partial \ln|\mathbf{A}|}{\partial a_{ij}} = \frac{(-1)^{i+j} |\mathbf{A}_{ij}|}{|\mathbf{A}|},$$

or, collecting terms,

$$\frac{\partial \ln|\mathbf{A}|}{\partial \mathbf{A}} = \mathbf{A}^{-1}.$$

Because the matrices for which we shall make use of this calculation will be symmetric in our applications, the transposition will be unnecessary.

### A.8.2 OPTIMIZATION

Consider finding the  $x$  where  $f(x)$  is maximized or minimized. Because  $f'(x)$  is the slope of  $f(x)$ , either optimum must occur where  $f'(x) = 0$ . Otherwise, the function will be increasing or decreasing at  $x$ . This result implies the **first-order or necessary condition for an optimum** (maximum or minimum):

$$\frac{dy}{dx} = 0. \quad (\text{A-134})$$

For a maximum, the function must be concave; for a minimum, it must be convex. The **sufficient condition for an optimum** is.

$$\begin{aligned} \text{For a maximum, } \frac{d^2y}{dx^2} &< 0; \\ \text{for a minimum, } \frac{d^2y}{dx^2} &> 0. \end{aligned} \quad (\text{A-135})$$

Some functions, such as the sine and cosine functions, have many **local optima**, that is, many minima and maxima. A function such as  $(\cos x)/(1 + x^2)$ , which is a damped cosine wave, does as well but differs in that although it has many local maxima, it has one, at  $x = 0$ , at which  $f(x)$  is greater than it is at any other point. Thus,  $x = 0$  is the **global maximum**, whereas the other maxima are only **local maxima**. Certain functions, such as a quadratic, have only a single optimum. These functions are **globally concave** if the optimum is a maximum and **globally convex** if it is a minimum.

For maximizing or minimizing a function of several variables, the first-order conditions are

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{0}. \quad (\text{A-136})$$

## 1080 PART VI ♦ Appendices

This result is interpreted in the same manner as the necessary condition in the univariate case. At the optimum, it must be true that no small change in any variable leads to an improvement in the function value. In the single-variable case,  $d^2y/dx^2$  must be positive for a minimum and negative for a maximum. The second-order condition for an optimum in the multivariate case is that, at the optimizing value,

$$\mathbf{H} = \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \quad (\text{A-137})$$

must be positive definite for a minimum and negative definite for a maximum.

In a single-variable problem, the second-order condition can usually be verified by inspection. This situation will not generally be true in the multivariate case. As discussed earlier, checking the definiteness of a matrix is, in general, a difficult problem. For most of the problems encountered in econometrics, however, the second-order condition will be implied by the structure of the problem. That is, the matrix  $\mathbf{H}$  will usually be of such a form that it is always definite.

For an example of the preceding, consider the problem

$$\text{maximize}_{\mathbf{x}} R = \mathbf{a}'\mathbf{x} - \mathbf{x}'\mathbf{A}\mathbf{x},$$

where

$$\mathbf{a}' = (5 \quad 4 \quad 2),$$

and

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 3 \\ 1 & 3 & 2 \\ 3 & 2 & 5 \end{bmatrix}.$$

Using some now familiar results, we obtain

$$\frac{\partial R}{\partial \mathbf{x}} = \mathbf{a} - 2\mathbf{A}\mathbf{x} = \begin{bmatrix} 5 \\ 4 \\ 2 \end{bmatrix} - \begin{bmatrix} 4 & 2 & 6 \\ 2 & 6 & 4 \\ 6 & 4 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{0}. \quad (\text{A-138})$$

The solutions are

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 6 \\ 2 & 6 & 4 \\ 6 & 4 & 10 \end{bmatrix}^{-1} \begin{bmatrix} 5 \\ 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 11.25 \\ 1.75 \\ -7.25 \end{bmatrix}.$$

The sufficient condition is that

$$\frac{\partial^2 R(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} = -2\mathbf{A} = \begin{bmatrix} -4 & -2 & -6 \\ -2 & -6 & -4 \\ -6 & -4 & -10 \end{bmatrix} \quad (\text{A-139})$$

must be negative definite. The three characteristic roots of this matrix are  $-15.746$ ,  $-4$ , and  $-0.25403$ . Because all three roots are negative, the matrix is negative definite, as required.

In the preceding, it was necessary to compute the characteristic roots of the Hessian to verify the sufficient condition. For a general matrix of order larger than 2, this will normally require a computer. Suppose, however, that  $\mathbf{A}$  is of the form

$$\mathbf{A} = \mathbf{B}'\mathbf{B},$$

where  $\mathbf{B}$  is some known matrix. Then, as shown earlier, we know that  $\mathbf{A}$  will always be positive definite (assuming that  $\mathbf{B}$  has full rank). In this case, it is not necessary to calculate the characteristic roots of  $\mathbf{A}$  to verify the sufficient conditions.

## APPENDIX A ♦ Matrix Algebra 1081

## A.8.3 CONSTRAINED OPTIMIZATION

It is often necessary to solve an optimization problem subject to some constraints on the solution. One method is merely to “solve out” the constraints. For example, in the maximization problem considered earlier, suppose that the constraint  $x_1 = x_2 - x_3$  is imposed on the solution. For a single constraint such as this one, it is possible merely to substitute the right-hand side of this equation for  $x_1$  in the objective function and solve the resulting problem as a function of the remaining two variables. For more general constraints, however, or when there is more than one constraint, the method of Lagrange multipliers provides a more straightforward method of solving the problem. We

$$\begin{aligned} \text{maximize}_{\mathbf{x}} f(\mathbf{x}) \text{ subject to } c_1(\mathbf{x}) &= 0, \\ c_2(\mathbf{x}) &= 0, \\ &\dots \\ c_J(\mathbf{x}) &= 0. \end{aligned} \quad (\text{A-140})$$

The Lagrangean approach to this problem is to find the stationary points—that is, the points at which the derivatives are zero—of

$$L^*(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{j=1}^J \lambda_j c_j(\mathbf{x}) = f(\mathbf{x}) + \boldsymbol{\lambda}' \mathbf{c}(\mathbf{x}). \quad (\text{A-141})$$

The solutions satisfy the equations

$$\begin{aligned} \frac{\partial L^*}{\partial \mathbf{x}} &= \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} + \frac{\partial \boldsymbol{\lambda}' \mathbf{c}(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{0} \quad (n \times 1), \\ \frac{\partial L^*}{\partial \boldsymbol{\lambda}} &= \mathbf{c}(\mathbf{x}) = \mathbf{0} \quad (J \times 1). \end{aligned} \quad (\text{A-142})$$

The second term in  $\partial L^*/\partial \mathbf{x}$  is

$$\frac{\partial \boldsymbol{\lambda}' \mathbf{c}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \mathbf{c}(\mathbf{x})' \boldsymbol{\lambda}}{\partial \mathbf{x}} = \left[ \frac{\partial \mathbf{c}(\mathbf{x})'}{\partial \mathbf{x}} \right] \boldsymbol{\lambda} = \mathbf{C}' \boldsymbol{\lambda}, \quad (\text{A-143})$$

where  $\mathbf{C}$  is the matrix of derivatives of the constraints with respect to  $\mathbf{x}$ . The  $j$ th row of the  $J \times n$  matrix  $\mathbf{C}$  is the vector of derivatives of the  $j$ th constraint,  $c_j(\mathbf{x})$ , with respect to  $\mathbf{x}'$ . Upon collecting terms, the first-order conditions are

$$\begin{aligned} \frac{\partial L^*}{\partial \mathbf{x}} &= \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} + \mathbf{C}' \boldsymbol{\lambda} = \mathbf{0}, \\ \frac{\partial L^*}{\partial \boldsymbol{\lambda}} &= \mathbf{c}(\mathbf{x}) = \mathbf{0}. \end{aligned} \quad (\text{A-144})$$

There is one very important aspect of the constrained solution to consider. In the unconstrained solution, we have  $\partial f(\mathbf{x})/\partial \mathbf{x} = \mathbf{0}$ . From (A-144), we obtain, for a constrained solution,

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = -\mathbf{C}' \boldsymbol{\lambda}, \quad (\text{A-145})$$

which will not equal  $\mathbf{0}$  unless  $\boldsymbol{\lambda} = \mathbf{0}$ . This result has two important implications:

- The constrained solution cannot be superior to the unconstrained solution. This is implied by the nonzero gradient at the constrained solution. (That is, unless  $\mathbf{C} = \mathbf{0}$  which could happen if the constraints were nonlinear. But, even if so, the solution is still no better than the unconstrained optimum.)
- If the Lagrange multipliers are zero, then the constrained solution will equal the unconstrained solution.

**1082 PART VI ♦ Appendices**

To continue the example begun earlier, suppose that we add the following conditions:

$$x_1 - x_2 + x_3 = 0,$$

$$x_1 + x_2 + x_3 = 0.$$

To put this in the format of the general problem, write the constraints as  $\mathbf{c}(\mathbf{x}) = \mathbf{Cx} = \mathbf{0}$ , where

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

The Lagrangean function is

$$R^*(\mathbf{x}, \lambda) = \mathbf{a}'\mathbf{x} - \mathbf{x}'\mathbf{Ax} + \lambda'\mathbf{Cx}.$$

Note the dimensions and arrangement of the various parts. In particular,  $\mathbf{C}$  is a  $2 \times 3$  matrix, with one row for each constraint and one column for each variable in the objective function. The vector of Lagrange multipliers thus has two elements, one for each constraint. The necessary conditions are

$$\mathbf{a} - 2\mathbf{Ax} + \mathbf{C}'\lambda = \mathbf{0} \quad (\text{three equations}), \tag{A-146}$$

and

$$\mathbf{Cx} = \mathbf{0} \quad (\text{two equations}).$$

These may be combined in the single equation

$$\begin{bmatrix} -2\mathbf{A} & \mathbf{C}' \\ \mathbf{C} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \lambda \end{bmatrix} = \begin{bmatrix} -\mathbf{a} \\ \mathbf{0} \end{bmatrix}.$$

Using the partitioned inverse of (A-74) produces the solutions

$$\lambda = -[\mathbf{CA}^{-1}\mathbf{C}']^{-1}\mathbf{CA}^{-1}\mathbf{a} \tag{A-147}$$

and

$$\mathbf{x} = \frac{1}{2}\mathbf{A}^{-1}[\mathbf{I} - \mathbf{C}'(\mathbf{CA}^{-1}\mathbf{C}')^{-1}\mathbf{CA}^{-1}]\mathbf{a}. \tag{A-148}$$

The two results, (A-147) and (A-148), yield analytic solutions for  $\lambda$  and  $\mathbf{x}$ . For the specific matrices and vectors of the example, these are  $\lambda = [-0.5 \ -7.5]'$ , and the constrained solution vector,  $\mathbf{x}^* = [1.5 \ 0 \ -1.5]'$ . Note that in computing the solution to this sort of problem, it is not necessary to use the rather cumbersome form of (A-148). Once  $\lambda$  is obtained from (A-147), the solution can be inserted in (A-146) for a much simpler computation. The solution

$$\mathbf{x} = \frac{1}{2}\mathbf{A}^{-1}\mathbf{a} + \frac{1}{2}\mathbf{A}^{-1}\mathbf{C}'\lambda$$

suggests a useful result for the constrained optimum:

$$\text{constrained solution} = \text{unconstrained solution} + [2\mathbf{A}]^{-1}\mathbf{C}'\lambda. \tag{A-149}$$

Finally, by inserting the two solutions in the original function, we find that  $R = 24.375$  and  $R^* = 2.25$ , which illustrates again that the constrained solution (in this *maximization* problem) is inferior to the unconstrained solution.

## APPENDIX A ♦ Matrix Algebra 1083

## A.8.4 TRANSFORMATIONS

If a function is strictly monotonic, then it is a **one-to-one function**. Each  $y$  is associated with exactly one value of  $x$ , and vice versa. In this case, an **inverse function** exists, which expresses  $x$  as a function of  $y$ , written

$$y = f(x)$$

and

$$x = f^{-1}(y).$$

An example is the inverse relationship between the log and the exponential functions.

The slope of the inverse function,

$$J = \frac{dx}{dy} = \frac{df^{-1}(y)}{dy} = f'^{-1}(y),$$

is the **Jacobian** of the transformation from  $y$  to  $x$ . For example, if

$$y = a + bx,$$

then

$$x = -\frac{a}{b} + \left[ \frac{1}{b} \right] y$$

is the inverse transformation and

$$J = \frac{dx}{dy} = \frac{1}{b}.$$

Looking ahead to the statistical application of this concept, we observe that if  $y = f(x)$  were *vertical*, then this would no longer be a functional relationship. The same  $x$  would be associated with more than one value of  $y$ . In this case, at this value of  $x$ , we would find that  $J = 0$ , indicating a singularity in the function.

If  $\mathbf{y}$  is a column vector of functions,  $\mathbf{y} = \mathbf{f}(\mathbf{x})$ , then

$$\mathbf{J} = \frac{\partial \mathbf{x}}{\partial \mathbf{y}'} = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & & & \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{bmatrix}.$$

Consider the set of linear functions  $\mathbf{y} = \mathbf{Ax} = \mathbf{f}(\mathbf{x})$ . The inverse transformation is  $\mathbf{x} = \mathbf{f}^{-1}(\mathbf{y})$ , which will be

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y},$$

if  $\mathbf{A}$  is nonsingular. If  $\mathbf{A}$  is singular, then there is no inverse transformation. Let  $\mathbf{J}$  be the matrix of partial derivatives of the inverse functions:

$$\mathbf{J} = \left[ \frac{\partial x_i}{\partial y_j} \right].$$

The absolute value of the determinant of  $\mathbf{J}$ ,

$$\text{abs}(|\mathbf{J}|) = \text{abs} \left( \det \left( \left[ \frac{\partial \mathbf{x}}{\partial \mathbf{y}'} \right] \right) \right),$$

is the **Jacobian** determinant of the transformation from  $\mathbf{y}$  to  $\mathbf{x}$ . In the nonsingular case,

$$\text{abs}(|\mathbf{J}|) = \text{abs}(|\mathbf{A}^{-1}|) = \frac{1}{\text{abs}(|\mathbf{A}|)}.$$

## 1084 PART VI ♦ Appendices

In the singular case, the matrix of partial derivatives will be singular and the determinant of the Jacobian will be zero. In this instance, the singular Jacobian implies that  $\mathbf{A}$  is singular or, equivalently, that the transformations from  $\mathbf{x}$  to  $\mathbf{y}$  are functionally dependent. The singular case is analogous to the single-variable case.

Clearly, if the vector  $\mathbf{x}$  is given, then  $\mathbf{y} = \mathbf{Ax}$  can be computed from  $\mathbf{x}$ . Whether  $\mathbf{x}$  can be deduced from  $\mathbf{y}$  is another question. Evidently, it depends on the Jacobian. If the Jacobian is not zero, then the inverse transformations exist, and we can obtain  $\mathbf{x}$ . If not, then we cannot obtain  $\mathbf{x}$ .

## APPENDIX B



# PROBABILITY AND DISTRIBUTION THEORY

## B.1 INTRODUCTION

This appendix reviews the distribution theory used later in the book. A previous course in statistics is assumed, so most of the results will be stated without proof. The more advanced results in the later sections will be developed in greater detail.

## B.2 RANDOM VARIABLES

We view our observation on some aspect of the economy as the **outcome** of a random process that is almost never under our (the analyst's) control. In the current literature, the descriptive (and perspective laden) term **data generating process**, or DGP is often used for this underlying mechanism. The observed (measured) outcomes of the process are assigned unique numeric values. The assignment is one to one; each outcome gets one value, and no two distinct outcomes receive the same value. This outcome variable,  $X$ , is a **random variable** because, until the data are actually observed, it is uncertain what value  $X$  will take. Probabilities are associated with outcomes to quantify this uncertainty. We usually use capital letters for the “name” of a random variable and lowercase letters for the values it takes. Thus, the probability that  $X$  takes a particular value  $x$  might be denoted  $\text{Prob}(X = x)$ .

A random variable is **discrete** if the set of outcomes is either finite in number or countably infinite. The random variable is **continuous** if the set of outcomes is infinitely divisible and, hence, not countable. These definitions will correspond to the types of data we observe in practice. Counts of occurrences will provide observations on discrete random variables, whereas measurements such as time or income will give observations on continuous random variables.

### B.2.1 PROBABILITY DISTRIBUTIONS

A listing of the values  $x$  taken by a random variable  $X$  and their associated probabilities is a **probability distribution**,  $f(x)$ . For a discrete random variable,

$$f(x) = \text{Prob}(X = x). \quad (\mathbf{B-1})$$

## APPENDIX B ♦ Probability and Distribution Theory **1085**

The axioms of probability require that

$$1. \quad 0 \leq \text{Prob}(X = x) \leq 1. \quad (\mathbf{B-2})$$

$$2. \quad \sum_x f(x) = 1. \quad (\mathbf{B-3})$$

For the continuous case, the probability associated with any particular point is zero, and we can only assign positive probabilities to intervals in the range of  $x$ . The **probability density function (pdf)** is defined so that  $f(x) \geq 0$  and

$$1. \quad \text{Prob}(a \leq x \leq b) = \int_a^b f(x) dx \geq 0. \quad (\mathbf{B-4})$$

This result is the area under  $f(x)$  in the range from  $a$  to  $b$ . For a continuous variable,

$$2. \quad \int_{-\infty}^{+\infty} f(x) dx = 1. \quad (\mathbf{B-5})$$

If the range of  $x$  is not infinite, then it is understood that  $f(x) = 0$  anywhere outside the appropriate range. Because the probability associated with any individual point is 0,

$$\begin{aligned} \text{Prob}(a \leq x \leq b) &= \text{Prob}(a \leq x < b) \\ &= \text{Prob}(a < x \leq b) \\ &= \text{Prob}(a < x < b). \end{aligned}$$

### B.2.2 CUMULATIVE DISTRIBUTION FUNCTION

For any random variable  $X$ , the probability that  $X$  is less than or equal to  $a$  is denoted  $F(a)$ .  $F(x)$  is the **cumulative distribution function (cdf)**. For a discrete random variable,

$$F(x) = \sum_{X \leq x} f(X) = \text{Prob}(X \leq x). \quad (\mathbf{B-6})$$

In view of the definition of  $f(x)$ ,

$$f(x_i) = F(x_i) - F(x_{i-1}). \quad (\mathbf{B-7})$$

For a continuous random variable,

$$F(x) = \int_{-\infty}^x f(t) dt, \quad (\mathbf{B-8})$$

and

$$f(x) = \frac{dF(x)}{dx}. \quad (\mathbf{B-9})$$

In both the continuous and discrete cases,  $F(x)$  must satisfy the following properties:

1.  $0 \leq F(x) \leq 1$ .
2. If  $x > y$ , then  $F(x) \geq F(y)$ .
3.  $F(+\infty) = 1$ .
4.  $F(-\infty) = 0$ .

From the definition of the cdf,

$$\text{Prob}(a < x \leq b) = F(b) - F(a). \quad (\mathbf{B-10})$$

Any valid pdf will imply a valid cdf, so there is no need to verify these conditions separately.

**1086 PART VI ♦ Appendices****B.3 EXPECTATIONS OF A RANDOM VARIABLE****DEFINITION B.1 Mean of a Random Variable**

*The mean, or expected value, of a random variable is*

$$E[x] = \begin{cases} \sum_x xf(x) & \text{if } x \text{ is discrete,} \\ \int_x xf(x) dx & \text{if } x \text{ is continuous.} \end{cases} \quad (\text{B-11})$$

The notation  $\sum_x$  or  $\int_x$ , used henceforth, means the sum or integral over the entire range of values of  $x$ . The mean is usually denoted  $\mu$ . It is a weighted average of the values taken by  $x$ , where the weights are the respective probabilities. It is not necessarily a value actually taken by the random variable. For example, the expected number of heads in one toss of a fair coin is  $\frac{1}{2}$ .

Other measures of central tendency are the median, which is the value  $m$  such that  $\text{Prob}(X \leq m) \geq \frac{1}{2}$  and  $\text{Prob}(X \geq m) \geq \frac{1}{2}$ , and the mode, which is the value of  $x$  at which  $f(x)$  takes its maximum. The first of these measures is more frequently used than the second. Loosely speaking, the median corresponds more closely than the mean to the middle of a distribution. It is unaffected by extreme values. In the discrete case, the modal value of  $x$  has the highest probability of occurring.

Let  $g(x)$  be a function of  $x$ . The function that gives the expected value of  $g(x)$  is denoted

$$E[g(x)] = \begin{cases} \sum_x g(x) \text{Prob}(X = x) & \text{if } X \text{ is discrete,} \\ \int_x g(x) f(x) dx & \text{if } X \text{ is continuous.} \end{cases} \quad (\text{B-12})$$

If  $g(x) = a + bx$  for constants  $a$  and  $b$ , then

$$E[a + bx] = a + bE[x].$$

An important case is the expected value of a constant  $a$ , which is just  $a$ .

**DEFINITION B.2 Variance of a Random Variable**

*The variance of a random variable is*

$$\begin{aligned} \text{Var}[x] &= E[(x - \mu)^2] \\ &= \begin{cases} \sum_x (x - \mu)^2 f(x) & \text{if } x \text{ is discrete,} \\ \int_x (x - \mu)^2 f(x) dx & \text{if } x \text{ is continuous.} \end{cases} \end{aligned} \quad (\text{B-13})$$

$\text{Var}[x]$ , which must be positive, is usually denoted  $\sigma^2$ . This function is a measure of the dispersion of a distribution. Computation of the variance is simplified by using the following

## APPENDIX B ♦ Probability and Distribution Theory **1087**

important result:

$$\text{Var}[x] = E[x^2] - \mu^2. \quad (\text{B-14})$$

A convenient corollary to (B-14) is

$$E[x^2] = \sigma^2 + \mu^2. \quad (\text{B-15})$$

By inserting  $y = a + bx$  in (B-13) and expanding, we find that

$$\text{Var}[a + bx] = b^2 \text{Var}[x], \quad (\text{B-16})$$

which implies, for any constant  $a$ , that

$$\text{Var}[a] = 0. \quad (\text{B-17})$$

To describe a distribution, we usually use  $\sigma$ , the positive square root, which is the **standard deviation** of  $x$ . The standard deviation can be interpreted as having the same units of measurement as  $x$  and  $\mu$ . For any random variable  $x$  and any positive constant  $k$ , the **Chebychev inequality** states that

$$\text{Prob}(\mu - k\sigma \leq x \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}. \quad (\text{B-18})$$

Two other measures often used to describe a probability distribution are

$$\text{skewness} = E[(x - \mu)^3],$$

and

$$\text{kurtosis} = E[(x - \mu)^4].$$

Skewness is a measure of the asymmetry of a distribution. For symmetric distributions,

$$f(\mu - x) = f(\mu + x),$$

and

$$\text{skewness} = 0.$$

For asymmetric distributions, the skewness will be positive if the “long tail” is in the positive direction. Kurtosis is a measure of the thickness of the tails of the distribution. A shorthand expression for other **central moments** is

$$\mu_r = E[(x - \mu)^r].$$

Because  $\mu_r$  tends to explode as  $r$  grows, the normalized measure,  $\mu_r/\sigma^r$ , is often used for description. Two common measures are

$$\text{skewness coefficient} = \frac{\mu_3}{\sigma^3},$$

and

$$\text{degree of excess} = \frac{\mu_4}{\sigma^4} - 3.$$

The second is based on the normal distribution, which has excess of zero.

For any two functions  $g_1(x)$  and  $g_2(x)$ ,

$$E[g_1(x) + g_2(x)] = E[g_1(x)] + E[g_2(x)]. \quad (\text{B-19})$$

For the general case of a possibly nonlinear  $g(x)$ ,

$$E[g(x)] = \int_x g(x) f(x) dx, \quad (\text{B-20})$$

## 1088 PART VI ♦ Appendices

and

$$\text{Var}[g(x)] = \int_x (g(x) - E[g(x)])^2 f(x) dx. \quad (\mathbf{B-21})$$

(For convenience, we shall omit the equivalent definitions for discrete variables in the following discussion and use the integral to mean either integration or summation, whichever is appropriate.)

A device used to approximate  $E[g(x)]$  and  $\text{Var}[g(x)]$  is the linear Taylor series approximation:

$$g(x) \approx [g(x^0) - g'(x^0)x^0] + g'(x^0)x = \beta_1 + \beta_2 x = g^*(x). \quad (\mathbf{B-22})$$

If the approximation is reasonably accurate, then the mean and variance of  $g^*(x)$  will be approximately equal to the mean and variance of  $g(x)$ . A natural choice for the expansion point is  $x^0 = \mu = E(x)$ . Inserting this value in (B-22) gives

$$g(x) \approx [g(\mu) - g'(\mu)\mu] + g'(\mu)x, \quad (\mathbf{B-23})$$

so that

$$E[g(x)] \approx g(\mu), \quad (\mathbf{B-24})$$

and

$$\text{Var}[g(x)] \approx [g'(\mu)]^2 \text{Var}[x]. \quad (\mathbf{B-25})$$

A point to note in view of (B-22) to (B-24) is that  $E[g(x)]$  will generally not equal  $g(E[x])$ . For the special case in which  $g(x)$  is concave—that is, where  $g''(x) < 0$ —we know from **Jensen's inequality** that  $E[g(x)] \leq g(E[x])$ . For example,  $E[\log(x)] \leq \log(E[x])$ .

## B.4 SOME SPECIFIC PROBABILITY DISTRIBUTIONS

Certain experimental situations naturally give rise to specific probability distributions. In the majority of cases in economics, however, the distributions used are merely models of the observed phenomena. Although the normal distribution, which we shall discuss at length, is the mainstay of econometric research, economists have used a wide variety of other distributions. A few are discussed here.<sup>1</sup>

### B.4.1 THE NORMAL DISTRIBUTION

The general form of the normal distribution with mean  $\mu$  and standard deviation  $\sigma$  is

$$f(x | \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2[(x-\mu)^2/\sigma^2]}. \quad (\mathbf{B-26})$$

This result is usually denoted  $x \sim N[\mu, \sigma^2]$ . The standard notation  $x \sim f(x)$  is used to state that “ $x$  has probability distribution  $f(x)$ .” Among the most useful properties of the normal distribution

---

<sup>1</sup> A much more complete listing appears in Maddala (1977a, Chapters 3 and 18) and in most mathematical statistics textbooks. See also Poirier (1995) and Stuart and Ord (1989). Another useful reference is Evans, Hastings, and Peacock (1993). Johnson et al. (1974, 1993, 1994, 1995, 1997) is an encyclopedic reference on the subject of statistical distributions.

**APPENDIX B ♦ Probability and Distribution Theory 1089**

is its preservation under linear transformation.

$$\text{If } x \sim N[\mu, \sigma^2], \text{ then } (a + bx) \sim N[a + b\mu, b^2\sigma^2]. \quad (\text{B-27})$$

One particularly convenient transformation is  $a = -\mu/\sigma$  and  $b = 1/\sigma$ . The resulting variable  $z = (x - \mu)/\sigma$  has the **standard normal distribution**, denoted  $N[0, 1]$ , with density

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}. \quad (\text{B-28})$$

The specific notation  $\phi(z)$  is often used for this distribution and  $\Phi(z)$  for its cdf. It follows from the definitions above that if  $x \sim N[\mu, \sigma^2]$ , then

$$f(x) = \frac{1}{\sigma} \phi\left[\frac{x - \mu}{\sigma}\right].$$

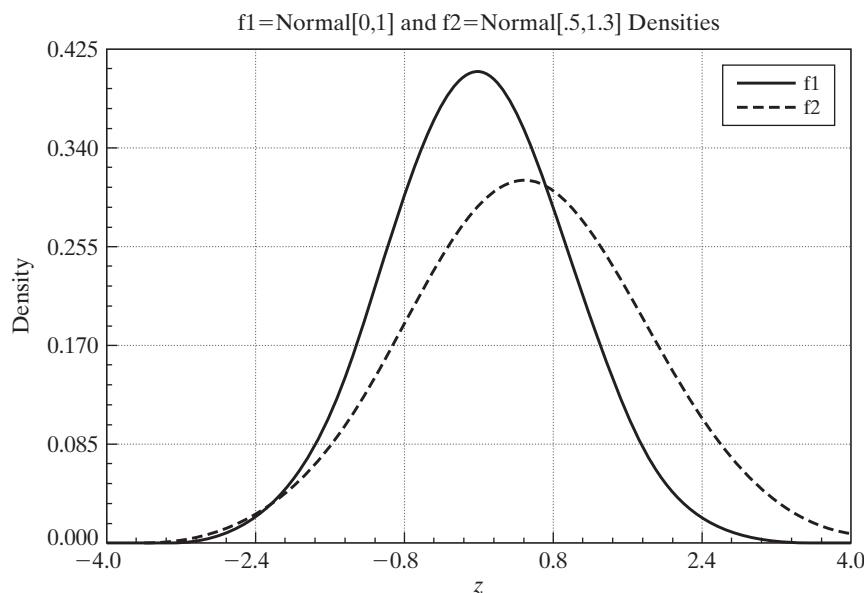
Figure B.1 shows the densities of the standard normal distribution and the normal distribution with mean 0.5, which shifts the distribution to the right, and standard deviation 1.3, which, it can be seen, scales the density so that it is shorter but wider. (The graph is a bit deceiving unless you look closely; both densities are symmetric.)

Tables of the standard normal cdf appear in most statistics and econometrics textbooks. Because the form of the distribution does not change under a linear transformation, it is not necessary to tabulate the distribution for other values of  $\mu$  and  $\sigma$ . For any normally distributed variable,

$$\text{Prob}(a \leq x \leq b) = \text{Prob}\left(\frac{a - \mu}{\sigma} \leq \frac{x - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right), \quad (\text{B-29})$$

which can always be read from a table of the standard normal distribution. In addition, because the distribution is symmetric,  $\Phi(-z) = 1 - \Phi(z)$ . Hence, it is not necessary to tabulate both the negative and positive halves of the distribution.

**FIGURE B.1** The Normal Distribution.



## 1090 PART VI ♦ Appendices

### B.4.2 THE CHI-SQUARED, $t$ , AND $F$ DISTRIBUTIONS

The chi-squared,  $t$ , and  $F$  distributions are derived from the normal distribution. They arise in econometrics as sums of  $n$  or  $n_1$  and  $n_2$  other variables. These three distributions have associated with them one or two “degrees of freedom” parameters, which for our purposes will be the number of variables in the relevant sum.

The first of the essential results is

- If  $z \sim N[0, 1]$ , then  $x = z^2 \sim \text{chi-squared}[1]$ —that is, **chi-squared** with one degree of freedom—denoted

$$z^2 \sim \chi^2[1]. \quad (\mathbf{B-30})$$

This distribution is a skewed distribution with mean 1 and variance 2. The second result is

- If  $x_1, \dots, x_n$  are  $n$  *independent* chi-squared[1] variables, then

$$\sum_{i=1}^n x_i \sim \text{chi-squared}[n]. \quad (\mathbf{B-31})$$

The mean and variance of a chi-squared variable with  $n$  degrees of freedom are  $n$  and  $2n$ , respectively. A number of useful corollaries can be derived using (B-30) and (B-31).

- If  $z_i, i = 1, \dots, n$ , are independent  $N[0, 1]$  variables, then

$$\sum_{i=1}^n z_i^2 \sim \chi^2[n]. \quad (\mathbf{B-32})$$

- If  $z_i, i = 1, \dots, n$ , are independent  $N[0, \sigma^2]$  variables, then

$$\sum_{i=1}^n (z_i/\sigma)^2 \sim \chi^2[n]. \quad (\mathbf{B-33})$$

- If  $x_1$  and  $x_2$  are independent chi-squared variables with  $n_1$  and  $n_2$  degrees of freedom, respectively, then

$$x_1 + x_2 \sim \chi^2[n_1 + n_2]. \quad (\mathbf{B-34})$$

This result can be generalized to the sum of an arbitrary number of independent chi-squared variables.

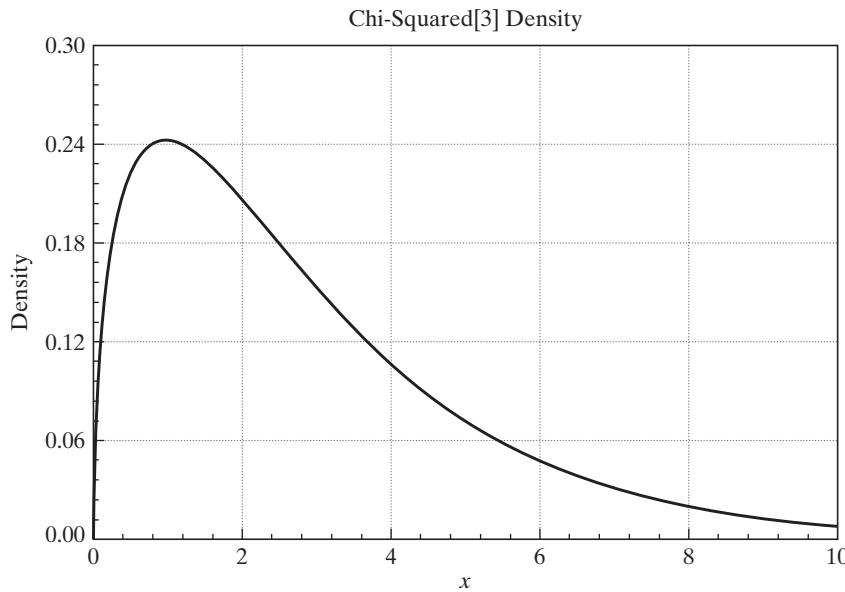
Figure B.2 shows the chi-squared density for three degrees of freedom. The amount of skewness declines as the number of degrees of freedom rises. Unlike the normal distribution, a separate table is required for the chi-squared distribution for each value of  $n$ . Typically, only a few percentage points of the distribution are tabulated for each  $n$ . Table G.3 in Appendix G of this book gives lower (left) tail areas for a number of values.

- If  $x_1$  and  $x_2$  are two *independent* chi-squared variables with degrees of freedom parameters  $n_1$  and  $n_2$ , respectively, then the ratio

$$F[n_1, n_2] = \frac{x_1/n_1}{x_2/n_2} \quad (\mathbf{B-35})$$

has the  **$F$  distribution** with  $n_1$  and  $n_2$  degrees of freedom.

The two degrees of freedom parameters  $n_1$  and  $n_2$  are the numerator and denominator degrees of freedom, respectively. Tables of the  $F$  distribution must be computed for each pair of values of  $(n_1, n_2)$ . As such, only one or two specific values, such as the 95 percent and 99 percent upper tail values, are tabulated in most cases.

APPENDIX B ♦ Probability and Distribution Theory **1091****FIGURE B.2** The Chi-Squared [3] Distribution.

- If  $z$  is an  $N[0, 1]$  variable and  $x$  is  $\chi^2[n]$  and is independent of  $z$ , then the ratio

$$t[n] = \frac{z}{\sqrt{x/n}} \quad (\text{B-36})$$

has the  **$t$  distribution** with  $n$  degrees of freedom.

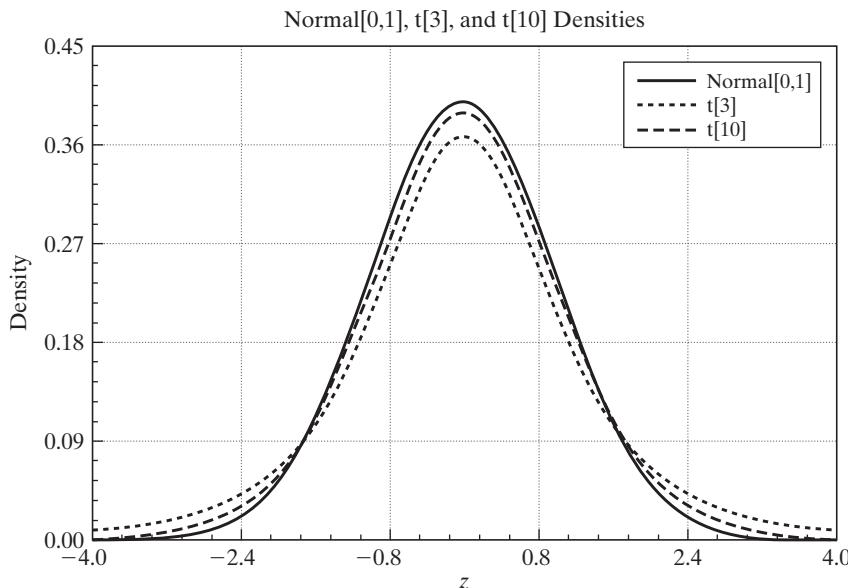
The  $t$  distribution has the same shape as the normal distribution but has thicker tails. Figure B.3 illustrates the  $t$  distributions with 3 and 10 degrees of freedom with the standard normal distribution. Two effects that can be seen in the figure are how the distribution changes as the degrees of freedom increases, and, overall, the similarity of the  $t$  distribution to the standard normal. This distribution is tabulated in the same manner as the chi-squared distribution, with several specific cutoff points corresponding to specified tail areas for various values of the degrees of freedom parameter.

Comparing (B-35) with  $n_1 = 1$  and (B-36), we see the useful relationship between the  $t$  and  $F$  distributions:

- If  $t \sim t[n]$ , then  $t^2 \sim F[1, n]$ .

If the numerator in (B-36) has a nonzero mean, then the random variable in (B-36) has a non-central  $t$  distribution and its square has a noncentral  $F$  distribution. These distributions arise in the  $F$  tests of linear restrictions [see (5-6)] when the restrictions do not hold as follows:

1. *Noncentral chi-squared distribution.* If  $z$  has a normal distribution with mean  $\mu$  and standard deviation 1, then the distribution of  $z^2$  is *noncentral* chi-squared with parameters 1 and  $\mu^2/2$ .
  - a. If  $\mathbf{z} \sim N[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$  with  $J$  elements, then  $\mathbf{z}'\boldsymbol{\Sigma}^{-1}\mathbf{z}$  has a noncentral chi-squared distribution with  $J$  degrees of freedom and noncentrality parameter  $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}/2$ , which we denote  $\chi_*^2[J, \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}/2]$ .
  - b. If  $\mathbf{z} \sim N[\boldsymbol{\mu}, \mathbf{I}]$  and  $\mathbf{M}$  is an idempotent matrix with rank  $J$ , then  $\mathbf{z}'\mathbf{M}\mathbf{z} \sim \chi_*^2[J, \boldsymbol{\mu}'\mathbf{M}\boldsymbol{\mu}/2]$ .

**1092 PART VI ♦ Appendices**


**FIGURE B.3** The Standard Normal,  $t[3]$ , and  $t[10]$  Distributions.

2. *Noncentral F distribution.* If  $X_1$  has a noncentral chi-squared distribution with noncentrality parameter  $\lambda$  and degrees of freedom  $n_1$  and  $X_2$  has a central chi-squared distribution with degrees of freedom  $n_2$  and is independent of  $X_1$ , then

$$F_* = \frac{X_1/n_1}{X_2/n_2}$$

has a noncentral  $F$  distribution with parameters  $n_1$ ,  $n_2$ , and  $\lambda$ .<sup>2</sup> Note that in each of these cases, the statistic and the distribution are the familiar ones, except that the effect of the nonzero mean, which induces the noncentrality, is to push the distribution to the right.

#### B.4.3 DISTRIBUTIONS WITH LARGE DEGREES OF FREEDOM

The chi-squared,  $t$ , and  $F$  distributions usually arise in connection with sums of sample observations. The degrees of freedom parameter in each case grows with the number of observations. We often deal with larger degrees of freedom than are shown in the tables. Thus, the standard tables are often inadequate. In all cases, however, there are **limiting distributions** that we can use when the degrees of freedom parameter grows large. The simplest case is the  $t$  distribution. The  $t$  distribution with infinite degrees of freedom is equivalent to the standard normal distribution. Beyond about 100 degrees of freedom, they are almost indistinguishable.

For degrees of freedom greater than 30, a reasonably good approximation for the distribution of the chi-squared variable  $x$  is

$$z = (2x)^{1/2} - (2n - 1)^{1/2}, \quad (\text{B-37})$$

which is approximately standard normally distributed. Thus,

$$\text{Prob}(\chi^2[n] \leq a) \approx \Phi[(2a)^{1/2} - (2n - 1)^{1/2}].$$

<sup>2</sup>The denominator chi-squared could also be noncentral, but we shall not use any statistics with doubly noncentral distributions.

## APPENDIX B ♦ Probability and Distribution Theory 1093

As used in econometrics, the  $F$  distribution with a large-denominator degrees of freedom is common. As  $n_2$  becomes infinite, the denominator of  $F$  converges identically to one, so we can treat the variable

$$x = n_1 F \quad (\mathbf{B-38})$$

as a chi-squared variable with  $n_1$  degrees of freedom. The numerator degree of freedom will typically be small, so this approximation will suffice for the types of applications we are likely to encounter.<sup>3</sup> If not, then the approximation given earlier for the chi-squared distribution can be applied to  $n_1 F$ .

### B.4.4 SIZE DISTRIBUTIONS: THE LOGNORMAL DISTRIBUTION

In modeling size distributions, such as the distribution of firm sizes in an industry or the distribution of income in a country, the **lognormal distribution**, denoted  $LN[\mu, \sigma^2]$ , has been particularly useful.<sup>4</sup>

$$f(x) = \frac{1}{\sqrt{2\pi} \sigma x} e^{-1/2[(\ln x - \mu)/\sigma]^2}, \quad x > 0.$$

A lognormal variable  $x$  has

$$E[x] = e^{\mu + \sigma^2/2},$$

and

$$\text{Var}[x] = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1).$$

The relation between the normal and lognormal distributions is

$$\text{If } y \sim LN[\mu, \sigma^2], \quad \ln y \sim N[\mu, \sigma^2].$$

A useful result for transformations is given as follows:

If  $x$  has a lognormal distribution with mean  $\theta$  and variance  $\lambda^2$ , then

$$\ln x \sim N(\mu, \sigma^2), \quad \text{where } \mu = \ln \theta^2 - \frac{1}{2} \ln(\theta^2 + \lambda^2) \quad \text{and} \quad \sigma^2 = \ln(1 + \lambda^2/\theta^2).$$

Because the normal distribution is preserved under linear transformation,

$$\text{if } y \sim LN[\mu, \sigma^2], \quad \text{then } \ln y^r \sim N[r\mu, r^2\sigma^2].$$

If  $y_1$  and  $y_2$  are independent lognormal variables with  $y_1 \sim LN[\mu_1, \sigma_1^2]$  and  $y_2 \sim LN[\mu_2, \sigma_2^2]$ , then

$$y_1 y_2 \sim LN[\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2].$$

### B.4.5 THE GAMMA AND EXPONENTIAL DISTRIBUTIONS

The **gamma distribution** has been used in a variety of settings, including the study of income distribution<sup>5</sup> and production functions.<sup>6</sup> The general form of the distribution is

$$f(x) = \frac{\lambda^P}{\Gamma(P)} e^{-\lambda x} x^{P-1}, \quad x \geq 0, \lambda > 0, P > 0. \quad (\mathbf{B-39})$$

Many familiar distributions are special cases, including the **exponential distribution** ( $P=1$ ) and chi-squared ( $\lambda = \frac{1}{2}$ ,  $P = \frac{n}{2}$ ). The **Erlang distribution** results if  $P$  is a positive integer. The mean is  $P/\lambda$ , and the variance is  $P/\lambda^2$ . The **inverse gamma distribution** is the distribution of  $1/x$ , where  $x$

<sup>3</sup>See Johnson, Kotz, and Balakrishnan (1994) for other approximations.

<sup>4</sup>A study of applications of the lognormal distribution appears in Aitchison and Brown (1969).

<sup>5</sup>Salem and Mount (1974).

<sup>6</sup>Greene (1980a).

## 1094 PART VI ♦ Appendices

has the gamma distribution. Using the change of variable,  $y = 1/x$ , the Jacobian is  $|dx/dy| = 1/y^2$ . Making the substitution and the change of variable, we find

$$f(y) = \frac{\lambda^P}{\Gamma(P)} e^{-\lambda/y} y^{-(P+1)}, \quad y \geq 0, \lambda > 0, P > 0.$$

The density is defined for positive  $P$ . However, the mean is  $\lambda/(P - 1)$  which is defined only if  $P > 1$  and the variance is  $\lambda^2/[(P - 1)^2(P - 2)]$  which is defined only for  $P > 2$ .

### B.4.6 THE BETA DISTRIBUTION

Distributions for models are often chosen on the basis of the range within which the random variable is constrained to vary. The lognormal distribution, for example, is sometimes used to model a variable that is always nonnegative. For a variable constrained between 0 and  $c > 0$ , the **beta distribution** has proved useful. Its density is

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{x}{c}\right)^{\alpha-1} \left(1 - \frac{x}{c}\right)^{\beta-1} \frac{1}{c}. \quad (\text{B-40})$$

This functional form is extremely flexible in the shapes it will accommodate. It is symmetric if  $\alpha = \beta$ , asymmetric otherwise, and can be hump-shaped or U-shaped. The mean is  $c\alpha/(\alpha + \beta)$ , and the variance is  $c^2\alpha\beta/[(\alpha + \beta + 1)(\alpha + \beta)^2]$ . The beta distribution has been applied in the study of labor force participation rates.<sup>7</sup>

### B.4.7 THE LOGISTIC DISTRIBUTION

The normal distribution is ubiquitous in econometrics. But researchers have found that for some microeconomic applications, there does not appear to be enough mass in the tails of the normal distribution; observations that a model based on normality would classify as “unusual” seem not to be very unusual at all. One approach has been to use thicker-tailed symmetric distributions. The **logistic distribution** is one candidate; the cdf for a logistic random variable is denoted

$$F(x) = \Lambda(x) = \frac{1}{1 + e^{-x}}.$$

The density is  $f(x) = \Lambda(x)[1 - \Lambda(x)]$ . The mean and variance of this random variable are zero and  $\pi^2/3$ .

### B.4.8 THE WISHART DISTRIBUTION

The Wishart distribution describes the distribution of a random matrix obtained as

$$\mathbf{W} = \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})',$$

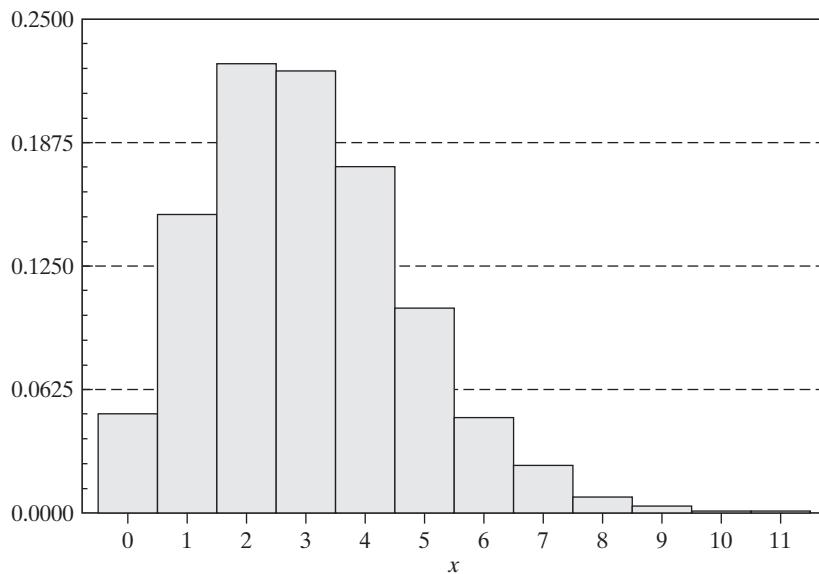
where  $\mathbf{x}_i$  is the  $i$ th of  $n$   $K$  element random vectors from the multivariate normal distribution with mean vector,  $\boldsymbol{\mu}$ , and covariance matrix,  $\boldsymbol{\Sigma}$ . This is a multivariate counterpart to the chi-squared distribution. The density of the Wishart random matrix is

$$f(\mathbf{W}) = \frac{\exp\left[-\frac{1}{2}\text{trace}(\boldsymbol{\Sigma}^{-1}\mathbf{W})\right] |\mathbf{W}|^{-\frac{1}{2}(n-K-1)}}{2^{nK/2} |\boldsymbol{\Sigma}|^{K/2} \pi^{K(K-1)/4} \prod_{j=1}^K \Gamma\left(\frac{n+1-j}{2}\right)}.$$

The mean matrix is  $n\boldsymbol{\Sigma}$ . For the individual pairs of elements in  $\mathbf{W}$ ,

$$\text{Cov}[w_{ij}, w_{rs}] = n(\sigma_{ir}\sigma_{js} + \sigma_{is}\sigma_{jr}).$$

<sup>7</sup>Heckman and Willis (1976).

APPENDIX B ♦ Probability and Distribution Theory **1095****FIGURE B.4** The Poisson [3] Distribution.**B.4.9 DISCRETE RANDOM VARIABLES**

Modeling in economics frequently involves random variables that take integer values. In these cases, the distributions listed thus far only provide approximations that are sometimes quite inappropriate. We can build up a class of models for discrete random variables from the **Bernoulli distribution** for a single binomial outcome (trial)

$$\text{Prob}(x = 1) = \alpha,$$

$$\text{Prob}(x = 0) = 1 - \alpha,$$

where  $0 \leq \alpha \leq 1$ . The modeling aspect of this specification would be the assumptions that the success probability  $\alpha$  is constant from one trial to the next and that successive trials are independent. If so, then the distribution for  $x$  successes in  $n$  trials is the **binomial distribution**,

$$\text{Prob}(X = x) = \binom{n}{x} \alpha^x (1 - \alpha)^{n-x}, \quad x = 0, 1, \dots, n.$$

The mean and variance of  $x$  are  $n\alpha$  and  $n\alpha(1 - \alpha)$ , respectively. If the number of trials becomes large at the same time that the success probability becomes small so that the mean  $n\alpha$  is stable, then, the limiting form of the binomial distribution is the **Poisson distribution**,

$$\text{Prob}(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}.$$

The Poisson distribution has seen wide use in econometrics in, for example, modeling patents, crime, recreation demand, and demand for health services. (See Chapter 18.) An example is shown in Figure B.4.

**B.5 THE DISTRIBUTION OF A FUNCTION OF A RANDOM VARIABLE**

We considered finding the expected value of a function of a random variable. It is fairly common to analyze the random variable itself, which results when we compute a function of some random variable. There are three types of transformation to consider. One discrete random variable may

## 1096 PART VI ♦ Appendices

be transformed into another, a continuous variable may be transformed into a discrete one, and one continuous variable may be transformed into another.

The simplest case is the first one. The probabilities associated with the new variable are computed according to the laws of probability. If  $y$  is derived from  $x$  and the function is one to one, then the probability that  $Y = y(x)$  equals the probability that  $X = x$ . If several values of  $x$  yield the same value of  $y$ , then  $\text{Prob}(Y = y)$  is the sum of the corresponding probabilities for  $x$ .

The second type of transformation is illustrated by the way individual data on income are typically obtained in a survey. Income in the population can be expected to be distributed according to some skewed, continuous distribution such as the one shown in Figure B.5.

Data are often reported categorically, as shown in the lower part of the figure. Thus, the random variable corresponding to observed income is a discrete transformation of the actual underlying continuous random variable. Suppose, for example, that the transformed variable  $y$  is the mean income in the respective interval. Then

$$\text{Prob}(Y = \mu_1) = P(-\infty < X \leq a),$$

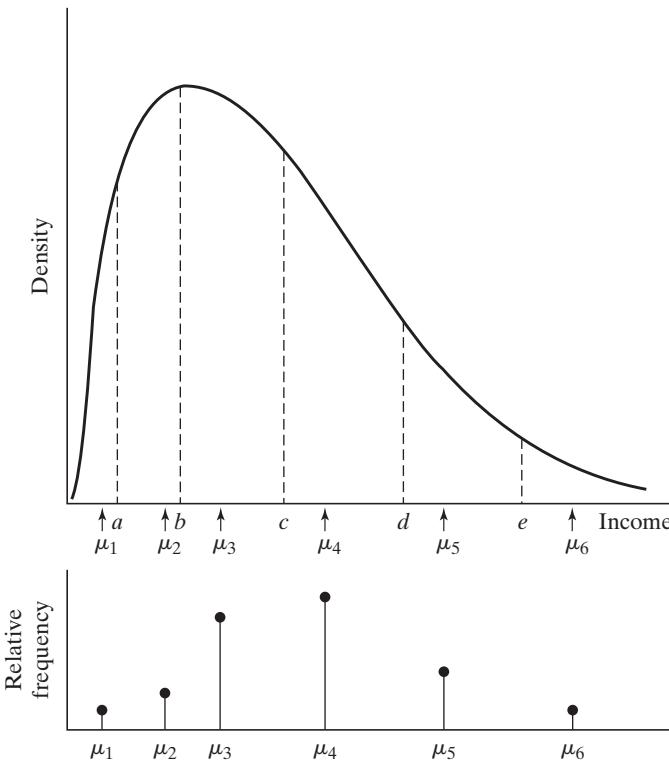
$$\text{Prob}(Y = \mu_2) = P(a < X \leq b),$$

$$\text{Prob}(Y = \mu_3) = P(b < X \leq c),$$

and so on, which illustrates the general procedure.

If  $x$  is a continuous random variable with pdf  $f_x(x)$  and if  $y = g(x)$  is a continuous monotonic function of  $x$ , then the density of  $y$  is obtained by using the change of variable technique to find

**FIGURE B.5** Censored Distribution.



**APPENDIX B ♦ Probability and Distribution Theory 1097**

the cdf of  $y$ :

$$\text{Prob}(y \leq b) = \int_{-\infty}^b f_x(g^{-1}(y))|g^{-1'}(y)| dy.$$

This equation can now be written as

$$\text{Prob}(y \leq b) = \int_{-\infty}^b f_y(y) dy.$$

Hence,

$$f_y(y) = f_x(g^{-1}(y))|g^{-1'}(y)|. \quad (\mathbf{B-41})$$

To avoid the possibility of a negative pdf if  $g(x)$  is decreasing, we use the absolute value of the derivative in the previous expression. The term  $|g^{-1'}(y)|$  must be nonzero for the density of  $y$  to be nonzero. In words, the probabilities associated with intervals in the range of  $y$  must be associated with intervals in the range of  $x$ . If the derivative is zero, the correspondence  $y = g(x)$  is vertical, and hence all values of  $y$  in the given range are associated with the same value of  $x$ . This single point must have probability zero.

One of the most useful applications of the preceding result is the linear transformation of a normally distributed variable. If  $x \sim N[\mu, \sigma^2]$ , then the distribution of

$$y = \frac{x - \mu}{\sigma}$$

is found using the preceding result. First, the derivative is obtained from the inverse transformation

$$y = \frac{x}{\sigma} - \frac{\mu}{\sigma} \Rightarrow x = \sigma y + \mu \Rightarrow f^{-1'}(y) = \frac{dx}{dy} = \sigma.$$

Therefore,

$$f_y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(\sigma y + \mu - \mu)^2/(2\sigma^2)} |\sigma| = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}.$$

This is the density of a normally distributed variable with mean zero and unit standard deviation one. This is the result which makes it unnecessary to have separate tables for the different normal distributions which result from different means and variances.

## B.6 REPRESENTATIONS OF A PROBABILITY DISTRIBUTION

The probability density function (pdf) is a natural and familiar way to formulate the distribution of a random variable. But, there are many other functions that are used to identify or characterize a random variable, depending on the setting. In each of these cases, we can identify some other function of the random variable that has a one-to-one relationship with the density. We have already used one of these quite heavily in the preceding discussion. For a random variable which has density function  $f(x)$ , the distribution function,  $F(x)$ , is an equally informative function that identifies the distribution; the relationship between  $f(x)$  and  $F(x)$  is defined in (B-6) for a discrete random variable and (B-8) for a continuous one. We now consider several other related functions.

For a continuous random variable, the **survival function** is  $S(x) = 1 - F(x) = \text{Prob}[X \geq x]$ . This function is widely used in epidemiology, where  $x$  is time until some transition, such as recovery

## 1098 PART VI ♦ Appendices

from a disease. The **hazard function** for a random variable is

$$h(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1 - F(x)}.$$

The hazard function is a conditional probability;

$$h(x) = \lim_{t \downarrow 0} \text{Prob}(X \leq x \leq X + t | X \geq x).$$

Hazard functions have been used in econometrics in studying the duration of spells, or conditions, such as unemployment, strikes, time until business failures, and so on. The connection between the hazard and the other functions is  $h(x) = -d \ln S(x)/dx$ . As an exercise, you might want to verify the interesting special case of  $h(x) = 1/\lambda$ , a constant—the only distribution which has this characteristic is the exponential distribution noted in Section B.4.5.

For the random variable  $X$ , with probability density function  $f(x)$ , if the function

$$M(t) = E[e^{tx}]$$

exists, then it is the **moment generating function**. Assuming the function exists, it can be shown that

$$d^r M(t)/dt^r|_{t=0} = E[x^r].$$

The moment generating function, like the survival and the hazard functions, is a unique characterization of a probability distribution. When it exists, the moment generating function (MGF) has a one-to-one correspondence with the distribution. Thus, for example, if we begin with some random variable and find that a transformation of it has a particular MGF, then we may infer that the function of the random variable has the distribution associated with that MGF. A convenient application of this result is the MGF for the normal distribution. The MGF for the standard normal distribution is  $M_z(t) = e^{t^2/2}$ .

A useful feature of MGFs is the following:

If  $x$  and  $y$  are independent, then the MGF of  $x + y$  is  $M_x(t)M_y(t)$ .

This result has been used to establish the **contagion** property of some distributions, that is, the property that sums of random variables with a given distribution have that same distribution. The normal distribution is a familiar example. This is usually not the case. It is for Poisson and chi-squared random variables.

One qualification of all of the preceding is that in order for these results to hold, the MGF must exist. It will for the distributions that we will encounter in our work, but in at least one important case, we cannot be sure of this. When computing sums of random variables which may have different distributions and whose specific distributions need not be so well behaved, it is likely that the MGF of the sum does not exist. However, the characteristic function,

$$\phi(t) = E[e^{itx}], i^2 = -1,$$

will always exist, at least for relatively small  $t$ . The characteristic function is the device used to prove that certain sums of random variables converge to a normally distributed variable—that is, the characteristic function is a fundamental tool in proofs of the central limit theorem.

**APPENDIX B ♦ Probability and Distribution Theory 1099**

## B.7 JOINT DISTRIBUTIONS

The **joint density function** for two random variables  $X$  and  $Y$  denoted  $f(x, y)$  is defined so that

$$\text{Prob}(a \leq x \leq b, c \leq y \leq d) = \begin{cases} \sum_{a \leq x \leq b} \sum_{c \leq y \leq d} f(x, y) & \text{if } x \text{ and } y \text{ are discrete,} \\ \int_a^b \int_c^d f(x, y) dy dx & \text{if } x \text{ and } y \text{ are continuous.} \end{cases} \quad (\text{B-42})$$

The counterparts of the requirements for a univariate probability density are

$$\begin{aligned} f(x, y) &\geq 0, \\ \sum_x \sum_y f(x, y) &= 1 \quad \text{if } x \text{ and } y \text{ are discrete,} \\ \int_x \int_y f(x, y) dy dx &= 1 \quad \text{if } x \text{ and } y \text{ are continuous.} \end{aligned} \quad (\text{B-43})$$

The cumulative probability is likewise the probability of a joint event:

$$\begin{aligned} F(x, y) &= \text{Prob}(X \leq x, Y \leq y) \\ &= \begin{cases} \sum_{X \leq x} \sum_{Y \leq y} f(x, y) & \text{in the discrete case} \\ \int_{-\infty}^x \int_{-\infty}^y f(t, s) ds dt & \text{in the continuous case.} \end{cases} \end{aligned} \quad (\text{B-44})$$

### B.7.1 MARGINAL DISTRIBUTIONS

A **marginal probability density** or marginal probability distribution is defined with respect to an individual variable. To obtain the marginal distributions from the joint density, it is necessary to sum or integrate out the other variable:

$$f_x(x) = \begin{cases} \sum_y f(x, y) & \text{in the discrete case} \\ \int_y f(x, s) ds & \text{in the continuous case,} \end{cases} \quad (\text{B-45})$$

and similarly for  $f_y(y)$ .

Two random variables are statistically independent if and only if their joint density is the product of the marginal densities:

$$f(x, y) = f_x(x) f_y(y) \Leftrightarrow x \text{ and } y \text{ are independent.} \quad (\text{B-46})$$

If (and only if)  $x$  and  $y$  are independent, then the cdf factors as well as the pdf:

$$F(x, y) = F_x(x) F_y(y), \quad (\text{B-47})$$

or

$$\text{Prob}(X \leq x, Y \leq y) = \text{Prob}(X \leq x) \text{Prob}(Y \leq y).$$

## 1100 PART VI ♦ Appendices

### B.7.2 EXPECTATIONS IN A JOINT DISTRIBUTION

The means, variances, and higher moments of the variables in a joint distribution are defined with respect to the marginal distributions. For the mean of  $x$  in a discrete distribution,

$$\begin{aligned} E[x] &= \sum_x x f_x(x) \\ &= \sum_x x \left[ \sum_y f(x, y) \right] \\ &= \sum_x \sum_y x f(x, y). \end{aligned} \tag{B-48}$$

The means of the variables in a continuous distribution are defined likewise, using integration instead of summation:

$$\begin{aligned} E[x] &= \int_x x f_x(x) dx \\ &= \int_x \int_y x f(x, y) dy dx. \end{aligned} \tag{B-49}$$

Variances are computed in the same manner:

$$\begin{aligned} \text{Var}[x] &= \sum_x (x - E[x])^2 f_x(x) \\ &= \sum_x \sum_y (x - E[x])^2 f(x, y). \end{aligned} \tag{B-50}$$

### B.7.3 COVARIANCE AND CORRELATION

For any function  $g(x, y)$ ,

$$E[g(x, y)] = \begin{cases} \sum_x \sum_y g(x, y) f(x, y) & \text{in the discrete case} \\ \int_x \int_y g(x, y) f(x, y) dy dx & \text{in the continuous case.} \end{cases} \tag{B-51}$$

The covariance of  $x$  and  $y$  is a special case:

$$\begin{aligned} \text{Cov}[x, y] &= E[(x - \mu_x)(y - \mu_y)] \\ &= E[xy] - \mu_x \mu_y \\ &= \sigma_{xy}. \end{aligned} \tag{B-52}$$

If  $x$  and  $y$  are independent, then  $f(x, y) = f_x(x) f_y(y)$  and

$$\begin{aligned} \sigma_{xy} &= \sum_x \sum_y f_x(x) f_y(y) (x - \mu_x)(y - \mu_y) \\ &= \sum_x (x - \mu_x) f_x(x) \sum_y (y - \mu_y) f_y(y) \\ &= E[x - \mu_x] E[y - \mu_y] \\ &= 0. \end{aligned}$$

## APPENDIX B ♦ Probability and Distribution Theory 1101

The sign of the covariance will indicate the direction of covariation of  $X$  and  $Y$ . Its magnitude depends on the scales of measurement, however. In view of this fact, a preferable measure is the correlation coefficient:

$$r[x, y] = \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \quad (\mathbf{B-53})$$

where  $\sigma_x$  and  $\sigma_y$  are the standard deviations of  $x$  and  $y$ , respectively. The correlation coefficient has the same sign as the covariance but is always between  $-1$  and  $1$  and is thus unaffected by any scaling of the variables.

Variables that are uncorrelated are not necessarily independent. For example, in the discrete distribution  $f(-1, 1) = f(0, 0) = f(1, 1) = \frac{1}{3}$ , the correlation is zero, but  $f(1, 1)$  does not equal  $f_x(1)f_y(1) = (\frac{1}{3})(\frac{2}{3})$ . An important exception is the joint normal distribution discussed subsequently, in which lack of correlation does imply independence.

Some general results regarding expectations in a joint distribution, which can be verified by applying the appropriate definitions, are

$$E[ax + by + c] = aE[x] + bE[y] + c, \quad (\mathbf{B-54})$$

$$\begin{aligned} \text{Var}[ax + by + c] &= a^2\text{Var}[x] + b^2\text{Var}[y] + 2ab\text{Cov}[x, y] \\ &= \text{Var}[ax + by], \end{aligned} \quad (\mathbf{B-55})$$

and

$$\text{Cov}[ax + by, cx + dy] = ac\text{Var}[x] + bd\text{Var}[y] + (ad + bc)\text{Cov}[x, y]. \quad (\mathbf{B-56})$$

If  $X$  and  $Y$  are uncorrelated, then

$$\begin{aligned} \text{Var}[x + y] &= \text{Var}[x - y] \\ &= \text{Var}[x] + \text{Var}[y]. \end{aligned} \quad (\mathbf{B-57})$$

For any two functions  $g_1(x)$  and  $g_2(y)$ , if  $x$  and  $y$  are independent, then

$$E[g_1(x)g_2(y)] = E[g_1(x)]E[g_2(y)]. \quad (\mathbf{B-58})$$

### B.7.4 DISTRIBUTION OF A FUNCTION OF BIVARIATE RANDOM VARIABLES

The result for a function of a random variable in (B-41) must be modified for a joint distribution. Suppose that  $x_1$  and  $x_2$  have a joint distribution  $f_x(x_1, x_2)$  and that  $y_1$  and  $y_2$  are two monotonic functions of  $x_1$  and  $x_2$ :

$$y_1 = y_1(x_1, x_2),$$

$$y_2 = y_2(x_1, x_2).$$

Because the functions are monotonic, the inverse transformations,

$$x_1 = x_1(y_1, y_2),$$

$$x_2 = x_2(y_1, y_2),$$

## 1102 PART VI ♦ Appendices

exist. The Jacobian of the transformations is the matrix of partial derivatives,

$$J = \begin{bmatrix} \partial x_1 / \partial y_1 & \partial x_1 / \partial y_2 \\ \partial x_2 / \partial y_1 & \partial x_2 / \partial y_2 \end{bmatrix} = \begin{bmatrix} \partial \mathbf{x} \\ \partial \mathbf{y}' \end{bmatrix}.$$

The joint distribution of  $y_1$  and  $y_2$  is

$$f_y(y_1, y_2) = f_x[x_1(y_1, y_2), x_2(y_1, y_2)] \text{abs}(|J|).$$

The determinant of the Jacobian must be nonzero for the transformation to exist. A zero determinant implies that the two transformations are functionally dependent.

Certainly the most common application of the preceding in econometrics is the linear transformation of a set of random variables. Suppose that  $x_1$  and  $x_2$  are independently distributed  $N[0, 1]$ , and the transformations are

$$\begin{aligned} y_1 &= \alpha_1 + \beta_{11}x_1 + \beta_{12}x_2, \\ y_2 &= \alpha_2 + \beta_{21}x_1 + \beta_{22}x_2. \end{aligned}$$

To obtain the joint distribution of  $y_1$  and  $y_2$ , we first write the transformations as

$$\mathbf{y} = \mathbf{a} + \mathbf{B}\mathbf{x}.$$

The inverse transformation is

$$\mathbf{x} = \mathbf{B}^{-1}(\mathbf{y} - \mathbf{a}),$$

so the absolute value of the determinant of the Jacobian is

$$\text{abs}|J| = \text{abs}|\mathbf{B}^{-1}| = \frac{1}{\text{abs}|\mathbf{B}|}.$$

The joint distribution of  $\mathbf{x}$  is the product of the marginal distributions since they are independent. Thus,

$$f_x(\mathbf{x}) = (2\pi)^{-1} e^{-(x_1^2 + x_2^2)/2} = (2\pi)^{-1} e^{-\mathbf{x}'\mathbf{x}/2}.$$

Inserting the results for  $\mathbf{x}(\mathbf{y})$  and  $J$  into  $f_y(y_1, y_2)$  gives

$$f_y(\mathbf{y}) = (2\pi)^{-1} \frac{1}{\text{abs}|\mathbf{B}|} e^{-(\mathbf{y}-\mathbf{a})'(\mathbf{B}\mathbf{B}')^{-1}(\mathbf{y}-\mathbf{a})/2}.$$

This **bivariate normal distribution** is the subject of Section B.9. Note that by formulating it as we did earlier, we can generalize easily to the multivariate case, that is, with an arbitrary number of variables.

Perhaps the more common situation is that in which it is necessary to find the distribution of one function of two (or more) random variables. A strategy that often works in this case is to form the joint distribution of the transformed variable and one of the original variables, then integrate (or sum) the latter out of the joint distribution to obtain the marginal distribution. Thus, to find the distribution of  $y_1(x_1, x_2)$ , we might formulate

$$y_1 = y_1(x_1, x_2)$$

$$y_2 = x_2.$$

## APPENDIX B ♦ Probability and Distribution Theory 1103

The absolute value of the determinant of the Jacobian would then be

$$J = \text{abs} \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ 0 & 1 \end{vmatrix} = \text{abs} \left| \left( \frac{\partial x_1}{\partial y_1} \right) \right|.$$

The density of  $y_1$  would then be

$$f_{y_1}(y_1) = \int_{y_2} f_x[x_1(y_1, y_2), y_2] \text{abs}|J| dy_2.$$

## B.8 CONDITIONING IN A BIVARIATE DISTRIBUTION

Conditioning and the use of conditional distributions play a pivotal role in econometric modeling. We consider some general results for a bivariate distribution. (All these results can be extended directly to the multivariate case.)

In a bivariate distribution, there is a **conditional distribution** over  $y$  for each value of  $x$ . The conditional densities are

$$f(y|x) = \frac{f(x,y)}{f_x(x)}, \quad (\text{B-59})$$

and

$$f(x|y) = \frac{f(x,y)}{f_y(y)}.$$

It follows from (B-46) that,

$$\text{If } x \text{ and } y \text{ are independent, then } f(y|x) = f_y(y) \quad \text{and} \quad f(x|y) = f_x(x). \quad (\text{B-60})$$

The interpretation is that if the variables are independent, the probabilities of events relating to one variable are unrelated to the other. The definition of conditional densities implies the important result

$$\begin{aligned} f(x,y) &= f(y|x)f_x(x) \\ &= f(x|y)f_y(y). \end{aligned} \quad (\text{B-61})$$

### B.8.1 REGRESSION: THE CONDITIONAL MEAN

A **conditional mean** is the mean of the conditional distribution and is defined by

$$E[y|x] = \begin{cases} \int_y yf(y|x) dy & \text{if } y \text{ is continuous} \\ \sum_y yf(y|x) & \text{if } y \text{ is discrete.} \end{cases} \quad (\text{B-62})$$

The conditional mean function  $E[y|x]$  is called the **regression** of  $y$  on  $x$ .

A random variable may always be written as

$$\begin{aligned} y &= E[y|x] + (y - E[y|x]) \\ &= E[y|x] + \varepsilon. \end{aligned}$$

## 1104 PART VI ♦ Appendices

### B.8.2 CONDITIONAL VARIANCE

A conditional variance is the variance of the conditional distribution:

$$\begin{aligned}\text{Var}[y|x] &= E[(y - E[y|x])^2 | x] \\ &= \int_y (y - E[y|x])^2 f(y|x) dy, \quad \text{if } y \text{ is continuous},\end{aligned}\tag{B-63}$$

or

$$\text{Var}[y|x] = \sum_y (y - E[y|x])^2 f(y|x), \quad \text{if } y \text{ is discrete.}\tag{B-64}$$

The computation can be simplified by using

$$\text{Var}[y|x] = E[y^2|x] - (E[y|x])^2.\tag{B-65}$$

The conditional variance is called the **scedastic function** and, like the regression, is generally a function of  $x$ . Unlike the conditional mean function, however, it is common for the conditional variance not to vary with  $x$ . We shall examine a particular case. This case does not imply, however, that  $\text{Var}[y|x]$  equals  $\text{Var}[y]$ , which will usually not be true. It implies only that the conditional variance is a constant. The case in which the conditional variance does not vary with  $x$  is called **homoscedasticity** (same variance).

### B.8.3 RELATIONSHIPS AMONG MARGINAL AND CONDITIONAL MOMENTS

Some useful results for the moments of a conditional distribution are given in the following theorems.

#### THEOREM B.1 Law of Iterated Expectations

$$E[y] = E_x[E[y|x]].\tag{B-66}$$

*The notation  $E_x[.]$  indicates the expectation over the values of  $x$ . Note that  $E[y|x]$  is a function of  $x$ .*

#### THEOREM B.2 Covariance

*In any bivariate distribution,*

$$\text{Cov}[x, y] = \text{Cov}_x[x, E[y|x]] = \int_x (x - E[x]) E[y|x] f_x(x) dx.\tag{B-67}$$

*(Note that this is the covariance of  $x$  and a function of  $x$ .)*

**APPENDIX B ♦ Probability and Distribution Theory 1105**

The preceding results provide an additional, extremely useful result for the special case in which the conditional mean function is linear in  $x$ .

**THEOREM B.3 Moments in a Linear Regression**

*If  $E[y|x] = \alpha + \beta x$ , then*

$$\alpha = E[y] - \beta E[x]$$

*and*

$$\beta = \frac{\text{Cov}[x, y]}{\text{Var}[x]}. \quad (\text{B-68})$$

*The proof follows from (B-66).*

The preceding theorems relate to the conditional mean in a bivariate distribution. The following theorems, which also appear in various forms in regression analysis, describe the conditional variance.

**THEOREM B.4 Decomposition of Variance**

*In a joint distribution,*

$$\text{Var}[y] = \text{Var}_x[E[y|x]] + E_x[\text{Var}[y|x]]. \quad (\text{B-69})$$

The notation  $\text{Var}_x[.]$  indicates the variance over the distribution of  $x$ . This equation states that in a bivariate distribution, the variance of  $y$  decomposes into the variance of the conditional mean function plus the expected variance around the conditional mean.

**THEOREM B.5 Residual Variance in a Regression**

*In any bivariate distribution,*

$$E_x[\text{Var}[y|x]] = \text{Var}[y] - \text{Var}_x[E[y|x]]. \quad (\text{B-70})$$

On average, conditioning reduces the variance of the variable subject to the conditioning. For example, if  $y$  is homoscedastic, then we have the unambiguous result that the variance of the conditional distribution(s) is less than or equal to the unconditional variance of  $y$ . Going a step further, we have the result that appears prominently in the bivariate normal distribution (Section B.9).

## 1106 PART VI ♦ Appendices

### THEOREM B.6 Linear Regression and Homoscedasticity

*In a bivariate distribution, if  $E[y|x] = \alpha + \beta x$  and if  $\text{Var}[y|x]$  is a constant, then*

$$\text{Var}[y|x] = \text{Var}[y](1 - \text{Corr}^2[y, x]) = \sigma_y^2(1 - \rho_{xy}^2). \quad (\text{B-71})$$

*The proof is straightforward using Theorems B.2 to B.4.*

#### B.8.4 THE ANALYSIS OF VARIANCE

The variance decomposition result implies that in a bivariate distribution, variation in  $y$  arises from two sources:

1. Variation because  $E[y|x]$  varies with  $x$ :

$$\text{regression variance} = \text{Var}_x[E[y|x]]. \quad (\text{B-72})$$

2. Variation because, in each conditional distribution,  $y$  varies around the conditional mean:

$$\text{residual variance} = E_x[\text{Var}[y|x]]. \quad (\text{B-73})$$

Thus,

$$\text{Var}[y] = \text{regression variance} + \text{residual variance}. \quad (\text{B-74})$$

In analyzing a regression, we shall usually be interested in which of the two parts of the total variance,  $\text{Var}[y]$ , is the larger one. A natural measure is the ratio

$$\text{coefficient of determination} = \frac{\text{regression variance}}{\text{total variance}}. \quad (\text{B-75})$$

In the setting of a linear regression, (B-75) arises from another relationship that emphasizes the interpretation of the correlation coefficient.

$$\text{If } E[y|x] = \alpha + \beta x, \text{ then the coefficient of determination} = \text{COD} = \rho^2, \quad (\text{B-76})$$

where  $\rho^2$  is the squared correlation between  $x$  and  $y$ . We conclude that the correlation coefficient (squared) is a measure of the proportion of the variance of  $y$  accounted for by variation in the mean of  $y$  given  $x$ . It is in this sense that correlation can be interpreted as a **measure of linear association** between two variables.

#### B.9 THE BIVARIATE NORMAL DISTRIBUTION

A bivariate distribution that embodies many of the features described earlier is the bivariate normal, which is the joint distribution of two normally distributed variables. The density is

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-1/2[(\varepsilon_x^2+\varepsilon_y^2-2\rho\varepsilon_x\varepsilon_y)/(1-\rho^2)]}, \quad (\text{B-77})$$

$$\varepsilon_x = \frac{x - \mu_x}{\sigma_x}, \quad \varepsilon_y = \frac{y - \mu_y}{\sigma_y}.$$

## APPENDIX B ♦ Probability and Distribution Theory 1107

The parameters  $\mu_x$ ,  $\sigma_x$ ,  $\mu_y$ , and  $\sigma_y$  are the means and standard deviations of the marginal distributions of  $x$  and  $y$ , respectively. The additional parameter  $\rho$  is the correlation between  $x$  and  $y$ . The covariance is

$$\sigma_{xy} = \rho\sigma_x\sigma_y. \quad (\text{B-78})$$

The density is defined only if  $\rho$  is not 1 or  $-1$ , which in turn requires that the two variables not be linearly related. If  $x$  and  $y$  have a bivariate normal distribution, denoted

$$(x, y) \sim N_2[\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho],$$

then

- The marginal distributions are normal:

$$\begin{aligned} f_x(x) &= N[\mu_x, \sigma_x^2], \\ f_y(y) &= N[\mu_y, \sigma_y^2]. \end{aligned} \quad (\text{B-79})$$

- The conditional distributions are normal:

$$\begin{aligned} f(y|x) &= N[\alpha + \beta x, \sigma_y^2(1 - \rho^2)], \\ \alpha &= \mu_y - \beta\mu_x, \quad \beta = \frac{\sigma_{xy}}{\sigma_x^2}, \end{aligned} \quad (\text{B-80})$$

and likewise for  $f(x|y)$ .

- $x$  and  $y$  are independent if and only if  $\rho = 0$ . The density factors into the product of the two marginal normal distributions if  $\rho = 0$ .

Two things to note about the conditional distributions beyond their normality are their linear regression functions and their constant conditional variances. The conditional variance is less than the unconditional variance, which is consistent with the results of the previous section.

## B.10 MULTIVARIATE DISTRIBUTIONS

The extension of the results for bivariate distributions to more than two variables is direct. It is made much more convenient by using matrices and vectors. The term **random vector** applies to a vector whose elements are random variables. The joint density is  $f(\mathbf{x})$ , whereas the cdf is

$$F(\mathbf{x}) = \int_{-\infty}^{x_n} \int_{-\infty}^{x_{n-1}} \cdots \int_{-\infty}^{x_1} f(\mathbf{t}) dt_1 \cdots dt_{n-1} dt_n. \quad (\text{B-81})$$

Note that the cdf is an  $n$ -fold integral. The marginal distribution of any one (or more) of the  $n$  variables is obtained by integrating or summing over the other variables.

### B.10.1 MOMENTS

The expected value of a vector or matrix is the vector or matrix of expected values. A mean vector is defined as

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} E[x_1] \\ E[x_2] \\ \vdots \\ E[x_n] \end{bmatrix} = E[\mathbf{x}]. \quad (\text{B-82})$$

## 1108 PART VI ♦ Appendices

Define the matrix

$$(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' = \begin{bmatrix} (x_1 - \mu_1)(x_1 - \mu_1) & (x_1 - \mu_1)(x_2 - \mu_2) & \cdots & (x_1 - \mu_1)(x_n - \mu_n) \\ (x_2 - \mu_2)(x_1 - \mu_1) & (x_2 - \mu_2)(x_2 - \mu_2) & \cdots & (x_2 - \mu_2)(x_n - \mu_n) \\ \vdots & \vdots & & \vdots \\ (x_n - \mu_n)(x_1 - \mu_1) & (x_n - \mu_n)(x_2 - \mu_2) & \cdots & (x_n - \mu_n)(x_n - \mu_n) \end{bmatrix}.$$

The expected value of each element in the matrix is the covariance of the two variables in the product. (The covariance of a variable with itself is its variance.) Thus,

$$E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix} = E[\mathbf{x}\mathbf{x}'] - \boldsymbol{\mu}\boldsymbol{\mu}', \quad (\mathbf{B}-83)$$

which is the **covariance matrix** of the random vector  $\mathbf{x}$ . Henceforth, we shall denote the covariance matrix of a random vector in boldface, as in

$$\text{Var}[\mathbf{x}] = \boldsymbol{\Sigma}.$$

By dividing  $\sigma_{ij}$  by  $\sigma_i \sigma_j$ , we obtain the **correlation matrix**:

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_{n1} & \rho_{n2} & \rho_{n3} & \cdots & 1 \end{bmatrix}.$$

### B.10.2 SETS OF LINEAR FUNCTIONS

Our earlier results for the mean and variance of a linear function can be extended to the multivariate case. For the mean,

$$\begin{aligned} E[a_1x_1 + a_2x_2 + \cdots + a_nx_n] &= E[\mathbf{a}'\mathbf{x}] \\ &= a_1E[x_1] + a_2E[x_2] + \cdots + a_nE[x_n] \\ &= a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n \\ &= \mathbf{a}'\boldsymbol{\mu}. \end{aligned} \quad (\mathbf{B}-84)$$

For the variance,

$$\begin{aligned} \text{Var}[\mathbf{a}'\mathbf{x}] &= E[(\mathbf{a}'\mathbf{x} - E[\mathbf{a}'\mathbf{x}])^2] \\ &= E[\{\mathbf{a}'(\mathbf{x} - E[\mathbf{x}])\}^2] \\ &= E[\mathbf{a}'(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\mathbf{a}] \end{aligned}$$

as  $E[\mathbf{x}] = \boldsymbol{\mu}$  and  $\mathbf{a}'(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})'\mathbf{a}$ . Because  $\mathbf{a}$  is a vector of constants,

$$\text{Var}[\mathbf{a}'\mathbf{x}] = \mathbf{a}'E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})']\mathbf{a} = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \sigma_{ij}. \quad (\mathbf{B}-85)$$

## APPENDIX B ♦ Probability and Distribution Theory 1109

It is the expected value of a square, so we know that a variance cannot be negative. As such, the preceding quadratic form is nonnegative, and the symmetric matrix  $\Sigma$  must be nonnegative definite.

In the set of linear functions  $\mathbf{y} = \mathbf{Ax}$ , the  $i$ th element of  $\mathbf{y}$  is  $y_i = \mathbf{a}_i \mathbf{x}$ , where  $\mathbf{a}_i$  is the  $i$ th row of  $\mathbf{A}$  [see result (A-14)]. Therefore,

$$E[y_i] = \mathbf{a}_i \boldsymbol{\mu}.$$

Collecting the results in a vector, we have

$$E[\mathbf{Ax}] = \mathbf{A}\boldsymbol{\mu}. \quad (\text{B-86})$$

For two row vectors  $\mathbf{a}_i$  and  $\mathbf{a}_j$ ,

$$\text{Cov}[\mathbf{a}_i \mathbf{x}, \mathbf{a}_j \mathbf{x}] = \mathbf{a}_i \Sigma \mathbf{a}'_j.$$

Because  $\mathbf{a}_i \Sigma \mathbf{a}'_j$  is the  $ij$ th element of  $\mathbf{A}\Sigma\mathbf{A}'$ ,

$$\text{Var}[\mathbf{Ax}] = \mathbf{A}\Sigma\mathbf{A}'. \quad (\text{B-87})$$

This matrix will be either nonnegative definite or positive definite, depending on the column rank of  $\mathbf{A}$ .

### B.10.3 NONLINEAR FUNCTIONS

Consider a set of possibly nonlinear functions of  $\mathbf{x}$ ,  $\mathbf{y} = \mathbf{g}(\mathbf{x})$ . Each element of  $\mathbf{y}$  can be approximated with a linear Taylor series. Let  $\mathbf{j}^i$  be the row vector of partial derivatives of the  $i$ th function with respect to the  $n$  elements of  $\mathbf{x}$ :

$$\mathbf{j}^i(\mathbf{x}) = \frac{\partial g_i(\mathbf{x})}{\partial \mathbf{x}'} = \frac{\partial y_i}{\partial \mathbf{x}'} . \quad (\text{B-88})$$

Then, proceeding in the now familiar way, we use  $\boldsymbol{\mu}$ , the mean vector of  $\mathbf{x}$ , as the expansion point, so that  $\mathbf{j}^i(\boldsymbol{\mu})$  is the row vector of partial derivatives evaluated at  $\boldsymbol{\mu}$ . Then

$$g_i(\mathbf{x}) \approx g_i(\boldsymbol{\mu}) + \mathbf{j}^i(\boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu}). \quad (\text{B-89})$$

From this we obtain

$$E[g_i(\mathbf{x})] \approx g_i(\boldsymbol{\mu}), \quad (\text{B-90})$$

$$\text{Var}[g_i(\mathbf{x})] \approx \mathbf{j}^i(\boldsymbol{\mu}) \Sigma \mathbf{j}^i(\boldsymbol{\mu})', \quad (\text{B-91})$$

and

$$\text{Cov}[g_i(\mathbf{x}), g_j(\mathbf{x})] \approx \mathbf{j}^i(\boldsymbol{\mu}) \Sigma \mathbf{j}^j(\boldsymbol{\mu})'. \quad (\text{B-92})$$

These results can be collected in a convenient form by arranging the row vectors  $\mathbf{j}^i(\boldsymbol{\mu})$  in a matrix  $\mathbf{J}(\boldsymbol{\mu})$ . Then, corresponding to the preceding equations, we have

$$E[\mathbf{g}(\mathbf{x})] \simeq \mathbf{g}(\boldsymbol{\mu}), \quad (\text{B-93})$$

$$\text{Var}[\mathbf{g}(\mathbf{x})] \simeq \mathbf{J}(\boldsymbol{\mu}) \Sigma \mathbf{J}(\boldsymbol{\mu})'. \quad (\text{B-94})$$

The matrix  $\mathbf{J}(\boldsymbol{\mu})$  in the last preceding line is  $\partial \mathbf{y} / \partial \mathbf{x}'$  evaluated at  $\mathbf{x} = \boldsymbol{\mu}$ .

## 1110 PART VI ♦ Appendices

### B.11 THE MULTIVARIATE NORMAL DISTRIBUTION

The foundation of most multivariate analysis in econometrics is the multivariate normal distribution. Let the vector  $(x_1, x_2, \dots, x_n)' = \mathbf{x}$  be the set of  $n$  random variables,  $\boldsymbol{\mu}$  their mean vector, and  $\boldsymbol{\Sigma}$  their covariance matrix. The general form of the joint density is

$$f(\mathbf{x}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} e^{(-1/2)(\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}. \quad (\text{B-95})$$

If  $\mathbf{R}$  is the correlation matrix of the variables and  $\mathbf{R}_{ij} = \sigma_{ij}/(\sigma_i \sigma_j)$ , then

$$f(\mathbf{x}) = (2\pi)^{-n/2} (\sigma_1 \sigma_2 \cdots \sigma_n)^{-1} |\mathbf{R}|^{-1/2} e^{(-1/2)\boldsymbol{\varepsilon}' \mathbf{R}^{-1} \boldsymbol{\varepsilon}}, \quad (\text{B-96})$$

where  $\boldsymbol{\varepsilon}_i = (x_i - \mu_i)/\sigma_i$ .<sup>8</sup>

Two special cases are of interest. If all the variables are uncorrelated, then  $\rho_{ij} = 0$  for  $i \neq j$ . Thus,  $\mathbf{R} = \mathbf{I}$ , and the density becomes

$$\begin{aligned} f(\mathbf{x}) &= (2\pi)^{-n/2} (\sigma_1 \sigma_2 \cdots \sigma_n)^{-1} e^{-\boldsymbol{\varepsilon}' \boldsymbol{\varepsilon}/2} \\ &= f(x_1) f(x_2) \cdots f(x_n) = \prod_{i=1}^n f(x_i). \end{aligned} \quad (\text{B-97})$$

As in the bivariate case, if normally distributed variables are uncorrelated, then they are independent. If  $\sigma_i = \sigma$  and  $\boldsymbol{\mu} = \mathbf{0}$ , then  $x_i \sim N[0, \sigma^2]$  and  $\boldsymbol{\varepsilon}_i = x_i/\sigma$ , and the density becomes

$$f(\mathbf{x}) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} e^{-\mathbf{x}' \mathbf{x}/(2\sigma^2)}. \quad (\text{B-98})$$

Finally, if  $\sigma = 1$ ,

$$f(\mathbf{x}) = (2\pi)^{-n/2} e^{-\mathbf{x}' \mathbf{x}/2}. \quad (\text{B-99})$$

This distribution is the **multivariate standard normal**, or **spherical normal distribution**.

#### B.11.1 MARGINAL AND CONDITIONAL NORMAL DISTRIBUTIONS

Let  $\mathbf{x}_1$  be any subset of the variables, including a single variable, and let  $\mathbf{x}_2$  be the remaining variables. Partition  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  likewise so that

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Then the marginal distributions are also normal. In particular, we have the following theorem.

#### THEOREM B.7 Marginal and Conditional Normal Distributions

If  $[\mathbf{x}_1, \mathbf{x}_2]$  have a joint multivariate normal distribution, then the marginal distributions are

$$\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}), \quad (\text{B-100})$$

<sup>8</sup>This result is obtained by constructing  $\Delta$ , the diagonal matrix with  $\sigma_i$  as its  $i$ th diagonal element. Then,  $\mathbf{R} = \Delta^{-1} \boldsymbol{\Sigma} \Delta^{-1}$ , which implies that  $\boldsymbol{\Sigma}^{-1} = \Delta^{-1} \mathbf{R}^{-1} \Delta^{-1}$ . Inserting this in (B-95) yields (B-96). Note that the  $i$ th element of  $\Delta^{-1}(\mathbf{x} - \boldsymbol{\mu})$  is  $(x_i - \mu_i)/\sigma_i$ .

## APPENDIX B ♦ Probability and Distribution Theory 1111

**THEOREM B.7 (Continued)**

and

$$\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}). \quad (\text{B-101})$$

The conditional distribution of  $\mathbf{x}_1$  given  $\mathbf{x}_2$  is normal as well:

$$\mathbf{x}_1 | \mathbf{x}_2 \sim N(\boldsymbol{\mu}_{1.2}, \boldsymbol{\Sigma}_{11.2}), \quad (\text{B-102})$$

where

$$\boldsymbol{\mu}_{1.2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2), \quad (\text{B-102a})$$

$$\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}. \quad (\text{B-102b})$$

**Proof:** We partition  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  as shown earlier and insert the parts in (B-95). To construct the density, we use (A-72) to partition the determinant,

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{22}| |\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}|,$$

and (A-74) to partition the inverse,

$$\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{11.2}^{-1} & -\boldsymbol{\Sigma}_{11.2}^{-1} \mathbf{B} \\ -\mathbf{B}' \boldsymbol{\Sigma}_{11.2}^{-1} & \boldsymbol{\Sigma}_{22}^{-1} + \mathbf{B}' \boldsymbol{\Sigma}_{11.2}^{-1} \mathbf{B} \end{bmatrix}.$$

For simplicity, we let

$$\mathbf{B} = \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1}.$$

Inserting these in (B-95) and collecting terms produces the joint density as a product of two terms:

$$f(\mathbf{x}_1, \mathbf{x}_2) = f_{1.2}(\mathbf{x}_1 | \mathbf{x}_2) f_2(\mathbf{x}_2).$$

The first of these is a normal distribution with mean  $\boldsymbol{\mu}_{1.2}$  and variance  $\boldsymbol{\Sigma}_{11.2}$ , whereas the second is the marginal distribution of  $\mathbf{x}_2$ .

The conditional mean vector in the multivariate normal distribution is a linear function of the unconditional mean and the conditioning variables, and the conditional covariance matrix is constant and is smaller (in the sense discussed in Section A.7.3) than the unconditional covariance matrix. Notice that the conditional covariance matrix is the inverse of the upper left block of  $\boldsymbol{\Sigma}^{-1}$ ; that is, this matrix is of the form shown in (A-74) for the partitioned inverse of a matrix.

### B.11.2 THE CLASSICAL NORMAL LINEAR REGRESSION MODEL

An important special case of the preceding is that in which  $\mathbf{x}_1$  is a single variable,  $y$ , and  $\mathbf{x}_2$  is  $K$  variables,  $\mathbf{x}$ . Then the conditional distribution is a multivariate version of that in (B-80) with  $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\sigma}_{xy}$ , where  $\boldsymbol{\sigma}_{xy}$  is the vector of covariances of  $y$  with  $\mathbf{x}_2$ . Recall that any random variable,  $y$ , can be written as its mean plus the deviation from the mean. If we apply this tautology to the multivariate normal, we obtain

$$y = E[y | \mathbf{x}] + (y - E[y | \mathbf{x}]) = \alpha + \boldsymbol{\beta}' \mathbf{x} + \varepsilon,$$

## 1112 PART VI ♦ Appendices

where  $\beta$  is given earlier,  $\alpha = \mu_y - \beta' \mu_x$ , and  $\varepsilon$  has a normal distribution. We thus have, in this multivariate normal distribution, the **classical normal linear regression model**.

### B.11.3 LINEAR FUNCTIONS OF A NORMAL VECTOR

Any linear function of a vector of joint normally distributed variables is also normally distributed. The mean vector and covariance matrix of  $\mathbf{Ax}$ , where  $\mathbf{x}$  is normally distributed, follow the general pattern given earlier. Thus,

$$\text{If } \mathbf{x} \sim N[\boldsymbol{\mu}, \boldsymbol{\Sigma}], \text{ then } \mathbf{Ax} + \mathbf{b} \sim N[\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}']. \quad (\text{B-103})$$

If  $\mathbf{A}$  does not have full rank, then  $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$  is singular and the density does not exist in the full dimensional space of  $\mathbf{x}$  although it does exist in the subspace of dimension equal to the rank of  $\boldsymbol{\Sigma}$ . Nonetheless, the individual elements of  $\mathbf{Ax} + \mathbf{b}$  will still be normally distributed, and the joint distribution of the full vector is still a multivariate normal.

### B.11.4 QUADRATIC FORMS IN A STANDARD NORMAL VECTOR

The earlier discussion of the chi-squared distribution gives the distribution of  $\mathbf{x}'\mathbf{x}$  if  $\mathbf{x}$  has a standard normal distribution. It follows from (A-36) that

$$\mathbf{x}'\mathbf{x} = \sum_{i=1}^n x_i^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n\bar{x}^2. \quad (\text{B-104})$$

We know from (B-32) that  $\mathbf{x}'\mathbf{x}$  has a chi-squared distribution. It seems natural, therefore, to invoke (B-34) for the two parts on the right-hand side of (B-104). It is not yet obvious, however, that either of the two terms has a chi-squared distribution or that the two terms are independent, as required. To show these conditions, it is necessary to derive the distributions of **idempotent quadratic forms** and to show when they are independent.

To begin, the second term is the square of  $\sqrt{n}\bar{x}$ , which can easily be shown to have a standard normal distribution. Thus, the second term is the square of a standard normal variable and has chi-squared distribution with one degree of freedom. But the first term is the sum of  $n$  nonindependent variables, and it remains to be shown that the two terms are independent.

#### DEFINITION B.3 Orthonormal Quadratic Form

*A particular case of (B-103) is the following:*

*If  $\mathbf{x} \sim N[\mathbf{0}, \mathbf{I}]$  and  $\mathbf{C}$  is a square matrix such that  $\mathbf{C}'\mathbf{C} = \mathbf{I}$ , then  $\mathbf{C}'\mathbf{x} \sim N[\mathbf{0}, \mathbf{I}]$ .*

Consider, then, a quadratic form in a standard normal vector  $\mathbf{x}$  with symmetric matrix  $\mathbf{A}$ :

$$q = \mathbf{x}'\mathbf{A}\mathbf{x}. \quad (\text{B-105})$$

Let the characteristic roots and vectors of  $\mathbf{A}$  be arranged in a diagonal matrix  $\Lambda$  and an orthogonal matrix  $\mathbf{C}$ , as in Section A.6.3. Then

$$q = \mathbf{x}'\mathbf{C}\Lambda\mathbf{C}'\mathbf{x}. \quad (\text{B-106})$$

## APPENDIX B ♦ Probability and Distribution Theory 1113

By definition,  $\mathbf{C}$  satisfies the requirement that  $\mathbf{C}'\mathbf{C} = \mathbf{I}$ . Thus, the vector  $\mathbf{y} = \mathbf{C}'\mathbf{x}$  has a standard normal distribution. Consequently,

$$q = \mathbf{y}'\Lambda\mathbf{y} = \sum_{i=1}^n \lambda_i y_i^2. \quad (\mathbf{B}-107)$$

If  $\lambda_i$  is always one or zero, then

$$q = \sum_{j=1}^J y_j^2, \quad (\mathbf{B}-108)$$

which has a chi-squared distribution. The sum is taken over the  $j = 1, \dots, J$  elements associated with the roots that are equal to one. A matrix whose characteristic roots are all zero or one is idempotent. Therefore, we have proved the next theorem.

### **THEOREM B.8 Distribution of an Idempotent Quadratic Form in a Standard Normal Vector**

*If  $\mathbf{x} \sim N[\mathbf{0}, \mathbf{I}]$  and  $\mathbf{A}$  is idempotent, then  $\mathbf{x}'\mathbf{A}\mathbf{x}$  has a chi-squared distribution with degrees of freedom equal to the number of unit roots of  $\mathbf{A}$ , which is equal to the rank of  $\mathbf{A}$ .*

The rank of a matrix is equal to the number of nonzero characteristic roots it has. Therefore, the degrees of freedom in the preceding chi-squared distribution equals  $J$ , the rank of  $\mathbf{A}$ .

We can apply this result to the earlier sum of squares. The first term is

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \mathbf{x}'\mathbf{M}^0\mathbf{x},$$

where  $\mathbf{M}^0$  was defined in (A-34) as the matrix that transforms data to mean deviation form:

$$\mathbf{M}^0 = \mathbf{I} - \frac{1}{n}\mathbf{i}\mathbf{i}'.$$

Because  $\mathbf{M}^0$  is idempotent, the sum of squared deviations from the mean has a chi-squared distribution. The degrees of freedom equals the rank  $\mathbf{M}^0$ , which is not obvious except for the useful result in (A-108), that

- The rank of an idempotent matrix is equal to its trace. (B-109)

Each diagonal element of  $\mathbf{M}^0$  is  $1 - (1/n)$ ; hence, the trace is  $n[1 - (1/n)] = n - 1$ . Therefore, we have an application of Theorem B.8.

- If  $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$ ,  $\sum_{i=1}^n (x_i - \bar{x})^2 \sim \chi^2[n - 1]$ . (B-110)

We have already shown that the second term in (B-104) has a chi-squared distribution with one degree of freedom. It is instructive to set this up as a quadratic form as well:

$$n\bar{x}^2 = \mathbf{x}' \left[ \frac{1}{n} \mathbf{i} \mathbf{i}' \right] \mathbf{x} = \mathbf{x}' [\mathbf{j} \mathbf{j}'] \mathbf{x}, \quad \text{where } \mathbf{j} = \left( \frac{1}{\sqrt{n}} \right) \mathbf{i}. \quad (\mathbf{B}-111)$$

The matrix in brackets is the outer product of a nonzero vector, which always has rank one. You can verify that it is idempotent by multiplication. Thus,  $\mathbf{x}'\mathbf{x}$  is the sum of two chi-squared variables,

## 1114 PART VI ♦ Appendices

one with  $n - 1$  degrees of freedom and the other with one. It is now necessary to show that the two terms are independent. To do so, we will use the next theorem.

### THEOREM B.9 Independence of Idempotent Quadratic Forms

If  $\mathbf{x} \sim N[\mathbf{0}, \mathbf{I}]$  and  $\mathbf{x}'\mathbf{A}\mathbf{x}$  and  $\mathbf{x}'\mathbf{B}\mathbf{x}$  are two idempotent quadratic forms in  $\mathbf{x}$ , then  $\mathbf{x}'\mathbf{A}\mathbf{x}$  and  $\mathbf{x}'\mathbf{B}\mathbf{x}$  are independent if  $\mathbf{AB} = \mathbf{0}$ . (B-112)

As before, we show the result for the general case and then specialize it for the example. Because both  $\mathbf{A}$  and  $\mathbf{B}$  are symmetric and idempotent,  $\mathbf{A} = \mathbf{A}'\mathbf{A}$  and  $\mathbf{B} = \mathbf{B}'\mathbf{B}$ . The quadratic forms are therefore

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{x} = \mathbf{x}'_1\mathbf{x}_1, \quad \text{where } \mathbf{x}_1 = \mathbf{Ax}, \quad \text{and} \quad \mathbf{x}'\mathbf{B}\mathbf{x} = \mathbf{x}'_2\mathbf{x}_2, \quad \text{where } \mathbf{x}_2 = \mathbf{Bx}. \quad (\text{B-113})$$

Both vectors have zero mean vectors, so the covariance matrix of  $\mathbf{x}_1$  and  $\mathbf{x}_2$  is

$$E(\mathbf{x}_1\mathbf{x}'_2) = \mathbf{A}\mathbf{I}\mathbf{B}' = \mathbf{AB} = \mathbf{0}.$$

Because  $\mathbf{Ax}$  and  $\mathbf{Bx}$  are linear functions of a normally distributed random vector, they are, in turn, normally distributed. Their zero covariance matrix implies that they are statistically independent,<sup>9</sup> which establishes the independence of the two quadratic forms. For the case of  $\mathbf{x}'\mathbf{x}$ , the two matrices are  $\mathbf{M}^0$  and  $[\mathbf{I} - \mathbf{M}^0]$ . You can show that  $\mathbf{M}^0[\mathbf{I} - \mathbf{M}^0] = \mathbf{0}$  just by multiplying it out.

### B.11.5 THE F DISTRIBUTION

The normal family of distributions (chi-squared,  $F$ , and  $t$ ) can all be derived as functions of idempotent quadratic forms in a standard normal vector. The  $F$  distribution is the ratio of two independent chi-squared variables, each divided by its respective degrees of freedom. Let  $\mathbf{A}$  and  $\mathbf{B}$  be two idempotent matrices with ranks  $r_a$  and  $r_b$ , and let  $\mathbf{AB} = \mathbf{0}$ . Then

$$\frac{\mathbf{x}'\mathbf{Ax}/r_a}{\mathbf{x}'\mathbf{Bx}/r_b} \sim F[r_a, r_b]. \quad (\text{B-114})$$

If  $\text{Var}[\mathbf{x}] = \sigma^2\mathbf{I}$  instead, then this is modified to

$$\frac{(\mathbf{x}'\mathbf{Ax}/\sigma^2)/r_a}{(\mathbf{x}'\mathbf{Bx}/\sigma^2)/r_b} \sim F[r_a, r_b]. \quad (\text{B-115})$$

### B.11.6 A FULL RANK QUADRATIC FORM

Finally, consider the general case,

$$\mathbf{x} \sim N[\boldsymbol{\mu}, \boldsymbol{\Sigma}].$$

We are interested in the distribution of

$$q = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \quad (\text{B-116})$$

---

<sup>9</sup>Note that both  $\mathbf{x}_1 = \mathbf{Ax}$  and  $\mathbf{x}_2 = \mathbf{Bx}$  have singular covariance matrices. Nonetheless, every element of  $\mathbf{x}_1$  is independent of every element  $\mathbf{x}_2$ , so the vectors are independent.

**APPENDIX B ♦ Probability and Distribution Theory 1115**

First, the vector can be written as  $\mathbf{z} = \mathbf{x} - \boldsymbol{\mu}$ , and  $\Sigma$  is the covariance matrix of  $\mathbf{z}$  as well as of  $\mathbf{x}$ . Therefore, we seek the distribution of

$$q = \mathbf{z}'\Sigma^{-1}\mathbf{z} = \mathbf{z}'(\text{Var}[\mathbf{z}])^{-1}\mathbf{z}, \quad (\mathbf{B}-117)$$

where  $\mathbf{z}$  is normally distributed with mean  $\mathbf{0}$ . This equation is a quadratic form, but not necessarily in an idempotent matrix.<sup>10</sup> Because  $\Sigma$  is positive definite, it has a square root. Define the symmetric matrix  $\Sigma^{1/2}$  so that  $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$ . Then

$$\Sigma^{-1} = \Sigma^{-1/2}\Sigma^{-1/2},$$

and

$$\begin{aligned} \mathbf{z}'\Sigma^{-1}\mathbf{z} &= \mathbf{z}'\Sigma^{-1/2}\Sigma^{-1/2}\mathbf{z} \\ &= (\Sigma^{-1/2}\mathbf{z})'(\Sigma^{-1/2}\mathbf{z}) \\ &= \mathbf{w}'\mathbf{w}. \end{aligned}$$

Now  $\mathbf{w} = \mathbf{Az}$ , so

$$E(\mathbf{w}) = \mathbf{A}E[\mathbf{z}] = \mathbf{0},$$

and

$$\text{Var}[\mathbf{w}] = \mathbf{A}\Sigma\mathbf{A}' = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = \Sigma^0 = \mathbf{I}.$$

This provides the following important result:

**THEOREM B.10 Distribution of a Standardized Normal Vector**

*If  $\mathbf{x} \sim N[\boldsymbol{\mu}, \Sigma]$ , then  $\Sigma^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \sim N[\mathbf{0}, \mathbf{I}]$ .*

The simplest special case is that in which  $\mathbf{x}$  has only one variable, so that the transformation is just  $(x - \mu)/\sigma$ . Combining this case with (B-32) concerning the sum of squares of standard normals, we have the following theorem.

**THEOREM B.11 Distribution of  $\mathbf{x}'\Sigma^{-1}\mathbf{x}$  When  $\mathbf{x}$  Is Normal**

*If  $\mathbf{x} \sim N[\boldsymbol{\mu}, \Sigma]$ , then  $(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2[n]$ .*

**B.11.7 INDEPENDENCE OF A LINEAR AND A QUADRATIC FORM**

The  $t$  distribution is used in many forms of hypothesis tests. In some situations, it arises as the ratio of a linear to a quadratic form in a normal vector. To establish the distribution of these statistics, we use the following result.

<sup>10</sup>It will be idempotent only in the special case of  $\Sigma = \mathbf{I}$ .

**1116 PART VI ♦ Appendices**
**THEOREM B.12 Independence of a Linear and a Quadratic Form**

A linear function  $\mathbf{L}\mathbf{x}$  and a symmetric idempotent quadratic form  $\mathbf{x}'\mathbf{A}\mathbf{x}$  in a standard normal vector are statistically independent if  $\mathbf{LA} = \mathbf{0}$ .

The proof follows the same logic as that for two quadratic forms. Write  $\mathbf{x}'\mathbf{A}\mathbf{x}$  as  $\mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{x} = (\mathbf{A}\mathbf{x})'(\mathbf{A}\mathbf{x})$ . The covariance matrix of the variables  $\mathbf{L}\mathbf{x}$  and  $\mathbf{A}\mathbf{x}$  is  $\mathbf{LA} = \mathbf{0}$ , which establishes the independence of these two random vectors. The independence of the linear function and the quadratic form follows because functions of independent random vectors are also independent.

The  $t$  distribution is defined as the ratio of a standard normal variable to the square root of a chi-squared variable divided by its degrees of freedom:

$$t[J] = \frac{N[0, 1]}{\{\chi^2[J]/J\}^{1/2}}.$$

A particular case is

$$t[n-1] = \frac{\sqrt{n}\bar{x}}{\left\{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2\right\}^{1/2}} = \frac{\sqrt{n}\bar{x}}{s},$$

where  $s$  is the standard deviation of the values of  $\mathbf{x}$ . The distribution of the two variables in  $t[n-1]$  was shown earlier; we need only show that they are independent. But

$$\sqrt{n}\bar{x} = \frac{1}{\sqrt{n}} \mathbf{i}'\mathbf{x} = \mathbf{j}'\mathbf{x},$$

and

$$s^2 = \frac{\mathbf{x}'\mathbf{M}^0\mathbf{x}}{n-1}.$$

It suffices to show that  $\mathbf{M}^0\mathbf{j} = \mathbf{0}$ , which follows from

$$\mathbf{M}^0\mathbf{i} = [\mathbf{I} - \mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}']\mathbf{i} = \mathbf{i} - \mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}(\mathbf{i}'\mathbf{i}) = \mathbf{0}.$$

---

**APPENDIX C**


## ESTIMATION AND INFERENCE

### C.1 INTRODUCTION

The probability distributions discussed in Appendix B serve as models for the underlying data generating processes that produce our observed data. The goal of statistical inference in econometrics is to use the principles of mathematical statistics to combine these theoretical distributions and the observed data into an empirical model of the economy. This analysis takes place in one of two frameworks, classical or Bayesian. The overwhelming majority of empirical study in

## APPENDIX C ♦ Estimation and Inference 1117

econometrics has been done in the classical framework. Our focus,  before, will be on classical methods of inference. Bayesian methods are discussed in Chapter 18.<sup>1</sup>

## C.2 SAMPLES AND RANDOM SAMPLING

The classical theory of statistical inference centers on rules for using the sampled data effectively. These rules, in turn, are based on the properties of samples and sampling distributions.

A sample of  $n$  observations on one or more variables, denoted  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ , is a **random sample** if the  $n$  observations are drawn independently from the same population, or probability distribution,  $f(\mathbf{x}_i, \theta)$ . The sample may be univariate if  $\mathbf{x}_i$  is a single random variable or multivariate if each observation contains several variables. A random sample of observations, denoted  $[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$  or  $\{\mathbf{x}_i\}_{i=1,\dots,n}$ , is said to be **independent, identically distributed**, which we denote *i.i.d.* The vector  $\theta$  contains one or more unknown parameters. Data are generally drawn in one of two settings. A **cross section** is a sample of a number of observational units all drawn at the same point in time. A **time series** is a set of observations drawn on the same observational unit at a number of (usually evenly spaced) points in time. Many recent studies have been based on time-series cross sections, which generally consist of the same cross-sectional units observed at several points in time. Because the typical data set of this sort consists of a large number of cross-sectional units observed at a few points in time, the common term **panel data set** is usually more fitting for this sort of study.

## C.3 DESCRIPTIVE STATISTICS

Before attempting to estimate parameters of a population or fit models to data, we normally examine the data themselves. In raw form, the sample data are a disorganized mass of information, so we will need some organizing principles to distill the information into something meaningful. Consider, first, examining the data on a single variable. In most cases, and particularly if the number of observations in the sample is large, we shall use some summary **statistics** to describe the sample data. Of most interest are measures of **location**—that is, the center of the data—and **scale**, or the dispersion of the data. A few measures of central tendency are as follows:

$$\begin{aligned} \text{mean: } \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i, \\ \text{median: } M &= \text{middle ranked observation,} \\ \text{sample midrange: } \text{midrange} &= \frac{\text{maximum} + \text{minimum}}{2}. \end{aligned} \tag{C-1}$$

The dispersion of the sample observations is usually measured by the

$$\text{standard deviation: } s_x = \left[ \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \right]^{1/2}. \tag{C-2}$$

Other measures, such as the average absolute deviation from the sample mean, are also used, although less frequently than the standard deviation. The shape of the distribution of values is often of interest as well. Samples of income or expenditure data, for example, tend to be highly

---

<sup>1</sup>An excellent reference is Leamer (1978). A summary of the results as they apply to econometrics is contained in Zellner (1971) and in Judge et al. (1985). See, as well, Poirier (1991, 1995). Recent textbooks on Bayesian econometrics include Koop (2003), Lancaster (2004) and Geweke (2005).

## 1118 PART VI ♦ Appendices

skewed while financial data such as asset returns and exchange rate movements are relatively more symmetrically distributed but are also more widely dispersed than other variables that might be observed. Two measures used to quantify these effects are the

$$\text{skewness} = \left[ \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{s_x^3 (n-1)} \right], \quad \text{and} \quad \text{kurtosis} = \left[ \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{s_x^4 (n-1)} \right].$$

(Benchmark values for these two measures are zero for a symmetric distribution, and three for one which is “normally” dispersed.) The skewness coefficient has a bit less of the intuitive appeal of the mean and standard deviation, and the kurtosis measure has very little at all. The box and whisker plot is a graphical device which is often used to capture a large amount of information about the sample in a simple visual display. This plot shows in a figure the median, the range of values contained in the 25th and 75th percentile, some limits that show the normal range of values expected, such as the median plus and minus two standard deviations, and in isolation values that could be viewed as outliers. A box and whisker plot is shown in Figure C.1 for the income variable in Example C.1.

If the sample contains data on more than one variable, we will also be interested in measures of association among the variables. A **scatter diagram** is useful in a bivariate sample if the sample contains a reasonable number of observations. Figure C.1 shows an example for a small data set. If the sample is a multivariate one, then the degree of linear association among the variables can be measured by the pairwise measures

$$\begin{aligned} \text{covariance: } s_{xy} &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}, \\ \text{correlation: } r_{xy} &= \frac{s_{xy}}{s_x s_y}. \end{aligned} \tag{C-3}$$

If the sample contains data on several variables, then it is sometimes convenient to arrange the covariances or correlations in a

$$\text{covariance matrix: } \mathbf{S} = [s_{ij}], \tag{C-4}$$

or

$$\text{correlation matrix: } \mathbf{R} = [r_{ij}].$$

Some useful algebraic results for any two variables  $(x_i, y_i), i = 1, \dots, n$ , and constants  $a$  and  $b$  are

$$s_x^2 = \frac{(\sum_{i=1}^n x_i^2) - n\bar{x}^2}{n-1}, \tag{C-5}$$

$$s_{xy} = \frac{(\sum_{i=1}^n x_i y_i) - n\bar{x}\bar{y}}{n-1}, \tag{C-6}$$

$$-1 \leq r_{xy} \leq 1,$$

$$r_{ax,by} = \frac{ab}{|ab|} r_{xy}, \quad a,b \neq 0, \tag{C-7}$$

$$s_{ax} = |a|s_x,$$

$$s_{ax,by} = (ab)s_{xy}. \tag{C-8}$$

**APPENDIX C ♦ Estimation and Inference 1119**

Note that these algebraic results parallel the theoretical results for bivariate probability distributions. [We note in passing, while the formulas in (C-2) and (C-5) are algebraically the same, (C-2) will generally be more accurate in practice, especially when the values in the sample are very widely dispersed.]

**Example C.1 Descriptive Statistics for a Random Sample**

Appendix Table FC.1 contains a (hypothetical) sample of observations on income and education. (The observations all appear in the calculations of the means below.) A scatter diagram appears in Figure C.1. It suggests a weak positive association between income and education in these data. The box and whisker plot for income at the left of the scatter plot shows the distribution of the income data as well.

$$\text{Means: } \bar{I} = \frac{1}{20} \left[ 20.5 + 31.5 + 47.7 + 26.2 + 44.0 + 8.28 + 30.8 + 17.2 + 19.9 + 9.96 + 55.8 + 25.2 + 29.0 + 85.5 + 15.1 + 28.5 + 21.4 + 17.7 + 6.42 + 84.9 \right] = 31.278,$$

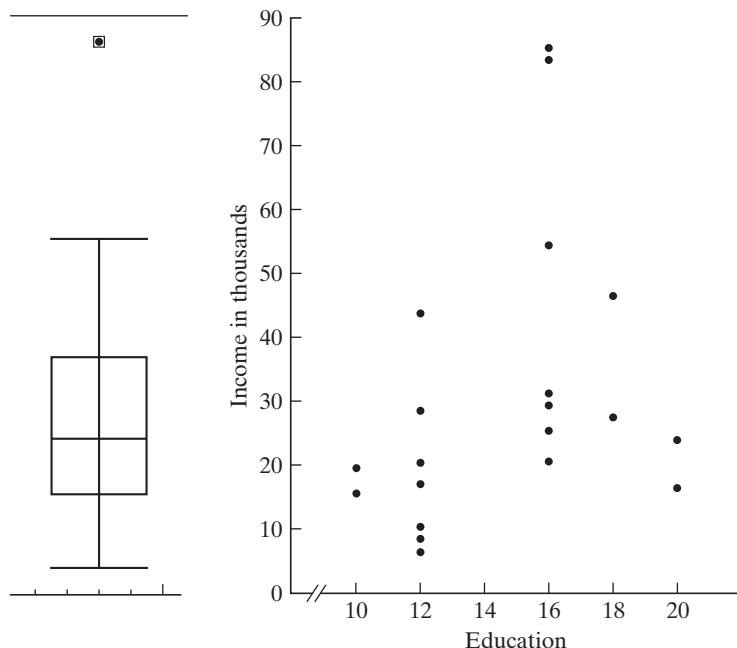
$$\bar{E} = \frac{1}{20} \left[ 12 + 16 + 18 + 16 + 12 + 12 + 16 + 12 + 10 + 12 + 16 + 20 + 12 + 16 + 10 + 18 + 16 + 20 + 12 + 16 \right] = 14.600.$$

*Standard deviations:*

$$s_I = \sqrt{\frac{1}{19} [(20.5 - 31.278)^2 + \dots + (84.9 - 31.278)^2]} = 22.376,$$

$$s_E = \sqrt{\frac{1}{19} [(12 - 14.6)^2 + \dots + (16 - 14.6)^2]} = 3.119.$$

**FIGURE C.1** Box and Whisker Plot for Income and Scatter Diagram for Income and Education.



## 1120 PART VI ♦ Appendices

Covariance:  $s_{IE} = \frac{1}{19}[20.5(12) + \dots + 84.9(16) - 20(31.28)(14.6)] = 23.597$ ,

Correlation:  $r_{IE} = \frac{23.597}{(22.376)(3.119)} = 0.3382$ .

The positive correlation is consistent with our observation in the scatter diagram.

The statistics just described will provide the analyst with a more concise description of the data than a raw tabulation. However, we have not, as yet, suggested that these measures correspond to some underlying characteristic of the process that generated the data. We do assume that there is an underlying mechanism, the data generating process, that produces the data in hand. Thus, these serve to do more than describe the data; they characterize that process, or population. Because we have assumed that there is an underlying probability distribution, it might be useful to produce a statistic that gives a broader view of the DGP. The **histogram** is a simple graphical device that produces this result—see Examples C.3 and C.4 for applications. For small samples or widely dispersed data, however, histograms tend to be rough and difficult to make informative. A burgeoning literature [see, e.g., Pagan and Ullah (1999) and Li and Racine (2007)] has demonstrated the usefulness of the **kernel density estimator** as a substitute for the histogram as a descriptive tool for the underlying distribution that produced a sample of data. The underlying theory of the kernel density estimator is fairly complicated, but the computations are surprisingly simple. The estimator is computed using

$$\hat{f}(x^*) = \frac{1}{nh} \sum_{i=1}^n K\left[\frac{x_i - x^*}{h}\right],$$

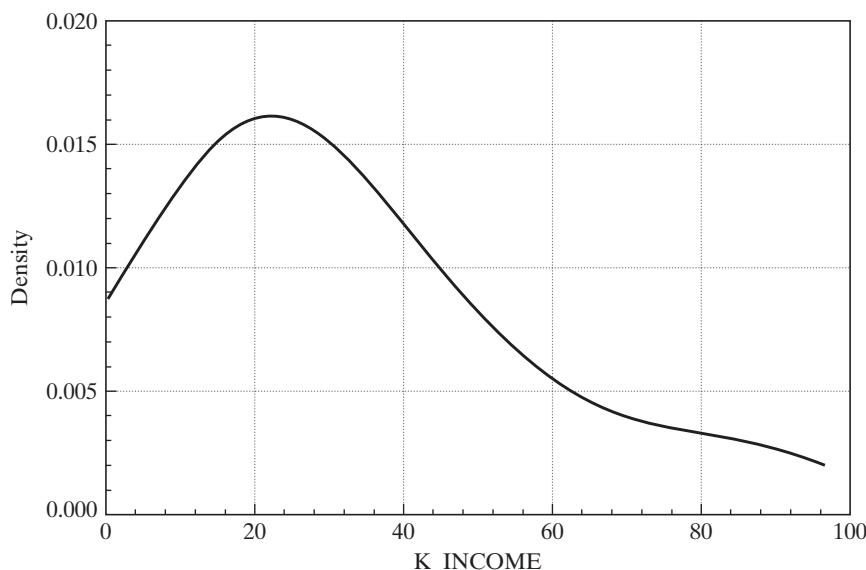
where  $x_1, \dots, x_n$  are the  $n$  observations in the sample,  $\hat{f}(x^*)$  denotes the estimated density function,  $x^*$  is the value at which we wish to evaluate the density, and  $h$  and  $K[\cdot]$  are the “bandwidth” and “kernel function” that we now consider. The density estimator is rather like a histogram, in which the bandwidth is the width of the intervals. The kernel function is a weight function which is generally chosen so that it takes large values when  $x^*$  is close to  $x_i$  and tapers off to zero in as they diverge in either direction. The weighting function used in the following example is the logistic density discussed in Section B.4.7. The bandwidth is chosen to be a function of  $1/n$  so that the intervals can become narrower as the sample becomes larger (and richer). The one used for Figure C.2 is  $h = 0.9\text{Min}(s, \text{range}/3)/n^{2/5}$ . (We will revisit this method of estimation in Chapter 12.) Example C.2 illustrates the computation for the income data used in Example C.1.

### **Example C.2 Kernel Density Estimator for the Income Data**

Figure C.2 suggests the large skew in the income data that is also suggested by the box and whisker plot (and the scatter plot) in Example C.1.

## C.4 STATISTICS AS ESTIMATORS—SAMPLING DISTRIBUTIONS

The measures described in the preceding section summarize the data in a random sample. Each measure has a counterpart in the population, that is, the distribution from which the data were drawn. Sample quantities such as the means and the correlation coefficient correspond to population expectations, whereas the kernel density estimator and the values in Table C.1 parallel the



**FIGURE C.2** Kernel Density Estimate for Income.

**TABLE C.1** Income Distribution

Range	Relative Frequency	Cumulative Frequency
<\$10,000	0.15	0.15
10,000–25,000	0.30	0.45
25,000–50,000	0.40	0.85
>50,000	0.15	1.00

population **pdf** and **cdf**. In the setting of a random sample, we expect these quantities to mimic the population, although not perfectly. The precise manner in which these quantities reflect the population values defines the sampling distribution of a sample statistic.

**DEFINITION C.1 Statistic**

*A statistic is any function computed from the data in a sample.*

If another sample were drawn under identical conditions, different values would be obtained for the observations, as each one is a random variable. Any statistic is a function of these random values, so it is also a random variable with a probability distribution called a **sampling distribution**. For example, the following shows an exact result for the sampling behavior of a widely used statistic.

**1122 PART VI ♦ Appendices**
**THEOREM C.1 Sampling Distribution of the Sample Mean**

If  $x_1, \dots, x_n$  are a random sample from a population with mean  $\mu$  and variance  $\sigma^2$ , then  $\bar{x}$  is a random variable with mean  $\mu$  and variance  $\sigma^2/n$ .

**Proof:**  $\bar{x} = (1/n)\sum_i x_i$ .  $E[\bar{x}] = (1/n)\sum_i \mu = \mu$ . The observations are independent, so  $\text{Var}[\bar{x}] = (1/n)^2 \text{Var}[\sum_i x_i] = (1/n^2) \sum_i \sigma^2 = \sigma^2/n$ .

Example C.3 illustrates the behavior of the sample mean in samples of four observations drawn from a chi-squared population with one degree of freedom. The crucial concepts illustrated in this example are, first, the mean and variance results in Theorem C.1 and, second, the phenomenon of **sampling variability**.

Notice that the fundamental result in Theorem C.1 does not assume a distribution for  $x_i$ . Indeed, looking back at Section C.3, nothing we have done so far has required any assumption about a particular distribution.

**Example C.3 Sampling Distribution of a Sample Mean**

Figure C.3 shows a frequency plot of the means of 1,000 random samples of four observations drawn from a chi-squared distribution with one degree of freedom, which has mean 1 and variance 2.

We are often interested in how a statistic behaves as the sample size increases. Example C.4 illustrates one such case. Figure C.4 shows two sampling distributions, one based on samples of three and a second, of the same statistic, but based on samples of six. The effect of increasing sample size in this figure is unmistakable. It is easy to visualize the behavior of this statistic if we extrapolate the experiment in Example C.4 to samples of, say, 100.

**Example C.4 Sampling Distribution of the Sample Minimum**

If  $x_1, \dots, x_n$  are a random sample from an exponential distribution with  $f(x) = \theta e^{-\theta x}$ , then the sampling distribution of the sample minimum in a sample of  $n$  observations, denoted  $x_{(1)}$ , is

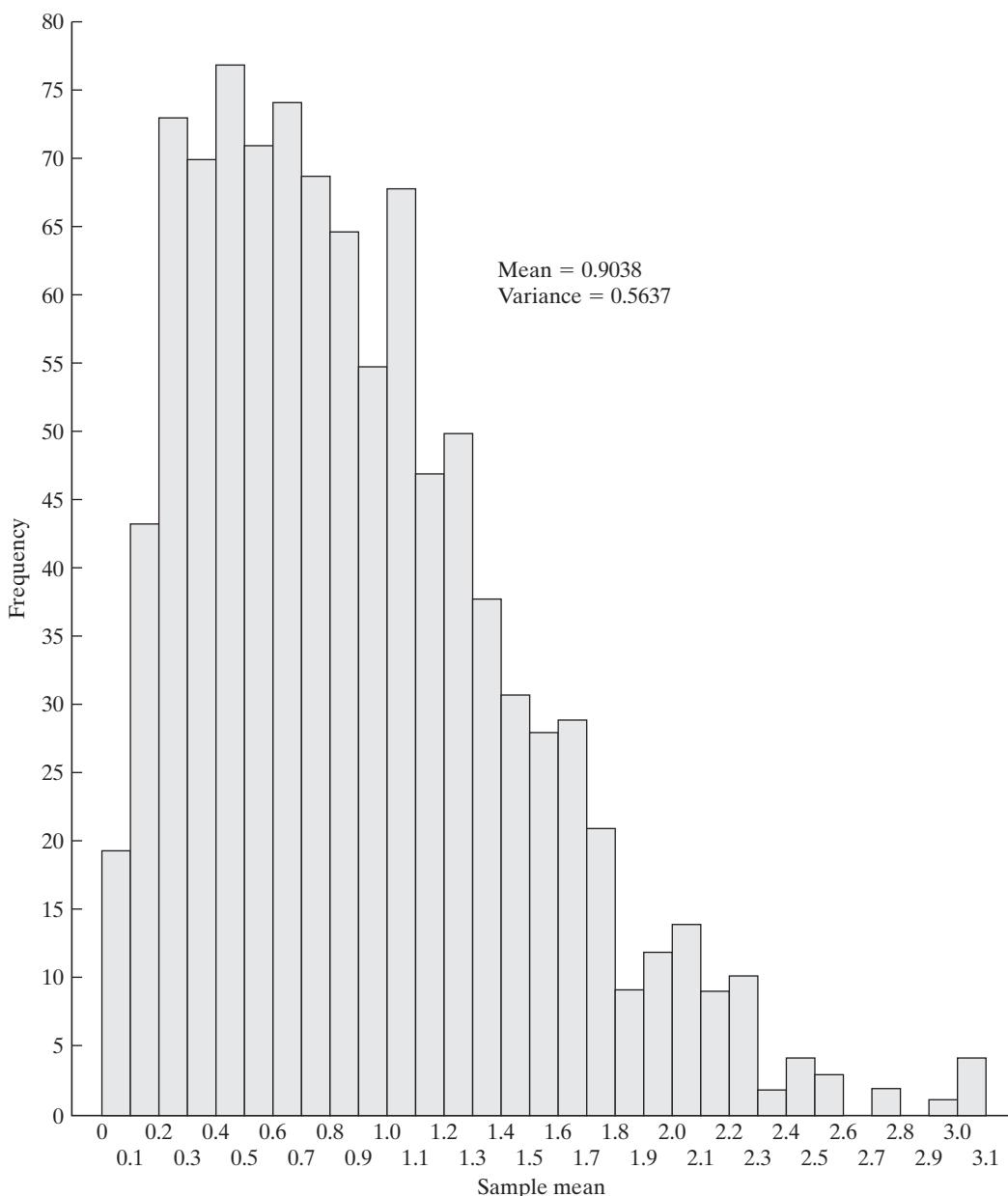
$$f(x_{(1)}) = (n\theta)e^{-(n\theta)x_{(1)}}.$$

Because  $E[x] = 1/\theta$  and  $\text{Var}[x] = 1/\theta^2$ , by analogy  $E[x_{(1)}] = 1/(n\theta)$  and  $\text{Var}[x_{(1)}] = 1/(n\theta)^2$ . Thus, in increasingly larger samples, the minimum will be arbitrarily close to 0. [The Chebychev inequality in Theorem D.2 can be used to prove this intuitively appealing result.]

Figure C.4 shows the results of a simple sampling experiment you can do to demonstrate this effect. It requires software that will allow you to produce pseudorandom numbers uniformly distributed in the range zero to one and that will let you plot a histogram and control the axes. (We used *NLOGIT*. This can be done with *Stata*, *Excel*, or several other packages.) The experiment consists of drawing 1,000 sets of nine random values,  $U_{ij}, i = 1, \dots, 1,000, j = 1, \dots, 9$ . To transform these uniform draws to exponential with parameter  $\theta$ —we used  $\theta = 1.5$ , use the inverse probability transform—see Section E.2.3. For an exponentially distributed variable, the transformation is  $z_{ij} = -(1/\theta) \log(1 - U_{ij})$ . We then created  $z_{(1)}|3$  from the first three draws and  $z_{(1)}|6$  from the other six. The two histograms show clearly the effect on the sampling distribution of increasing sample size from just 3 to 6.

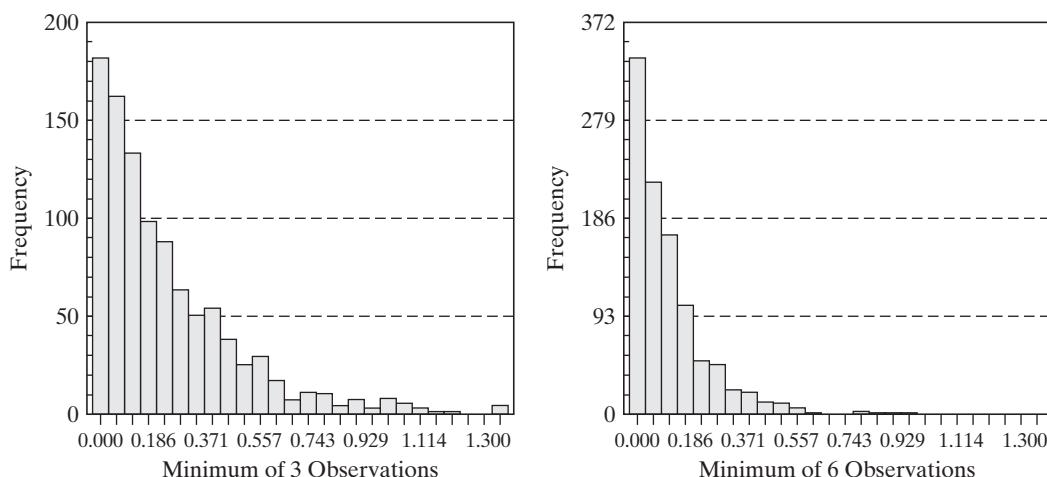
Sampling distributions are used to make inferences about the population. To consider a perhaps obvious example, because the sampling distribution of the mean of a set of normally distributed observations has mean  $\mu$ , the sample mean is a natural candidate for an estimate of  $\mu$ . The observation that the sample “mimics” the population is a statement about the sampling

## APPENDIX C ♦ Estimation and Inference 1123



**FIGURE C.3** Sampling Distribution of Means of 1,000 Samples of Size 4 from Chi-Squared [1].

## 1124 PART VI ♦ Appendices



**FIGURE C.4** Histograms of the Sample Minimum of 3 and 6 Observations.

distributions of the sample statistics. Consider, for example, the sample data collected in Figure C.3. The sample mean of four observations clearly has a sampling distribution, which appears to have a mean roughly equal to the population mean. Our theory of parameter estimation departs from this point.

## C.5 POINT ESTIMATION OF PARAMETERS

Our objective is to use the sample data to infer the value of a parameter or set of parameters, which we denote  $\theta$ . A **point estimate** is a statistic computed from a sample that gives a single value for  $\theta$ . The **standard error** of the estimate is the standard deviation of the sampling distribution of the statistic; the square of this quantity is the **sampling variance**. An **interval estimate** is a range of values that will contain the true parameter with a preassigned probability. There will be a connection between the two types of estimates; generally, if  $\hat{\theta}$  is the point estimate, then the interval estimate will be  $\hat{\theta} \pm$  a measure of sampling error.

An **estimator** is a rule or strategy for using the data to estimate the parameter. It is defined before the data are drawn. Obviously, some estimators are better than others. To take a simple example, your intuition should convince you that the sample mean would be a better estimator of the population mean than the sample minimum; the minimum is almost certain to underestimate the mean. Nonetheless, the minimum is not entirely without virtue; it is easy to compute, which is occasionally a relevant criterion. The search for good estimators constitutes much of econometrics. Estimators are compared on the basis of a variety of attributes. **Finite sample properties** of estimators are those attributes that can be compared regardless of the sample size. Some estimation problems involve characteristics that are not known in finite samples. In these instances, estimators are compared on the basis on their large sample, or **asymptotic properties**. We consider these in turn.

### C.5.1 ESTIMATION IN A FINITE SAMPLE

The following are some finite sample estimation criteria for estimating a single parameter. The extensions to the multiparameter case are direct. We shall consider them in passing where necessary.

**DEFINITION C.2 Unbiased Estimator**

An estimator of a parameter  $\theta$  is **unbiased** if the mean of its sampling distribution is  $\theta$ . Formally,

$$E[\hat{\theta}] = \theta$$

or

$$E[\hat{\theta} - \theta] = \text{Bias}[\hat{\theta} | \theta] = 0$$

implies that  $\hat{\theta}$  is unbiased. Note that this implies that the expected sampling error is zero. If  $\boldsymbol{\theta}$  is a vector of parameters, then the estimator is unbiased if the expected value of every element of  $\hat{\boldsymbol{\theta}}$  equals the corresponding element of  $\boldsymbol{\theta}$ .

If samples of size  $n$  are drawn repeatedly and  $\hat{\theta}$  is computed for each one, then the average value of these estimates will tend to equal  $\theta$ . For example, the average of the 1,000 sample means underlying Figure C.2 is 0.9038, which is reasonably close to the population mean of one. The sample minimum is clearly a biased estimator of the mean; it will almost always underestimate the mean, so it will do so on average as well.

Unbiasedness is a desirable attribute, but it is rarely used by itself as an estimation criterion. One reason is that there are many unbiased estimators that are poor uses of the data. For example, in a sample of size  $n$ , the first observation drawn is an unbiased estimator of the mean that clearly wastes a great deal of information. A second criterion used to choose among unbiased estimators is efficiency.

**DEFINITION C.3 Efficient Unbiased Estimator**

An unbiased estimator  $\hat{\theta}_1$  is more **efficient** than another unbiased estimator  $\hat{\theta}_2$  if the sampling variance of  $\hat{\theta}_1$  is less than that of  $\hat{\theta}_2$ . That is,

$$\text{Var}[\hat{\theta}_1] < \text{Var}[\hat{\theta}_2].$$

In the multiparameter case, the comparison is based on the covariance matrices of the two estimators;  $\hat{\boldsymbol{\theta}}_1$  is more efficient than  $\hat{\boldsymbol{\theta}}_2$  if  $\text{Var}[\hat{\boldsymbol{\theta}}_2] - \text{Var}[\hat{\boldsymbol{\theta}}_1]$  is a positive definite matrix.

By this criterion, the sample mean is obviously to be preferred to the first observation as an estimator of the population mean. If  $\sigma^2$  is the population variance, then

$$\text{Var}[x_1] = \sigma^2 > \text{Var}[\bar{x}] = \frac{\sigma^2}{n}.$$

In discussing efficiency, we have restricted the discussion to unbiased estimators. Clearly, there are biased estimators that have smaller variances than the unbiased ones we have considered. Any constant has a variance of zero. Of course, using a constant as an estimator is not likely to be an effective use of the sample data. Focusing on unbiasedness may still preclude a tolerably biased estimator with a much smaller variance, however. A criterion that recognizes this possible tradeoff is the mean squared error.

**1126 PART VI ♦ Appendices**
**DEFINITION C.4 Mean Squared Error**

The mean squared error of an estimator is

$$\begin{aligned} \text{MSE}[\hat{\theta} | \theta] &= E[(\hat{\theta} - \theta)^2] \\ &= \text{Var}[\hat{\theta}] + (\text{Bias}[\hat{\theta} | \theta])^2 && \text{if } \theta \text{ is a scalar,} \\ \text{MSE}[\hat{\theta} | \theta] &= \text{Var}[\hat{\theta}] + \text{Bias}[\hat{\theta} | \theta]\text{Bias}[\hat{\theta} | \theta]' && \text{if } \theta \text{ is a vector.} \end{aligned} \quad (\text{C-9})$$

Figure C.5 illustrates the effect. In this example, on average, the biased estimator will be closer to the true parameter than will the unbiased estimator.

Which of these criteria should be used in a given situation depends on the particulars of that setting and our objectives in the study. Unfortunately, the MSE criterion is rarely operational; minimum mean squared error estimators, when they exist at all, usually depend on unknown parameters. Thus, we are usually less demanding. A commonly used criterion is **minimum variance unbiasedness**.

**Example C.5 Mean Squared Error of the Sample Variance**

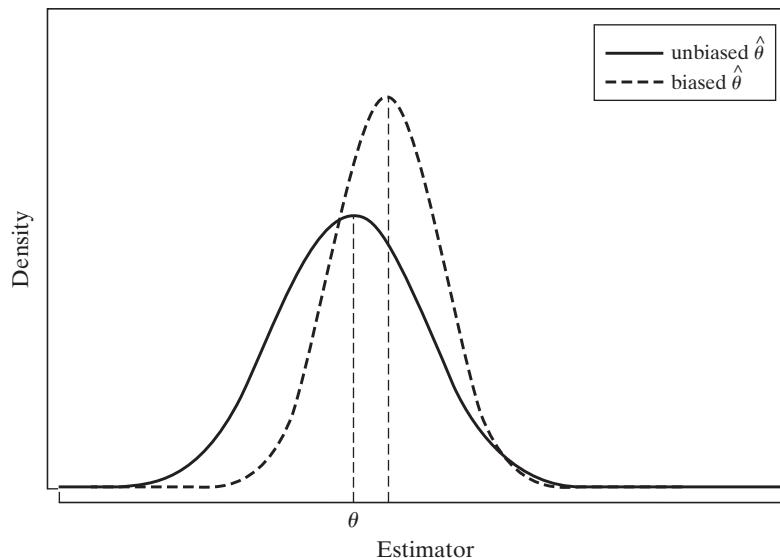
In sampling from a normal distribution, the most frequently used estimator for  $\sigma^2$  is

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}.$$

It is straightforward to show that  $s^2$  is unbiased, so

$$\text{Var}[s^2] = \frac{2\sigma^4}{n - 1} = \text{MSE}[s^2 | \sigma^2].$$

**FIGURE C.5** Sampling Distributions.



## APPENDIX C ♦ Estimation and Inference 1127

[A proof is based on the distribution of the idempotent quadratic form  $(\mathbf{x} - \mathbf{i}\mu)' \mathbf{M}^0 (\mathbf{x} - \mathbf{i}\mu)$ , which we discussed in Section B11.4.] A less frequently used estimator is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = [(n-1)/n] s^2.$$

This estimator is slightly biased downward:

$$E[\hat{\sigma}^2] = \frac{(n-1)E(s^2)}{n} = \frac{(n-1)\sigma^2}{n},$$

so its bias is

$$E[\hat{\sigma}^2 - \sigma^2] = \text{Bias}[\hat{\sigma}^2 | \sigma^2] = \frac{-1}{n}\sigma^2.$$

But it has a smaller variance than  $s^2$ :

$$\text{Var}[\hat{\sigma}^2] = \left[ \frac{n-1}{n} \right]^2 \left[ \frac{2\sigma^4}{n-1} \right] < \text{Var}[s^2].$$

To compare the two estimators, we can use the difference in their mean squared errors:

$$\text{MSE}[\hat{\sigma}^2 | \sigma^2] - \text{MSE}[s^2 | \sigma^2] = \sigma^4 \left[ \frac{2n-1}{n^2} - \frac{2}{n-1} \right] < 0.$$

The biased estimator is a bit more precise. The difference will be negligible in a large sample, but, for example, it is about 1.2 percent in a sample of 16.

### C.5.2 EFFICIENT UNBIASED ESTIMATION

In a random sample of  $n$  observations, the density of each observation is  $f(x_i, \theta)$ . Because the  $n$  observations are independent, their joint density is

$$\begin{aligned} f(x_1, x_2, \dots, x_n, \theta) &= f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta) \\ &= \prod_{i=1}^n f(x_i, \theta) = L(\theta | x_1, x_2, \dots, x_n). \end{aligned} \tag{C-10}$$

This function, denoted  $L(\theta | \mathbf{X})$ , is called the likelihood function for  $\theta$  given the data  $\mathbf{X}$ . It is frequently abbreviated to  $L(\theta)$ . Where no ambiguity can arise, we shall abbreviate it further to  $L$ .

#### **Example C.6 Likelihood Functions for Exponential and Normal Distributions**

If  $x_1, \dots, x_n$  are a sample of  $n$  observations from an exponential distribution with parameter  $\theta$ , then

$$L(\theta) = \prod_{i=1}^n \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^n x_i}.$$

If  $x_1, \dots, x_n$  are a sample of  $n$  observations from a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ , then

$$\begin{aligned} L(\mu, \sigma) &= \prod_{i=1}^n (2\pi\sigma^2)^{-1/2} e^{-(1/(2\sigma^2))(x_i - \mu)^2} \\ &= (2\pi\sigma^2)^{-n/2} e^{-(1/(2\sigma^2))\sum_i (x_i - \mu)^2}. \end{aligned} \tag{C-11}$$

## 1128 PART VI ♦ Appendices

The likelihood function is the cornerstone for most of our theory of parameter estimation. An important result for efficient estimation is the following.

### THEOREM C.2 Cramér–Rao Lower Bound

*Assuming that the density of  $x$  satisfies certain regularity conditions, the variance of an unbiased estimator of a parameter  $\theta$  will always be at least as large as*

$$[I(\theta)]^{-1} = \left( -E\left[ \frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \right] \right)^{-1} = \left( E\left[ \left( \frac{\partial \ln L(\theta)}{\partial \theta} \right)^2 \right] \right)^{-1}. \quad (\text{C-12})$$

*The quantity  $I(\theta)$  is the information number for the sample. We will prove the result that the negative of the expected second derivative equals the expected square of the first derivative in Chapter 14. Proof of the main result of the theorem is quite involved. See, for example, Stuart and Ord (1989).*

The regularity conditions are technical in nature. (See Section 14.4.1.) Loosely, they are conditions imposed on the density of the random variable that appears in the likelihood function; these conditions will ensure that the Lindeberg–Levy central limit theorem will apply to moments of the sample of observations on the random vector  $\mathbf{y} = \partial \ln f(x_i | \theta) / \partial \theta, i = 1, \dots, n$ . Among the conditions are finite moments of  $x$  up to order 3. An additional condition normally included in the set is that the range of the random variable be independent of the parameters.

In some cases, the second derivative of the log likelihood is a constant, so the Cramér–Rao bound is simple to obtain. For instance, in sampling from an exponential distribution, from Example C.6,

$$\begin{aligned} \ln L &= n \ln \theta - \theta \sum_{i=1}^n x_i, \\ \frac{\partial \ln L}{\partial \theta} &= \frac{n}{\theta} - \sum_{i=1}^n x_i, \end{aligned}$$

so  $\partial^2 \ln L / \partial \theta^2 = -n/\theta^2$  and the variance bound is  $[I(\theta)]^{-1} = \theta^2/n$ . In many situations, the second derivative is a random variable with a distribution of its own. The following examples show two such cases.

#### Example C.7 Variance Bound for the Poisson Distribution

For the Poisson distribution,

$$f(x) = \frac{e^{-\theta} \theta^x}{x!},$$

$$\ln L = -n\theta + \left( \sum_{i=1}^n x_i \right) \ln \theta - \sum_{i=1}^n \ln(x_i!),$$

$$\frac{\partial \ln L}{\partial \theta} = -n + \frac{\sum_{i=1}^n x_i}{\theta},$$

$$\frac{\partial^2 \ln L}{\partial \theta^2} = \frac{-\sum_{i=1}^n x_i}{\theta^2}.$$

## APPENDIX C ♦ Estimation and Inference 1129

The sum of  $n$  identical Poisson variables has a Poisson distribution with parameter equal to  $n$  times the parameter of the individual variables. Therefore, the actual distribution of the first derivative will be that of a linear function of a Poisson distributed variable. Because  $E[\sum_{i=1}^n x_i] = nE[x_i] = n\theta$ , the variance bound for the Poisson distribution is  $[I(\theta)]^{-1} = \theta/n$ . (Note also that the same result implies that  $E[\partial \ln L / \partial \theta] = 0$ , which is a result we will use in Chapter 14. The same result holds for the exponential distribution.)

Consider, finally, a multivariate case. If  $\theta$  is a vector of parameters, then  $\mathbf{I}(\theta)$  is the **information matrix**. The Cramér–Rao theorem states that the difference between the covariance matrix of any unbiased estimator and the inverse of the information matrix,

$$[\mathbf{I}(\theta)]^{-1} = \left( -E \left[ \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} \right] \right)^{-1} = \left\{ E \left[ \left( \frac{\partial \ln L(\theta)}{\partial \theta} \right) \left( \frac{\partial \ln L(\theta)}{\partial \theta'} \right) \right] \right\}^{-1}, \quad (\text{C-13})$$

will be a nonnegative definite matrix.

In many settings, numerous estimators are available for the parameters of a distribution. The usefulness of the Cramér–Rao bound is that if one of these is known to attain the variance bound, then there is no need to consider any other to seek a more efficient estimator. Regarding the use of the variance bound, we emphasize that if an unbiased estimator attains it, then that estimator is efficient. If a given estimator does not attain the variance bound, however, then we do not know, except in a few special cases, whether this estimator is efficient or not. It may be that no unbiased estimator can attain the Cramér–Rao bound, which can leave the question of whether a given unbiased estimator is efficient or not unanswered.

We note, finally, that in some cases we further restrict the set of estimators to linear functions of the data.

### **DEFINITION C.5 Minimum Variance Linear Unbiased Estimator (MVLUE)**

*An estimator is the minimum variance linear unbiased estimator or best linear unbiased estimator (BLUE) if it is a linear function of the data and has minimum variance among linear unbiased estimators.*

In a few instances, such as the normal mean, there will be an efficient linear unbiased estimator;  $\bar{x}$  is efficient among all unbiased estimators, both linear and nonlinear. In other cases, such as the normal variance, there is no linear unbiased estimator. This criterion is useful because we can sometimes find an MVLUE without having to specify the distribution at all. Thus, by limiting ourselves to a somewhat restricted class of estimators, we free ourselves from having to assume a particular distribution.

## C.6 INTERVAL ESTIMATION

Regardless of the properties of an estimator, the estimate obtained will vary from sample to sample, and there is some probability that it will be quite erroneous. A point estimate will not provide any information on the likely range of error. The logic behind an **interval estimate** is that we use the sample data to construct an interval, [lower ( $\mathbf{X}$ ), upper ( $\mathbf{X}$ )], such that we can expect this interval to contain the true parameter in some specified proportion of samples, or

## 1130 PART VI ♦ Appendices

equivalently, with some desired level of confidence. Clearly, the wider the interval, the more confident we can be that it will, in any given sample, contain the parameter being estimated.

The theory of interval estimation is based on a **pivotal quantity**, which is a function of both the parameter and a point estimate that has a known distribution. Consider the following examples.

### **Example C.8 Confidence Intervals for the Normal Mean**

In sampling from a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ ,

$$z = \frac{\sqrt{n}(\bar{x} - \mu)}{s} \sim t[n - 1],$$

and

$$c = \frac{(n - 1)s^2}{\sigma^2} \sim \chi^2[n - 1].$$

Given the pivotal quantity, we can make probability statements about events involving the parameter and the estimate. Let  $p(g, \theta)$  be the constructed random variable, for example,  $z$  or  $c$ . Given a prespecified **confidence level**,  $1 - \alpha$ , we can state that

$$\text{Prob}(\text{lower} \leq p(g, \theta) \leq \text{upper}) = 1 - \alpha, \quad (\text{C-14})$$

where lower and upper are obtained from the appropriate table. This statement is then manipulated to make equivalent statements about the endpoints of the intervals. For example, the following statements are equivalent:

$$\text{Prob}\left(-z \leq \frac{\sqrt{n}(\bar{x} - \mu)}{s} \leq z\right) = 1 - \alpha,$$

$$\text{Prob}\left(\bar{x} - \frac{zs}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{zs}{\sqrt{n}}\right) = 1 - \alpha.$$

The second of these is a statement about the interval, not the parameter; that is, it is the interval that is random, not the parameter. We attach a probability, or  $100(1 - \alpha)$  percent confidence level, to the interval itself; in repeated sampling, an interval constructed in this fashion will contain the true parameter  $100(1 - \alpha)$  percent of the time.

In general, the interval constructed by this method will be of the form

$$\text{lower}(\mathbf{X}) = \hat{\theta} - e_1,$$

$$\text{upper}(\mathbf{X}) = \hat{\theta} + e_2,$$

where  $\mathbf{X}$  is the sample data,  $e_1$  and  $e_2$  are sampling errors, and  $\hat{\theta}$  is a point estimate of  $\theta$ . It is clear from the preceding example that if the sampling distribution of the pivotal quantity is either  $t$  or standard normal, which will be true in the vast majority of cases we encounter in practice, then the confidence interval will be

$$\hat{\theta} \pm C_{1-\alpha/2}[\text{se}(\hat{\theta})], \quad (\text{C-15})$$

where  $\text{se}(\cdot)$  is the (known or estimated) standard error of the parameter estimate and  $C_{1-\alpha/2}$  is the value from the  $t$  or standard normal distribution that is exceeded with probability  $1 - \alpha/2$ . The usual values for  $\alpha$  are 0.10, 0.05, or 0.01. The theory does not prescribe exactly how to choose the endpoints for the confidence interval. An obvious criterion is to minimize the width of the interval. If the sampling distribution is symmetric, then the symmetric interval is the best one. If the sampling distribution is not symmetric, however, then this procedure will not be optimal.

**Example C.9 Estimated Confidence Intervals for a Normal Mean and Variance**

In a sample of 25,  $\bar{x} = 1.63$  and  $s = 0.51$ . Construct a 95 percent confidence interval for  $\mu$ . Assuming that the sample of 25 is from a normal distribution,

$$\text{Prob}\left(-2.064 \leq \frac{5(\bar{x} - \mu)}{s} \leq 2.064\right) = 0.95,$$

where 2.064 is the critical value from a  $t$  distribution with 24 degrees of freedom. Thus, the confidence interval is  $1.63 \pm [2.064(0.51)/5]$  or  $[1.4195, 1.8405]$ .

**Remark:** Had the parent distribution not been specified, it would have been natural to use the standard normal distribution instead, perhaps relying on the central limit theorem. But a sample size of 25 is small enough that the more conservative  $t$  distribution might still be preferable.

The chi-squared distribution is used to construct a confidence interval for the variance of a normal distribution. Using the data from Example C.9, we find that the usual procedure would use

$$\text{Prob}\left(12.4 \leq \frac{24s^2}{\sigma^2} \leq 39.4\right) = 0.95,$$

where 12.4 and 39.4 are the 0.025 and 0.975 cutoff points from the chi-squared (24) distribution. This procedure leads to the 95 percent confidence interval  $[0.1581, 0.5032]$ . By making use of the asymmetry of the distribution, a narrower interval can be constructed. Allocating 4 percent to the left-hand tail and 1 percent to the right instead of 2.5 percent to each, the two cutoff points are 13.4 and 42.9, and the resulting 95 percent confidence interval is  $[0.1455, 0.4659]$ .

Finally, the confidence interval can be manipulated to obtain a confidence interval for a function of a parameter. For example, based on the preceding, a 95 percent confidence interval for  $\sigma$  would be  $[\sqrt{0.1581}, \sqrt{0.5032}] = [0.3976, 0.7094]$ .

## C.7 HYPOTHESIS TESTING

The second major group of statistical inference procedures is hypothesis tests. The classical testing procedures are based on constructing a statistic from a random sample that will enable the analyst to decide, with reasonable confidence, whether or not the data in the sample would have been generated by a hypothesized population. The formal procedure involves a statement of the hypothesis, usually in terms of a “null” or maintained hypothesis and an “alternative,” conventionally denoted  $H_0$  and  $H_1$ , respectively. The procedure itself is a rule, stated in terms of the data, that dictates whether the null hypothesis should be rejected or not. For example, the hypothesis might state a parameter is equal to a specified value. The decision rule might state that the hypothesis should be rejected if a sample estimate of that parameter is too far away from that value (where “far” remains to be defined). The classical, or Neyman–Pearson, methodology involves partitioning the sample space into two regions. If the observed data (i.e., the test statistic) fall in the **rejection region** (sometimes called the **critical region**), then the null hypothesis is rejected; if they fall in the **acceptance region**, then it is not.

### C.7.1 CLASSICAL TESTING PROCEDURES

Since the sample is random, the test statistic, however defined, is also random. The same test procedure can lead to different conclusions in different samples. As such, there are two ways such a procedure can be in error:

1. **Type I error.** The procedure may lead to rejection of the null hypothesis when it is true.
2. **Type II error.** The procedure may fail to reject the null hypothesis when it is false.

## 1132 PART VI ♦ Appendices

To continue the previous example, there is some probability that the estimate of the parameter will be quite far from the hypothesized value, even if the hypothesis is true. This outcome might cause a type I error.

### DEFINITION C.6 Size of a Test

*The probability of a type I error is the size of the test. This is conventionally denoted  $\alpha$  and is also called the significance level.*

The size of the test is under the control of the analyst. It can be changed just by changing the decision rule. Indeed, the type I error could be eliminated altogether just by making the rejection region very small, but this would come at a cost. By eliminating the probability of a type I error—that is, by making it unlikely that the hypothesis is rejected—we must increase the probability of a type II error. Ideally, we would like both probabilities to be as small as possible. It is clear, however, that there is a tradeoff between the two. The best we can hope for is that for a given probability of type I error, the procedure we choose will have as small a probability of type II error as possible.

### DEFINITION C.7 Power of a Test

*The power of a test is the probability that it will correctly lead to rejection of a false null hypothesis:*

$$\text{power} = 1 - \beta = 1 - \text{Prob}(\text{type II error}). \quad (\text{C-16})$$

For a given significance level  $\alpha$ , we would like  $\beta$  to be as small as possible. Because  $\beta$  is defined in terms of the alternative hypothesis, it depends on the value of the parameter.

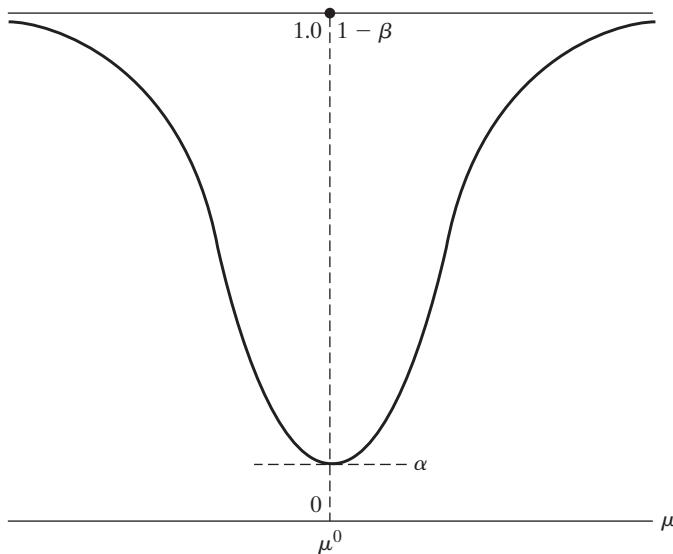
#### Example C.10 Testing a Hypothesis About a Mean

For testing  $H_0: \mu = \mu^0$  in a normal distribution with known variance  $\sigma^2$ , the decision rule is to reject the hypothesis if the absolute value of the z statistic,  $\sqrt{n}(\bar{x} - \mu^0)/\sigma$ , exceeds the predetermined critical value. For a test at the 5 percent significance level, we set the critical value at 1.96. The power of the test, therefore, is the probability that the absolute value of the test statistic will exceed 1.96 given that the true value of  $\mu$  is, in fact, not  $\mu^0$ . This value depends on the alternative value of  $\mu$ , as shown in Figure C.6. Notice that for this test the power is equal to the size at the point where  $\mu$  equals  $\mu^0$ . As might be expected, the test becomes more powerful the farther the true mean is from the hypothesized value.

Testing procedures, like estimators, can be compared using a number of criteria.

### DEFINITION C.8 Most Powerful Test

*A test is most powerful if it has greater power than any other test of the same size.*



**FIGURE C.6** Power Function for a Test.

This requirement is very strong. Because the power depends on the alternative hypothesis, we might require that the test be **uniformly most powerful (UMP)**, that is, have greater power than any other test of the same size for all admissible values of the parameter. There are few situations in which a UMP test is available. We usually must be less stringent in our requirements. Nonetheless, the criteria for comparing hypothesis testing procedures are generally based on their respective power functions. A common and very modest requirement is that the test be unbiased.

**DEFINITION C.9 Unbiased Test**

*A test is **unbiased** if its power  $(1 - \beta)$  is greater than or equal to its size  $\alpha$  for all values of the parameter.*

If a test is **biased**, then, for some values of the parameter, we are more likely to accept the null hypothesis when it is false than when it is true.

The use of the term *unbiased* here is unrelated to the concept of an unbiased estimator. Fortunately, there is little chance of confusion. Tests and estimators are clearly connected, however. The following criterion derives, in general, from the corresponding attribute of a parameter estimate.

**DEFINITION C.10 Consistent Test**

*A test is **consistent** if its power goes to one as the sample size grows to infinity.*

## 1134 PART VI ♦ Appendices

### **Example C.11 Consistent Test About a Mean**

A confidence interval for the mean of a normal distribution is  $\bar{x} \pm t_{1-\alpha/2}(s/\sqrt{n})$ , where  $\bar{x}$  and  $s$  are the usual consistent estimators for  $\mu$  and  $\sigma$  (see Section D.2.1),  $n$  is the sample size, and  $t_{1-\alpha/2}$  is the correct critical value from the  $t$  distribution with  $n - 1$  degrees of freedom. For testing  $H_0: \mu = \mu_0$  versus  $H_1: \mu \neq \mu_0$ , let the procedure be to reject  $H_0$  if the confidence interval does not contain  $\mu_0$ . Because  $\bar{x}$  is consistent for  $\mu$ , one can discern if  $H_0$  is false as  $n \rightarrow \infty$ , with probability 1, because  $\bar{x}$  will be arbitrarily close to the true  $\mu$ . Therefore, this test is consistent.

As a general rule, a test will be consistent if it is based on a consistent estimator of the parameter.

### **C.7.2 TESTS BASED ON CONFIDENCE INTERVALS**

There is an obvious link between interval estimation and the sorts of hypothesis tests we have been discussing here. The confidence interval gives a range of plausible values for the parameter. Therefore, it stands to reason that if a hypothesized value of the parameter does not fall in this range of plausible values, then the data are not consistent with the hypothesis, and it should be rejected. Consider, then, testing

$$H_0: \theta = \theta_0,$$

$$H_1: \theta \neq \theta_0.$$

We form a confidence interval based on  $\hat{\theta}$  as described earlier:

$$\hat{\theta} - C_{1-\alpha/2}[\text{se}(\hat{\theta})] < \theta < \hat{\theta} + C_{1-\alpha/2}[\text{se}(\hat{\theta})].$$

$H_0$  is rejected if  $\theta_0$  exceeds the upper limit or is less than the lower limit. Equivalently,  $H_0$  is rejected if

$$\left| \frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})} \right| > C_{1-\alpha/2}.$$

In words, the hypothesis is rejected if the estimate is too far from  $\theta_0$ , where the distance is measured in standard error units. The critical value is taken from the  $t$  or standard normal distribution, whichever is appropriate.

### **Example C.12 Testing a Hypothesis About a Mean with a Confidence Interval**

For the results in Example C.8, test  $H_0: \mu = 1.98$  versus  $H_1: \mu \neq 1.98$ , assuming sampling from a normal distribution:

$$t = \left| \frac{\bar{x} - 1.98}{s/\sqrt{n}} \right| = \left| \frac{1.63 - 1.98}{0.102} \right| = 3.43.$$

The 95 percent critical value for  $t(24)$  is 2.064. Therefore, reject  $H_0$ . If the critical value for the standard normal table of 1.96 is used instead, then the same result is obtained.

If the test is one-sided, as in

$$H_0: \theta \geq \theta_0,$$

$$H_1: \theta < \theta_0,$$

then the critical region must be adjusted. Thus, for this test,  $H_0$  will be rejected if a point estimate of  $\theta$  falls sufficiently below  $\theta_0$ . (Tests can usually be set up by departing from the decision criterion, “What sample results are inconsistent with the hypothesis?”)

## APPENDIX D ♦ Large-Sample Distribution Theory 1135

**Example C.13 One-Sided Test About a Mean**

A sample of 25 from a normal distribution yields  $\bar{x} = 1.63$  and  $s = 0.51$ . Test

$$H_0: \mu \leq 1.5,$$

$$H_1: \mu > 1.5.$$

Clearly, no observed  $\bar{x}$  less than or equal to 1.5 will lead to rejection of  $H_0$ . Using the borderline value of 1.5 for  $\mu$ , we obtain

$$\text{Prob}\left(\frac{\sqrt{n}(\bar{x} - 1.5)}{s} > \frac{5(1.63 - 1.5)}{0.51}\right) = \text{Prob}(t_{24} > 1.27).$$

This is approximately 0.11. This value is not unlikely by the usual standards. Hence, at a significant level of 0.11, we would not reject the hypothesis.

### C.7.3 SPECIFICATION TESTS

The hypothesis testing procedures just described are known as “classical” testing procedures. In each case, the null hypothesis tested came in the form of a restriction on the alternative. You can verify that in each application we examined, the parameter space assumed under the null hypothesis is a subspace of that described by the alternative. For that reason, the models implied are said to be “nested.” The null hypothesis is contained within the alternative. This approach suffices for most of the testing situations encountered in practice, but there are common situations in which two competing models cannot be viewed in these terms. For example, consider a case in which there are two completely different, competing theories to explain the same observed data. Many models for censoring and truncation discussed in Chapter 19 rest upon a fragile assumption of normality, for example. Testing of this nature requires a different approach from the classical procedures discussed here. These are discussed at various points throughout the book, for example, in Chapter 19, where we study the difference between fixed and random effects models.

## APPENDIX D

---

# LARGE-SAMPLE DISTRIBUTION THEORY

## D.1 INTRODUCTION

Most of this book is about parameter estimation. In studying that subject, we will usually be interested in determining how best to use the observed data when choosing among competing estimators. That, in turn, requires us to examine the sampling behavior of estimators. In a few cases, such as those presented in Appendix C and the least squares estimator considered in Chapter 4, we can make broad statements about sampling distributions that will apply regardless of the size of the sample. But, in most situations, it will only be possible to make approximate statements about estimators, such as whether they improve as the sample size increases and what can be said about their sampling distributions in large samples as an approximation to the finite

## APPENDIX D ♦ Large-Sample Distribution Theory 1135

**Example C.13 One-Sided Test About a Mean**

A sample of 25 from a normal distribution yields  $\bar{x} = 1.63$  and  $s = 0.51$ . Test

$$H_0: \mu \leq 1.5,$$

$$H_1: \mu > 1.5.$$

Clearly, no observed  $\bar{x}$  less than or equal to 1.5 will lead to rejection of  $H_0$ . Using the borderline value of 1.5 for  $\mu$ , we obtain

$$\text{Prob}\left(\frac{\sqrt{n}(\bar{x} - 1.5)}{s} > \frac{5(1.63 - 1.5)}{0.51}\right) = \text{Prob}(t_{24} > 1.27).$$

This is approximately 0.11. This value is not unlikely by the usual standards. Hence, at a significant level of 0.11, we would not reject the hypothesis.

### C.7.3 SPECIFICATION TESTS

The hypothesis testing procedures just described are known as “classical” testing procedures. In each case, the null hypothesis tested came in the form of a restriction on the alternative. You can verify that in each application we examined, the parameter space assumed under the null hypothesis is a subspace of that described by the alternative. For that reason, the models implied are said to be “nested.” The null hypothesis is contained within the alternative. This approach suffices for most of the testing situations encountered in practice, but there are common situations in which two competing models cannot be viewed in these terms. For example, consider a case in which there are two completely different, competing theories to explain the same observed data. Many models for censoring and truncation discussed in Chapter 19 rest upon a fragile assumption of normality, for example. Testing of this nature requires a different approach from the classical procedures discussed here. These are discussed at various points throughout the book, for example, in Chapter 19, where we study the difference between fixed and random effects models.

## APPENDIX D

---

# LARGE-SAMPLE DISTRIBUTION THEORY

## D.1 INTRODUCTION

Most of this book is about parameter estimation. In studying that subject, we will usually be interested in determining how best to use the observed data when choosing among competing estimators. That, in turn, requires us to examine the sampling behavior of estimators. In a few cases, such as those presented in Appendix C and the least squares estimator considered in Chapter 4, we can make broad statements about sampling distributions that will apply regardless of the size of the sample. But, in most situations, it will only be possible to make approximate statements about estimators, such as whether they improve as the sample size increases and what can be said about their sampling distributions in large samples as an approximation to the finite

## 1136 PART VI ♦ Appendices

samples we actually observe. This appendix will collect most of the formal, fundamental theorems and results needed for this analysis. A few additional results will be developed in the discussion of time-series analysis later in the book.

### D.2 LARGE-SAMPLE DISTRIBUTION THEORY<sup>1</sup>

In most cases, whether an estimator is exactly unbiased or what its exact sampling variance is in samples of a given size will be unknown. But we may be able to obtain approximate results about the behavior of the distribution of an estimator as the sample becomes large. For example, it is well known that the distribution of the mean of a sample tends to approximate normality as the sample size grows, regardless of the distribution of the individual observations. Knowledge about the limiting behavior of the distribution of an estimator can be used to infer an approximate distribution for the estimator in a finite sample. To describe how this is done, it is necessary, first, to present some results on convergence of random variables.

#### D.2.1 CONVERGENCE IN PROBABILITY

Limiting arguments in this discussion will be with respect to the sample size  $n$ . Let  $x_n$  be a sequence random variable indexed by the sample size.

#### DEFINITION D.1 Convergence in Probability

*The random variable  $x_n$  converges in probability to a constant  $c$  if  $\lim_{n \rightarrow \infty} \text{Prob}(|x_n - c| > \varepsilon) = 0$  for any positive  $\varepsilon$ .*

Convergence in probability implies that the values that the variable may take that are not close to  $c$  become increasingly unlikely as  $n$  increases. To consider one example, suppose that the random variable  $x_n$  takes two values, zero and  $n$ , with probabilities  $1 - (1/n)$  and  $(1/n)$ , respectively. As  $n$  increases, the second point will become ever more remote from any constant but, at the same time, will become increasingly less probable. In this example,  $x_n$  converges in probability to zero. The crux of this form of convergence is that all the mass of the probability distribution becomes concentrated at points close to  $c$ . If  $x_n$  converges in probability to  $c$ , then we write

$$\text{plim } x_n = c. \quad (\text{D-1})$$

We will make frequent use of a special case of convergence in probability, **convergence in mean square** or **convergence in quadratic mean**.

#### THEOREM D.1 Convergence in Quadratic Mean

*If  $x_n$  has mean  $\mu_n$  and variance  $\sigma_n^2$  such that the ordinary limits of  $\mu_n$  and  $\sigma_n^2$  are  $c$  and 0, respectively, then  $x_n$  converges in mean square to  $c$ , and*

$$\text{plim } x_n = c.$$

<sup>1</sup>A comprehensive summary of many results in large-sample theory appears in White (2001). The results discussed here will apply to samples of independent observations. Time-series cases in which observations are correlated are analyzed in Chapters 20 through 23.

**APPENDIX D ♦ Large-Sample Distribution Theory 1137**

A proof of Theorem D.1 can be based on another useful theorem.

**THEOREM D.2 Chebychev's Inequality**

If  $x_n$  is a random variable and  $c$  and  $\varepsilon$  are constants, then  $\text{Prob}(|x_n - c| > \varepsilon) \leq E[(x_n - c)^2]/\varepsilon^2$ .

To establish the Chebychev inequality, we use another result [see Goldberger (1991, p. 31)].

**THEOREM D.3 Markov's Inequality**

If  $y_n$  is a nonnegative random variable and  $\delta$  is a positive constant, then  $\text{Prob}[y_n \geq \delta] \leq E[y_n]/\delta$ .

**Proof:**  $E[y_n] = \text{Prob}[y_n < \delta]E[y_n | y_n < \delta] + \text{Prob}[y_n \geq \delta]E[y_n | y_n \geq \delta]$ . Because  $y_n$  is non-negative, both terms must be nonnegative, so  $E[y_n] \geq \text{Prob}[y_n \geq \delta]E[y_n | y_n \geq \delta]$ . Because  $E[y_n | y_n \geq \delta]$  must be greater than or equal to  $\delta$ ,  $E[y_n] \geq \text{Prob}[y_n \geq \delta]\delta$ , which is the result.

Now, to prove Theorem D.1, let  $y_n$  be  $(x_n - c)^2$  and  $\delta$  be  $\varepsilon^2$  in Theorem D.3. Then,  $(x_n - c)^2 > \delta$  implies that  $|x_n - c| > \varepsilon$ . Finally, we will use a special case of the Chebychev inequality, where  $c = \mu_n$ , so that we have

$$\text{Prob}(|x_n - \mu_n| > \varepsilon) \leq \sigma_n^2/\varepsilon^2. \quad (\mathbf{D-2})$$

Taking the limits of  $\mu_n$  and  $\sigma_n^2$  in (D-2), we see that if

$$\lim_{n \rightarrow \infty} E[x_n] = c, \quad \text{and} \quad \lim_{n \rightarrow \infty} \text{Var}[x_n] = 0, \quad (\mathbf{D-3})$$

then

$$\text{plim } x_n = c.$$

We have shown that convergence in mean square implies convergence in probability. Mean-square convergence implies that the distribution of  $x_n$  collapses to a spike at  $\text{plim } x_n$ , as shown in Figure D.1.

**Example D.1 Mean Square Convergence of the Sample Minimum in Exponential Sampling**

As noted in Example C.4, in sampling of  $n$  observations from an exponential distribution, for the sample minimum  $x_{(1)}$ ,

$$\lim_{n \rightarrow \infty} E[x_{(1)}] = \lim_{n \rightarrow \infty} \frac{1}{n\theta} = 0$$

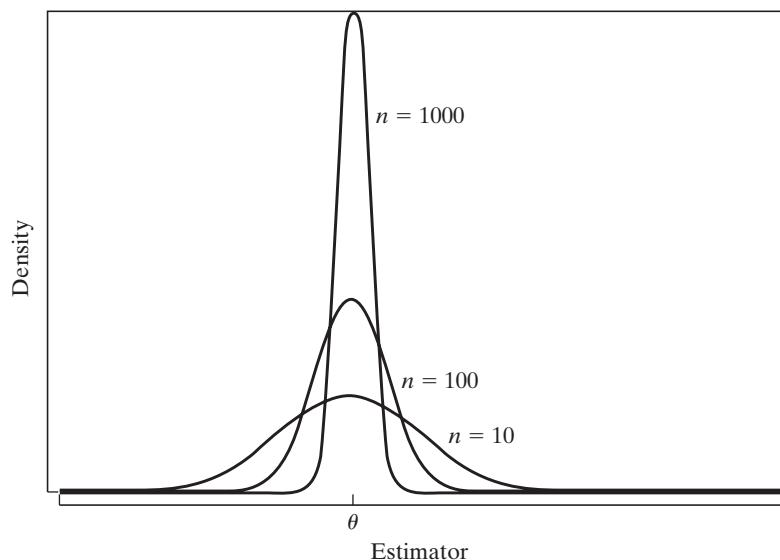
and

$$\lim_{n \rightarrow \infty} \text{Var}[x_{(1)}] = \lim_{n \rightarrow \infty} \frac{1}{(n\theta)^2} = 0.$$

Therefore,

$$\text{plim } x_{(1)} = 0.$$

Note, in particular, that the variance is divided by  $n^2$ . Thus, this estimator converges very rapidly to 0.

**1138 PART VI ♦ Appendices**


**FIGURE D.1** Quadratic Convergence to a Constant,  $\theta$ .

Convergence in probability does not imply convergence in mean square. Consider the simple example given earlier in which  $x_n$  equals either zero or  $n$  with probabilities  $1 - (1/n)$  and  $(1/n)$ . The exact expected value of  $x_n$  is 1 for all  $n$ , which is not the probability limit. Indeed, if we let  $\text{Prob}(x_n = n^2) = (1/n)$  instead, the mean of the distribution explodes, but the probability limit is still zero. Again, the point  $x_n = n^2$  becomes ever more extreme but, at the same time, becomes ever less likely.

The conditions for convergence in mean square are usually easier to verify than those for the more general form. Fortunately, we shall rarely encounter circumstances in which it will be necessary to show convergence in probability in which we cannot rely upon convergence in mean square. Our most frequent use of this concept will be in formulating consistent estimators.

**DEFINITION D.2 Consistent Estimator**

An estimator  $\hat{\theta}_n$  of a parameter  $\theta$  is a **consistent estimator** of  $\theta$  if and only if

$$\text{plim } \hat{\theta}_n = \theta. \quad (\text{D-4})$$

**THEOREM D.4 Consistency of the Sample Mean**

The mean of a random sample from any population with finite mean  $\mu$  and finite variance  $\sigma^2$  is a consistent estimator of  $\mu$ .

**Proof:**  $E[\bar{x}_n] = \mu$  and  $\text{Var}[\bar{x}_n] = \sigma^2/n$ . Therefore,  $\bar{x}_n$  converges in mean square to  $\mu$ , or  $\text{plim } \bar{x}_n = \mu$ .

**APPENDIX D ♦ Large-Sample Distribution Theory 1139**

Theorem D.4 is broader than it might appear at first.

**COROLLARY TO THEOREM D.4 Consistency of a Mean of Functions**

*In random sampling, for any function  $g(x)$ , if  $E[g(x)]$  and  $\text{Var}[g(x)]$  are finite constants, then*

$$\text{plim } \frac{1}{n} \sum_{i=1}^n g(x_i) = E[g(x)]. \quad (\mathbf{D-5})$$

**Proof:** Define  $y_i = g(x_i)$  and use Theorem D.4.

**Example D.2 Estimating a Function of the Mean**

In sampling from a normal distribution with mean  $\mu$  and variance 1,  $E[e^x] = e^{\mu+1/2}$  and  $\text{Var}[e^x] = e^{2\mu+2} - e^{2\mu+1}$ . (See Section B.4.4 on the lognormal distribution.) Hence,

$$\text{plim } \frac{1}{n} \sum_{i=1}^n e^{x_i} = e^{\mu+1/2}.$$

**D.2.2 OTHER FORMS OF CONVERGENCE AND LAWS OF LARGE NUMBERS**

Theorem D.4 and the corollary just given are particularly narrow forms of a set of results known as **laws of large numbers** that are fundamental to the theory of parameter estimation. Laws of large numbers come in two forms depending on the type of convergence considered. The simpler of these are “weak laws of large numbers” which rely on convergence in probability as we defined it above. “Strong laws” rely on a broader type of convergence called **almost sure convergence**. Overall, the law of large numbers is a statement about the behavior of an average of a large number of random variables.

**THEOREM D.5 Khinchine’s Weak Law of Large Numbers**

*If  $x_i, i = 1, \dots, n$  is a random (i.i.d.) sample from a distribution with finite mean  $E[x_i] = \mu$ , then*

$$\text{plim } \bar{x}_n = \mu.$$

*Proofs of this and the theorem below are fairly intricate. Rao (1973) provides one.*

Notice that this is already broader than Theorem D.4, as it does not require that the variance of the distribution be finite. On the other hand, it is not broad enough, because most of the situations we encounter where we will need a result such as this will not involve i.i.d. random sampling. A broader result is

**1140 PART VI ♦ Appendices**
**THEOREM D.6 Chebychev's Weak Law of Large Numbers**

If  $x_i, i = 1, \dots, n$  is a sample of observations such that  $E[x_i] = \mu_i < \infty$  and  $\text{Var}[x_i] = \sigma_i^2 < \infty$  such that  $\bar{\sigma}_n^2/n = (1/n^2)\sum_i \sigma_i^2 \rightarrow 0$  as  $n \rightarrow \infty$ , then  $\text{plim}(\bar{x}_n - \bar{\mu}_n) = 0$ .

There is a subtle distinction between these two theorems that you should notice. The Chebychev theorem does not state that  $\bar{x}_n$  converges to  $\bar{\mu}_n$ , or even that it converges to a constant at all. That would require a precise statement about the behavior of  $\bar{\mu}_n$ . The theorem states that as  $n$  increases without bound, these two quantities will be arbitrarily close to each other—that is, the difference between them converges to a constant, zero. This is an important notion that enters the derivation when we consider statistics that converge to random variables, instead of to constants. What we do have with these two theorems are extremely broad conditions under which a sample mean will converge in probability to its population counterpart. The more important difference between the Khinchine and Chebychev theorems is that the second allows for heterogeneity in the distributions of the random variables that enter the mean.

In analyzing time-series data, the sequence of outcomes is itself viewed as a random event. Consider, then, the sample mean,  $\bar{x}_n$ . The preceding results concern the behavior of this statistic as  $n \rightarrow \infty$  for a particular realization of the sequence  $\bar{x}_1, \dots, \bar{x}_n$ . But, if the sequence, itself, is viewed as a random event, then limit to which  $\bar{x}_n$  converges may be also. The stronger notion of almost sure convergence relates to this possibility.

**DEFINITION D.3 Almost Sure Convergence**

The random variable  $x_n$  converges almost surely to the constant  $c$  if and only if

$$\text{Prob}\left(\lim_{n \rightarrow \infty} x_n = c\right) = 1.$$

This is denoted  $x_n \xrightarrow{a.s.} c$ . It states that the probability of observing a sequence that does not converge to  $c$  ultimately vanishes. Intuitively, it states that once the sequence  $x_n$  becomes close to  $c$ , it stays close to  $c$ .

Almost sure convergence is used in a stronger form of the law of large numbers:

**THEOREM D.7 Kolmogorov's Strong Law of Large Numbers**

If  $x_i, i = 1, \dots, n$  is a sequence of independently distributed random variables such that  $E[x_i] = \mu_i < \infty$  and  $\text{Var}[x_i] = \sigma_i^2 < \infty$  such that  $\sum_{i=1}^{\infty} \sigma_i^2/i^2 < \infty$  as  $n \rightarrow \infty$  then  $\bar{x}_n - \bar{\mu}_n \xrightarrow{a.s.} 0$ .

**THEOREM D.8** **Markov's Strong Law of Large Numbers**

If  $\{z_i\}$  is a sequence of independent random variables with  $E[z_i] = \mu_i < \infty$  and if for some  $\delta > 0$ ,  $\sum_{i=1}^{\infty} E[|z_i - \mu_i|^{1+\delta}] / i^{1+\delta} < \infty$ , then  $\bar{z}_n - \bar{\mu}_n$  converges almost surely to 0, which we denote  $\bar{z}_n - \bar{\mu}_n \xrightarrow{a.s.} 0$ .<sup>2</sup>

The variance condition is satisfied if every variance in the sequence is finite, but this is not strictly required; it only requires that the variances in the sequence increase at a slow enough rate that the sequence of variances as defined is bounded. The theorem allows for heterogeneity in the means and variances. If we return to the conditions of the Khinchine theorem, i.i.d. sampling, we have a corollary:

**COROLLARY TO THEOREM D.8 (Kolmogorov)**

If  $x_i, i = 1, \dots, n$  is a sequence of independent and identically distributed random variables such that  $E[x_i] = \mu < \infty$  and  $E[|x_i|] < \infty$ , then  $\bar{x}_n - \mu \xrightarrow{a.s.} 0$ .

Note that the corollary requires identically distributed observations while the theorem only requires independence. Finally, another form of convergence encountered in the analysis of time-series data is convergence in  $r$ th mean:

**DEFINITION D.4** **Convergence in  $r$ th Mean**

If  $x_n$  is a sequence of random variables such that  $E[|x_n|^r] < \infty$  and  $\lim_{n \rightarrow \infty} E[|x_n - c|^r] = 0$ , then  $x_n$  converges in  $r$ th mean to  $c$ . This is denoted  $x_n \xrightarrow{r.m.} c$ .

Surely the most common application is the one we met earlier, convergence in means square, which is convergence in the second mean. Some useful results follow from this definition:

**THEOREM D.9** **Convergence in Lower Powers**

If  $x_n$  converges in  $r$ th mean to  $c$ , then  $x_n$  converges in  $s$ th mean to  $c$  for any  $s < r$ . The proof uses Jensen's Inequality, Theorem D.13. Write  $E[|x_n - c|^s] = E[(|x_n - c|^r)^{s/r}] \leq \{E[|x_n - c|^r]\}^{s/r}$  and the inner term converges to zero so the full function must also.

<sup>2</sup>The use of the expected absolute deviation differs a bit from the expected squared deviation that we have used heretofore to characterize the spread of a distribution. Consider two examples. If  $z \sim N[0, \sigma^2]$ , then  $E[|z|] = \text{Prob}[z < 0]E[-z|z < 0] + \text{Prob}[z \geq 0]E[z|z \geq 0] = 0.7979\sigma$ . (See Theorem 18.2.) So, finite expected absolute value is the same as finite second moment for the normal distribution. But if  $z$  takes values  $[0, n]$  with probabilities  $[1 - 1/n, 1/n]$ , then the variance of  $z$  is  $(n - 1)$ , but  $E[|z - \mu_z|]$  is  $2 - 2/n$ . For this case, finite expected absolute value occurs without finite expected second moment. These are different characterizations of the spread of the distribution.

**1142 PART VI ♦ Appendices**
**THEOREM D.10 Generalized Chebychev's Inequality**

If  $x_n$  is a random variable and  $c$  is a constant such that with  $E[|x_n - c|^r] < \infty$  and  $\varepsilon$  is a positive constant, then  $\text{Prob}(|x_n - c| > \varepsilon) \leq E[|x_n - c|^r]/\varepsilon^r$ .

We have considered two cases of this result already, when  $r = 1$  which is the Markov inequality, Theorem D.3, and when  $r = 2$ , which is the Chebychev inequality we looked at first in Theorem D.2.

**THEOREM D.11 Convergence in  $r$ th mean and Convergence in Probability**

If  $x_n \xrightarrow{r.m.} c$ , for some  $r > 0$ , then  $x_n \xrightarrow{P} c$ . The proof relies on Theorem D.10. By assumption,  $\lim_{n \rightarrow \infty} E[|x_n - c|^r] = 0$  so for some  $n$  sufficiently large,  $E[|x_n - c|^r] < \infty$ . By Theorem D.10, then,  $\text{Prob}(|x_n - c| > \varepsilon) \leq E[|x_n - c|^r]/\varepsilon^r$  for any  $\varepsilon > 0$ . The denominator of the fraction is a fixed constant and the numerator converges to zero by our initial assumption, so  $\lim_{n \rightarrow \infty} \text{Prob}(|x_n - c| > \varepsilon) = 0$ , which completes the proof.

One implication of Theorem D.11 is that although convergence in mean square is a convenient way to prove convergence in probability, it is actually stronger than necessary, as we get the same result for any positive  $r$ .

Finally, we note that we have now shown that both almost sure convergence and convergence in  $r$ th mean are stronger than convergence in probability; each implies the latter. But they, themselves, are different notions of convergence, and neither implies the other.

**DEFINITION D.5 Convergence of a Random Vector or Matrix**

Let  $\mathbf{x}_n$  denote a random vector and  $\mathbf{X}_n$  a random matrix, and  $\mathbf{c}$  and  $\mathbf{C}$  denote a vector and matrix of constants with the same dimensions as  $\mathbf{x}_n$  and  $\mathbf{X}_n$ , respectively. All of the preceding notions of convergence can be extended to  $(\mathbf{x}_n, \mathbf{c})$  and  $(\mathbf{X}_n, \mathbf{C})$  by applying the results to the respective corresponding elements.

**D.2.3 CONVERGENCE OF FUNCTIONS**

A particularly convenient result is the following.

**THEOREM D.12 Slutsky Theorem**

For a continuous function  $g(x_n)$  that is not a function of  $n$ ,

$$\text{plim } g(x_n) = g(\text{plim } x_n). \quad (\mathbf{D-6})$$

The generalization of Theorem D.12 to a function of several random variables is direct, as illustrated in the next example.

## APPENDIX D ♦ Large-Sample Distribution Theory **1143**

**Example D.3 Probability Limit of a Function of  $\bar{x}$  and  $s^2$**

In random sampling from a population with mean  $\mu$  and variance  $\sigma^2$ , the exact expected value of  $\bar{x}_n^2/s_n^2$  will be difficult, if not impossible, to derive. But, by the Slutsky theorem,

$$\text{plim } \frac{\bar{x}_n^2}{s_n^2} = \frac{\mu^2}{\sigma^2}.$$

An application that highlights the difference between expectation and probability is suggested by the following useful relationships.

**THEOREM D.13 Inequalities for Expectations**

**Jensen's Inequality.** If  $g(x_n)$  is a concave function of  $x_n$ , then  $g(E[x_n]) \geq E[g(x_n)]$ .

**Cauchy-Schwarz Inequality.** For two random variables,

$$E[|xy|] \leq \{E[x^2]\}^{1/2} \{E[y^2]\}^{1/2}.$$

Although the expected value of a function of  $x_n$  may not equal the function of the expected value—it exceeds it if the function is concave—the probability limit of the function is equal to the function of the probability limit.

The Slutsky theorem highlights a comparison between the expectation of a random variable and its probability limit. Theorem D.12 extends directly in two important directions. First, though stated in terms of convergence in probability, the same set of results applies to convergence in  $r$ th mean and almost sure convergence. Second, so long as the functions are continuous, the Slutsky theorem can be extended to vector or matrix valued functions of random scalars, vectors, or matrices. The following describe some specific applications. Some implications of the Slutsky theorem are now summarized.

**THEOREM D.14 Rules for Probability Limits**

If  $x_n$  and  $y_n$  are random variables with  $\text{plim } x_n = c$  and  $\text{plim } y_n = d$ , then

$$\text{plim}(x_n + y_n) = c + d, \quad (\text{sum rule}) \tag{D-7}$$

$$\text{plim } x_n y_n = cd, \quad (\text{product rule}) \tag{D-8}$$

$$\text{plim } x_n/y_n = c/d \quad \text{if } d \neq 0. \quad (\text{ratio rule}) \tag{D-9}$$

If  $\mathbf{W}_n$  is a matrix whose elements are random variables and if  $\text{plim } \mathbf{W}_n = \Omega$ , then

$$\text{plim } \mathbf{W}_n^{-1} = \Omega^{-1}. \quad (\text{matrix inverse rule}) \tag{D-10}$$

If  $\mathbf{X}_n$  and  $\mathbf{Y}_n$  are random matrices with  $\text{plim } \mathbf{X}_n = \mathbf{A}$  and  $\text{plim } \mathbf{Y}_n = \mathbf{B}$ , then

$$\text{plim } \mathbf{X}_n \mathbf{Y}_n = \mathbf{AB}. \quad (\text{matrix product rule}) \tag{D-11}$$

### D.2.4 CONVERGENCE TO A RANDOM VARIABLE

The preceding has dealt with conditions under which a random variable converges to a constant, for example, the way that a sample mean converges to the population mean. To develop a theory

## 1144 PART VI ♦ Appendices

for the behavior of estimators, as a prelude to the discussion of limiting distributions, we now consider cases in which a random variable converges not to a constant, but to another random variable. These results will actually subsume those in the preceding section, as a constant may always be viewed as a degenerate random variable, that is one with zero variance.

### DEFINITION D.6 Convergence in Probability to a Random Variable

*The random variable  $x_n$  converges in probability to the random variable  $x$  if  $\lim_{n \rightarrow \infty} \text{Prob}(|x_n - x| > \varepsilon) = 0$  for any positive  $\varepsilon$ .*

As before, we write  $\text{plim } x_n = x$  to denote this case. The interpretation (at least the intuition) of this type of convergence is different when  $x$  is a random variable. The notion of closeness defined here relates not to the concentration of the mass of the probability mechanism generating  $x_n$  at a point  $c$ , but to the closeness of that probability mechanism to that of  $x$ . One can think of this as a convergence of the CDF of  $x_n$  to that of  $x$ .

### DEFINITION D.7 Almost Sure Convergence to a Random Variable

*The random variable  $x_n$  converges almost surely to the random variable  $x$  if and only if  $\lim_{n \rightarrow \infty} \text{Prob}(|x_i - x| > \varepsilon \text{ for all } i \geq n) = 0$  for all  $\varepsilon > 0$ .*

### DEFINITION D.8 Convergence in $r$ th Mean to a Random Variable

*The random variable  $x_n$  converges in  $r$ th mean to the random variable  $x$  if and only if  $\lim_{n \rightarrow \infty} E[|x_n - x|^r] = 0$ . This is labeled  $x_n \xrightarrow{r.m.} x$ . As before, the case  $r = 2$  is labeled convergence in mean square.*

Once again, we have to revise our understanding of convergence when convergence is to a random variable.

### THEOREM D.15 Convergence of Moments

*Suppose  $x_n \xrightarrow{r.m.} x$  and  $E[|x|^r]$  is finite. Then,  $\lim_{n \rightarrow \infty} E[|x_n|^r] = E[|x|^r]$ .*

Theorem D.15 raises an interesting question. Suppose we let  $r$  grow, and suppose that  $x_n \xrightarrow{r.m.} x$  and, in addition, all moments are finite. If this holds for any  $r$ , do we conclude that these random variables have the same distribution? The answer to this longstanding problem in probability theory—the problem of the sequence of moments—is no. The sequence of moments does not uniquely determine the distribution. Although convergence in  $r$ th mean and almost surely still both imply convergence in probability, it remains true, even with convergence to a random variable instead of a constant, that these are different forms of convergence.

## APPENDIX D ♦ Large-Sample Distribution Theory **1145**

### D.2.5 CONVERGENCE IN DISTRIBUTION: LIMITING DISTRIBUTIONS

A second form of convergence is **convergence in distribution**. Let  $x_n$  be a sequence of random variables indexed by the sample size, and assume that  $x_n$  has cdf  $F_n(x_n)$ .

#### **DEFINITION D.9** Convergence in Distribution

$x_n$  converges in distribution to a random variable  $x$  with CDF  $F(x)$  if  $\lim_{n \rightarrow \infty} |F_n(x_n) - F(x)| = 0$  at all continuity points of  $F(x)$ .

This statement is about the probability distribution associated with  $x_n$ ; it does not imply that  $x_n$  converges at all. To take a trivial example, suppose that the exact distribution of the random variable  $x_n$  is

$$\text{Prob}(x_n = 1) = \frac{1}{2} + \frac{1}{n+1}, \quad \text{Prob}(x_n = 2) = \frac{1}{2} - \frac{1}{n+1}.$$

As  $n$  increases without bound, the two probabilities converge to  $\frac{1}{2}$ , but  $x_n$  does not converge to a constant.

#### **DEFINITION D.10** Limiting Distribution

If  $x_n$  converges in distribution to  $x$ , where  $F_n(x_n)$  is the CDF of  $x_n$ , then  $F(x)$  is the **limiting distribution** of  $x_n$ . This is written

$$x_n \xrightarrow{d} x.$$

The limiting distribution is often given in terms of the pdf, or simply the parametric family. For example, “the limiting distribution of  $x_n$  is standard normal.”

Convergence in distribution can be extended to random vectors and matrices, although not in the element by element manner that we extended the earlier convergence forms. The reason is that convergence in distribution is a property of the CDF of the random variable, not the variable itself. Thus, we can obtain a convergence result analogous to that in Definition D.9 for vectors or matrices by applying definition to the joint CDF for the elements of the vector or matrices. Thus,  $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$  if  $\lim_{n \rightarrow \infty} |F_n(\mathbf{x}_n) - F(\mathbf{x})| = 0$  and likewise for a random matrix.

#### **Example D.4** Limiting Distribution of $t_{n-1}$

Consider a sample of size  $n$  from a standard normal distribution. A familiar inference problem is the test of the hypothesis that the population mean is zero. The test statistic usually used is the  $t$  statistic:

$$t_{n-1} = \frac{\bar{x}_n}{s_n / \sqrt{n}},$$

where

$$s_n^2 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{n-1}.$$

## 1146 PART VI ♦ Appendices

The exact distribution of the random variable  $t_{n-1}$  is  $t$  with  $n - 1$  degrees of freedom. The density is different for every  $n$ :

$$f(t_{n-1}) = \frac{\Gamma(n/2)}{\Gamma[(n-1)/2]} [(n-1)\pi]^{-1/2} \left[ 1 + \frac{t_{n-1}^2}{n-1} \right]^{-n/2}, \quad (\text{D-12})$$

as is the CDF,  $F_{n-1}(t) = \int_{-\infty}^t f_{n-1}(x) dx$ . This distribution has mean zero and variance  $(n-1)/(n-3)$ . As  $n$  grows to infinity,  $t_{n-1}$  converges to the standard normal, which is written

$$t_{n-1} \xrightarrow{d} N[0, 1].$$

### DEFINITION D.11 Limiting Mean and Variance

*The limiting mean and variance of a random variable are the mean and variance of the limiting distribution, assuming that the limiting distribution and its moments exist.*

For the random variable with  $t[n]$  distribution, the exact mean and variance are zero and  $n/(n-2)$ , whereas the limiting mean and variance are zero and one. The example might suggest that the limiting mean and variance are zero and one; that is, that the moments of the limiting distribution are the ordinary limits of the moments of the finite sample distributions. This situation is almost always true, but it need not be. It is possible to construct examples in which the exact moments do not even exist, even though the moments of the limiting distribution are well defined.<sup>3</sup> Even in such cases, we can usually derive the mean and variance of the limiting distribution.

Limiting distributions, like probability limits, can greatly simplify the analysis of a problem. Some results that combine the two concepts are as follows.<sup>4</sup>

### THEOREM D.16 Rules for Limiting Distributions

1. If  $x_n \xrightarrow{d} x$  and  $\text{plim } y_n = c$ , then

$$x_n y_n \xrightarrow{d} cx, \quad (\text{D-13})$$

which means that the limiting distribution of  $x_n y_n$  is the distribution of  $cx$ . Also,

$$x_n + y_n \xrightarrow{d} x + c, \quad (\text{D-14})$$

$$x_n / y_n \xrightarrow{d} x/c, \quad \text{if } c \neq 0. \quad (\text{D-15})$$

2. If  $x_n \xrightarrow{d} x$  and  $g(x_n)$  is a continuous function, then

$$g(x_n) \xrightarrow{d} g(x). \quad (\text{D-16})$$

This result is analogous to the Slutsky theorem for probability limits. For an example, consider the  $t_n$  random variable discussed earlier. The exact distribution of  $t_n^2$  is  $F[1, n]$ . But as  $n \rightarrow \infty$ ,  $t_n$  converges to a standard normal variable. According to this result, the limiting distribution of  $t_n^2$  will be that of the square of a standard normal, which is chi-squared with one

<sup>3</sup>See, for example, Maddala (1977a, p. 150).

<sup>4</sup>For proofs and further discussion, see, for example, Greenberg and Webster (1983).

APPENDIX D ♦ Large-Sample Distribution Theory **1147****THEOREM D.16 (Continued)**

*degree of freedom. We conclude, therefore, that*

$$F[1, n] \xrightarrow{d} \text{chi-squared}[1]. \quad (\text{D-17})$$

*We encountered this result in our earlier discussion of limiting forms of the standard normal family of distributions.*

3. *If  $y_n$  has a limiting distribution and  $\text{plim } (x_n - y_n) = 0$ , then  $x_n$  has the same limiting distribution as  $y_n$ .*

The third result in Theorem D.16 combines convergence in distribution and in probability. The second result can be extended to vectors and matrices.

**Example D.5 The F Distribution**

Suppose that  $\mathbf{t}_{1,n}$  and  $\mathbf{t}_{2,n}$  are a  $K \times 1$  and an  $M \times 1$  random vector of variables whose components are independent with each distributed as  $t$  with  $n$  degrees of freedom. Then, as we saw in the preceding, for any component in either random vector, the limiting distribution is standard normal, so for the entire vector,  $\mathbf{t}_{j,n} \xrightarrow{d} \mathbf{z}_j$ , a vector of independent standard normally distributed variables. The results so far show that  $\frac{(\mathbf{t}_{1,n}, \mathbf{t}_{1,n})/K}{(\mathbf{t}_{2,n}, \mathbf{t}_{2,n})/M} \xrightarrow{d} F[K, M]$ .

Finally, a specific case of result 2 in Theorem D.16 produces a tool known as the Cramér–Wold device.

**THEOREM D.17 Cramér–Wold Device**

*If  $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$ , then  $\mathbf{c}'\mathbf{x}_n \xrightarrow{d} \mathbf{c}'\mathbf{x}$  for all conformable vectors  $\mathbf{c}$  with real valued elements.*

By allowing  $\mathbf{c}$  to be a vector with just a one in a particular position and zeros elsewhere, we see that convergence in distribution of a random vector  $\mathbf{x}_n$  to  $\mathbf{x}$  does imply that each component does likewise.

**D.2.6 CENTRAL LIMIT THEOREMS**

We are ultimately interested in finding a way to describe the statistical properties of estimators when their exact distributions are unknown. The concepts of consistency and convergence in probability are important. But the theory of limiting distributions given earlier is not yet adequate. We rarely deal with estimators that are not consistent for something, though perhaps not always the parameter we are trying to estimate. As such,

$$\text{if } \text{plim } \hat{\theta}_n = \theta, \text{ then } \hat{\theta}_n \xrightarrow{d} \theta.$$

That is, the limiting distribution of  $\hat{\theta}_n$  is a spike. This is not very informative, nor is it at all what we have in mind when we speak of the statistical properties of an estimator. (To endow our finite sample estimator  $\hat{\theta}_n$  with the zero sampling variance of the spike at  $\theta$  would be optimistic in the extreme.)

As an intermediate step, then, to a more reasonable description of the statistical properties of an estimator, we use a **stabilizing transformation** of the random variable to one that does have

## 1148 PART VI ♦ Appendices

a well-defined limiting distribution. To jump to the most common application, whereas

$$\text{plim } \hat{\theta}_n = \theta,$$

we often find that

$$z_n = \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} f(z),$$

where  $f(z)$  is a well-defined distribution with a mean and a positive variance. An estimator which has this property is said to be **root-n consistent**. The single most important theorem in econometrics provides an application of this proposition. A basic form of the theorem is as follows.

### THEOREM D.18 Lindeberg–Levy Central Limit Theorem (Univariate)

If  $x_1, \dots, x_n$  are a random sample from a probability distribution with finite mean  $\mu$  and finite variance  $\sigma^2$  and  $\bar{x}_n = (1/n) \sum_{i=1}^n x_i$ , then

$$\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} N[0, \sigma^2],$$

A proof appears in Rao (1973, p. 127).

The result is quite remarkable as it holds regardless of the form of the parent distribution. For a striking example, return to Figure C.2. The distribution from which the data were drawn in that figure does not even remotely resemble a normal distribution. In samples of only four observations the force of the central limit theorem is clearly visible in the sampling distribution of the means. The sampling experiment Example D.6 shows the effect in a systematic demonstration of the result.

The Lindeberg–Levy theorem is one of several forms of this extremely powerful result. For our purposes, an important extension allows us to relax the assumption of equal variances. The Lindeberg–Feller form of the central limit theorem is the centerpiece of most of our analysis in econometrics.

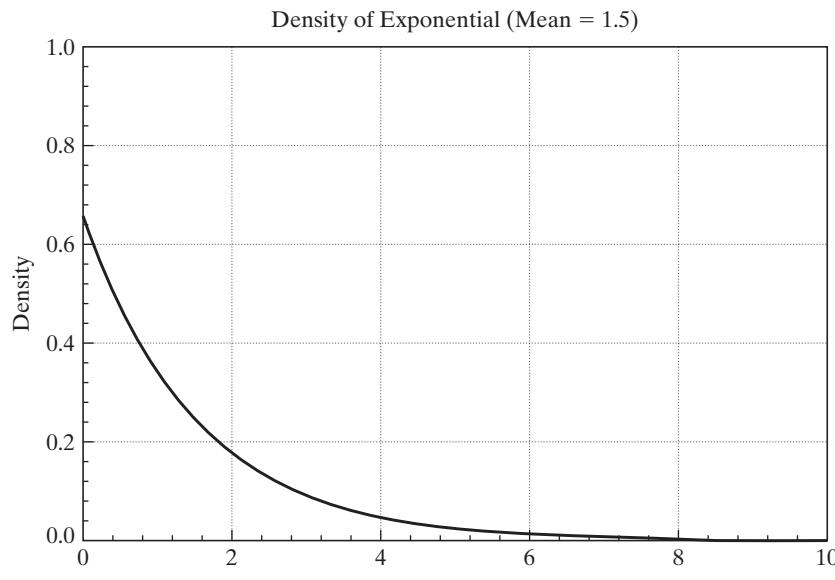
### THEOREM D.19 Lindeberg–Feller Central Limit Theorem (with Unequal Variances)

Suppose that  $\{x_i\}$ ,  $i = 1, \dots, n$ , is a sequence of independent random variables with finite means  $\mu_i$  and finite positive variances  $\sigma_i^2$ . Let

$$\bar{\mu}_n = \frac{1}{n}(\mu_1 + \mu_2 + \dots + \mu_n), \quad \text{and} \quad \bar{\sigma}_n^2 = \frac{1}{n}(\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2).$$

If no single term dominates this average variance, which we could state as  $\lim_{n \rightarrow \infty} \max(\sigma_i)/(n\bar{\sigma}_n) = 0$ , and if the average variance converges to a finite constant,  $\bar{\sigma}^2 = \lim_{n \rightarrow \infty} \bar{\sigma}_n^2$ , then

$$\sqrt{n}(\bar{x}_n - \bar{\mu}_n) \xrightarrow{d} N[0, \bar{\sigma}^2].$$

APPENDIX D ♦ Large-Sample Distribution Theory **1149****FIGURE D.2** The Exponential Distribution.

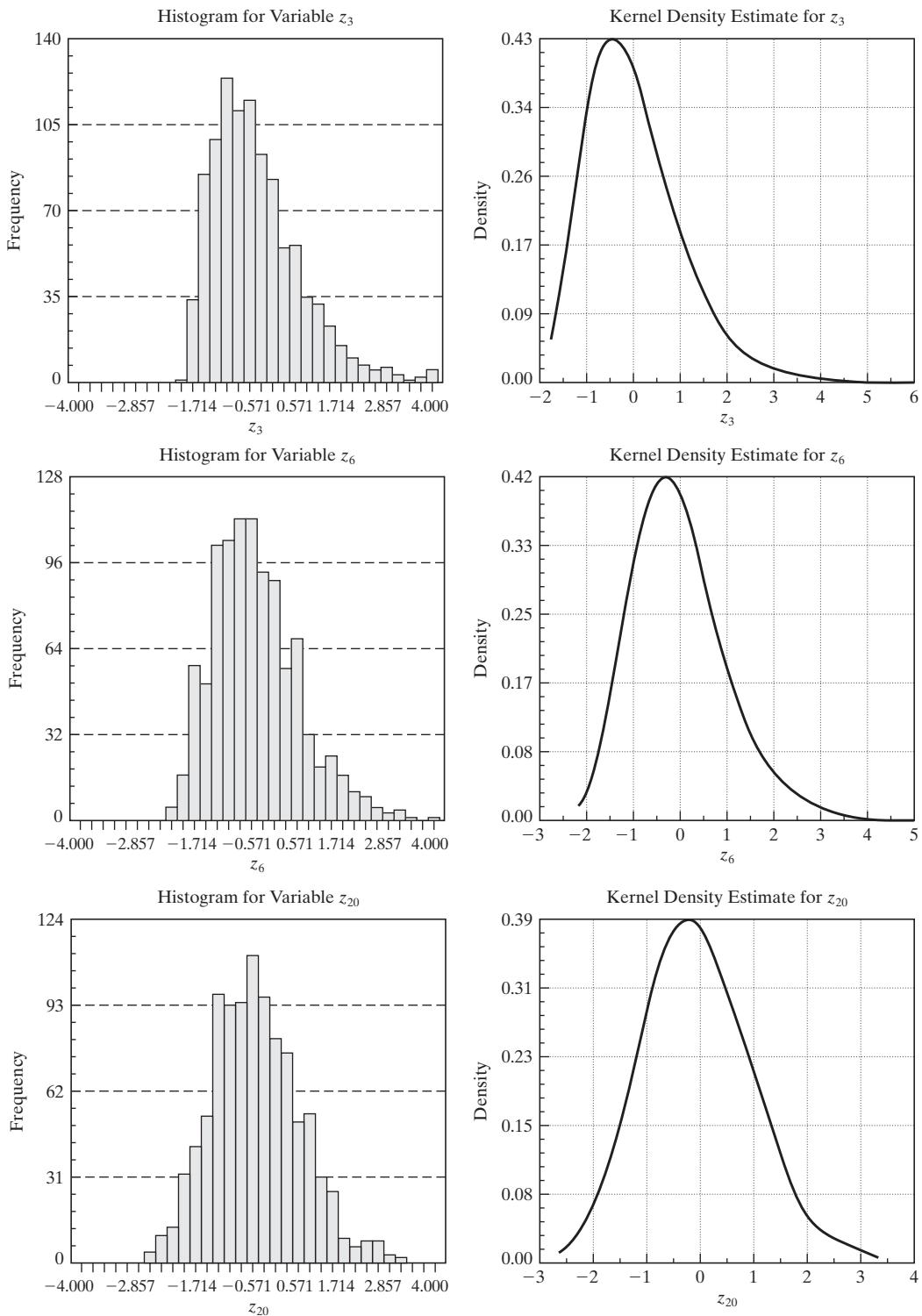
In practical terms, the theorem states that sums of random variables, regardless of their form, will tend to be normally distributed. The result is yet more remarkable in that *it does not require the variables in the sum to come from the same underlying distribution. It requires, essentially, only that the mean be a mixture of many random variables, none of which is large compared with their sum.* Because nearly all the estimators we construct in econometrics fall under the purview of the central limit theorem, it is obviously an important result.

**Example D.6 The Lindeberg–Levy Central Limit Theorem**

We'll use a sampling experiment to demonstrate the operation of the central limit theorem. Consider random sampling from the exponential distribution with mean 1.5—this is the setting used in Example C.4. The density is shown in Figure D.2.

We've drawn 1,000 samples of 3, 6, and 20 observations from this population and computed the sample means for each. For each mean, we then computed  $z_{in} = \sqrt{n}(\bar{x}_{in} - \mu)$ , where  $i = 1, \dots, 1,000$  and  $n$  is 3, 6 or 20. The three rows of figures in Figure D.3 show histograms of the observed samples of sample means and kernel density estimates of the underlying distributions for the three samples of transformed means.

Proof of the Lindeberg–Feller theorem requires some quite intricate mathematics [see, e.g., Loeve (1977)] that are well beyond the scope of our work here. We do note an important consideration in this theorem. The result rests on a condition known as the Lindeberg condition. The sample mean computed in the theorem is a mixture of random variables from possibly different distributions. The Lindeberg condition, in words, states that the contribution of the tail areas of these underlying distributions to the variance of the sum must be negligible in the limit. The condition formalizes the assumption in Theorem D.19 that the average variance be positive and not be dominated by any single term. [For an intuitively crafted mathematical discussion of this condition, see White (2001, pp. 117–118).] The condition is essentially impossible to verify in practice, so it is useful to have a simpler version of the theorem that encompasses it.



**FIGURE D.3** The Central Limit Theorem.

APPENDIX D ♦ Large-Sample Distribution Theory **1151****THEOREM D.20 Liapounov Central Limit Theorem**

Suppose that  $\{x_i\}$  is a sequence of independent random variables with finite means  $\mu_i$  and finite positive variances  $\sigma_i^2$  such that  $E[|x_i - \mu_i|^{2+\delta}]$  is finite for some  $\delta > 0$ . If  $\bar{\sigma}_n$  is positive and finite for all  $n$  sufficiently large, then

$$\sqrt{n}(\bar{x}_n - \bar{\mu}_n)/\bar{\sigma}_n \xrightarrow{d} N[0, 1].$$

This version of the central limit theorem requires only that moments slightly larger than two be finite.

Note the distinction between the laws of large numbers in Theorems D.5 and D.6 and the central limit theorems. Neither asserts that sample means tend to normality. Sample means (i.e., the distributions of them) converge to spikes at the true mean. It is the transformation of the mean,  $\sqrt{n}(\bar{x}_n - \mu)/\sigma$ , that converges to standard normality. To see this at work, if you have access to the necessary software, you might try reproducing Example D.6 using the raw means,  $\bar{x}_{in}$ . What do you expect to observe?

For later purposes, we will require multivariate versions of these theorems. Proofs of the following may be found, for example, in Greenberg and Webster (1983) or Rao (1973) and references cited there.

**THEOREM D.18A Multivariate Lindeberg–Levy Central Limit Theorem**

If  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are a random sample from a multivariate distribution with finite mean vector  $\boldsymbol{\mu}$  and finite positive definite covariance matrix  $\mathbf{Q}$ , then

$$\sqrt{n}(\bar{\mathbf{x}}_n - \boldsymbol{\mu}) \xrightarrow{d} N[\mathbf{0}, \mathbf{Q}],$$

where

$$\bar{\mathbf{x}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i.$$

To get from D.18 to D.18A (and D.19 to D.19A) we need to add a step. Theorem D.18 applies to the individual elements of the vector. A vector has a multivariate normal distribution if the individual elements are normally distributed and if every linear combination is normally distributed. We can use Theorem D.18 (D.19) for the individual terms and Theorem D.17 to establish that linear combinations behave likewise. This establishes the extensions.

The extension of the Lindeberg–Feller theorem to unequal covariance matrices requires some intricate mathematics. The following is an informal statement of the relevant conditions. Further discussion and references appear in Fomby, Hill, and Johnson (1984) and Greenberg and Webster (1983).

**1152 PART VI ♦ Appendices**
**THEOREM D.19A Multivariate Lindeberg–Feller Central Limit Theorem**

Suppose that  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are a sample of random vectors such that  $E[\mathbf{x}_i] = \boldsymbol{\mu}_i$ ,  $\text{Var}[\mathbf{x}_i] = \mathbf{Q}_i$ , and all mixed third moments of the multivariate distribution are finite. Let

$$\bar{\boldsymbol{\mu}}_n = \frac{1}{n} \sum_{i=1}^n \boldsymbol{\mu}_i,$$

$$\bar{\mathbf{Q}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{Q}_i.$$

We assume that

$$\lim_{n \rightarrow \infty} \bar{\mathbf{Q}}_n = \mathbf{Q},$$

where  $\mathbf{Q}$  is a finite, positive definite matrix, and that for every  $i$ ,

$$\lim_{n \rightarrow \infty} (n\bar{\mathbf{Q}}_n)^{-1} \mathbf{Q}_i = \lim_{n \rightarrow \infty} \left( \sum_{i=1}^n \mathbf{Q}_i \right)^{-1} \mathbf{Q}_i = \mathbf{0}.$$

We allow the means of the random vectors to differ, although in the cases that we will analyze, they will generally be identical. The second assumption states that individual components of the sum must be finite and diminish in significance. There is also an implicit assumption that the sum of matrices is nonsingular. Because the limiting matrix is nonsingular, the assumption must hold for large enough  $n$ , which is all that concerns us here. With these in place, the result is

$$\sqrt{n}(\bar{\mathbf{x}}_n - \bar{\boldsymbol{\mu}}_n) \xrightarrow{d} N[\mathbf{0}, \mathbf{Q}].$$

**D.2.7 THE DELTA METHOD**

At several points in Appendix C, we used a linear Taylor series approximation to analyze the distribution and moments of a random variable. We are now able to justify this usage. We complete the development of Theorem D.12 (probability limit of a function of a random variable), Theorem D.16 (2) (limiting distribution of a function of a random variable), and the central limit theorems, with a useful result that is known as the **delta method**. For a single random variable (sample mean or otherwise), we have the following theorem.

**THEOREM D.21 Limiting Normal Distribution of a Function**

If  $\sqrt{n}(z_n - \mu) \xrightarrow{d} N[0, \sigma^2]$  and if  $g(z_n)$  is a continuous and continuously differentiable function with  $g'(\mu)$  not equal to zero and not involving  $n$ , then

$$\sqrt{n}[g(z_n) - g(\mu)] \xrightarrow{d} N[0, \{g'(\mu)\}^2 \sigma^2]. \quad (\mathbf{D-18})$$

**APPENDIX D ♦ Large-Sample Distribution Theory 1153**

Notice that the mean and variance of the limiting distribution are the mean and variance of the linear Taylor series approximation:

$$g(z_n) \simeq g(\mu) + g'(\mu)(z_n - \mu).$$

The multivariate version of this theorem will be used at many points in the text.

**THEOREM D.21A Limiting Normal Distribution of a Set of Functions**

If  $\mathbf{z}_n$  is a  $K \times 1$  sequence of vector-valued random variables such that  $\sqrt{n}(\mathbf{z}_n - \boldsymbol{\mu}) \xrightarrow{d} N[\mathbf{0}, \Sigma]$  and if  $\mathbf{c}(\mathbf{z}_n)$  is a set of  $J$  continuous and continuously differentiable functions of  $\mathbf{z}_n$  with  $\mathbf{C}(\boldsymbol{\mu})$  not equal to zero, not involving  $n$ , then

$$\sqrt{n}[\mathbf{c}(\mathbf{z}_n) - \mathbf{c}(\boldsymbol{\mu})] \xrightarrow{d} N[\mathbf{0}, \mathbf{C}(\boldsymbol{\mu})\Sigma\mathbf{C}(\boldsymbol{\mu})'], \quad (\text{D-19})$$

where  $\mathbf{C}(\boldsymbol{\mu})$  is the  $J \times K$  matrix  $\partial\mathbf{c}(\boldsymbol{\mu})/\partial\boldsymbol{\mu}'$ . The  $j$ th row of  $\mathbf{C}(\boldsymbol{\mu})$  is the vector of partial derivatives of the  $j$ th function with respect to  $\boldsymbol{\mu}'$ .

### D.3 ASYMPTOTIC DISTRIBUTIONS

The theory of limiting distributions is only a means to an end. We are interested in the behavior of the estimators themselves. The limiting distributions obtained through the central limit theorem all involve unknown parameters, generally the ones we are trying to estimate. Moreover, our samples are always finite. Thus, we depart from the limiting distributions to derive the asymptotic distributions of the estimators.

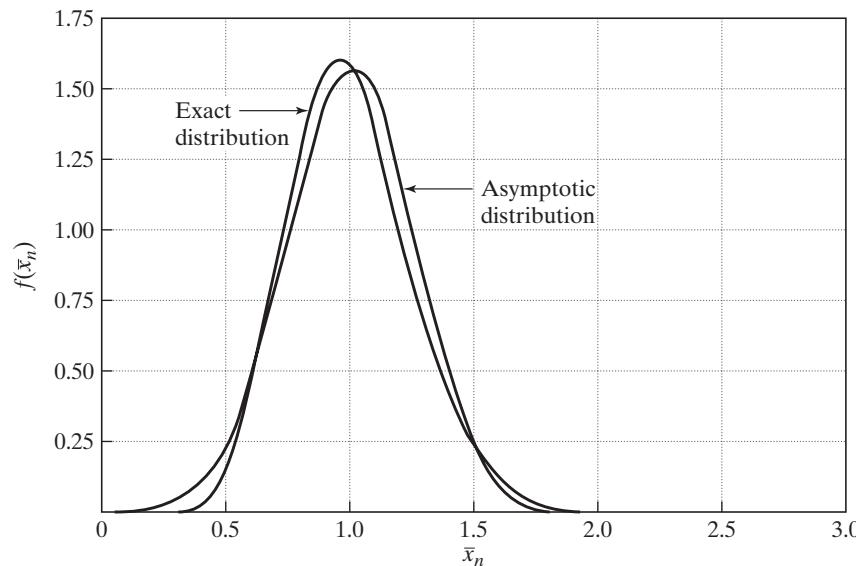
**DEFINITION D.12 Asymptotic Distribution**

An asymptotic distribution is a distribution that is used to approximate the true finite sample distribution of a random variable.<sup>5</sup>

By far the most common means of formulating an asymptotic distribution (at least by econometricians) is to construct it from the known limiting distribution of a function of the random variable. If

$$\sqrt{n}[(\bar{x}_n - \mu)/\sigma] \xrightarrow{d} N[0, 1],$$

<sup>5</sup>We depart somewhat from some other treatments [e.g., White (2001), Hayashi (2000, p. 90)] at this point, because they make no distinction between an asymptotic distribution and the limiting distribution, although the treatments are largely along the lines discussed here. In the interest of maintaining consistency of the discussion, we prefer to retain the sharp distinction and derive the asymptotic distribution of an estimator,  $\mathbf{t}$  by first obtaining the *limiting* distribution of  $\sqrt{n}(\mathbf{t} - \boldsymbol{\theta})$ . By our construction, the *limiting* distribution of  $\mathbf{t}$  is degenerate, whereas the *asymptotic* distribution of  $\sqrt{n}(\mathbf{t} - \boldsymbol{\theta})$  is not useful.

**1154 PART VI ♦ Appendices**


**FIGURE D.4** True Versus Asymptotic Distribution.

then approximately, or asymptotically,  $\bar{x}_n \sim N[\mu, \sigma^2/n]$ , which we write as

$$\bar{x} \xrightarrow{a} N[\mu, \sigma^2/n].$$

The statement “ $\bar{x}_n$  is asymptotically normally distributed with mean  $\mu$  and variance  $\sigma^2/n$ ” says only that this normal distribution provides an approximation to the true distribution, not that the true distribution is exactly normal.

**Example D.7 Asymptotic Distribution of the Mean of an Exponential Sample**

In sampling from an exponential distribution with parameter  $\theta$ , the exact distribution of  $\bar{x}_n$  is that of  $\theta/(2n)$  times a chi-squared variable with  $2n$  degrees of freedom. The asymptotic distribution is  $N[\theta, \theta^2/n]$ . The exact and asymptotic distributions are shown in Figure D.4 for the case of  $\theta = 1$  and  $n = 16$ .

Extending the definition, suppose that  $\hat{\theta}_n$  is an estimator of the parameter vector  $\theta$ . The asymptotic distribution of the vector  $\hat{\theta}_n$  is obtained from the limiting distribution:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N[0, \mathbf{V}] \quad (\text{D-20})$$

implies that

$$\hat{\theta}_n \xrightarrow{a} N\left[\theta, \frac{1}{n}\mathbf{V}\right]. \quad (\text{D-21})$$

This notation is read “ $\hat{\theta}_n$  is asymptotically normally distributed, with mean vector  $\theta$  and covariance matrix  $(1/n)\mathbf{V}$ .” The covariance matrix of the asymptotic distribution is the **asymptotic covariance matrix** and is denoted

$$\text{Asy. Var}[\hat{\theta}_n] = \frac{1}{n}\mathbf{V}.$$

## APPENDIX D ♦ Large-Sample Distribution Theory **1155**

Note, once again, the logic used to reach the result; (D-20) holds exactly as  $n \rightarrow \infty$ . We assume that it holds approximately for finite  $n$ , which leads to (D-21).

### **DEFINITION D.13 Asymptotic Normality and Asymptotic Efficiency**

*An estimator  $\hat{\theta}_n$  is asymptotically normal if (D-20) holds. The estimator is asymptotically efficient if the covariance matrix of any other consistent, asymptotically normally distributed estimator exceeds  $(1/n)\mathbf{V}$  by a nonnegative definite matrix.*

For most estimation problems, these are the criteria used to choose an estimator.

#### **Example D.8 Asymptotic Inefficiency of the Median in Normal Sampling**

In sampling from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , both the mean  $\bar{x}_n$  and the median  $M_n$  of the sample are consistent estimators of  $\mu$ . The limiting distributions of both estimators are spikes at  $\mu$ , so they can only be compared on the basis of their asymptotic properties. The necessary results are

$$\bar{x}_n \xrightarrow{a} N[\mu, \sigma^2/n], \quad \text{and} \quad M_n \xrightarrow{a} N[\mu, (\pi/2)\sigma^2/n]. \quad (\text{D-22})$$

Therefore, the mean is more efficient by a factor of  $\pi/2$ . (But, see Example 15.7 for a finite sample result.)

#### **D.3.1 ASYMPTOTIC DISTRIBUTION OF A NONLINEAR FUNCTION**

Theorems D.12 and D.14 for functions of a random variable have counterparts in asymptotic distributions.

### **THEOREM D.22 Asymptotic Distribution of a Nonlinear Function**

*If  $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N[0, \sigma^2]$  and if  $g(\theta)$  is a continuous and continuously differentiable function with  $g'(\theta)$  not equal to zero and not involving  $n$ , then  $g(\hat{\theta}_n) \xrightarrow{a} N[g(\theta), (1/n)\{g'(\theta)\}^2\sigma^2]$ .*

*If  $\hat{\theta}_n$  is a vector of parameter estimators such that  $\hat{\theta}_n \xrightarrow{a} N[\theta, (1/n)\mathbf{V}]$  and if  $\mathbf{c}(\theta)$  is a set of  $J$  continuous functions not involving  $n$ , then  $\mathbf{c}(\hat{\theta}_n) \xrightarrow{a} N[\mathbf{c}(\theta), (1/n)\mathbf{C}(\theta)\mathbf{V}\mathbf{C}(\theta)']$ , where  $\mathbf{C}(\theta) = \partial\mathbf{c}(\theta)/\partial\theta'$ .*

#### **Example D.9 Asymptotic Distribution of a Function of Two Estimators**

Suppose that  $b_n$  and  $t_n$  are estimators of parameters  $\beta$  and  $\theta$  such that

$$\begin{bmatrix} b_n \\ t_n \end{bmatrix} \xrightarrow{a} N\left[\begin{pmatrix} \beta \\ \theta \end{pmatrix}, \begin{pmatrix} \sigma_{\beta\beta} & \sigma_{\beta\theta} \\ \sigma_{\theta\beta} & \sigma_{\theta\theta} \end{pmatrix}\right].$$

Find the asymptotic distribution of  $c_n = b_n/(1-t_n)$ . Let  $\gamma = \beta/(1-\theta)$ . By the Slutsky theorem,  $c_n$  is consistent for  $\gamma$ . We shall require

$$\frac{\partial\gamma}{\partial\beta} = \frac{1}{1-\theta} = \gamma_\beta, \quad \frac{\partial\gamma}{\partial\theta} = \frac{\beta}{(1-\theta)^2} = \gamma_\theta.$$

## 1156 PART VI ♦ Appendices

Let  $\Sigma$  be the  $2 \times 2$  asymptotic covariance matrix given previously. Then the asymptotic variance of  $c_n$  is

$$\text{Asy. Var}[c_n] = (\gamma_\beta \ \gamma_\theta) \Sigma \begin{pmatrix} \gamma_\beta \\ \gamma_\theta \end{pmatrix} = \gamma_\beta^2 \sigma_{\beta\beta} + \gamma_\theta^2 \sigma_{\theta\theta} + 2\gamma_\beta \gamma_\theta \sigma_{\beta\theta},$$

which is the variance of the linear Taylor series approximation:

$$\hat{\gamma}_n \simeq \gamma + \gamma_\beta(b_n - \beta) + \gamma_\theta(t_n - \theta).$$

### D.3.2 ASYMPTOTIC EXPECTATIONS

The asymptotic mean and variance of a random variable are usually the mean and variance of the asymptotic distribution. Thus, for an estimator with the limiting distribution defined in

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N[\mathbf{0}, \mathbf{V}],$$

the asymptotic expectation is  $\theta$  and the asymptotic variance is  $(1/n)\mathbf{V}$ . This statement implies, among other things, that the estimator is “asymptotically unbiased.”

At the risk of clouding the issue a bit, it is necessary to reconsider one aspect of the previous description. We have deliberately avoided the use of consistency even though, in most instances, that is what we have in mind. The description thus far might suggest that consistency and asymptotic unbiasedness are the same. Unfortunately (because it is a source of some confusion), they are not. They are if the estimator is consistent and asymptotically normally distributed, or CAN. They may differ in other settings, however. There are at least three possible definitions of asymptotic unbiasedness:

1. The mean of the limiting distribution of  $\sqrt{n}(\hat{\theta}_n - \theta)$  is 0.
  2.  $\lim_{n \rightarrow \infty} E[\hat{\theta}_n] = \theta$ .
  3.  $\text{plim } \theta_n = \theta$ .
- (D-23)

In most cases encountered in practice, the estimator in hand will have all three properties, so there is no ambiguity. It is not difficult to construct cases in which the left-hand sides of all three definitions are different, however.<sup>6</sup> There is no general agreement among authors as to the precise meaning of asymptotic unbiasedness, perhaps because the term is misleading at the outset; *asymptotic* refers to an approximation, whereas *unbiasedness* is an exact result.<sup>7</sup> Nonetheless, the majority view seems to be that (2) is the proper definition of asymptotic unbiasedness.<sup>8</sup> Note, though, that this definition relies on quantities that are generally unknown and that may not exist.

A similar problem arises in the definition of the asymptotic variance of an estimator. One common definition is<sup>9</sup>

$$\text{Asy. Var}[\hat{\theta}_n] = \frac{1}{n} \lim_{n \rightarrow \infty} E \left[ \left\{ \sqrt{n}(\hat{\theta}_n - \lim_{n \rightarrow \infty} E[\hat{\theta}_n]) \right\}^2 \right]. \quad (\text{D-24})$$

---

<sup>6</sup>See, for example, Maddala (1977a, p. 150).

<sup>7</sup>See, for example, Theil (1971, p. 377).

<sup>8</sup>Many studies of estimators analyze the “asymptotic bias” of, say,  $\hat{\theta}_n$  as an estimator of a parameter  $\theta$ . In most cases, the quantity of interest is actually  $\text{plim } [\hat{\theta}_n - \theta]$ . See, for example, Greene (1980b) and another example in Johnston (1984, p. 312).

<sup>9</sup>Kmenta (1986, p.165).

## APPENDIX D ♦ Large-Sample Distribution Theory 1157

This result is a **leading term approximation**, and it will be sufficient for nearly all applications. Note, however, that like definition 2 of asymptotic unbiasedness, it relies on unknown and possibly nonexistent quantities.

### **Example D.10 Asymptotic Moments of the Sample Variance**

The exact expected value and variance of the variance estimator

$$m_2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (\text{D-25})$$

are

$$E[m_2] = \frac{(n-1)\sigma^2}{n}, \quad (\text{D-26})$$

and

$$\text{Var}[m_2] = \frac{\mu_4 - \sigma^4}{n} - \frac{2(\mu_4 - 2\sigma^4)}{n^2} + \frac{\mu_4 - 3\sigma^4}{n^3}, \quad (\text{D-27})$$

where  $\mu_4 = E[(x - \mu)^4]$ . [See Goldberger (1964, pp. 97–99).] The leading term approximation would be

$$\text{Asy. Var}[m_2] = \frac{1}{n}(\mu_4 - \sigma^4).$$

## D.4 SEQUENCES AND THE ORDER OF A SEQUENCE

This section has been concerned with sequences of constants, denoted, for example,  $c_n$ , and random variables, such as  $x_n$ , that are indexed by a sample size,  $n$ . An important characteristic of a sequence is the rate at which it converges (or diverges). For example, as we have seen, the mean of a random sample of  $n$  observations from a distribution with finite mean,  $\mu$ , and finite variance,  $\sigma^2$ , is itself a random variable with variance  $\gamma_n^2 = \sigma^2/n$ . We see that as long as  $\sigma^2$  is a finite constant,  $\gamma_n^2$  is a sequence of constants that converges to zero. Another example is the random variable  $x_{(1),n}$ , the minimum value in a random sample of  $n$  observations from the exponential distribution with mean  $1/\theta$  defined in Example C.4. It turns out that  $x_{(1),n}$  has variance  $1/(n\theta)^2$ . Clearly, this variance also converges to zero, but, intuition suggests, faster than  $\sigma^2/n$  does. On the other hand, the sum of the integers from one to  $n$ ,  $S_n = n(n+1)/2$ , obviously diverges as  $n \rightarrow \infty$ , albeit faster (one might expect) than the log of the likelihood function for the exponential distribution in Example C.6, which is  $\ln L(\theta) = n(\ln \theta - \theta \bar{x}_n)$ . As a final example, consider the downward bias of the maximum likelihood estimator of the variance of the normal distribution,  $c_n = (n-1)/n$ , which is a constant that converges to one. (See Example C.5.)

We will define the rate at which a sequence converges or diverges in terms of the order of the sequence.

### **DEFINITION D.14 Order $n^\delta$**

*A sequence  $c_n$  is of order  $n^\delta$ , denoted  $O(n^\delta)$ , if and only if  $\text{plim}(1/n^\delta)c_n$  is a finite nonzero constant.*

**1158 PART VI ♦ Appendices****DEFINITION D.15 Order less than  $n^\delta$** 

*A sequence  $c_n$ , is of order less than  $n^\delta$ , denoted  $o(n^\delta)$ , if and only if  $\text{plim}(1/n^\delta)c_n$  equals zero.*

Thus, in our examples,  $\gamma_n^2$  is  $O(n^{-1})$ ,  $\text{Var}[x_{(1),n}]$  is  $O(n^{-2})$  and  $o(n^{-1})$ ,  $S_n$  is  $O(n^2)$  ( $\delta$  equals +2 in this case),  $\ln L(\theta)$  is  $O(n)$  ( $\delta$  equals +1), and  $c_n$  is  $O(1)$  ( $\delta = 0$ ). Important particular cases that we will encounter repeatedly in our work are sequences for which  $\delta = 1$  or  $-1$ .

The notion of order of a sequence is often of interest in econometrics in the context of the variance of an estimator. Thus, we see in Section D.3 that an important element of our strategy for forming an asymptotic distribution is that the variance of the limiting distribution of  $\sqrt{n}(\bar{x}_n - \mu)/\sigma$  is  $O(1)$ . In Example D.10 the variance of  $m_2$  is the sum of three terms that are  $O(n^{-1})$ ,  $O(n^{-2})$ , and  $O(n^{-3})$ . The sum is  $O(n^{-1})$ , because  $n \text{Var}[m_2]$  converges to  $\mu_4 - \sigma^4$ , the numerator of the first, or *leading term*, whereas the second and third terms converge to zero. This term is also the *dominant term* of the sequence. Finally, consider the two divergent examples in the preceding list.  $S_n$  is simply a deterministic function of  $n$  that explodes. However,  $\ln L(\theta) = n \ln \theta - \theta \sum_i x_i$  is the sum of a constant that is  $O(n)$  and a random variable with variance equal to  $n/\theta$ . The random variable “diverges” in the sense that its variance grows without bound as  $n$  increases.

**APPENDIX E****COMPUTATION AND  
OPTIMIZATION****E.1 INTRODUCTION**

The computation of empirical estimates by econometricians involves using digital computers and software written either by the researchers themselves or by others.<sup>1</sup> It is also a surprisingly balanced mix of art and science. It is important for software users to be aware of how results are obtained, not only to understand routine computations, but also to be able to explain the occasional strange and contradictory results that do arise. This appendix will describe some of the basic elements of computing and a number of tools that are used by econometricians.<sup>2</sup> Section E.2

<sup>1</sup>It is one of the interesting aspects of the development of econometric methodology that the adoption of certain classes of techniques has proceeded in discrete jumps with the development of software. Noteworthy examples include the appearance, both around 1970, of G. K. Joreskog's LISREL [Joreskog and Sorbom (1981)] program, which spawned a still-growing industry in linear structural modeling, and TSP [Hall (1982)], which was among the first computer programs to accept symbolic representations of econometric models and which provided a significant advance in econometric practice with its LSQ procedure for systems of equations. An extensive survey of the evolution of econometric software is given in Renfro (2007).

<sup>2</sup>This discussion is not intended to teach the reader how to write computer programs. For those who expect to do so, there are whole libraries of useful sources. Three very useful works are Kennedy and Gentle (1980), Abramovitz and Stegun (1971), and especially Press et al. (1986). The third of these provides a wealth of expertly written programs and a large amount of information about how to do computation efficiently and accurately. A recent survey of many areas of computation is Judd (1998).

## APPENDIX E ♦ Computation and Optimization 1159

then describes some techniques for computing certain integrals and derivatives that are recurrent in econometric applications. Section E.3 presents methods of optimization of functions. Some examples are given in Section E.4.

### E.2 COMPUTATION IN ECONOMETRICS

This section will discuss some methods of computing integrals that appear frequently in econometrics.

#### E.2.1 COMPUTING INTEGRALS

One advantage of computers is their ability rapidly to compute approximations to complex functions such as logs and exponents. The basic functions, such as these, trigonometric functions, and so forth, are standard parts of the libraries of programs that accompany all scientific computing installations.<sup>3</sup> But one of the very common applications that often requires some high-level creativity by econometricians is the evaluation of integrals that do not have simple closed forms and that do not typically exist in “system libraries.” We will consider several of these in this section. We will not go into detail on the nuts and bolts of how to compute integrals with a computer; rather, we will turn directly to the most common applications in econometrics.

#### E.2.2 THE STANDARD NORMAL CUMULATIVE DISTRIBUTION FUNCTION

The standard normal cumulative distribution function (cdf) is ubiquitous in econometric models. Yet this most homely of applications must be computed by approximation. There are a number of ways to do so.<sup>4</sup> Recall that what we desire is

$$\Phi(x) = \int_{-\infty}^x \phi(t) dt, \quad \text{where } \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

One way to proceed is to use a Taylor series:

$$\Phi(x) \approx \sum_{i=0}^M \frac{1}{i!} \frac{d^i \Phi(x_0)}{dx_0^i} (x - x_0)^i.$$

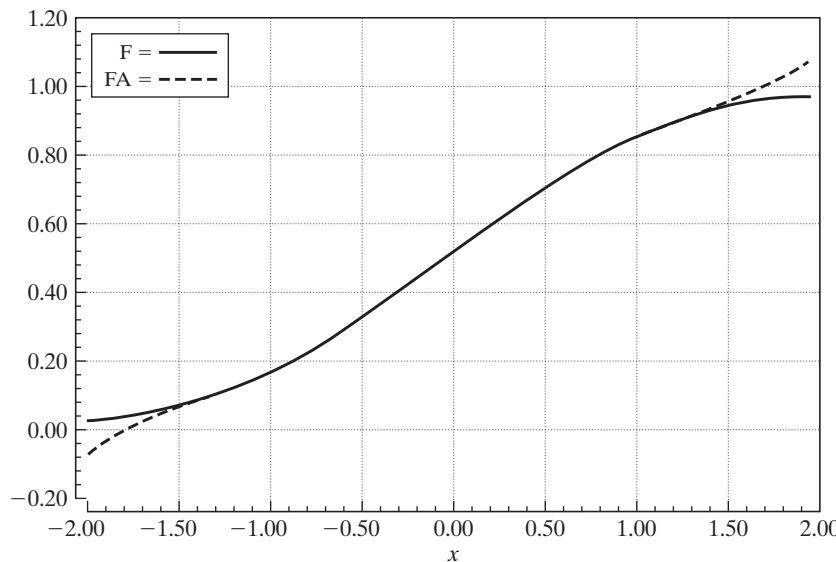
The normal cdf has some advantages for this approach. First, the derivatives are simple and not integrals. Second, the function is **analytic**; as  $M \rightarrow \infty$ , the approximation converges to the true value. Third, the derivatives have a simple form; they are the **Hermite polynomials** and they can be computed by a simple recursion. The 0th term in the preceding expansion is  $\Phi(x)$  evaluated at the expansion point. The first derivative of the cdf is the pdf, so the terms from 2 onward are the derivatives of  $\phi(x)$ , once again evaluated at  $x_0$ . The derivatives of the standard normal pdf obey the recursion

$$\phi^i / \phi(x) = -x \phi^{i-1} / \phi(x) - (i-1) \phi^{i-2} / \phi(x),$$

where  $\phi^i$  is  $d^i \phi(x) / dx^i$ . The zero and one terms in the sequence are one and  $-x$ . The next term is  $x^2 - 1$ , followed by  $3x - x^3$  and  $x^4 - 6x^2 + 3$ , and so on. The approximation can be made

<sup>3</sup>Of course, at some level, these must have been programmed as approximations by someone.

<sup>4</sup>Many system libraries provide a related function, the *error function*,  $\text{erf}(x) = (2/\sqrt{\pi}) \int_0^x e^{-t^2} dt$ . If this is available, then the normal cdf can be obtained from  $\Phi(x) = \frac{1}{2} + \frac{1}{2} \text{erf}(x/\sqrt{2})$ ,  $x \geq 0$  and  $\Phi(x) = 1 - \Phi(-x)$ ,  $x \leq 0$ .

**1160 PART VI ♦ Appendices**


**FIGURE E.1** Approximation to Normal cdf.

more accurate by adding terms. Consider using a fifth-order Taylor series approximation around the point  $x = 0$ , where  $\Phi(0) = 0.5$  and  $\phi(0) = 0.3989423$ . Evaluating the derivatives at zero and assembling the terms produces the approximation

$$\Phi(x) \approx \frac{1}{2} + 0.3989423[x - x^3/6 + x^5/40].$$

[Some of the terms (every other one, in fact) will conveniently drop out.] Figure E.1 shows the actual values ( $F$ ) and approximate values ( $FA$ ) over the range  $-2$  to  $2$ . The figure shows two important points. First, the approximation is remarkably good over most of the range. Second, as is usually true for Taylor series approximations, the quality of the approximation deteriorates as one gets far from the expansion point.

Unfortunately, it is the tail areas of the standard normal distribution that are usually of interest, so the preceding is likely to be problematic. An alternative approach that is used much more often is a polynomial approximation reported by Abramovitz and Stegun (1971, p. 932):

$$\Phi(-|x|) = \phi(x) \sum_{i=1}^5 a_i t^i + \varepsilon(x), \quad \text{where } t = 1/[1 + a_0|x|].$$

(The complement is taken if  $x$  is positive.) The error of approximation is less than  $\pm 7.5 \times 10^{-8}$  for all  $x$ . (Note that the error exceeds the function value at  $|x| > 5.7$ , so this is the operational limit of this approximation.)

### E.2.3 THE GAMMA AND RELATED FUNCTIONS

The standard normal cdf is probably the most common application of numerical integration of a function in econometrics. Another very common application is the class of gamma functions. For

## APPENDIX E ♦ Computation and Optimization 1161

positive constant  $P$ , the gamma function is

$$\Gamma(P) = \int_0^\infty t^{P-1} e^{-t} dt.$$

The gamma function obeys the recursion  $\Gamma(P) = (P - 1)\Gamma(P - 1)$ , so for integer values of  $P$ ,  $\Gamma(P) = (P - 1)!$  This result suggests that the gamma function can be viewed as a generalization of the factorial function for noninteger values. Another convenient value is  $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ . By making a change of variable, it can be shown that for positive constants  $a, c$ , and  $P$ ,

$$\int_0^\infty t^{P-1} e^{-at^c} dt = \int_0^\infty t^{-(P+1)} e^{-a/t^c} dt = \left(\frac{1}{c}\right) a^{-P/c} \Gamma\left(\frac{P}{c}\right). \quad (\text{E-1})$$

As a generalization of the factorial function, the gamma function will usually overflow for the sorts of values of  $P$  that normally appear in applications. The log of the function should normally be used instead. The function  $\ln \Gamma(P)$  can be approximated remarkably accurately with only a handful of terms and is very easy to program. A number of approximations appear in the literature; they are generally modifications of **Stirling's approximation** to the factorial function  $P! \approx (2\pi P)^{1/2} P^P e^{-P}$ , so

$$\ln \Gamma(P) \approx (P - 0.5)\ln P - P + 0.5 \ln(2\pi) + C + \varepsilon(P),$$

where  $C$  is the correction term [see, e.g., Abramovitz and Stegun (1971, p. 257), Press et al. (1986, p. 157), or Rao (1973, p. 59)] and  $\varepsilon(P)$  is the approximation error.<sup>5</sup>

The derivatives of the gamma function are

$$\frac{d^r \Gamma(P)}{dP^r} = \int_0^\infty (\ln t)^r t^{P-1} e^{-t} dt.$$

The first two derivatives of  $\ln \Gamma(P)$  are denoted  $\Psi(P) = \Gamma'/\Gamma$  and  $\Psi'(P) = (\Gamma\Gamma'' - \Gamma'^2)/\Gamma^2$  and are known as the **digamma** and **trigamma** functions.<sup>6</sup> The **beta function**, denoted  $\beta(a, b)$ ,

$$\beta(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

is related.

### E.2.4 APPROXIMATING INTEGRALS BY QUADRATURE

The digamma and trigamma functions, and the gamma function for noninteger values of  $P$  and values that are not integers plus  $\frac{1}{2}$ , do not exist in closed form and must be approximated. Most other applications will also involve integrals for which no simple computing function exists. The simplest approach to approximating

$$F(x) = \int_{L(x)}^{U(x)} f(t) dt$$

---

<sup>5</sup>For example, one widely used formula is  $C = z^{-1}/12 - z^{-3}/360 - z^{-5}/1260 + z^{-7}/1680 - q$ , where  $z = P$  and  $q = 0$  if  $P > 18$ , or  $z = P + J$  and  $q = \ln[P(P+1)(P+2)\cdots(P+J-1)]$ , where  $J = 18 - \text{INT}(P)$ , if not. Note, in the approximation, we write  $\Gamma(P) = (P!)/P + \text{a correction}$ .

<sup>6</sup>Tables of specific values for the gamma, digamma, and trigamma functions appear in Abramovitz and Stegun (1971). Most contemporary econometric programs have built-in functions for these common integrals, so the tables are not generally needed.

## 1162 PART VI ♦ Appendices

is likely to be a variant of Simpson's rule, or the trapezoid rule. For example, one approximation [see Press et al. (1986, p. 108)] is

$$F(x) \approx \Delta \left[ \frac{1}{3} f_1 + \frac{4}{3} f_2 + \frac{2}{3} f_3 + \frac{4}{3} f_4 + \cdots + \frac{2}{3} f_{N-2} + \frac{4}{3} f_{N-1} + \frac{1}{3} f_N \right],$$

where  $f_j$  is the function evaluated at  $N$  equally spaced points in  $[L, U]$  including the endpoints and  $\Delta = (U - L)/(N - 1)$ . There are a number of problems with this method, most notably that it is difficult to obtain satisfactory accuracy with a moderate number of points.

**Gaussian quadrature** is a popular method of computing integrals. The general approach is to use an approximation of the form

$$\int_L^U W(x) f(x) dx \approx \sum_{j=1}^M w_j f(a_j),$$

where  $W(x)$  is viewed as a “weighting” function for integrating  $f(x)$ ,  $w_j$  is the **quadrature weight**, and  $a_j$  is the **quadrature abscissa**. Different weights and abscissas have been derived for several weighting functions. Two weighting functions common in econometrics are

$$W(x) = x^c e^{-x}, \quad x \in [0, \infty),$$

for which the computation is called **Gauss–Laguerre quadrature**, and

$$W(x) = e^{-x^2}, \quad x \in (-\infty, \infty),$$

for which the computation is called **Gauss–Hermite quadrature**. The theory for deriving weights and abscissas is given in Press et al. (1986, pp. 121–125). Tables of weights and abscissas for many values of  $M$  are given by Abramovitz and Stegun (1971). Applications of the technique appear in Chapters 14 and 17.

### E.3 OPTIMIZATION

Nonlinear optimization (e.g., maximizing log-likelihood functions) is an intriguing practical problem. Theory provides few hard and fast rules, and there are relatively few cases in which it is obvious how to proceed. This section introduces some of the terminology and underlying theory of nonlinear optimization.<sup>7</sup> We begin with a general discussion on how to search for a solution to a nonlinear optimization problem and describe some specific commonly used methods. We then consider some practical problems that arise in optimization. An example is given in the final section.

Consider maximizing the quadratic function

$$F(\boldsymbol{\theta}) = a + \mathbf{b}'\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}'\mathbf{C}\boldsymbol{\theta},$$

where  $\mathbf{C}$  is a positive definite matrix. The first-order condition for a maximum is

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{b} - \mathbf{C}\boldsymbol{\theta} = \mathbf{0}. \tag{E-2}$$

This set of *linear* equations has the unique solution

$$\boldsymbol{\theta} = \mathbf{C}^{-1}\mathbf{b}. \tag{E-3}$$

---

<sup>7</sup>There are numerous excellent references that offer a more complete exposition. Among these are Quandt (1983), Bazaraa and Shetty (1979), Fletcher (1980), and Judd (1998).

## APPENDIX E ♦ Computation and Optimization 1163

This is a linear optimization problem. Note that it has a **closed-form solution**; for any  $a$ ,  $\mathbf{b}$ , and  $\mathbf{C}$ , the solution can be computed directly.<sup>8</sup> In the more typical situation,

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \quad (\text{E-4})$$

is a set of nonlinear equations that cannot be solved explicitly for  $\boldsymbol{\theta}$ .<sup>9</sup> The techniques considered in this section provide systematic means of searching for a solution.

We now consider the general problem of maximizing a function of several variables:

$$\text{maximize}_{\boldsymbol{\theta}} F(\boldsymbol{\theta}), \quad (\text{E-5})$$

where  $F(\boldsymbol{\theta})$  may be a log-likelihood or some other function. Minimization of  $F(\boldsymbol{\theta})$  is handled by maximizing  $-F(\boldsymbol{\theta})$ . Two special cases are

$$F(\boldsymbol{\theta}) = \sum_{i=1}^n f_i(\boldsymbol{\theta}), \quad (\text{E-6})$$

which is typical for maximum likelihood problems, and the **least squares problem**,<sup>10</sup>

$$f_i(\boldsymbol{\theta}) = -(y_i - f(\mathbf{x}_i, \boldsymbol{\theta}))^2. \quad (\text{E-7})$$

We treated the nonlinear least squares problem in detail in Chapter 7. An obvious way to search for the  $\boldsymbol{\theta}$  that maximizes  $F(\boldsymbol{\theta})$  is by trial and error. If  $\boldsymbol{\theta}$  has only a single element and it is known approximately where the optimum will be found, then a **grid search** will be a feasible strategy. An example is a common time-series problem in which a one-dimensional search for a correlation coefficient is made in the interval  $(-1, 1)$ . The grid search can proceed in the obvious fashion—that is,  $\dots, -0.1, 0, 0.1, 0.2, \dots$ , then  $\hat{\theta}_{\max} - 0.1$  to  $\hat{\theta}_{\max} + 0.1$  in increments of 0.01, and so on—until the desired precision is achieved.<sup>11</sup> If  $\boldsymbol{\theta}$  contains more than one parameter, then a grid search is likely to be extremely costly, particularly if little is known about the parameter vector at the outset. Nonetheless, relatively efficient methods have been devised. Quandt (1983) and Fletcher (1980) contain further details.

There are also systematic, derivative-free methods of searching for a function optimum that resemble in some respects the algorithms that we will examine in the next section. The **downhill simplex** (and other simplex) methods<sup>12</sup> have been found to be very fast and effective for some problems. A recent entry in the econometrics literature is the method of **simulated annealing**.<sup>13</sup> These derivative-free methods, particularly the latter, are often very effective in problems with many variables in the objective function, but they usually require far more function evaluations than the methods based on derivatives that are considered below. Because the problems typically analyzed in econometrics involve relatively few parameters but often quite complex functions involving large numbers of terms in a summation, on balance, the gradient methods are usually going to be preferable.<sup>14</sup>

<sup>8</sup>Notice that the constant  $a$  is irrelevant to the solution. Many maximum likelihood problems are presented with the preface “neglecting an irrelevant constant.” For example, the log-likelihood for the normal linear regression model contains a term  $-(n/2) \ln(2\pi)$ —that can be discarded.

<sup>9</sup>See, for example, the normal equations for the nonlinear least squares estimators of Chapter 7.

<sup>10</sup>Least squares is, of course, a minimization problem. The negative of the criterion is used to maintain consistency with the general formulation.

<sup>11</sup>There are more efficient methods of carrying out a one-dimensional search, for example, the **golden section** method. See Press et al. (1986, Chap. 10).

<sup>12</sup>See Nelder and Mead (1965) and Press et al. (1986).

<sup>13</sup>See Goffe, Ferrier, and Rodgers (1994) and Press et al. (1986, pp. 326–334).

<sup>14</sup>Goffe, Ferrier, and Rodgers (1994) did find that the method of simulated annealing was quite adept at finding the best among multiple solutions. This problem is common for derivative-based methods, because they usually have no method of distinguishing between a local optimum and a global one.

## 1164 PART VI ♦ Appendices

### E.3.1 ALGORITHMS

A more effective means of solving most nonlinear maximization problems is by an **iterative algorithm**:

Beginning from initial value  $\theta_0$ , at entry to iteration  $t$ , if  $\theta_t$  is not the optimal value for  $\theta$ , compute direction vector  $\Delta_t$ , step size  $\lambda_t$ , then

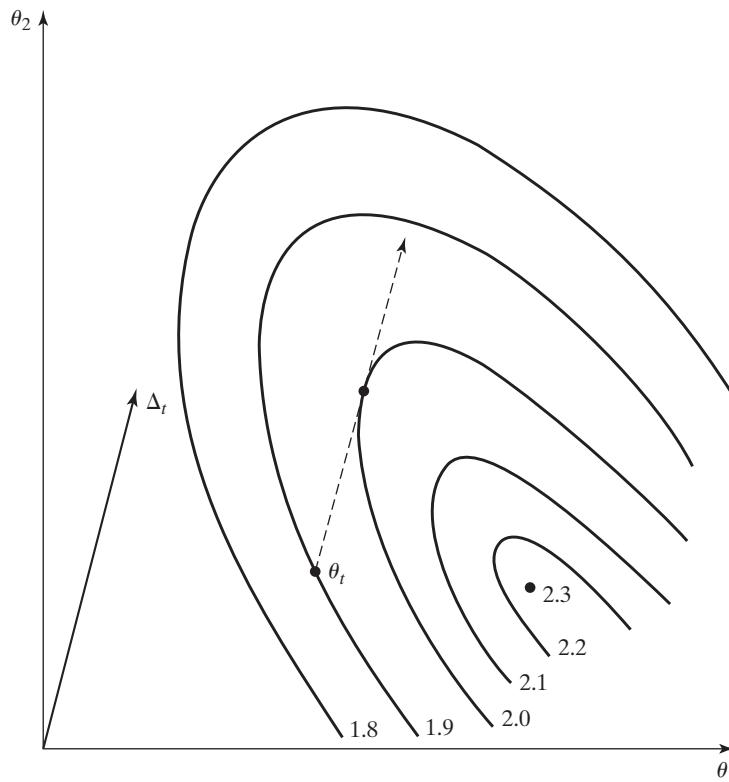
$$\theta_{t+1} = \theta_t + \lambda_t \Delta_t. \quad (\text{E-8})$$

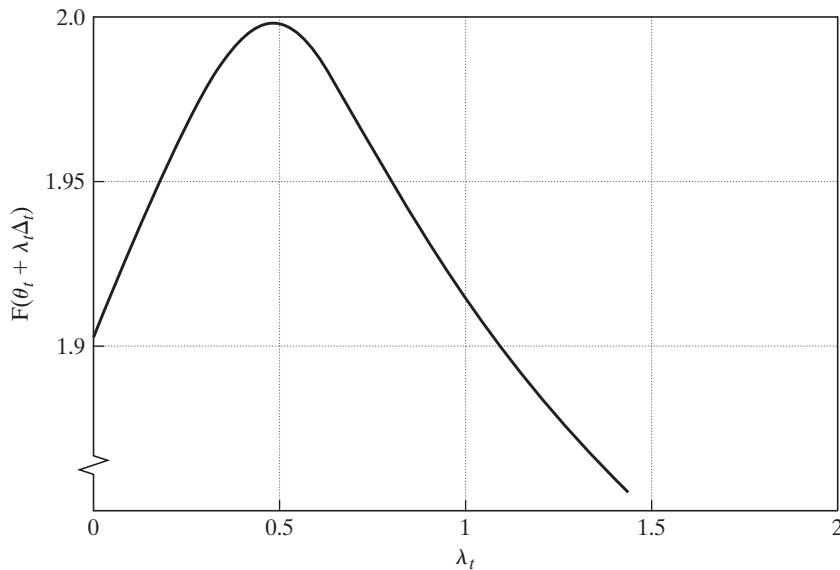
Figure E.2 illustrates the structure of an iteration for a hypothetical function of two variables. The direction vector  $\Delta_t$  is shown in the figure with  $\theta_t$ . The dashed line is the set of points  $\theta_t + \lambda_t \Delta_t$ . Different values of  $\lambda_t$  lead to different contours; for this  $\theta_t$  and  $\Delta_t$ , the best value of  $\lambda_t$  is about 0.5.

Notice in Figure E.2 that for a given direction vector  $\Delta_t$  and current parameter vector  $\theta_t$ , a secondary optimization is required to find the best  $\lambda_t$ . Translating from Figure E.2, we obtain the form of this problem as shown in Figure E.3. This subsidiary search is called a **line search**, as we search along the line  $\theta_t + \lambda_t \Delta_t$  for the optimal value of  $F(\cdot)$ . The formal solution to the line search problem would be the  $\lambda_t$  that satisfies

$$\frac{\partial F(\theta_t + \lambda_t \Delta_t)}{\partial \lambda_t} = \mathbf{g}(\theta_t + \lambda_t \Delta_t)' \Delta_t = 0, \quad (\text{E-9})$$

**FIGURE E.2** Iteration.





**FIGURE E.3** Line Search.

where  $\mathbf{g}$  is the vector of partial derivatives of  $F(\cdot)$  evaluated at  $\boldsymbol{\theta}_t + \lambda_t \Delta_t$ . In general, this problem will also be a nonlinear one. In most cases, adding a formal search for  $\lambda_t$  will be too expensive, as well as unnecessary. Some approximate or ad hoc method will usually be chosen. It is worth emphasizing that finding the  $\lambda_t$  that maximizes  $F(\boldsymbol{\theta}_t + \lambda_t \Delta_t)$  at a given iteration does not generally lead to the overall solution in that iteration. This situation is clear in Figure E.3, where the optimal value of  $\lambda_t$  leads to  $F(\cdot) = 2.0$ , at which point we reenter the iteration.

### E.3.2 COMPUTING DERIVATIVES

For certain functions, the programming of derivatives may be quite difficult. Numeric approximations can be used, although it should be borne in mind that analytic derivatives obtained by formally differentiating the functions involved are to be preferred. First derivatives can be approximated by using

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \theta_i} \approx \frac{F(\cdots \theta_i + \varepsilon \cdots) - F(\cdots \theta_i - \varepsilon \cdots)}{2\varepsilon}.$$

The choice of  $\varepsilon$  is a remaining problem. Extensive discussion may be found in Quandt (1983).

There are three drawbacks to this means of computing derivatives compared with using the analytic derivatives. A possible major consideration is that it may substantially increase the amount of computation needed to obtain a function and its gradient. In particular,  $K+1$  function evaluations (the criterion and  $K$  derivatives) are replaced with  $2K+1$  functions. The latter may be more burdensome than the former, depending on the complexity of the partial derivatives compared with the function itself. The comparison will depend on the application. But in most settings, careful programming that avoids superfluous or redundant calculation can make the advantage of the analytic derivatives substantial. Second, the choice of  $\varepsilon$  can be problematic. If it is chosen too large, then the approximation will be inaccurate. If it is chosen too small, then there may be insufficient variation in the function to produce a good estimate of the derivative.

## 1166 PART VI ♦ Appendices

A compromise that is likely to be effective is to compute  $\varepsilon_i$  separately for each parameter, as in

$$\varepsilon_i = \text{Max}[\alpha|\theta_i|, \gamma]$$

[see Goldfeld and Quandt (1971)]. The values  $\alpha$  and  $\gamma$  should be relatively small, such as  $10^{-5}$ . Third, although numeric derivatives computed in this fashion are likely to be reasonably accurate, in a sum of a large number of terms, say, several thousand, enough approximation error can accumulate to cause the numerical derivatives to differ significantly from their analytic counterparts. Second derivatives can also be computed numerically. In addition to the preceding problems, however, it is generally not possible to ensure negative definiteness of a Hessian computed in this manner. Unless the choice of  $\varepsilon$  is made extremely carefully, an indefinite matrix is a possibility. In general, the use of numeric derivatives should be avoided if the analytic derivatives are available.

### E.3.3 GRADIENT METHODS

The most commonly used algorithms are **gradient methods**, in which

$$\Delta_t = \mathbf{W}_t \mathbf{g}_t, \quad (\text{E-10})$$

where  $\mathbf{W}_t$  is a positive definite matrix and  $\mathbf{g}_t$  is the **gradient** of  $F(\theta_t)$ :

$$\mathbf{g}_t = \mathbf{g}(\theta_t) = \frac{\partial F(\theta_t)}{\partial \theta_t}. \quad (\text{E-11})$$

These methods are motivated partly by the following. Consider a linear Taylor series approximation to  $F(\theta_t + \lambda_t \Delta_t)$  around  $\lambda_t = 0$ :

$$F(\theta_t + \lambda_t \Delta_t) \simeq F(\theta_t) + \lambda_t \mathbf{g}(\theta_t)' \Delta_t. \quad (\text{E-12})$$

Let  $F(\theta_t + \lambda_t \Delta_t)$  equal  $F_{t+1}$ . Then,

$$F_{t+1} - F_t \simeq \lambda_t \mathbf{g}_t' \Delta_t.$$

If  $\Delta_t = \mathbf{W}_t \mathbf{g}_t$ , then

$$F_{t+1} - F_t \simeq \lambda_t \mathbf{g}_t' \mathbf{W}_t \mathbf{g}_t.$$

If  $\mathbf{g}_t$  is not  $\mathbf{0}$  and  $\lambda_t$  is small enough, then  $F_{t+1} - F_t$  must be positive. Thus, if  $F(\theta)$  is not already at its maximum, then we can always find a step size such that a gradient-type iteration will lead to an increase in the function. (Recall that  $\mathbf{W}_t$  is assumed to be positive definite.)

In the following, we will omit the iteration index  $t$ , except where it is necessary to distinguish one vector from another. The following are some commonly used algorithms.<sup>15</sup>

**Steepest Ascent** The simplest algorithm to employ is the **steepest ascent** method, which uses

$$\mathbf{W} = \mathbf{I} \text{ so that } \Delta = \mathbf{g}. \quad (\text{E-13})$$

As its name implies, the direction is the one of greatest increase of  $F(\cdot)$ . Another virtue is that the line search has a straightforward solution; at least near the maximum, the optimal  $\lambda$  is

$$\lambda = \frac{-\mathbf{g}' \mathbf{g}}{\mathbf{g}' \mathbf{H} \mathbf{g}}, \quad (\text{E-14})$$

---

<sup>15</sup>A more extensive catalog may be found in Judge et al. (1985, Appendix B). Those mentioned here are some of the more commonly used ones and are chosen primarily because they illustrate many of the important aspects of nonlinear optimization.

## APPENDIX E ♦ Computation and Optimization 1167

where

$$\mathbf{H} = \frac{\partial^2 F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'},$$

Therefore, the steepest ascent iteration is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\mathbf{g}' \mathbf{g}_t}{\mathbf{g}' \mathbf{H}_t \mathbf{g}_t} \mathbf{g}_t. \quad (\text{E-15})$$

Computation of the second derivatives matrix may be extremely burdensome. Also, if  $\mathbf{H}_t$  is not negative definite, which is likely if  $\boldsymbol{\theta}_t$  is far from the maximum, the iteration may diverge. A systematic line search can bypass this problem. This algorithm usually converges very slowly, however, so other techniques are usually used.

**Newton's Method** The template for most gradient methods in common use is Newton's method. The basis for **Newton's method** is a linear Taylor series approximation. Expanding the first-order conditions,

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0},$$

equation by equation, in a linear Taylor series around an arbitrary  $\boldsymbol{\theta}^0$  yields

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \simeq \mathbf{g}^0 + \mathbf{H}^0(\boldsymbol{\theta} - \boldsymbol{\theta}^0) = \mathbf{0}, \quad (\text{E-16})$$

where the superscript indicates that the term is evaluated at  $\boldsymbol{\theta}^0$ . Solving for  $\boldsymbol{\theta}$  and then equating  $\boldsymbol{\theta}$  to  $\boldsymbol{\theta}_{t+1}$  and  $\boldsymbol{\theta}^0$  to  $\boldsymbol{\theta}_t$ , we obtain the iteration

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{H}_t^{-1} \mathbf{g}_t. \quad (\text{E-17})$$

Thus, for Newton's method,

$$\mathbf{W} = -\mathbf{H}^{-1}, \quad \Delta = -\mathbf{H}^{-1} \mathbf{g}, \quad \lambda = 1. \quad (\text{E-18})$$

Newton's method will converge very rapidly in many problems. If the function is quadratic, then this method will reach the optimum in one iteration from any starting point. If the criterion function is globally concave, as it is in a number of problems that we shall examine in this text, then it is probably the best algorithm available. This method is very well suited to maximum likelihood estimation.

**Alternatives to Newton's Method** Newton's method is very effective in some settings, but it can perform very poorly in others. If the function is not approximately quadratic or if the current estimate is very far from the maximum, then it can cause wide swings in the estimates and even fail to converge at all. A number of algorithms have been devised to improve upon Newton's method. An obvious one is to include a line search at each iteration rather than use  $\lambda = 1$ . Two problems remain, however. At points distant from the optimum, the second derivatives matrix may not be negative definite, and, in any event, the computational burden of computing  $\mathbf{H}$  may be excessive.

The **quadratic hill-climbing method** proposed by Goldfeld, Quandt, and Trotter (1966) deals directly with the first of these problems. In any iteration, if  $\mathbf{H}$  is not negative definite, then it is replaced with

$$\mathbf{H}_\alpha = \mathbf{H} - \alpha \mathbf{I}, \quad (\text{E-19})$$

## 1168 PART VI ♦ Appendices

where  $\alpha$  is a positive number chosen large enough to ensure the negative definiteness of  $\mathbf{H}_\alpha$ . Another suggestion is that of Greenstadt (1967), which uses, at every iteration,

$$\mathbf{H}_\pi = - \sum_{i=1}^n |\pi_i| \mathbf{c}_i \mathbf{c}'_i, \quad (\text{E-20})$$

where  $\pi_i$  is the  $i$ th characteristic root of  $\mathbf{H}$  and  $\mathbf{c}_i$  is its associated characteristic vector. Other proposals have been made to ensure the negative definiteness of the required matrix at each iteration.<sup>16</sup>

**Quasi-Newton Methods: Davidon–Fletcher–Powell** A very effective class of algorithms has been developed that eliminates second derivatives altogether and has excellent convergence properties, even for ill-behaved problems. These are the **quasi-Newton methods**, which form

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mathbf{E}_t,$$

where  $\mathbf{E}_t$  is a positive definite matrix.<sup>17</sup> As long as  $\mathbf{W}_0$  is positive definite— $\mathbf{I}$  is commonly used— $\mathbf{W}_t$  will be positive definite at every iteration. In the **Davidon–Fletcher–Powell (DFP) method**, after a sufficient number of iterations,  $\mathbf{W}_{t+1}$  will be an approximation to  $-\mathbf{H}^{-1}$ . Let

$$\delta_t = \lambda_t \Delta_t, \quad \text{and} \quad \gamma_t = \mathbf{g}(\theta_{t+1}) - \mathbf{g}(\theta_t). \quad (\text{E-21})$$

The DFP **variable metric algorithm** uses

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \frac{\delta_t \delta'_t}{\delta'_t \gamma_t} + \frac{\mathbf{W}_t \gamma_t \gamma'_t \mathbf{W}_t}{\gamma'_t \mathbf{W}_t \gamma_t}. \quad (\text{E-22})$$

Notice that in the DFP algorithm, the change in the first derivative vector is used in  $\mathbf{W}$ ; an estimate of the inverse of the second derivatives matrix is being accumulated.

The variable metric algorithms are those that update  $\mathbf{W}$  at each iteration while preserving its definiteness. For the DFP method, the accumulation of  $\mathbf{W}_{t+1}$  is of the form

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mathbf{a}\mathbf{a}' + \mathbf{b}\mathbf{b}' = \mathbf{W}_t + [\mathbf{a} \quad \mathbf{b}][\mathbf{a} \quad \mathbf{b}]'.$$

The two-column matrix  $[\mathbf{a} \quad \mathbf{b}]$  will have rank two; hence, DFP is called a **rank two update** or **rank two correction**. The **Broyden–Fletcher–Goldfarb–Shanno (BFGS)** method is a rank three correction that subtracts  $v\mathbf{d}\mathbf{d}'$  from the **DFP** update, where  $v = (\gamma'_t \mathbf{W}_t \gamma_t)$  and

$$\mathbf{d}_t = \left( \frac{1}{\delta'_t \gamma_t} \right) \delta_t - \left( \frac{1}{\gamma'_t \mathbf{W}_t \gamma_t} \right) \mathbf{W}_t \gamma_t.$$

There is some evidence that this method is more efficient than DFP. Other methods, such as **Broyden's method**, involve a rank one correction instead. Any method that is of the form

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mathbf{Q}\mathbf{Q}'$$

will preserve the definiteness of  $\mathbf{W}$  regardless of the number of columns in  $\mathbf{Q}$ .

The DFP and BFGS algorithms are extremely effective and are among the most widely used of the gradient methods. An important practical consideration to keep in mind is that although  $\mathbf{W}_t$  accumulates an estimate of the negative inverse of the second derivatives matrix for both algorithms, in maximum likelihood problems it rarely converges to a very good estimate of the covariance matrix of the estimator and should generally not be used as one.

<sup>16</sup>See, for example, Goldfeld and Quandt (1971).

<sup>17</sup>See Fletcher (1980).

### E.3.4 ASPECTS OF MAXIMUM LIKELIHOOD ESTIMATION

Newton's method is often used for maximum likelihood problems. For solving a maximum likelihood problem, the **method of scoring** replaces  $\mathbf{H}$  with

$$\bar{\mathbf{H}} = E[\mathbf{H}(\boldsymbol{\theta})], \quad (\text{E-23})$$

which will be recognized as the asymptotic covariance of the maximum likelihood estimator. There is some evidence that where it can be used, this method performs better than Newton's method. The exact form of the expectation of the Hessian of the log likelihood is rarely known, however.<sup>18</sup> Newton's method, which uses actual instead of expected second derivatives, is generally used instead.

**One-Step Estimation** A convenient variant of Newton's method is the **one-step maximum likelihood estimator**. It has been shown that if  $\boldsymbol{\theta}^0$  is *any* consistent initial estimator of  $\boldsymbol{\theta}$  and  $\mathbf{H}^*$  is  $\mathbf{H}$ ,  $\bar{\mathbf{H}}$ , or any other asymptotically equivalent estimator of  $\text{Var}[\mathbf{g}(\hat{\boldsymbol{\theta}}_{\text{MLE}})]$ , then

$$\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 - (\mathbf{H}^*)^{-1} \mathbf{g}^0 \quad (\text{E-24})$$

is an estimator of  $\boldsymbol{\theta}$  that has the same asymptotic properties as the maximum likelihood estimator.<sup>19</sup> (Note that it is *not* the maximum likelihood estimator. As such, for example, it should not be used as the basis for likelihood ratio tests.)

**Covariance Matrix Estimation** In computing maximum likelihood estimators, a commonly used method of estimating  $\mathbf{H}$  simultaneously simplifies the calculation of  $\mathbf{W}$  and solves the occasional problem of indefiniteness of the Hessian. The method of Berndt et al. (1974) replaces  $\mathbf{W}$  with

$$\hat{\mathbf{W}} = \left[ \sum_{i=1}^n \mathbf{g}_i \mathbf{g}'_i \right]^{-1} = (\mathbf{G}' \mathbf{G})^{-1}, \quad (\text{E-25})$$

where

$$\mathbf{g}_i = \frac{\partial \ln f(y_i | \mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \quad (\text{E-26})$$

Then,  $\mathbf{G}$  is the  $n \times K$  matrix with  $i$ th row equal to  $\mathbf{g}'_i$ . Although  $\hat{\mathbf{W}}$  and other suggested estimators of  $(-\mathbf{H})^{-1}$  are asymptotically equivalent,  $\hat{\mathbf{W}}$  has the additional virtues that it is always nonnegative definite, and it is only necessary to differentiate the log-likelihood once to compute it.

**The Lagrange Multiplier Statistic** The use of  $\hat{\mathbf{W}}$  as an estimator of  $(-\mathbf{H})^{-1}$  brings another intriguing convenience in maximum likelihood estimation. When testing restrictions on parameters estimated by maximum likelihood, one approach is to use the **Lagrange multiplier** statistic. We will examine this test at length at various points in this book, so we need only sketch it briefly here. The logic of the LM test is as follows. The gradient  $\mathbf{g}(\boldsymbol{\theta})$  of the log-likelihood function equals  $\mathbf{0}$  at the unrestricted maximum likelihood estimators (that is, at least to within the precision of the computer program in use). If  $\hat{\boldsymbol{\theta}}_r$  is an MLE that is computed subject to some restrictions on  $\boldsymbol{\theta}$ , then we know that  $\mathbf{g}(\hat{\boldsymbol{\theta}}_r) \neq \mathbf{0}$ . The LM test is used to test whether, at  $\hat{\boldsymbol{\theta}}_r$ ,  $\mathbf{g}_r$  is significantly different from  $\mathbf{0}$  or whether the deviation of  $\mathbf{g}_r$  from  $\mathbf{0}$  can be viewed as sampling variation. The covariance matrix of the gradient of the log-likelihood is  $-\mathbf{H}$ , so the Wald statistic for testing this hypothesis is  $W = \mathbf{g}'(-\mathbf{H})^{-1} \mathbf{g}$ . Now, suppose that we use  $\hat{\mathbf{W}}$  to estimate  $-\mathbf{H}^{-1}$ . Let  $\mathbf{G}$  be the  $n \times K$  matrix with  $i$ th row equal to  $\mathbf{g}'_i$ , and let  $\mathbf{i}$  denote an  $n \times 1$  column of ones. Then the LM statistic can be

<sup>18</sup>Amemiya (1981) provides a number of examples.

<sup>19</sup>See, for example, Rao (1973).

## 1170 PART VI ♦ Appendices

computed as

$$\text{LM} = \mathbf{i}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{i}.$$

Because  $\mathbf{i}'\mathbf{i} = n$ ,

$$\text{LM} = n[\mathbf{i}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{i}/n] = nR_i^2,$$

where  $R_i^2$  is the *uncentered R*<sup>2</sup> in a regression of a column of ones on the derivatives of the log-likelihood function.

**The Concentrated Log-Likelihood** Many problems in maximum likelihood estimation can be formulated in terms of a partitioning of the parameter vector  $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]$  such that at the solution to the optimization problem,  $\boldsymbol{\theta}_{2,\text{ML}}$ , can be written as an explicit function of  $\boldsymbol{\theta}_{1,\text{ML}}$ . When the solution to the likelihood equation for  $\boldsymbol{\theta}_2$  produces

$$\boldsymbol{\theta}_{2,\text{ML}} = \mathbf{t}(\boldsymbol{\theta}_{1,\text{ML}}),$$

then, if it is convenient, we may “concentrate” the log-likelihood function by writing

$$F^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = F[\boldsymbol{\theta}_1, \mathbf{t}(\boldsymbol{\theta}_1)] = F_c(\boldsymbol{\theta}_1).$$

The unrestricted solution to the problem  $\text{Max}_{\boldsymbol{\theta}_1} F_c(\boldsymbol{\theta}_1)$  provides the full solution to the optimization problem. Once the optimizing value of  $\boldsymbol{\theta}_1$  is obtained, the optimizing value of  $\boldsymbol{\theta}_2$  is simply  $\mathbf{t}(\hat{\boldsymbol{\theta}}_{1,\text{ML}})$ . Note that  $F^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$  is a subset of the set of values of the log-likelihood function, namely those values at which the second parameter vector satisfies the first-order conditions.<sup>20</sup>

### E.3.5 OPTIMIZATION WITH CONSTRAINTS

Occasionally, some or all the parameters of a model are constrained, for example, to be positive in the case of a variance or to be in a certain range, such as a correlation coefficient. Optimization subject to constraints is often yet another art form. The elaborate literature on the general problem provides some guidance—see, for example, Appendix B in Judge et al. (1985)—but applications still, as often as not, require some creativity on the part of the analyst. In this section, we will examine a few of the most common forms of constrained optimization as they arise in econometrics.

Parametric constraints typically come in two forms, which may occur simultaneously in a problem. Equality constraints can be written  $\mathbf{c}(\boldsymbol{\theta}) = \mathbf{0}$ , where  $c_j(\boldsymbol{\theta})$  is a continuous and differentiable function. Typical applications include linear constraints on slope vectors, such as a requirement that a set of elasticities in a log-linear model add to one; exclusion restrictions, which are often cast in the form of interesting hypotheses about whether or not a variable should appear in a model (i.e., whether a coefficient is zero or not); and equality restrictions, such as the symmetry restrictions in a translog model, which require that parameters in two different equations be equal to each other. Inequality constraints, in general, will be of the form  $a_j \leq c_j(\boldsymbol{\theta}) \leq b_j$ , where  $a_j$  and  $b_j$  are known constants (either of which may be infinite). Once again, the typical application in econometrics involves a restriction on a single parameter, such as  $\sigma > 0$  for a variance parameter,  $-1 \leq \rho \leq 1$  for a correlation coefficient, or  $\beta_j \geq 0$  for a particular slope coefficient in a model. We will consider the two cases separately.

In the case of equality constraints, for practical purposes of optimization, there are usually two strategies available. One can use a Lagrangean multiplier approach. The new optimization problem is

$$\text{Max}_{\boldsymbol{\theta}, \lambda} L(\boldsymbol{\theta}, \lambda) = F(\boldsymbol{\theta}) + \lambda' \mathbf{c}(\boldsymbol{\theta}).$$

---

<sup>20</sup>A formal proof that this is a valid way to proceed is given by Amemiya (1985, pp. 125–127).

## APPENDIX E ♦ Computation and Optimization 1171

The necessary conditions for an optimum are

$$\frac{\partial L(\theta, \lambda)}{\partial \theta} = \mathbf{g}(\theta) + \mathbf{C}(\theta)' \lambda = \mathbf{0},$$

$$\frac{\partial L(\theta, \lambda)}{\partial \lambda} = \mathbf{c}(\theta) = \mathbf{0},$$

where  $\mathbf{g}(\theta)$  is the familiar gradient of  $F(\theta)$  and  $\mathbf{C}(\theta)$  is a  $J \times K$  matrix of derivatives with  $j$ th row equal to  $\partial c_j / \partial \theta'$ . The joint solution will provide the constrained optimizer, as well as the Lagrange multipliers, which are often interesting in their own right. The disadvantage of this approach is that it increases the dimensionality of the optimization problem. An alternative strategy is to eliminate some of the parameters by either imposing the constraints directly on the function or by solving out the constraints. For exclusion restrictions, which are usually of the form  $\theta_j = 0$ , this step usually means dropping a variable from a model. Other restrictions can often be imposed just by building them into the model. For example, in a function of  $\theta_1$ ,  $\theta_2$ , and  $\theta_3$ , if the restriction is of the form  $\theta_3 = \theta_1 \theta_2$ , then  $\theta_3$  can be eliminated from the model by a direct substitution.

Inequality constraints are more difficult. For the general case, one suggestion is to transform the constrained problem into an unconstrained one by imposing some sort of penalty function into the optimization criterion that will cause a parameter vector that violates the constraints, or nearly does so, to be an unattractive choice. For example, to force a parameter  $\theta_j$  to be nonzero, one might maximize the augmented function  $F(\theta) - |1/\theta_j|$ . This approach is feasible, but it has the disadvantage that because the penalty is a function of the parameters, different penalty functions will lead to different solutions of the optimization problem. For the most common problems in econometrics, a simpler approach will usually suffice. One can often reparameterize a function so that the new parameter is unconstrained. For example, the “method of squaring” is sometimes used to force a parameter to be positive. If we require  $\theta_j$  to be positive, then we can define  $\theta_j = \alpha^2$  and substitute  $\alpha^2$  for  $\theta_j$  wherever it appears in the model. Then an unconstrained solution for  $\alpha$  is obtained. An alternative reparameterization for a parameter that must be positive that is often used is  $\theta_j = \exp(\alpha)$ . To force a parameter to be between zero and one, we can use the function  $\theta_j = 1/[1 + \exp(\alpha)]$ . The range of  $\alpha$  is now unrestricted. Experience suggests that a third, less orthodox approach works very well for many problems. When the constrained optimization is begun, there is a starting value  $\theta^0$  that begins the iterations. Presumably,  $\theta^0$  obeys the restrictions. (If not, and none can be found, then the optimization process must be terminated immediately.) The next iterate,  $\theta^1$ , is a step away from  $\theta^0$ , by  $\theta^1 = \theta^0 + \lambda_0 \delta^0$ . Suppose that  $\theta^1$  violates the constraints. By construction, we know that there is some value  $\theta_*^1$  between  $\theta^0$  and  $\theta^1$  that does not violate the constraint, where “between” means only that a shorter step is taken. Therefore, the next value for the iteration can be  $\theta_*^1$ . The logic is true at every iteration, so a way to proceed is to alter the iteration so that the step length is shortened when necessary when a parameter violates the constraints.

### E.3.6 SOME PRACTICAL CONSIDERATIONS

The reasons for the good performance of many algorithms, including DFP, are unknown. Moreover, different algorithms may perform differently in given settings. Indeed, for some problems, one algorithm may fail to converge whereas another will succeed in finding a solution without great difficulty. In view of this, computer programs such as GQOPT,<sup>21</sup> Gauss, and MatLab that offer a menu of different preprogrammed algorithms can be particularly useful. It is sometimes worth the effort to try more than one algorithm on a given problem.

---

<sup>21</sup>Goldfeld and Quandt (1972).

## 1172 PART VI ♦ Appendices

**Step Sizes** Except for the steepest ascent case, an optimal line search is likely to be infeasible or to require more effort than it is worth in view of the potentially large number of function evaluations required. In most cases, the choice of a step size is likely to be rather ad hoc. But within limits, the most widely used algorithms appear to be robust to inaccurate line searches. For example, one method employed by the widely used TSP computer program<sup>22</sup> is the method of *squeezing*, which tries  $\lambda = 1, \frac{1}{2}, \frac{1}{4}$ , and so on until an improvement in the function results. Although this approach is obviously a bit unorthodox, it appears to be quite effective when used with the Gauss–Newton method for nonlinear least squares problems. (See Chapter 7.) A somewhat more elaborate rule is suggested by Berndt et al. (1974). Choose an  $\varepsilon$  between 0 and  $\frac{1}{2}$ , and then find a  $\lambda$  such that

$$\varepsilon < \frac{F(\boldsymbol{\theta} + \lambda\Delta) - F(\boldsymbol{\theta})}{\lambda g'\Delta} < 1 - \varepsilon. \quad (\text{E-27})$$

Of course, which value of  $\varepsilon$  to choose is still open, so the choice of  $\lambda$  remains ad hoc. Moreover, in neither of these cases is there any optimality to the choice; we merely find a  $\lambda$  that leads to a function improvement. Other authors have devised relatively efficient means of searching for a step size without doing the full optimization at each iteration.<sup>23</sup>

**Assessing Convergence** Ideally, the iterative procedure should terminate when the gradient is zero. In practice, this step will not be possible, primarily because of accumulated rounding error in the computation of the function and its derivatives. Therefore, a number of alternative convergence criteria are used. Most of them are based on the relative changes in the function or the parameters. There is considerable variation in those used in different computer programs, and there are some pitfalls that should be avoided. A critical absolute value for the elements of the gradient or its norm will be affected by any scaling of the function, such as normalizing it by the sample size. Similarly, stopping on the basis of small absolute changes in the parameters can lead to premature convergence when the parameter vector approaches the maximizer. It is probably best to use several criteria simultaneously, such as the proportional change in both the function and the parameters. Belsley (1980) discusses a number of possible stopping rules. One that has proved useful and is immune to the scaling problem is to base convergence on  $\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$ .

**Multiple Solutions** It is possible for a function to have several local extrema. It is difficult to know a priori whether this is true of the one at hand. But if the function is not globally concave, then it may be a good idea to attempt to maximize it from several starting points to ensure that the maximum obtained is the global one. Ideally, a starting value near the optimum can facilitate matters; in some settings, this can be obtained by using a consistent estimate of the parameter for the starting point. The method of moments, if available, is sometimes a convenient device for doing so.

**No Solution** Finally, it should be noted that in a nonlinear setting the iterative algorithm can break down, even in the absence of constraints, for at least two reasons. The first possibility is that the problem being solved may be so numerically complex as to defy solution. The second possibility, which is often neglected, is that the proposed model may simply be inappropriate for the data. In a linear setting, a low  $R^2$  or some other diagnostic test may suggest that the model and data are mismatched, but as long as the full rank condition is met by the regressor matrix, a linear regression can *always* be computed. Nonlinear models are not so forgiving. The failure of an iterative algorithm to find a maximum of the criterion function may be a warning that the model is not appropriate for this body of data.

<sup>22</sup>Hall (1982, p. 147).

<sup>23</sup>See, for example, Joreskog and Gruvaeus (1970), Powell (1964), Quandt (1983), and Hall (1982).

### E.3.7 THE EM ALGORITHM

The latent class model can be characterized as a **missing data model**. Consider the mixture model we used for DocVis in Chapter 14, which we will now generalize to allow more than two classes:

$$f(y_{it} | \mathbf{x}_{it}, \text{class}_i = j) = \theta_{it,j}(1 - \theta_{it,j})^{y_{it}}, \theta_{it,j} = 1/(1 + \lambda_{it,j}), \lambda_{it,j} = \exp(\mathbf{x}'_{it}\boldsymbol{\beta}_j), y_{it} = 0, 1, \dots$$

$$\text{Prob}(\text{class}_i = j | \mathbf{z}_i) = \frac{\exp(\mathbf{z}'_i\boldsymbol{\alpha}_j)}{\sum_{j=1}^J \exp(\mathbf{z}'_i\boldsymbol{\alpha}_j)}, j = 1, 2, \dots, J.$$

With all parts incorporated, the log-likelihood for this latent class model is

$$\begin{aligned} \ln L_M &= \sum_{i=1}^n \ln L_{i,M} \\ &= \sum_{i=1}^n \ln \left\{ \sum_{j=1}^J \frac{\exp(\mathbf{z}'_i\boldsymbol{\alpha}_j)}{\sum_{m=1}^J \exp(\mathbf{z}'_i\boldsymbol{\alpha}_m)} \prod_{t=1}^{T_i} \left( \frac{1}{1 + \exp(\mathbf{x}'_{it}\boldsymbol{\beta}_j)} \right)^{(1-y_{it})} \left( \frac{\exp(\mathbf{x}'_{it}\boldsymbol{\beta}_j)}{1 + \exp(\mathbf{x}'_{it}\boldsymbol{\beta}_j)} \right)^{y_{it}} \right\}. \end{aligned} \quad (\text{E-28})$$

Suppose the actual class memberships were known (i.e., observed). Then, the class probabilities in  $\ln L_M$  would be unnecessary. The appropriate **complete data log-likelihood** for this case would be

$$\begin{aligned} \ln L_C &= \sum_{i=1}^n \ln L_{i,C} \\ &= \sum_{i=1}^n \ln \left\{ \sum_{j=1}^J D_{ij} \prod_{t=1}^{T_i} \left( \frac{1}{1 + \exp(\mathbf{x}'_{it}\boldsymbol{\beta}_j)} \right)^{(1-y_{it})} \left( \frac{\exp(\mathbf{x}'_{it}\boldsymbol{\beta}_j)}{1 + \exp(\mathbf{x}'_{it}\boldsymbol{\beta}_j)} \right)^{y_{it}} \right\}, \end{aligned} \quad (\text{E-29})$$

where  $D_{ij}$  is an observed dummy variable that equals one if individual  $i$  is from class  $j$ , and zero otherwise. With this specification, the log-likelihood breaks into  $J$  separate log-likelihoods, one for each (now known) class. The maximum likelihood estimates of  $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_J$  would be obtained simply by separating the sample into the respective subgroups and estimating the appropriate model for each group using maximum likelihood. The method we have used to estimate the parameters of the full model is to replace the  $D_{ij}$  variables with their unconditional expectations,  $\text{Prob}(\text{class}_i = j | \mathbf{z}_i)$ , then maximize the resulting log-likelihood function. This is the essential logic of the **EM** (expectation–maximization) **algorithm** [Dempster et al. (1977)]; however, the method uses the conditional (posterior) class probabilities instead of the unconditional probabilities. The iterative steps of the EM algorithm are

- (E step) Form the expectation of the missing data log-likelihood, conditional on the previous parameter estimates and the data in the sample;
- (M step) Maximize the expected log-likelihood function. Then either return to the E step or exit if the estimates have converged.

The EM algorithm can be used in a variety of settings. [See McLachlan and Krishnan (1997).] It has a particularly appealing form for estimating latent class models. The iterative steps for the latent class model are as follows:

- (E step) Form the conditional (posterior) class probabilities,  $\pi_{ij} | \mathbf{z}_i$ , based on the current estimates. These are based on the likelihood function.

## 1174 PART VI ♦ Appendices

- (M step) For each class, estimate the class-specific parameters by maximizing a weighted log-likelihood,

$$\ln L_{M \text{ step}, j} = \sum_{i=1}^{n_c} \pi_{ij} \ln L_i | \text{class} = j.$$

The parameters of the class probability model are also reestimated, as shown later, when there are variables in  $\mathbf{z}_i$  other than a constant term.

This amounts to a simple weighted estimation. For example, in the latent class linear regression model, the M step would amount to nothing more than weighted least squares. For nonlinear models such as the geometric model above, the M step involves maximizing a weighted log-likelihood function.

For the preceding geometric model, the precise steps are as follows: First, obtain starting values for  $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_J, \boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_J$ . Recall,  $\boldsymbol{\alpha}_J = \mathbf{0}$ . Then;

1. Form the contributions to the likelihood function using (E-28),

$$\begin{aligned} L_i &= \sum_{j=1}^J \pi_{ij} \prod_{t=1}^{T_i} f(y_{it} | \mathbf{x}_{it}, \boldsymbol{\beta}_j, \text{class}_i = j) \\ &= \sum_{j=1}^J L_i | \text{class} = j. \end{aligned} \quad (\text{E-30})$$

2. Form the conditional probabilities,  $w_{ij} = \frac{L_i | \text{class} = j}{\sum_{m=1}^J L_i | \text{class} = m}$ .
3. For each  $j$ , now maximize the weighted log likelihood functions (one at a time),

$$\ln L_{j,M}(\boldsymbol{\beta}_j) = \sum_{i=1}^n w_{ij} \ln \prod_{t=1}^{T_i} \left( \frac{1}{1 + \exp(\mathbf{x}'_{it} \boldsymbol{\beta}_j)} \right)^{(1-y_{it})} \left( \frac{\exp(\mathbf{x}'_{it} \boldsymbol{\beta}_j)}{1 + \exp(\mathbf{x}'_{it} \boldsymbol{\beta}_j)} \right)^{y_{it}} \quad (\text{E-32})$$

4. To update the  $\boldsymbol{\alpha}_j$  parameters, maximize the following log-likelihood function

$$\ln L(\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_J) = \sum_{i=1}^n \sum_{j=1}^J w_{ij} \ln \frac{\exp(\mathbf{z}'_i \boldsymbol{\alpha}_j)}{\sum_{j=1}^J \exp(\mathbf{z}'_i \boldsymbol{\alpha}_j)}, \quad \boldsymbol{\alpha}_J = \mathbf{0}. \quad (\text{E-33})$$

Step 4 defines a multinomial logit model (with “grouped”) data. If the class probability model does not contain any variables in  $\mathbf{z}_i$ , other than a constant, then the solutions to this optimization will be

$$\hat{\pi}_j = \frac{\sum_{i=1}^n w_{ij}}{\sum_{i=1}^n \sum_{j=1}^J w_{ij}}, \text{ then } \hat{\alpha}_j = \ln \frac{\hat{\pi}_j}{\hat{\pi}_J}. \quad (\text{E-34})$$

(Note that this preserves the restriction  $\hat{\alpha}_J = 0$ .) With these in hand, we return to steps 1 and 2 to rebuild the weights, then perform steps 3 and 4. The process is iterated until the estimates of  $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_J$  converge. Step 1 is constructed in a generic form. For a different model, it is necessary only to change the density that appears at the end of the expression in (E-32). For a cross section instead of a panel, the product term in step 1 becomes simply the log of the single term.

The EM algorithm has an intuitive appeal in this (and other) settings. In practical terms, it is often found to be a very slow algorithm. It can take many iterations to converge. (The estimates in Example 14.17 were computed using a gradient method, not the EM algorithm.) In its favor,

**APPENDIX E ♦ Computation and Optimization 1175**

the EM method is very stable. It has been shown [Dempster, Laird, and Rubin (1977)] that the algorithm always climbs uphill. The log-likelihood improves with each iteration. Applications differ widely in the methods used to estimate latent class models. Adding to the variety are the very many Bayesian applications, none of which use either of the methods discussed here.

## E.4 EXAMPLES

To illustrate the use of gradient methods, we consider some simple problems.

### E.4.1 FUNCTION OF ONE PARAMETER

First, consider maximizing a function of a single variable,  $f(\theta) = \ln(\theta) - 0.1\theta^2$ . The function is shown in Figure E.4. The first and second derivatives are

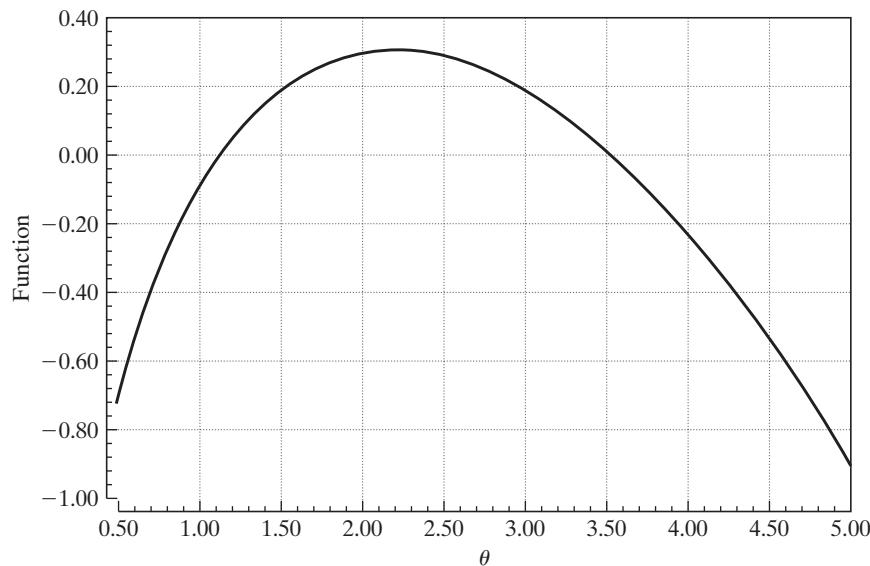
$$f'(\theta) = \frac{1}{\theta} - 0.2\theta,$$

$$f''(\theta) = \frac{-1}{\theta^2} - 0.2.$$

Equating  $f'$  to zero yields the solution  $\theta = \sqrt{5} = 2.236$ . At the solution,  $f'' = -0.4$ , so this solution is indeed a maximum. To demonstrate the use of an iterative method, we solve this problem using Newton's method. Observe, first, that the second derivative is always negative for any admissible (positive)  $\theta$ .<sup>24</sup> Therefore, it should not matter where we start the iterations; we shall eventually find the maximum. For a single parameter, Newton's method is

$$\theta_{t+1} = \theta_t - [f'_t/f''_t].$$

**FIGURE E.4** Function of One Variable Parameter.



<sup>24</sup>In this problem, an inequality restriction,  $\theta > 0$ , is required. As is common, however, for our first attempt we shall neglect the constraint.

## 1176 PART VI ♦ Appendices

**TABLE E.1** Iterations for Newton's Method

Iteration	$\theta$	$f$	$f'$	$f''$
0	5.00000	-0.890562	-0.800000	-0.240000
1	1.66667	0.233048	0.266667	-0.560000
2	2.14286	0.302956	0.030952	-0.417778
3	2.23404	0.304718	0.000811	-0.400363
4	2.23607	0.304719	0.0000004	-0.400000

The sequence of values that results when 5 is used as the starting value is given in Table E.1. The path of the iterations is also shown in the table.

### E.4.2 FUNCTION OF TWO PARAMETERS: THE GAMMA DISTRIBUTION

For random sampling from the gamma distribution,

$$f(y_i, \beta, \rho) = \frac{\beta^\rho}{\Gamma(\rho)} e^{-\beta y_i} y_i^{\rho-1}.$$

The log-likelihood is  $\ln L(\beta, \rho) = n\rho \ln \beta - n \ln \Gamma(\rho) - \beta \sum_{i=1}^n y_i + (\rho - 1) \sum_{i=1}^n \ln y_i$ . (See Section 14.6.4 and Example 13.5.) It is often convenient to scale the log-likelihood by the sample size. Suppose, as well, that we have a sample with  $\bar{y} = 3$  and  $\ln \bar{y} = 1$ . Then the function to be maximized is  $F(\beta, \rho) = \rho \ln \beta - \ln \Gamma(\rho) - 3\beta + \rho - 1$ . The derivatives are

$$\begin{aligned} \frac{\partial F}{\partial \beta} &= \frac{\rho}{\beta} - 3, & \frac{\partial F}{\partial \rho} &= \ln \beta - \frac{\Gamma'}{\Gamma} + 1 = \ln \beta - \Psi(\rho) + 1, \\ \frac{\partial^2 F}{\partial \beta^2} &= -\frac{\rho}{\beta^2}, & \frac{\partial^2 F}{\partial \rho^2} &= \frac{-(\Gamma \Gamma'' - \Gamma'^2)}{\Gamma^2} = -\Psi'(\rho), & \frac{\partial^2 F}{\partial \beta \partial \rho} &= \frac{1}{\beta}. \end{aligned}$$

Finding a good set of starting values is often a difficult problem. Here we choose three starting points somewhat arbitrarily:  $(\rho^0, \beta^0) = (4, 1)$ ,  $(8, 3)$ , and  $(2, 7)$ . The solution to the problem is  $(5.233, 1.743)$ . We used Newton's method and DFP with a line search to maximize this function.<sup>25</sup> For Newton's method,  $\lambda = 1$ . The results are shown in Table E.2. The two methods were essentially the same when starting from a good starting point (trial 1), but they differed substantially when starting from a poorer one (trial 2). Note that DFP and Newton approached the solution from different directions in trial 2. The third starting point shows the value of a line search. At this

**TABLE E.2** Iterative Solutions to  $\text{Max}(\rho, \beta) \rho \ln \beta - \ln \Gamma(\rho) - 3\beta + \rho - 1$

Iter.	Trial 1				Trial 2				Trial 3			
	DFP		Newton		DFP		Newton		DFP		Newton	
	$\rho$	$\beta$	$\rho$	$\beta$	$\rho$	$\beta$	$\rho$	$\beta$	$\rho$	$\beta$	$\rho$	$\beta$
0	4.000	1.000	4.000	1.000	8.000	3.000	8.000	3.000	2.000	7.000	2.000	7.000
1	3.981	1.345	3.812	1.203	7.117	2.518	2.640	0.615	6.663	2.027	-47.7	-233.
2	4.005	1.324	4.795	1.577	7.144	2.372	3.203	0.931	6.195	2.075	—	—
3	5.217	1.743	5.190	1.728	7.045	2.389	4.257	1.357	5.239	1.731	—	—
4	5.233	1.744	5.231	1.744	5.114	1.710	5.011	1.656	5.251	1.754	—	—
5	—	—	—	—	5.239	1.747	5.219	1.740	5.233	1.744	—	—
6	—	—	—	—	5.233	1.744	5.233	1.744	—	—	—	—

<sup>25</sup>The one used is described in Joreskog and Gruvaeus (1970).

## APPENDIX E ♦ Computation and Optimization 1177

starting value, the Hessian is extremely large, and the second value for the parameter vector with Newton's method is  $(-47.671, -233.35)$ , at which point  $F$  cannot be computed and this method must be abandoned. Beginning with  $\mathbf{H} = \mathbf{I}$  and using a line search, DFP reaches the point  $(6.63, 2.03)$  at the first iteration, after which convergence occurs routinely in three more iterations. At the solution, the Hessian is  $[(-1.72038, 0.191153)', (0.191153, -0.210579)']$ . The diagonal elements of the Hessian are negative and its determinant is 0.32574, so it is negative definite. (The two characteristic roots are  $-1.7442$  and  $-0.18675$ ). Therefore, this result is indeed the maximum of the function.

### E.4.3 A CONCENTRATED LOG-LIKELIHOOD FUNCTION

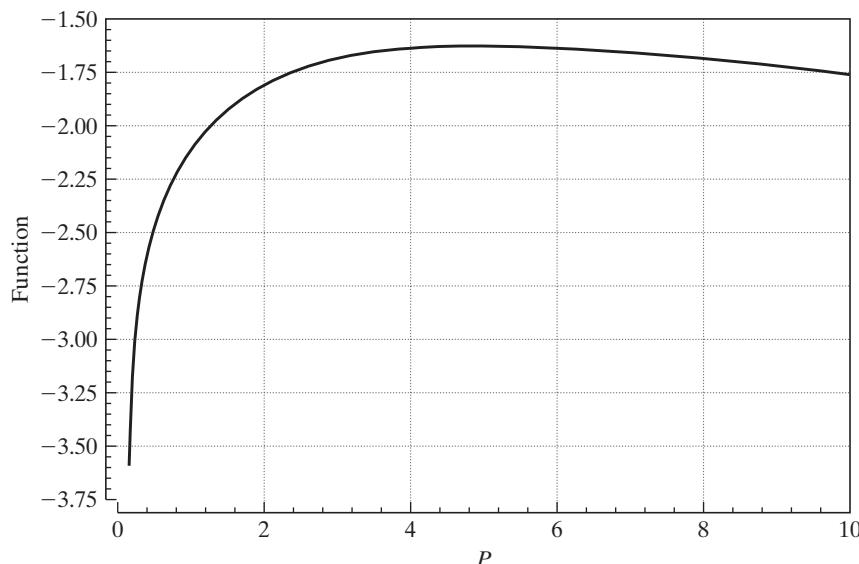
There is another way that the preceding problem might have been solved. The first of the necessary conditions implies that at the joint solution for  $(\beta, \rho)$ ,  $\beta$  will equal  $\rho/3$ . Suppose that we impose this requirement on the function we are maximizing. The **concentrated** (over  $\beta$ ) **log-likelihood function** is then produced:

$$\begin{aligned} F_c(\rho) &= \rho \ln(\rho/3) - \ln \Gamma(\rho) - 3(\rho/3) + \rho - 1 \\ &= \rho \ln(\rho/3) - \ln \Gamma(\rho) - 1. \end{aligned}$$

This function could be maximized by an iterative search or by a simple one-dimensional grid search. Figure E.5 shows the behavior of the function. As expected, the maximum occurs at  $\rho = 5.233$ . The value of  $\beta$  is found as  $5.23/3 = 1.743$ .

The concentrated log-likelihood is a useful device in many problems. (See Section 14.9.6.d for an application.) Note the interpretation of the function plotted in Figure E.5. The original function of  $\rho$  and  $\beta$  is a surface in three dimensions. The curve in Figure E.5 is a projection of that function; it is a plot of the function values above the line  $\beta = \rho/3$ . By virtue of the first-order condition, we know that one of these points will be the maximizer of the function. Therefore, we may restrict our search for the overall maximum of  $F(\beta, \rho)$  to the points on this line.

**FIGURE E.5** Concentrated Log-Likelihood.



**1178 PART VI ♦ Appendices****APPENDIX F**

## **DATA SETS USED IN APPLICATIONS**

The following data sets are used in the examples and applications in the text. With the exception of the Bertschek and Lechner file, the data sets themselves can be downloaded either from the Web site for this text, [pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm](http://pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm), or from the URLs to the publicly accessible archives indicated as “*Location*.” The points in the text where the data are used for examples or suggested exercises are noted as “*Uses*.”

**TABLE F1.1 Consumption and Income, 10 Yearly Observations, 2000–2009**

*Source:* *Economic Report of the President*, 1987, Council of Economic Advisors

*Location:* Text Web site

*Use:* Example 1.2

**TABLE F2.1 Consumption and Income, 11 Yearly Observations, 1940–1950**

*Source:* *Economic Report of the President*, U.S. Government Printing Office, Washington, D.C., 1983

*Location:* Text Web site

*Uses:* Examples 2.1, 3.2, 16.3

**TABLE F2.2 The U.S. Gasoline Market, 52 Yearly Observations 1953–2004**

*Source:* The data were compiled by Professor Chris Bell, Department of Economics, University of North Carolina, Asheville. *Sources:* [www.bea.gov](http://www.bea.gov) and [www.bls.gov](http://www.bls.gov).

*Location:* Text Web site

*Uses:* Examples 2.3, 4.2, 4.4, 4.8, 4.9, 6.9, 15.4, 20.2, 20.6, 21.1, 21.3

Sections 20.9.2

Applications 4.1, 5.3, 7.6, 7.7

**TABLE F3.1 Investment, 15 Yearly Observations, 1968–1982**

*Source:* *Economic Report of the President*, U.S. Government Printing Office, Washington, D.C., 1983

*Location:* Text Web site

*Uses:* Examples 3.1, 3.3

Section 3.2.2

Exercise 3.12

**TABLE F3.2 Koop and Tobias Labor Market Experience, 17,919 Observations**

*Source:* Koop and Tobias (2004)

*Location:* *Journal of Applied Econometrics* data archive,

<http://www.econ.queensu.ca/jae/2004-v19.7/koop-tobias/>.

*Uses:* Example 15.16

Applications 3.1, 5.1, 6.1, 6.2

**APPENDIX F ♦ Data Sets Used in Applications 1179****TABLE F4.1** Auction Data for Monet Paintings, 430 Observations

*Source:* Author  
*Location:* Text Web site  
*Uses:* Examples 4.5, 4.10, 5.8, 6.2, 11.2  
 Section 4.7.6  
 Exercise 4.17

**TABLE F4.2** The Longley Data, 15 Yearly Observations, 1947–1962

*Source:* Longley (1967)  
*Location:* Text Web site  
*Use:* Example 4.11

**TABLE F4.3** Movie Buzz Data, 62 Observations

*Source:* Author  
*Location:* Text Web site  
*Uses:* Examples 4.12, 6.3

**TABLE F4.4** Cost Function, 158 1970 Cross-Section Firm Level Observations

*Note:* The file contains 158 observations. Christensen and Greene used the first 123. The extras are the holding companies. Use only the first 123 observations to replicate Christensen and Greene.  
*Source:* Christensen and Greene (1976)  
*Location:* Text Web site  
*Uses:* Examples 7.11, 7.12  
 Applications 4.2, 5.2, 7.5, 10.1, 19.4

**TABLE F5.1** Labor Supply Data from Mroz (1987), 753 Observations

*Source:* 1976 Panel Study of Income Dynamics, Mroz (1987)  
*Location:* Text Web site  
*Uses:* Examples 5.2, 5.5, 6.1, 17.1, 17.8, 17.10, 19.11

**TABLE F5.2** Macroeconomics Data Set, Quarterly, 1950I to 2000IV

*Source:* Department of Commerce, BEA Web site, and [www.economagic.com](http://www.economagic.com)  
*Location:* Text Web site  
*Uses:* Examples 5.3, 5.6, 5.7, 7.4, 7.8, 8.7, 8.10, 14.7, 16.3, 20.1, 20.3, 20.4, 21.2, 21.4, 21.5, 23.1, 23.2, 23.3, 23.4, 23.5  
 Applications 5.4, 10.3, 20.1, 20.3, 21.1, 23.1, 23.2, 23.3  
 Sections 21.5.1, 21.6.8.e, 23.2.4

**TABLE F5.3** Production for SIC 33: Primary Metals, 27 Statewide Observations

*Note:* Data are per establishment, labor is a measure of labor input, and capital is the gross value of plant and equipment. A scale factor used to normalize the capital figure in the original study has been omitted. Further details on construction of the data are given in Aigner et al. (1977).

*Source:* Hildebrand and Liu (1957)

*Location:* Text Web site

*Uses:* Example 5.4

Application 7.1

## 1180 PART VI ♦ Appendices

**TABLE F6.1** Costs for U.S. Airlines, 90 Total Observations on 6 Firms for 1970–1984

*Note:* These data are a subset of a larger data set provided to the author by Professor Moshe Kim.

*Source:* Christensen Associates of Madison, Wisconsin

*Location:* Text Web site

*Uses:* Examples 6.4, 9.4, 14.6

Applications 9.3, 11.2

**TABLE F6.2** Cost Function, 145 U.S. Electricity Producers, Nerlove's 1955 Data

*Note:* The data file contains several extra observations that are aggregates of commonly owned firms. Use only the first 145 for analysis.

*Sources:* Nerlove (1960) and Christensen and Greene (1976)

*Location:* Text Web site

*Uses:* Example 6.6

Section 10.5.1

**TABLE F6.3** World Health Organization Panel Data, 840 Total Observations

*Note:* Variables marked \* were updated with more recent sources in Greene (2004a). Missing values for some of the variables in this data set are filled by using fitted values from a linear regression.

*Sources:* The World Health Organization [Evans et al. (2000) and www.who.int]

*Location:* Text Web site

*Uses:* Examples 6.10, 11.4

**TABLE F6.4** Solow's Technological Change Data, 41 Yearly Observations, 1909–1949

*Source:* Solow (1957, p. 314). Several variables are omitted

*Location:* Text Web site

*Use:* Application 6.3

**TABLE F7.1** German Health Care Data, Unbalanced Panel, 7,293 Individuals, 27,326 Observations

*Notes:* In the applications in the text, the following additional variables are used:

*NUMOBS* = Number of observations for this person. Repeated in each row of data.

*NEWHSAT* = *HSAT*; 40 observations on *HSAT* recorded between 6 and 7 were changed to 7.

Frequencies are 1 = 1525, 2 = 1079, 3 = 825, 4 = 926, 5 = 1051, 6 = 1000, 7 = 887.

*Source:* Riphahn et al. (2003)

*Location:* *Journal of Applied Econometrics* data archive,

<http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/>

*Uses:* Examples 7.6, 11.16, 11.17, 13.7, 14.5, 14.9, 14.10, 14.14, 14.17, 17.4, 17.5, 17.7, 17.11, 17.13, 17.15, 17.16, 17.17, 17.18, 17.19, 17.20, 18.6, 18.7, 18.10, 18.12, 19.13

Section 14.9.5

Applications 14.1, 18.2, 18.3, 18.4

**TABLE F7.2** Statewide Data on Transportation Equipment Manufacturing, 25 Observations

*Note:* “Value added,” “Capital,” and “Labor” in millions of 1957 dollars. Data used in regression examples are per establishment. Totals are used for the stochastic frontier application in Chapter 19.

*Source:* Zellner and Revankar (1970, p. 249)

*Location:* Text Web site

*Uses:* Example 7.9

Applications 7.2., 7.3

**APPENDIX F ♦ Data Sets Used in Applications 1181****TABLE F7.3 Expenditure and Default Data, 13,999 Observations***Source:* Greene (1992)*Location:* Text Web site*Uses:* Examples 7.10, 9.1, 17.9, 17.22, 18.8, 18.11**TABLE F8.1 Cornwell and Rupert, Labor Market Data, 595 Individuals, 7 years***Source:* See Cornwell and Rupert (1988)*Location:* Web site for Baltagi (2005), <http://www.wiley.com/legacy/wileychi/baltagi/supp/WAGES.xls>*Location (ASCII form):* Text Web site*Uses:* Examples 8.5, 8.6, 8.8, 11.1, 11.3, 11.5, 11.6, 11.7, 11.8, 11.9, 11.11, 11.15, 14.11, 15.6, 15.12**TABLE F9.1 Income and Expenditure Data. 100 Cross-Section Observations***Source:* Greene (1992)*Location:* Text Web site*Uses:* Examples 9.1, 9.2, 9.3**TABLE F9.2 Baltagi and Griffin Gasoline Data, 18 OECD Countries, 19 Years***Source:* See Baltagi and Griffin (1983) and Baltagi (2005)*Location:* Web site for Baltagi (2005), <http://www.wiley.com/legacy/wileychi/baltagi/supp/Gasoline.dat>*Uses:* Example 9.5

Application 9.2

**TABLE F10.1 Munnell Productivity Data, 48 Continental U.S. States, 17 years, 1970–1986***Sources:* Munnell (1990) and Baltagi (2005)*Location:* Web site for Baltagi (2005), <http://www.wiley.com/legacy/wileychi/baltagi/supp/PRODUC.prn>*Uses:* Examples 10.1, 11.9, 14.12, 15.13, 15.15, 20.5**TABLE F10.2 Manufacturing Costs, U.S. Economy, 25 Yearly Observations, 1947–1971***Source:* Berndt and Wood (1975)*Location:* Text Web site*Use:* Example 10.3**TABLE F10.3 Klein's Model I, 22 Yearly Observations, 1920–1941***Source:* Klein (1950)*Location:* Text Web site*Use:* Examples 10.6**TABLE F10.4 Grunfeld Investment Data, 200 Yearly Observations on 10 Firms for 1935–1954***Sources:* Grunfeld (1958) and Boot and deWitt (1960)*Location:* Text Web site*Uses:* Example 14.8

Applications 10.2., 11.1

## 1182 PART VI ♦ Appendices

**TABLE F13.1** Dahlberg and Johanssen Expenditure Data, 265 Municipalities, 9 Years

*Location:* *Journal of Applied Econometrics* data archive  
<http://qed.econ.queensu.ca/jae/2000-v15.4/dahlberg-johansson/dj-data.zip>

*Uses:* Examples 13.10, 21.7

**TABLE F14.1** Program Effectiveness, 32 Cross-Section Observations

*Source:* Spector and Mazzeo (1980)

*Location:* Text Web site

*Uses:* Examples 14.15, 14.16, 17.3

**TABLE F15.1** Bertschek and Lechner Binary Choice Data, Balanced Panel, 5 years, 1,270 firms

*Source:* Bertschek and Lechner (1998)

*Location:* These data are proprietary and may not be redistributed

*Uses:* Examples 15.17, 17.23

Section 12.4.1

**TABLE F17.1** Burnett Analysis of Liberal Arts College Gender Economics Courses, 132 Observations

*Source:* Burnett (1997). Data provided by the author

*Location:* Text Web site

*Use:* Example 17.21

**TABLE F17.2** Fair, *Redbook* Survey on Extramarital Affairs, 6,366 Observations

*Source:* Fair (1978), data provided by the author.

*Location:* Text Web site

*Uses:* Example 19.6

Applications 17.1, 18.1, 18.2, 19.2, 19.3

**TABLE F18.1** Fair's (1977) Extramarital Affairs Data, 601 Cross-Section Observations

*Note:* Several variables not used are denoted  $X_1, \dots, X_5$ .

*Source:* Fair (1977)

*Location:* <http://fairmodel.econ.yale.edu/rayfair/pdf/1978ADAT.ZIP>

*Location:* Text Web site

*Uses:* Examples 18.1, 18.9

Application 19.1

**TABLE F18.2** Data Used to Study Travel Mode Choice, 840 Observations, on 4 Modes for 210 Individuals

*Source:* Greene and Hensher (1997)

*Location:* Text Web site

*Uses:* Sections 18.2.9, 18.2.10

**TABLE F18.3** Ship Accidents, 40 Observations, 5 Types, 4 Vintages, and 2 Service Periods

*Source:* McCullagh and Nelder (1983)

*Location:* Text Web site

*Use:* Application 18.5

APPENDIX F ♦ Data Sets Used in Applications **1183****TABLE F19.1** Filippini, Farsi, Greene, Swiss Railroads Data, Unbalanced Panel  
50 Firms, 605 Observations

*Source:* Authors  
*Location:* Text Web site  
*Use:* Example 19.3

**TABLE F19.2** Strike Duration Data, 63 Observations in 9 Years, 1968–1976

*Source:* Kennan (1985)  
*Location:* Text Web site  
*Use:* Example 19.8

**TABLE F19.3** LaLonde (1986) Earnings Data, 2,490 Control Observations and  
185 Treatment Observations

*Note:* We also scaled all earnings variables by 10,000 before beginning the analysis.  
*Source:* LaLonde (1986)  
*Location:* <http://www.nber.org/%7Erdehejia/nswdata.htm>. The two specific subsamples are in  
[http://www.nber.org/%7Erdehejia//psid\\_controls.txt](http://www.nber.org/%7Erdehejia//psid_controls.txt) and  
[http://www.nber.org/%7Erdehejia/nswre74\\_treated.txt](http://www.nber.org/%7Erdehejia/nswre74_treated.txt)  
*Use:* Example 19.15

**TABLE F20.1** Bollerslev and Ghysels Exchange Rate Data, 1974 Daily  
Observations

*Source:* Bollerslev (1986)  
*Location:* Text Web site  
*Uses:* Examples 20.7, 20.8

**TABLE F22.1** Bond Yield, Moody's AAA Rated, Monthly, 60 Observations,  
1990–1994

*Sources:* *National Income and Product Accounts*, U.S. Department of Commerce, Bureau of Economic  
Analysis, *Survey of Current Business: Business Statistics*  
*Location:* Text Web site  
*Use:* Example 22.1

**TABLE F23.1** Money, Output, Price Deflator Data, 136 Quarterly Observations,  
1950–1983

*Sources:* *National Income and Product Accounts*, U.S. Department of Commerce, Bureau of Economic  
Analysis, *Survey of Current Business: Business Statistics*  
*Location:* Text Web site  
*Uses:* Examples 23.1, 23.5

**TABLE FC.1** Observations on Income and Education, 20 Observations

*Source:* Data are artificial  
*Location:* Text Web site  
*Uses:* Examples 13.5, 15.17, C.1, C.2

**1184** PART VI ♦ Appendices

## **APPENDIX G**



# STATISTICAL TABLES

**TABLE G.1** Cumulative Normal Distribution. Table Entry Is  $\Phi(z) = \text{Prob}[Z \leq z]$

APPENDIX G ♦ Statistical Tables **1185****TABLE G.2** Percentiles of the Student's *t* Distribution. Table Entry Is *x* Such That  $\text{Prob}[t_n \leq x] = P$ 

<i>n</i>	.750	.900	.950	.975	.990	.995
1	1.000	3.078	6.314	12.706	31.821	63.657
2	.816	1.886	2.920	4.303	6.965	9.925
3	.765	1.638	2.353	3.182	4.541	5.841
4	.741	1.533	2.132	2.776	3.747	4.604
5	.727	1.476	2.015	2.571	3.365	4.032
6	.718	1.440	1.943	2.447	3.143	3.707
7	.711	1.415	1.895	2.365	2.998	3.499
8	.706	1.397	1.860	2.306	2.896	3.355
9	.703	1.383	1.833	2.262	2.821	3.250
10	.700	1.372	1.812	2.228	2.764	3.169
11	.697	1.363	1.796	2.201	2.718	3.106
12	.695	1.356	1.782	2.179	2.681	3.055
13	.694	1.350	1.771	2.160	2.650	3.012
14	.692	1.345	1.761	2.145	2.624	2.977
15	.691	1.341	1.753	2.131	2.602	2.947
16	.690	1.337	1.746	2.120	2.583	2.921
17	.689	1.333	1.740	2.110	2.567	2.898
18	.688	1.330	1.734	2.101	2.552	2.878
19	.688	1.328	1.729	2.093	2.539	2.861
20	.687	1.325	1.725	2.086	2.528	2.845
21	.686	1.323	1.721	2.080	2.518	2.831
22	.686	1.321	1.717	2.074	2.508	2.819
23	.685	1.319	1.714	2.069	2.500	2.807
24	.685	1.318	1.711	2.064	2.492	2.797
25	.684	1.316	1.708	2.060	2.485	2.787
26	.684	1.315	1.706	2.056	2.479	2.779
27	.684	1.314	1.703	2.052	2.473	2.771
28	.683	1.313	1.701	2.048	2.467	2.763
29	.683	1.311	1.699	2.045	2.462	2.756
30	.683	1.310	1.697	2.042	2.457	2.750
35	.682	1.306	1.690	2.030	2.438	2.724
40	.681	1.303	1.684	2.021	2.423	2.704
45	.680	1.301	1.679	2.014	2.412	2.690
50	.679	1.299	1.676	2.009	2.403	2.678
60	.679	1.296	1.671	2.000	2.390	2.660
70	.678	1.294	1.667	1.994	2.381	2.648
80	.678	1.292	1.664	1.990	2.374	2.639
90	.677	1.291	1.662	1.987	2.368	2.632
100	.677	1.290	1.660	1.984	2.364	2.626
$\infty$	.674	1.282	1.645	1.960	2.326	2.576

**1186 PART VI ♦ Appendices**
**TABLE G.3** Percentiles of the Chi-Squared Distribution. Table Entry Is  $c$  Such That  $\text{Prob}[\chi_n^2 \leq c] = P$ 

<i>n</i>	.005	.010	.025	.050	.100	.250	.500	.750	.900	.950	.975	.990	.995
1	.00004	.0002	.001	.004	.02	.10	.45	1.32	2.71	3.84	5.02	6.63	7.88
2	.01	.02	.05	.10	.21	.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	.07	.11	.22	.35	.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	.21	.30	.48	.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	.41	.55	.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	.68	.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
35	17.19	18.51	20.57	22.47	24.80	29.05	34.34	40.22	46.06	49.80	53.20	57.34	60.27
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77
45	24.31	25.90	28.37	30.61	33.35	38.29	44.34	50.98	57.51	61.66	65.41	69.96	73.17
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49

APPENDIX G ♦ Statistical Tables **1187****TABLE G.4** 95th Percentiles of the *F* Distribution. Table Entry Is *f* Such That  
 $\text{Prob}[F_{n_1, n_2} \leq f] = .95$ 

<i>n<sub>1</sub></i> = Degrees of Freedom for the Numerator									
<i>n<sub>2</sub></i>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
$\infty$	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88
<i>n<sub>2</sub></i>	<b>10</b>	<b>12</b>	<b>15</b>	<b>20</b>	<b>30</b>	<b>40</b>	<b>50</b>	<b>60</b>	$\infty$
1	241.88	243.91	245.95	248.01	250.10	251.14	252.20	252.20	254.19
2	19.40	19.41	19.43	19.45	19.46	19.47	19.48	19.48	19.49
3	8.79	8.74	8.70	8.66	8.62	8.59	8.57	8.57	8.53
4	5.96	5.91	5.86	5.80	5.75	5.72	5.69	5.69	5.63
5	4.74	4.68	4.62	4.56	4.50	4.46	4.43	4.43	4.37
6	4.06	4.00	3.94	3.87	3.81	3.77	3.74	3.74	3.67
7	3.64	3.57	3.51	3.44	3.38	3.34	3.30	3.30	3.23
8	3.35	3.28	3.22	3.15	3.08	3.04	3.01	3.01	2.93
9	3.14	3.07	3.01	2.94	2.86	2.83	2.79	2.79	2.71
10	2.98	2.91	2.85	2.77	2.70	2.66	2.62	2.62	2.54
15	2.54	2.48	2.40	2.33	2.25	2.20	2.16	2.16	2.07
20	2.35	2.28	2.20	2.12	2.04	1.99	1.95	1.95	1.85
25	2.24	2.16	2.09	2.01	1.92	1.87	1.82	1.82	1.72
30	2.16	2.09	2.01	1.93	1.84	1.79	1.74	1.74	1.63
40	2.08	2.00	1.92	1.84	1.74	1.69	1.64	1.64	1.52
50	2.03	1.95	1.87	1.78	1.69	1.63	1.58	1.58	1.45
70	1.97	1.89	1.81	1.72	1.62	1.57	1.50	1.50	1.36
100	1.93	1.85	1.77	1.68	1.57	1.52	1.45	1.45	1.30
$\infty$	1.83	1.75	1.67	1.57	1.46	1.39	1.34	1.31	1.30

**1188 PART VI ♦ Appendices**
**TABLE G.5** 99th Percentiles of the *F* Distribution. Table Entry Is *f* Such That  
 $\text{Prob}[F_{n_1, n_2} \leq f] = .99$ 

<i>n</i> <sub>1</sub> = Degrees of Freedom for the Numerator									
<i>n</i> <sub>2</sub>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59
$\infty$	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43
<i>n</i> <sub>2</sub>	<b>10</b>	<b>12</b>	<b>15</b>	<b>20</b>	<b>30</b>	<b>40</b>	<b>50</b>	<b>60</b>	$\infty$
1	6055.85	6106.32	6157.28	6208.73	6260.65	6286.78	6313.03	6313.03	6362.68
2	99.40	99.42	99.43	99.45	99.47	99.47	99.48	99.48	99.50
3	27.23	27.05	26.87	26.69	26.50	26.41	26.32	26.32	26.14
4	14.55	14.37	14.20	14.02	13.84	13.75	13.65	13.65	13.47
5	10.05	9.89	9.72	9.55	9.38	9.29	9.20	9.20	9.03
6	7.87	7.72	7.56	7.40	7.23	7.14	7.06	7.06	6.89
7	6.62	6.47	6.31	6.16	5.99	5.91	5.82	5.82	5.66
8	5.81	5.67	5.52	5.36	5.20	5.12	5.03	5.03	4.87
9	5.26	5.11	4.96	4.81	4.65	4.57	4.48	4.48	4.32
10	4.85	4.71	4.56	4.41	4.25	4.17	4.08	4.08	3.92
15	3.80	3.67	3.52	3.37	3.21	3.13	3.05	3.05	2.88
20	3.37	3.23	3.09	2.94	2.78	2.69	2.61	2.61	2.43
25	3.13	2.99	2.85	2.70	2.54	2.45	2.36	2.36	2.18
30	2.98	2.84	2.70	2.55	2.39	2.30	2.21	2.21	2.02
40	2.80	2.66	2.52	2.37	2.20	2.11	2.02	2.02	1.82
50	2.70	2.56	2.42	2.27	2.10	2.01	1.91	1.91	1.70
70	2.59	2.45	2.31	2.15	1.98	1.89	1.78	1.78	1.56
100	2.50	2.37	2.22	2.07	1.89	1.80	1.69	1.69	1.45
$\infty$	2.34	2.20	2.06	1.90	1.72	1.61	1.50	1.50	1.16

APPENDIX G ♦ Statistical Tables **1189****TABLE G.6** Durbin–Watson Statistic: 5 Percent Significance Points of  $d_L$  and  $d_U$ 

<i>n</i>	<i>k</i> = 1		<i>k</i> = 2		<i>k</i> = 3		<i>k</i> = 4		<i>k</i> = 5		<i>k</i> = 10		<i>k</i> = 15	
	<i>dL</i>	<i>dU</i>	<i>dL</i>	<i>dU</i>	<i>dL</i>	<i>dU</i>								
15	1.08	1.36	.95	1.54	.82	1.75	.69	1.97	.56	2.21				
16	1.10	1.37	.98	1.54	.86	1.73	.74	1.93	.62	2.15	.16	3.30		
17	1.13	1.38	1.02	1.54	.90	1.71	.78	1.90	.67	2.10	.20	3.18		
18	1.16	1.39	1.05	1.53	.93	1.69	.82	1.87	.71	2.06	.24	3.07		
19	1.18	1.40	1.08	1.53	.97	1.68	.86	1.85	.75	2.02	.29	2.97		
20	1.20	1.41	1.10	1.54	1.00	1.68	.90	1.83	.79	1.99	.34	2.89	.06	3.68
21	1.22	1.42	1.13	1.54	1.03	1.67	.93	1.81	.83	1.96	.38	2.81	.09	3.58
22	1.24	1.43	1.15	1.54	1.05	1.66	.96	1.80	.86	1.94	.42	2.73	.12	3.55
23	1.26	1.44	1.17	1.54	1.08	1.66	.99	1.79	.90	1.92	.47	2.67	.15	3.41
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	.93	1.90	.51	2.61	.19	3.33
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	.95	1.89	.54	2.57	.22	3.25
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	.98	1.88	.58	2.51	.26	3.18
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86	.62	2.47	.29	3.11
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85	.65	2.43	.33	3.05
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84	.68	2.40	.36	2.99
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83	.71	2.36	.39	2.94
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83	.74	2.33	.43	2.99
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82	.77	2.31	.46	2.84
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81	.80	2.28	.49	2.80
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81	.82	2.26	.52	2.75
35	1.40	1.52	1.34	1.53	1.28	1.65	1.22	1.73	1.16	1.80	.85	2.24	.55	2.72
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80	.87	2.22	.58	2.68
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80	.89	2.20	.60	2.65
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79	.91	2.18	.63	2.61
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79	.93	2.16	.65	2.59
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79	.95	2.15	.68	2.56
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78	1.04	2.09	.79	2.44
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77	1.11	2.04	.88	2.35
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77	1.17	2.01	.96	2.28
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77	1.22	1.98	1.03	2.23
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77	1.27	1.96	1.09	2.18
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77	1.30	1.95	1.14	2.15
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77	1.34	1.94	1.18	2.12
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77	1.37	1.93	1.22	2.09
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77	1.40	1.92	1.26	2.07
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78	1.42	1.91	1.29	2.06
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78	1.44	1.90	1.32	2.04
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78	1.46	1.90	1.35	2.03

Source: Extracted from N.E. Savin and K.J. White, "The Durbin–Watson Test for Serial Correlation with Extreme Sample Sizes and Many Regressors," *Econometrica*, 45 (8), Nov. 1977, pp. 1992–1995.

Note: *k* is the number of regressors excluding the intercept.