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 synonomous [*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

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 plateform 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? [Hanuschek (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).]

Macroeconometrics 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

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, . . . 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.[1] The final paragraph asserts that the average propensity to consume (APC), C/X, falls as income rises, or $d(C/X)/dX = (\text{MPC} - APC)/X < 0$. 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 β lies between zero and one and if α 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

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

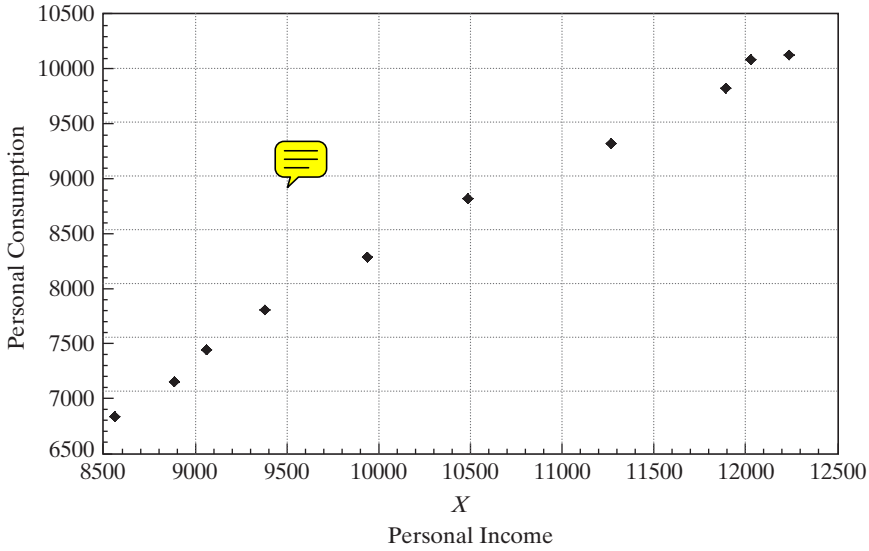

FIGURE 1.1 Aggregate U.S. Consumption and Income Data, 2000–2009.

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 α is positive and β 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

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.[2]

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

[2]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.[3] 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

[3]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 macroeconometrics that includes both econometric theory and applications. The numerous more focused studies include books such as Bårdsen, G., Eitrheim, ∅., 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. Zoega, eds), pp.399–414. Reykjavik: University of Iceland Press, 2000 also at http://www.ecomod.net/conferences/ecomod2001/papers_web/Wallis_Iceland.pdf

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 microeconometrics and macroeconometrics, 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. Microeconometric Methods
Chapters 17 through 19 are about microeconometrics, 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 macroeconometrics.

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/~wgreene/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 provided 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 $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$ 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

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, $\boldsymbol{\beta}' = [\beta_1, \beta_2, \ldots, \beta_K]$. The product of a row and a column vector will always be denoted in a form such as $\boldsymbol{\beta}'\mathbf{x} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K$. The elements in a matrix, $\mathbf{X}$, form a set of vectors. In terms of its columns, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K]$—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 ith row of $\mathbf{X}$ are the row vector, $\mathbf{x}_i' = [x_{i1}, x_{i2}, \ldots, x_{iK}]$. 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 ith 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 ith 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.