A Guide to Econometrics

Peter Kennedy

Simon Fraser University

The MIT Press

Cambridge, Massachusetts

Preface

This book is a *supplement* to econometrics texts at either the introductory or advanced level. As such its purpose is to enhance the ability of the text to communicate the subject of econometrics to its readers. Unlike most such supplements, it does not do this through new examples, applications or exercises; instead it provides an overview of the subject and an intuitive feel for its concepts and techniques, without the usual clutter of notation and technical detail that necessarily characterize an econometrics textbook.

It is often said of econometrics textbooks that their readers miss the forest for the trees. This is inevitable – the terminology and techniques that must be taught do not allow the text to convey a proper intuitive sense of 'What's it all about?' and 'How does it all fit together?' All econometrics textbooks fail to provide this overview. This is not from lack of trying – most textbooks have excellent passages containing the relevant insights and interpretations. They make good sense to instructors, but do *not* make the expected impact on the students. Why? Because these insights and interpretations are broken up, appearing throughout the book, mixed with the technical details. In their struggle to keep up with notation and to learn these technical details, students miss the overview so essential to a real understanding of those details.

This guide was reviewed in manuscript form by a large number of instructors and used by several classes at both the introductory and advanced levels. It is interesting that the majority of the instructors did not feel that this book was needed whereas the students were unanimously of the opposite opinion. To the instructors, this guide presented nothing they did not feel was already in textbooks. To the students, this guide provided an overview of the subject and non-technical descriptions of its landmarks, without distracting notation, formulae, proofs or technical details; it provided the uninitiated with a perspective from which it was possible to assimilate the detail of the textbook more easily.

A brief introductory chapter sets the stage. The second chapter discusses at some length the criteria for choosing estimators, and in doing so develops many of the basic concepts used throughout the book. The third chapter structures the basic overview of the subject

matter, presenting the five assumptions of the classical linear regression model and explaining how most problems encountered in econometrics can be interpreted as a violation of one of these assumptions. The fourth chapter discusses some concepts of inference to provide a foundation for later chapters. Chapters 5 to 9 take each of the five assumptions of the classical linear regression model in turn and describe possible violations of these assumptions, implications of these violations, and means of resolving the resulting estimation problems. A final chapter examines some selected topics.

To minimize readers' distractions, there are no footnotes. All references, peripheral points and details worthy of comment are relegated to a section at the end of each chapter entitled 'General Notes'. The few technical discussions that appear in the book are placed in end-of-chapter sections entitled 'Technical Notes'. Students are advised to wait until a second or third reading of a chapter before attempting to integrate the material in the General or Technical Notes with the body of the chapter. A glossary explains common econometric terms not found in the body of this book.

No author is without debts. Thanks are due to Judy Alexander, Sandy Christensen, Larry Clark, Charles Conrod, John Cuddington, Art Goldberger, Dick Holmes, Teh Hu, Jon Nelson, Angus Oliver, Rod Peterson, Alan Sleeman and numerous students for comments both general and detailed. Expert typing was provided by Carol Eddy, Patricia Lyles and Donna Wilson. None of these people can be held responsible for any shortcomings or errors this book may contain; only the author can be blamed.

Dedication

To ANNA and RED

who, until they discovered what an econometrician was, were very impressed that their son might become one. With apologies to K. A. C. Manderville, I draw their attention to the following, adapted from *The Undoing of Lamia Gurdleneck.*

'You haven't told me yet,' said Lady Nuttal, 'what it is your fiancé does for a living.'

'He's an econometrician,' replied Lamia, with an annoying sense of being on the defensive.

Lady Nuttal was obviously taken aback. It had not occurred to her that econometricians entered into normal social relationships. The species, she would have surmised, was perpetuated in some collateral manner, like mules.

'But Aunt Sara, it's a very interesting profession,' said Lamia warmly.

'I don't doubt it,' said her aunt, who obviously doubted it very much. 'To express anything important in mere figures is so plainly impossible that there must be endless scope for well-paid advice on how to do it. But don't you think that life with an econometrician would be rather, shall we say, humdrum?'

Lamia was silent. She felt reluctant to discuss the surprising depth of emotional possibility which she had discovered below Edward's numerical veneer.

'It's not the figures themselves,' she said finally, 'it's what you do with them that matters.'

1. Introduction

1.1 What is Econometrics?

Strange as it may seem, there does not exist a generally accepted answer to this question. Responses vary from the silly 'Econometrics is what econometricians do' to the staid 'Econometrics is the study of the application of statistical methods to the analysis of economic phenomena', with sufficient disagreements to warrant an entire journal article devoted to this question (Tintner, 1953).

This confusion stems from the fact that econometricians wear many different hats. At times they are *mathematicians*, formulating economic theory in ways that make it appropriate for statistical testing. At times they are *accountants*, concerned with the problem of finding and collecting economic data and relating theoretical economic variables to observable ones. At times they are *applied statisticians*, spending hours with the computer trying to estimate economic relationships or predict economic events. And at times they are *theoretical statisticians*, applying their skills to the development of statistical techniques appropriate to the empirical problems characterizing the science of economics. It is to the last of these roles that the term 'econometric theory' applies, and it is on this aspect of econometrics that most textbooks on the subject focus. This guide is accordingly devoted to this 'econometric theory' dimension of econometrics, discussing the empirical problems typical of economics and the statistical techniques used to overcome these problems.

What distinguishes an econometrician from a statistician is the former's preoccupation with problems caused by violations of statisticians' standard assumptions; due to the nature of economic relationships and the lack of controlled experimentation these assumptions are seldom met. Patching up statistical methods to deal with situations frequently encountered in empirical work in economics has created a large battery of extremely sophisticated statistical techniques. In fact, econometricians are often accused of using sledgehammers to crack open peanuts while turning a blind eye to data deficiencies and the many questionable assumptions required for the

successful application of these techniques. Valavanis has expressed this feeling forcefully:

> Econometric theory is like an exquisitely balanced French recipe, spelling out precisely with how many turns to mix the sauce, how many carats of spice to add, and for how many milliseconds to bake the mixture at exactly 474 degrees of temperature. But when the statistical cook turns to raw materials, he finds that hearts of cactus fruit are unavailable, so he substitutes chunks of cantaloupe; where the recipe calls for vermicelli he uses shredded wheat; and he substitutes green garment dye for curry, ping-pong balls for turtle's eggs, and, for Chalifougnac vintage 1883, a can of turpentine. [1959, p. 83]

Criticisms of econometrics along these lines are not uncommon. Rebuttals cite improvements in data collecting, extol the fruits of the computer revolution and provide examples of improvements in estimation due to advanced techniques. It remains a fact, though, that in practice good results depend as much on the input of sound and imaginative economic theory as on the application of correct statistical methods. The skill of the econometrician lies in judiciously mixing these two essential ingredients; in the words of Malinvaud:

> The art of the econometrician consists in finding the set of assumptions which are both sufficiently specific and sufficiently realistic to allow him to take the best possible advantage of the data available to him. [1966, p. 514]

Modern econometrics texts try to infuse this art into students by providing a large number of detailed examples of empirical applications. This important dimension of econometrics texts lies beyond the scope of this book. Readers should keep this in mind as they use this guide to improve their understanding of the purely statistical methods of econometrics.

1.2 The Disturbance Term

A major distinction between economists and econometricians is the latter's concern with disturbance terms. An economist will specify, for example, that consumption is a function of income, and write $C = f(Y)$ where C is consumption and Y is income. An econometrician will claim that this relationship must also include a *disturbance* (or *error*) term, and may alter the equation to read

$C = f(Y) + \varepsilon$ where ε (epsilon) is a disturbance term. Without the disturbance term the relationship is said to be *exact* or *deterministic*; with the disturbance term it is said to be *stochastic*.

The word 'stochastic' comes from the Greek 'stokhos' meaning a target or bull's eye. A stochastic relationship is not always right on target in the sense that it predicts the precise value of the variable being explained, just as a dart thrown at a target seldom hits the bull's eye. The disturbance term is used to capture explicitly the size of these 'misses' or 'errors'. The existence of the disturbance term is justified in three main ways. (Note these are not mutually exclusive.)

(1) *Omission of the influence of innumerable chance events.* Although income might be the major determinant of the level of consumption, it is not the only determinant. Other variables, such as the interest rate of liquid asset holdings, may have a systematic influence on consumption. Their omission constitutes one type of *specification error*: the nature of the economic relationship is not correctly specified. In addition to these systematic influences, however, are innumerable less-systematic influences such as weather variations, taste changes, earthquakes, epidemics and postal strikes. Although some of these variables may have a significant impact on consumption, and thus should definitely be included in the specified relationship, many have only a very slight, irregular influence; the disturbance is often viewed as representing the net influence of a large number of such small and independent causes.

(2) *Measurement error.* It may be the case that the variable being explained cannot be measured accurately, either because of data collection difficulties or because it is inherently unmeasurable and a proxy variable must be used in its stead. The disturbance term can in these circumstances be thought of as representing this measurement error. But note that errors in measuring the explaining variable(s) (as opposed to the variable being explained) are treated differently under the heading *errors in variables*. The terminology *errors in equations* is sometimes used to denote errors or disturbances in the context in which they are being discussed here.

(3) *Human indeterminancy.* Some people believe that human behaviour is such that actions taken under identical circumstances will differ in a random way. The disturbance term can be thought of as representing this inherent randomness in human behaviour.

Associated with any explanatory relationship are unknown con-

stants, called *parameters*, which tie the relevant variables into an equation. For example, the relationship between consumption and income could be specified as

$$C = \beta_1 + \beta_2 Y + \varepsilon$$

where β_1 and β_2 are the parameters characterizing this consumption function. Economists are often keenly interested in learning the values of these unknown parameters.

The existence of the disturbance term, coupled with the fact that its magnitude is unknown, makes calculation of these parameter values impossible. Instead they must be *estimated.* It is on this task, the estimation of parameter values, that the bulk of econometric theory focuses. The success of econometricians' methods of estimating parameter values depends in large part on the nature of the disturbance term; statistical assumptions concerning the characteristics of the disturbance term, and means of testing these assumptions, therefore play a prominent role in econometric theory.

1.3 Estimates and Estimators

In their mathematical notation, econometricians usually employ Greek letters to represent the true, unknown values of parameters. The Greek letter most often used in this context is beta (β). Thus, throughout this book, β is used as the parameter value that the econometrician is seeking to learn. Of course, no one ever actually learns the value of β, but it can be estimated: via statistical techniques empirical data can be used to take an educated guess at β. In any particular application, an estimate of β is simply a number. For example, β might be estimated as 16.2. But in general econometricians are seldom interested in estimating a single parameter; economic relationships are usually sufficiently complex as to require more than one parameter, and because these parameters occur in the same relationship better estimates of these parameters can be obtained if they are estimated together (i.e., the influence of one explaining variable is more accurately captured if the influence of the other explaining variables is simultaneously accounted for). As a result, β seldom refers to a single parameter value; it almost always refers to a set of parameter values, individually called $\beta_1, \beta_2, \ldots, \beta_k$ where k is the number of different parameters in the set. β is then referred to as a vector and is written as

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

In any particular application, an estimate of β will be a set of numbers. For example, if three parameters are being estimated (i.e., the dimension of β is three), β might be estimated as

$$\begin{bmatrix} 0.8 \\ 1.2 \\ -4.6 \end{bmatrix}$$

In general, econometric theory focuses not on the estimate itself, but on the *estimator* – the formula or 'recipe' by which the data are transformed into an actual estimate. The reason for this is that the justification of an estimate computed from a particular sample rests on a justification of the estimation method (the estimator). The econometrician has no way of knowing the actual values of the disturbances inherent in a sample of data; depending on these disturbances, an estimate calculated from that sample could be quite inaccurate. It is therefore impossible to justify the estimate itself. However, it may be the case that the econometrician can justify the estimator by showing, for example, that that estimator 'usually' produces an estimate that is 'quite close' to the true parameter value regardless of the particular sample chosen. (The meaning of this sentence is discussed at length in the next chapter.) Thus an estimate of β from a particular sample is defended by justifying the estimator.

Because attention is focused on estimators of β, a convenient way of denoting those estimators is required. An easy way of doing this is to place a mark over the β or a superscript on it. Thus $\hat{\beta}$ (beta-hat) and β^* (beta-star) are often used to denote estimators of beta. One estimator, the ordinary least squares (OLS) estimator, is very popular in econometrics; the notation β^{OLS} is used throughout this book to represent it. Alternative estimators are denoted by $\hat{\beta}$, β^*, or something similar.

1.4 Good and Preferred Estimators

Any fool can produce an estimator of β, since literally an infinite

number of them exists, i.e., there exists an infinite number of different ways in which a sample of data can be used to produce an estimate of β, all but a few of these ways producing 'bad' estimates. What distinguishes an econometrician is the ability to produce 'good' estimators, which in turn produce 'good' estimates. One of these 'good' estimators could be chosen as the 'best' or 'preferred' estimator and be used to generate the 'preferred' estimate of β. What further distinguishes an econometrician is his ability to provide 'good' estimators in a variety of different estimating contexts. The set of 'good' estimators (and the choice of 'preferred' estimator) is not the same in all estimating problems. In fact, a 'good' estimator in one estimating situation could be a 'bad' estimator in another situation.

The study of econometrics revolves around how to generate a 'good' or the 'preferred' estimator in a given estimating situation. But before the 'how to' can be explained, the meaning of 'good' and 'preferred' must be made clear. This takes the discussion into the subjective realm: the meaning of 'good' or 'preferred' estimator depends upon the subjective values of the person doing the estimating. The best the econometrician can do under these circumstances is to recognize the more popular criteria used in this regard and generate estimators that meet one or more of these criteria. Estimators meeting certain of these criteria could be called 'good' estimators. The ultimate choice of the 'preferred' estimator, however, lies in the hands of the person doing the estimating, for it is his or her value judgments that determine which of these criteria is the most important. This value judgment may well be influenced by the purpose for which the estimate is sought, in addition to the subjective prejudices of the individual.

Clearly, our investigation of the subject of econometrics can go no further until the possible criteria for a 'good' estimator are discussed. This is the purpose of the next chapter.

General Notes

1.1

- The term 'econometrics' first came into prominence with the formation in the early 1930s of the Econometric Society and the founding of the journal *Econometrica*. The introduction of Dowling and Glahe (1970) surveys briefly the landmark publications related to econometrics.
- Brunner (1973), Rubner (1970) and Streissler (1970) are good sources of cynical views of econometrics. More comments appear in this book in section 8A.2 on errors in variables and section 10.5 on prediction. Fair (1973) and Fromm and Schink (1973) are examples of studies defending the use of sophisticated econometric techniques.
- Critics might choose to paraphrase the Malinvaud quote as 'The art of drawing a crooked line from an unproved assumption to a foregone conclusion'. The importance of a proper understanding of econometric techniques in the face of a potential inferiority of econometrics to inspired economic theorizing is captured nicely by Samuelson (1965) p. 9: 'Even if a scientific regularity were less accurate than the intuitive hunches of a virtuoso, the fact that it can be put into operation by thousands of people who are not virtuosos gives it a transcendental importance.' This guide is designed for those of us who are not virtuosos!
- There exist several books addressed to the empirical applications dimension of econometrics. Examples are Bridge (1971), Cramer (1971), Desai (1976), Pindyck and Rubinfeld (1976), Wallis (1973) and Wynn and Holden (1974). Evans (1969) is also useful in this regard. In addition, the econometric theory texts of Gujarati (1978), Intriligator (1978), Koutsoyiannis (1977) and Maddala (1977) all contain numerous examples of empirical applications. Zellner (1968) contains many well-known articles addressed to this empirical dimension of econometrics.

1.2

- The error term associated with a relationship need not necessarily be additive, as it is in the example cited. For some non-linear functions it is often convenient to specify the error term in a multiplicative form, for example, as noted later in section 5.3 on non-linear functional forms. In other instances it may be appropriate to build the stochastic element into the relationship by specifying the parameters to be random variables rather than constants. (This is called the random-coefficients model.) These are usually treated as special topics in econometrics.
- Econometricians usually do not know the actual form of the economic relationship being studied and consequently commit a specification error when formulating the estimation problem, either by omitting variables that should be included or by adopting an incorrect functional form. In these cases the econometrician often views the disturbance term as incorporating this specification error in addition to the other types of error cited. Following this operational procedure creates an error term with unusual properties; estimation under these circumstances is discussed under the heading of specification error (chapter 5).
- Estimation of parameter values is not the only purpose of econometrics. Two other major themes can be identified: testing of hypotheses and economic forecasting. Because both these problems are intimately related to the estimation of parameter values, it is not misleading to characterize econometrics as being primarily concerned with parameter estimation.
- In terms of the throwing-darts-at-a-target analogy, characterizing disturbance terms refers to describing the nature of the misses: are the darts distributed uniformly around the bull's eye? Is the average miss large or small? Does the average miss depend on who is throwing the darts? Is a

miss to the right likely to be followed by another miss to the right? In later chapters the statistical specification of these characteristics and the related terminology (such as 'homoskedasticity' and 'autocorrelated errors') are explained in considerable detail.

1.3

- An estimator is simply an algebraic function of a potential sample of data; once the sample is drawn, this function creates an actual numerical estimate.
- Chapter 2 discusses in detail the means whereby an estimator is 'justified' and compared to alternative estimators.

1.4

- The terminology 'preferred' estimator is used instead of the term 'best' estimator because the latter has a specific meaning in econometrics. This is explained in chapter 2.

2. Criteria for Estimators

2.1 Introduction

Chapter 1 posed the question 'What is a "good" estimator?' The aim of this chapter is to answer that question by describing a number of criteria that econometricians feel are measures of 'goodness'. These criteria are discussed under the following headings:

1. Computational cost
2. Least squares
3. Highest R^2
4. Unbiasedness
5. Best unbiased
6. Mean square error
7. Asymptotic properties
8. Maximum likelihood

Since econometrics can be characterized as a search for estimators satisfying one or more of these criteria, care is taken in the discussion of these criteria to ensure that the reader understands fully the meaning of the different criteria and the terminology associated with them. Many fundamental ideas of econometrics, critical to the question 'What's econometrics all about?', are presented in this chapter.

2.2 Computational Cost

To anyone, but particularly to economists, the extra benefit associated with choosing one estimator over another must be compared to its extra cost, where cost refers to expenditure of both money and effort. Thus the computational ease and cost of using one estimator rather than another must be taken into account whenever selecting an estimator. Fortunately, the existence and ready availability of high-speed computers, along with standard packaged routines for most of the popular estimators, has made computational cost very low. As a result, this criterion does not play as strong a role as it once did. Its

influence is now felt only when dealing with two kinds of estimators. One is the case of an atypical estimation procedure for which there does not exist a readily available packaged computer program and for which the cost of programming is high. The second is an estimation method for which the cost of running a packaged program is high because it needs large quantities of computer time; the so-called systems methods of simultaneous equation estimation, involving the simultaneous estimation of an entire set of βs, fall into this latter category.

2.3 Least Squares

For any set of values of the parameters characterizing a relationship, estimated values of the dependent variable (the variable being explained) can be calculated using the values of the independent variables (the explaining variables) in the data set. These estimated values (called $\hat{y}$) of the dependent variable can be subtracted from the actual values (y) of the dependent variable in the data set to produce what are called the *residuals* ($y - \hat{y}$). These residuals could be thought of as estimates of the unknown disturbances inherent in the data set. This is illustrated in Fig. 2.1. The line labelled $\hat{y}$ is the estimated

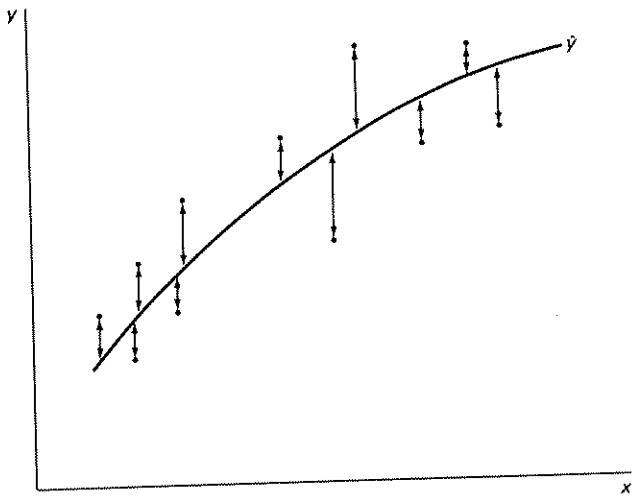

Fig. 2.1

relationship corresponding to a specific set of values of the unknown parameters. The dots represent actual observations on the dependent variable y and the independent variable x. Each observation is a certain vertical distance away from the estimated line, as pictured by the double-ended arrows. The lengths of these double-ended arrows measure the residuals. A different set of specific values of the parameters would create a different estimating line and thus a different set of residuals.

It seems natural to ask that a 'good' estimator be one that generates a set of estimates of the parameters that makes these residuals 'small'. Controversy arises, however, over the appropriate definition of 'small'. Although it is agreed that the estimator should be chosen to minimize a weighted sum of all these residuals, full agreement as to what the weights should be does not exist. For example, those feeling that all residuals should be weighted equally advocate choosing the estimator that minimizes the sum of the absolute values of these residuals. Those feeling that large residuals should be avoided advocate weighting larger residuals more heavily by choosing the estimator that minimizes the sum of the squared values of these residuals. Those worried about misplaced decimals and other data errors advocate placing a constant (sometimes zero) weight on the squared values of particularly large residuals. Those concerned only with whether or not a residual is bigger than some specified value suggest placing a zero weight on residuals smaller than this critical value and a weight equal to the inverse of the residual on residuals larger than this value. Clearly a large number of alternative definitions could be proposed, each with appealing features.

By far the most popular of these definitions of 'small' is the minimization of the sum of squared residuals. The estimator generating the set of values of the parameters that minimizes the sum of squared residuals is called the *ordinary least squares* estimator. It is referred to as the OLS estimator and is denoted by β^{OLS} in this book. This estimator is probably the most popular estimator among researchers doing empirical work. The reason for this popularity, however, does *not* stem from the fact that it makes the residuals 'small' by minimizing the sum of squared residuals. Many econometricians are leery of this criterion because minimizing the sum of squared residuals does not say anything specific about the relationship of the estimator to the true parameter value β that it is estimating. In fact, it is possible to be too successful in minimizing the sum of squared residuals, accounting for so many unique features of that *particular sample* that the estimator loses its general validity, in

the sense that, were that estimator applied to a new sample, poor estimates would result. The great popularity of the OLS estimator comes from the fact that in some estimating problems (but not all!) it scores well on some of the other criteria, described below, that are thought to be of greater importance. A secondary reason for its popularity is its computational ease; all computer packages include the OLS estimator for linear relationships, and many have routines for non-linear cases.

Because the OLS estimator is used so much in econometrics, the characteristics of this estimator in different estimating problems are explored very thoroughly by all econometrics texts. The OLS estimator *always* minimizes the sum of squared residuals; but it does *not* always meet other criteria that econometricians feel are more important. As will become clear in the next chapter, the subject of econometrics can be characterized as an attempt to find alternative estimators to the OLS estimator for situations in which the OLS estimator does not meet the estimating criterion considered to be of greatest importance in the problem at hand.

2.4 Highest R^2

A statistic that appears frequently in econometrics is the coefficient of determination, R^2. It is supposed to represent the proportion of the variation in the dependent variable 'explained' by variation in the independent variables. It does this in a meaningful sense in the case of a linear relationship estimated by OLS. In this case it happens that the sum of squared deviations of the dependent variable about its mean (the 'total' variation in the dependent variable) can be broken into two parts, called the 'explained' variation (the sum of squared deviations of the estimated values of the dependent variable around their mean) and the 'unexplained' variation (the sum of squared residuals). R^2 is measured either as the ratio of the 'explained' variation to the 'total' variation or, equivalently, as 1 minus the ratio of the 'unexplained' variation to the 'total' variation, and thus represents the percentage of variation in the dependent variable 'explained' by variation in the independent variables.

Because the OLS estimator minimizes the sum of squared residuals (the 'unexplained' variation), it automatically maximizes R^2. Thus maximization of R^2, as a criterion for an estimator, is formally identical to the least squares criterion, and as such it really does not

deserve a separate section in this chapter. It is given a separate section for two reasons. The first is that the formal identity between the highest R^2 criterion and the least squares criterion is worthy of emphasis. And the second is to distinguish clearly the difference between applying R^2 as a criterion in the context of searching for a 'good' estimator when the functional form and included independent variables are known, as is the case in the present discussion, and using R^2 to help determine the proper functional form and the appropriate independent variables to be included. This latter use of R^2 (and its misuse) are discussed later in the book (in sections 4.5, 5.2 and 5.3).

2.5 Unbiasedness

Suppose we perform the conceptual experiment of taking what is called a *repeated* sample: keeping the values of the independent variables unchanged, we obtain new observations for the dependent variable by drawing a new set of disturbances. This could be repeated, say, 2000 times, obtaining 2000 of these repeated samples. For each of these repeated samples we could use an estimator β^* to calculate an estimate of β. Because the samples differ, these 2000 estimates will not be the same. The manner in which these estimates are distributed is called the *sampling distribution* of β^*. This is illustrated for the one-dimensional case in Fig. 2.2, where the sampling distribution of the estimator is labelled $f(\beta^*)$. It is simply the probability density function of β^*, approximated by using the 2000 estimates of β to construct a histogram, which in turn is used to approximate the relative frequencies of different estimates of β from the estimator β^*. The sampling distribution of an alternative estimator, $\hat{\beta}$, is also shown in Fig. 2.2.

This concept of a sampling distribution, the distribution of estimates produced by an estimator in repeated sampling, is crucial to an understanding of econometrics. Most estimators are adopted because their sampling distributions have 'good' properties; the criteria discussed in this and the following three sections are directly concerned with the nature of an estimator's sampling distribution.

The first of these properties is unbiasedness. An estimator β^* is said to be an *unbiased* estimator of β if the mean of its sampling distribution is equal to β, i.e., if the average value of β^* in repeated sampling is β. The mean of the sampling distribution of β^* is called the expected value of β^* and is written $E\beta^*$; the bias of β^* is the difference between $E\beta^*$ and β. In Fig. 2.2, β^* is seen to be unbiased, whereas $\hat{\beta}$ has a bias

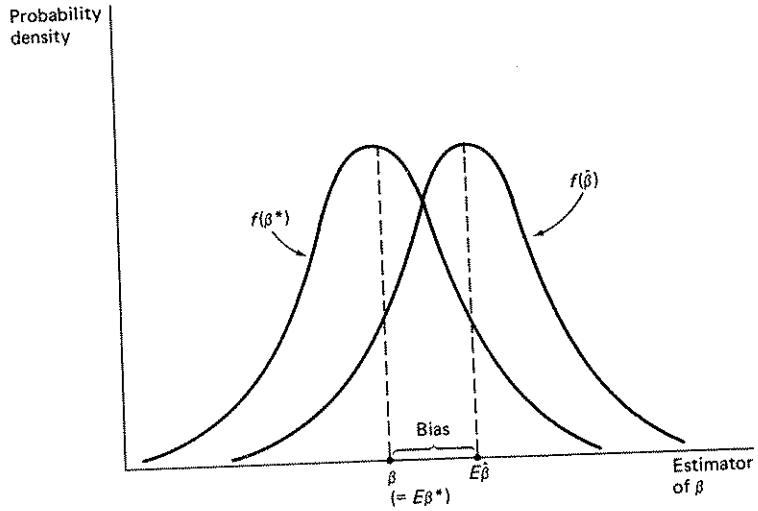

Fig. 2.2

of size $(E\hat{\beta} - \beta)$. The property of unbiasedness does not mean that $\beta^* = \beta$; it says only that if we could undertake repeated sampling an infinite number of times, we would get the correct estimate 'on the average'.

The OLS criterion can be applied with no information concerning how the data were generated. This is not the case for the unbiasedness criterion (and all other criteria related to the sampling distribution) since this knowledge is required to construct the sampling distribution. Econometricians have therefore developed a standard set of assumptions (discussed in chapter 3) concerning the way in which observations are generated. The general, but not the specific, way in which the disturbances are distributed is an important component of this. These assumptions are sufficient to allow the basic nature of the sampling distribution of many estimators to be calculated, either by mathematical means (part of the technical skill of an econometrician) or, failing that, by an empirical means called a Monte Carlo study (discussed in the general notes to this section).

Although the mean of a distribution is not necessarily the ideal measure of its location (the median or mode in some circumstances might be considered superior), most econometricians consider unbiasedness a desirable property for an estimator to have. This preference for an unbiased estimator stems from the *hope* that a

particular estimate (i.e., from the sample at hand) will be close to the mean of the estimator's sampling distribution. Having to justify a particular estimate on a 'hope' is not especially satisfactory, however. As a result, econometricians have recognized that being centred over the parameter to be estimated is only *one* good property that the sampling distribution of an estimator can have. The variance of the sampling distribution, discussed next, is also of great importance.

2.6 Best Unbiased

In some econometric problems it is impossible to find an unbiased estimator. But whenever one unbiased estimator can be found, it is usually the case that a large number of other unbiased estimators can also be found. In this circumstance the unbiased estimator whose sampling distribution has the smallest variance is considered the most desirable of these unbiased estimators; it is called the *best unbiased* estimator, or the most *efficient* estimator among all unbiased estimators. Why it is considered the most desirable of all unbiased estimators is easy to visualize. In Fig. 2.3 the sampling distributions of

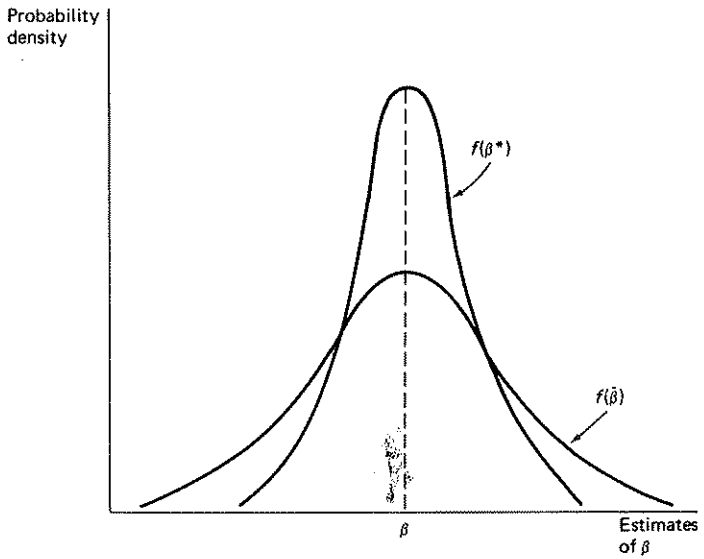

Fig. 2.3

two unbiased estimators are drawn. The sampling distribution of the estimator $\hat{\beta}$, denoted $f(\hat{\beta})$, is drawn 'flatter' or 'wider' than the sampling distribution of β^*, reflecting the larger variance of $\hat{\beta}$. Although both estimators would produce estimates in repeated samples whose average would be β, the estimates from $\hat{\beta}$ would range more widely and thus would be less desirable. A researcher using $\hat{\beta}$ would be less certain that his estimate was close to β than would a researcher using β^*.

Sometimes reference is made to a criterion called 'minimum variance'. This criterion, by itself, is meaningless. Consider the estimator $\beta^* = 5.2$ (i.e., whenever a sample is taken, estimate β by 5.2, ignoring the sample). This estimator has a variance of zero, the smallest possible variance, but no one would use this estimator because it performs so poorly on other criteria such as unbiasedness. (It is interesting to note, however, that it performs exceptionally well on the computational cost criterion!) Thus, whenever the minimum variance, or 'efficiency', criterion is mentioned, there must exist some additional constraint, such as unbiasedness, accompanying that criterion. When the additional constraint accompanying the minimum variance criterion is that the estimators under consideration be unbiased, the estimator is referred to as the *best unbiased* estimator. Unfortunately, in many cases it is impossible to determine mathematically which estimator, of all unbiased estimators, has the smallest variance. Because of this problem, econometricians frequently add the further restriction that the estimator be *linear*, i.e., that it be a linear function of the observations on the dependent variable. This reduces the task of finding the efficient estimator to mathematically manageable proportions. An estimator that is linear, unbiased and has minimum variance among all linear unbiased estimators is called the *Best Linear Unbiased Estimator* (BLUE). The BLUE is very popular among econometricians.

This discussion of minimum variance or efficiency has been implicitly undertaken in the context of a unidimensional estimator, i.e., the case in which β is a single number rather than a vector containing several numbers. In the multidimensional case the variance of $\hat{\beta}$ becomes a matrix called the variance–covariance matrix of $\hat{\beta}$. This creates special problems in determining which estimator has the smallest variance. The technical notes to this section discuss this further.

2.7 Mean Square Error (MSE)

Using the best unbiased criterion allows unbiasedness to play an extremely strong role in determining the choice of an estimator, since only unbiased estimators are considered. It may well be the case that by restricting attention to only unbiased estimators we are ignoring estimators that are only slightly biased but have considerably lower variances. This phenomenon is illustrated in Fig. 2.4. The sampling distribution of $\hat{\beta}$, the best unbiased estimator, is labelled $f(\hat{\beta})$. β^* is a biased estimator with sampling distribution $f(\beta^*)$. It is apparent from Fig. 2.4 that although $f(\beta^*)$ is not centred over β, reflecting the bias of β^*, it is 'narrower' than $f(\hat{\beta})$, indicating a smaller variance. It should be clear from the diagram that most researchers would probably choose the biased estimator β^* in preference to the best unbiased estimator $\hat{\beta}$.

This trade-off between low bias and low variance is formalized by using as a criterion the minimization of a weighted average of the bias and the variance (i.e., choose the estimator that minimizes this weighted average). This is not a viable formalization, however, because the bias could be negative. One way to correct for this is to use the absolute value of the bias; a more popular way is to use its square. When the estimator is chosen so as to minimize a weighted

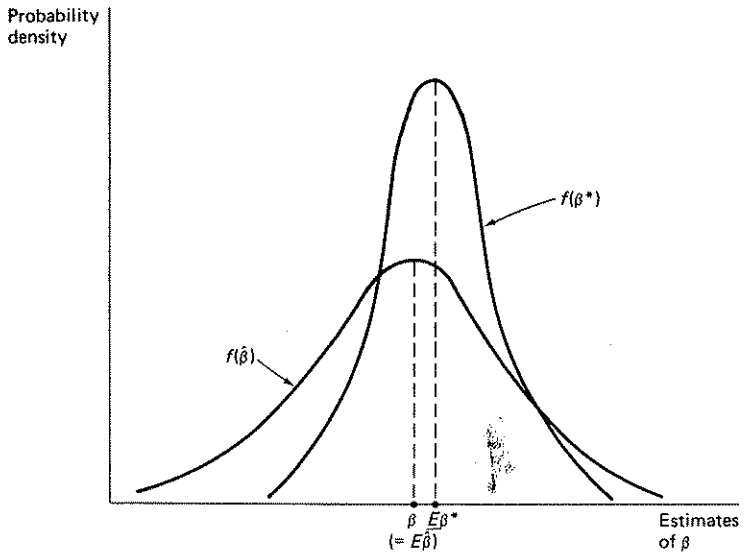

Fig. 2.4

average of the variance and the square of the bias, the estimator is said to be chosen on the *weighted square error* criterion. When the weights are equal, the criterion is the popular mean square error (MSE) criterion. The popularity of the mean square error criterion comes from an alternative derivation of this criterion: it happens that the expected value of a loss function consisting of the square of the difference between β and its estimate (i.e., the square of the estimation error) is the same as the sum of the variance and the squared bias. Minimization of the expected value of this loss function makes good intuitive sense as a criterion for choosing an estimator.

In practice, the MSE criterion is not usually adopted unless the best unbiased criterion is unable to produce estimates with small variances. The problem of multicollinearity, discussed in chapter 9, is an example of such a situation.

2.8 Asymptotic Properties

The estimator properties discussed in sections 2.5, 2.6 and 2.7 above relate to the nature of an estimator's sampling distribution. An unbiased estimator, for example, is one whose sampling distribution is centred over the true value of the parameter being estimated. These properties do not depend on the size of the sample of data at hand: an unbiased estimator, for example, is unbiased in both small and large samples. In many econometric problems, however, it is impossible to find estimators possessing these desirable sampling distribution properties in small samples. When this happens, as it frequently does, econometricians justify an estimator on the basis of its *asymptotic* properties – the nature of the estimator's sampling distribution in extremely large samples.

The sampling distribution of most estimators changes as the sample size changes. The sample mean statistic, for example, has a sampling distribution that is centred over the population mean but whose variance becomes smaller as the sample size becomes larger. In many cases it happens that a biased estimator becomes less and less biased as the sample size becomes larger and larger – as the sample size becomes larger its sampling distribution changes, such that the mean of its sampling distribution shifts closer to the true value of the parameter being estimated. Econometricians have formalized their study of these phenomena by structuring the concept of a *limiting* or *asymptotic distribution* and defining desirable asymptotic or 'large-

sample' properties of an estimator in terms of the character of its limiting distribution.

Consider the sequence of sampling distributions of an estimator $\hat{\beta}$, formed by calculating the sampling distribution of $\hat{\beta}$ for successively larger sample sizes. If the distributions in this sequence become more and more similar in form to some specific distribution (such as a normal distribution) as the sample size becomes extremely large, this specific distribution is called the *limiting* or *asymptotic distribution* of $\hat{\beta}$. Two basic estimator properties are defined in terms of the limiting distribution.

(1) If this limiting distribution of $\hat{\beta}$ tends to become concentrated on a particular value k as the sample size approaches infinity, k is said to be the *probability limit* of $\hat{\beta}$ and is written $\operatorname{plim} \hat{\beta} = k$; if $\operatorname{plim} \hat{\beta} = \beta$, then $\hat{\beta}$ is said to be *consistent.*

(2) The variance of the limiting distribution of $\hat{\beta}$ is called the asymptotic variance of $\hat{\beta}$; if $\hat{\beta}$ is consistent and its asymptotic variance is smaller than the asymptotic variance of all other consistent estimators, $\hat{\beta}$ is said to be *asymptotically efficient.*

A third, and weaker, asymptotic property that is often encountered relates to an estimator's asymptotic expectation. Rather than being defined in terms of the character of the limiting distribution, however, it is defined in terms of the limit of the expectation of $\hat{\beta}$. The *asymptotic expectation* of $\hat{\beta}$ is defined as

$$\lim_{T \to \infty} E(\hat{\beta});$$

if this asymptotic expectation is equal to β then $\hat{\beta}$ is said to be *asymptotically unbiased.*

There exists considerable confusion concerning the exact definitions of these three asymptotic properties, their interrelationships and means of undertaking calculations related to them. Several technical notes to this section attempt to clarify these problems. For our intents and purposes, however, we can usefully conceptualize asymptotic unbiasedness as being the large-sample equivalent of unbiasedness. Similarly, consistency can be crudely conceptualized as the large-sample equivalent of the minimum mean square error property, since a consistent estimator can be (loosely speaking) thought of as having, in the limit, zero bias and a zero variance. Asymptotic efficiency is the large-sample equivalent of best unbiasedness: the variance of an asymptotically efficient estimator goes to zero faster than the variance of any other consistent estimator.

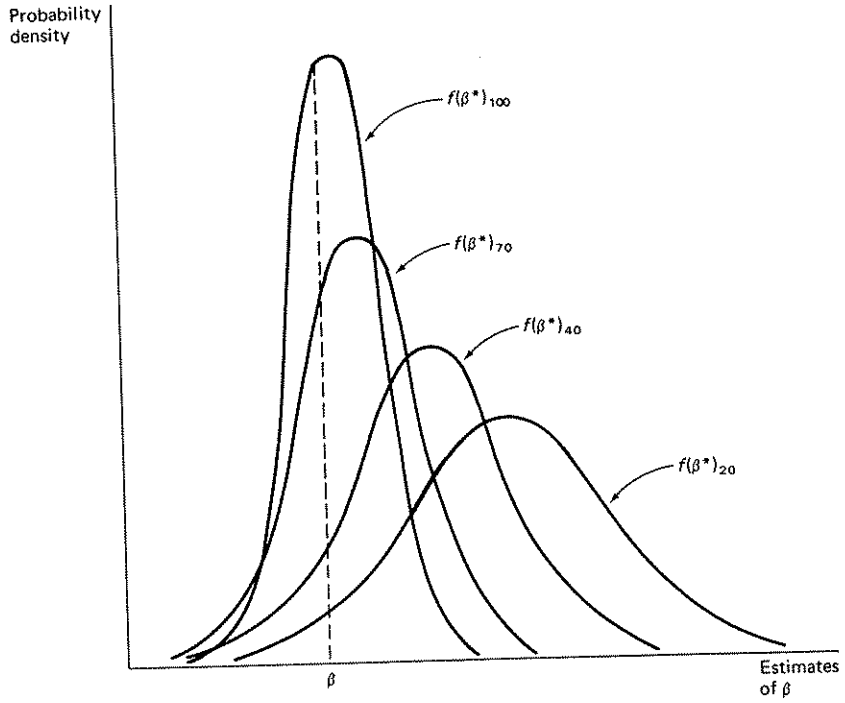

Fig. 2.5

Fig. 2.5 illustrates the basic appeal of asymptotic properties. For sample size 20 the sampling distribution of β^* is shown as $(\beta)_{20}$. Since this sampling distribution is not centred over β, the estimator β^* is biased. As shown in Fig. 2.5, however, as the sample size increases to 40, then 70 and then 100, the sampling distribution of β^* shifts so as to be more closely centred over β (i.e., it becomes less biased) and it becomes less spread out (i.e., its variance becomes smaller). If β^* were consistent, as the sample size increased to infinity the sampling distribution would shrink in width to a single vertical line, of infinite height, placed exactly at the point β.

It must be emphasized that these asymptotic criteria are only employed in situations in which estimators with the traditional desirable small-sample properties, such as unbiasedness, best unbiasedness and minimum mean square error, cannot be found. Since econometricians quite often must work with small samples, defending estimators on the basis of their asymptotic properties is legitimate only if

it is the case that estimators with desirable asymptotic properties have more desirable small-sample properties than do estimators without desirable asymptotic properties. Monte Carlo studies have shown that in general this supposition is warranted.

2.9 Maximum Likelihood

The maximum likelihood principle of estimation is based on the idea that the sample of data at hand is more likely to have come from a 'real world' characterized by one particular set of parameter values than from a 'real world' characterized by any other set of parameter values. The maximum likelihood estimate (MLE) of a vector of parameter values β is simply the particular vector β^{MLE} which gives the greatest probability of obtaining the observed data.

This idea is illustrated in Fig. 2.6. Each of the dots represents an observation on x drawn at random from a population with mean μ and variance σ^2. Pair A of parameter values, μ^A and $(\sigma^2)^A$, gives rise in Fig. 2.6 to the probability density function A for x, while the pair B, μ^B and $(\sigma^2)^B$, gives rise to probability density function B. Inspection of the diagram should reveal that the probability of having obtained the sample in question if the parameter values were μ^A and $(\sigma^2)^A$ is very low compared to the probability of having obtained the sample if the parameter values were μ^B and $(\sigma^2)^B$. On the maximum likelihood principle pair B is preferred to pair A as an estimate of μ and σ^2. The

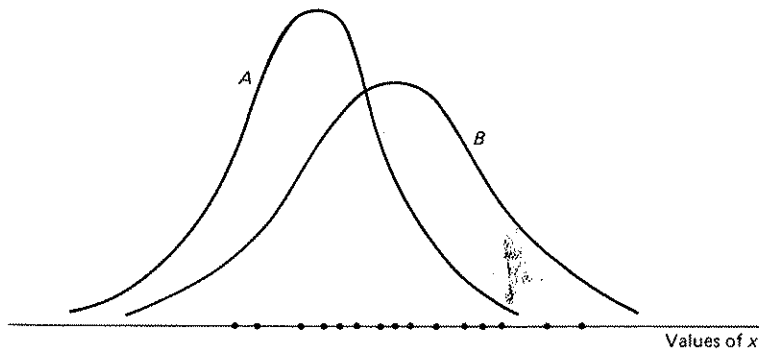

Fig. 2.6

maximum likelihood estimate is the particular pair of values μ^{MLE} and $(\sigma^2)^{MLE}$ that creates the greatest probability of having obtained the sample in question, i.e., no other pair of values would be preferred to this maximum likelihood pair, in the sense that pair B is preferred to pair A. The means by which the econometrician finds this maximum likelihood estimate is discussed briefly in the technical notes to this section.

In addition to its intuitive appeal, the maximum likelihood estimator has several desirable asymptotic properties. It is asymptotically unbiased, it is consistent, it is asymptotically efficient (its asymptotic variance is in fact given by the Cramer–Rao lower bound, discussed in the general notes to section 2.6) and it is distributed asymptotically normally.

These properties make maximum likelihood estimation very appealing for situations in which it is impossible to find estimators with desirable small-sample properties; in fact, as is apparent later in this book, maximum likelihood estimators are often proposed to deal with these kinds of situations. Although the technique has desirable theoretical properties, it is not widely adopted in these situations, however. One reason for this is that, in order to calculate the MLE, the econometrician must assume a *specific* (e.g., normal) distribution for the disturbance term. A second, more forceful reason is that this technique can sometimes be computationally difficult.

2.10 Adding Up

Because in most estimating situations there does not exist a 'superestimator' that is better than all other estimators on all or even most of these (or other) criteria, the ultimate choice of estimator is made by forming an 'overall judgment' of the desirableness of each available estimator by combining the degree to which an estimator meets each of these criteria with a subjective (on the part of the econometrician) evaluation of the importance of each of these criteria. Sometimes an econometrician will hold a particular criterion in very high esteem and this will determine the estimator chosen (if an estimator meeting this criterion can be found). More typically, other criteria also play a role in the econometrician's choice of estimator, so that, for example, only estimators with reasonable computational cost are considered. Among these major criteria, most attention seems to

be paid to the best unbiased criterion, with occasional deference to the mean square error criterion in estimating situations in which all unbiased estimators have variances that are considered too large. If estimators meeting these criteria cannot be found, as is often the case, asymptotic criteria are adopted.

A major skill of econometricians is the ability to determine estimator properties with regard to the criteria discussed in this chapter. This is done either through theoretical derivations using mathematics, part of the technical expertise of the econometrician, or through Monte Carlo studies. To derive estimator properties by either of these means, the mechanism generating the observations must be known; changing the way in which the observations are generated creates a new estimating problem, in which old estimators may have new properties and for which new estimators may have to be developed.

The OLS estimator has a special place in all this. When faced with any estimating problem, the econometric theorist usually tests the OLS estimator first, determining whether or not it has desirable properties. As seen in the next chapter, in some circumstances it does have desirable properties and is chosen as the 'preferred' estimator, but in many other circumstances it does not have desirable properties and a replacement must be found. The econometrician must investigate the circumstances under which the OLS estimate is desirable, and suggest alternative estimators for situations in which the OLS estimator is not acceptable. The next chapter explains how the econometrician orders this investigation.

General Notes

2.2

- Many recently developed programs have been designed specifically for econometricians, as their names (such as ESP, Econometric Software Package, and RAPE, Regression Analysis Program for Economists) indicate. Further, they are becoming more and more comprehensive, encompassing most of the econometric techniques discussed in textbooks. These packages should only be used by those well versed in econometric theory, however. Misleading or even erroneous results can easily be produced if these packages are used without a full understanding of the circumstances in which they are applicable, their inherent assumptions and the nature of their output; sound research cannot be produced merely

can no longer have the same meaning, however, and could possibly lie outside the 0–1 interval. The zero intercept case is discussed at length in Aigner (1971a) pp. 85–90. An alternative R^2 measure, in which the variations in y and $\hat{y}$ are measured as deviations from zero rather than their means, is suggested.

- R^2 is sensitive to the range of variation of the dependent variable, so that comparisons of R^2s must be undertaken with care. The favourite example used to illustrate this is the case of the consumption function versus the savings function. If savings is defined as income less consumption, income will do exactly as well in explaining variations in consumption as in explaining variations in saving, in the sense that the sum of squared residuals, the unexplained variation, will be exactly the same for each case. But in *percentage* terms, the unexplained variation will be a higher percentage of the variation in saving than of the variation in consumption because the latter are larger numbers. Thus the R^2 in the savings function case will be lower than in the consumption function case. This reflects the finding of Barrett (1974) that R^2 depends on the size of β.

2.5

- Problems in the use and interpretation of the unbiasedness criterion arise in contexts in which it is illegitimate even to conceptualize the independent variables as being fixed in repeated samples, since that criterion requires that the independent variables be fixed in repeated samples. Examples, such as cases in which the independent variables are themselves random variables, are not uncommon in econometrics. See chapters 8A and 8B.
- In contrast to the OLS and R^2 criteria, the unbiasedness criterion (and the other criteria related to the sampling distribution) say something specific about the relationship of the estimator to β, the parameter being estimated.
- Many econometricians are not impressed with the unbiasedness criterion, as our later discussion of the mean square error criterion will attest. Savage (1954) p. 244 goes so far as to say 'A serious reason to prefer unbiased estimates seems never to have been proposed.' This feeling probably stems from the fact that it is possible to have an 'unlucky' sample and thus a bad estimate, with only cold comfort from the knowledge that had all possible samples of that size been taken the correct estimate would have been hit on average. This is especially the case whenever a crucial outcome, such as in the case of a matter of life or death, or a decision to undertake a huge capital expenditure, hinges on a single correct estimate.
- A Monte Carlo study is an empirical construction of an estimator's sampling distribution, undertaken to learn the nature of the estimator's sampling distribution when mathematical techniques are unable to do so. Using a specific value for β and a representative set of values for the independent variables, a set of observations on the dependent variable is created by drawing a set of disturbances. Keeping the value of β and the set of values of the independent variables unchanged, this process is repeated, say, 100 times. This gives 100 'repeated samples' from each of

which a value for $\hat{\beta}$ can be calculated. The mean and variance of these calculated values are computed to estimate the mean and variance of the sampling distribution of $\hat{\beta}$, and sometimes these calculated values are formed into a histogram to give a graphical approximation to $\hat{\beta}$'s sampling distribution. Because β is *known*, conclusions can be drawn concerning the bias of $\hat{\beta}$, as well as other characteristics of $\hat{\beta}$ related to its sampling distribution.

- Frequent reference will be made throughout this book to Monte Carlo studies, since more often than not Monte Carlo studies are the only means through which information on an estimator's sampling distribution properties can be obtained. It was not until the high-speed computer was developed that Monte Carlo studies became common. Smith (1971) and Sowey (1973) have useful surveys of Monte Carlo studies in econometrics.

2.6

- The BLUE is simply the linear unbiased estimator that minimizes the variance of $\hat{\beta}$, i.e., minimizes the expected value of the square of the deviation of $\hat{\beta}$ from its expected value β. Minimizing the square of this deviation is not the only alternative. Some researchers might prefer to minimize the expectation of the absolute value of this deviation, for reasons similar to those given for the least absolute error estimator discussed in the notes to section 2.3. See, for example, Kadiyala (1972).
- The efficiency property can be sensitive to sample size. An estimator that is efficient for one sample size may not be for another. This is particularly the case when small samples are compared to very large samples.
- Linear estimators are not suitable for all estimating problems. For example, in estimating the variance σ^2 of the disturbance term, quadratic estimators are more appropriate. The traditional formula $SSE/(T-K)$, where K is the number of explanatory variables (including a constant), is under general conditions the best quadratic unbiased estimator of σ^2. Note that this formula is sometimes written as $SSE/(T-K-1)$. Use of this version occurs when K, the number of explanatory variables, does not include the constant (intercept) term.
- Although in many instances it is mathematically impossible to determine the best unbiased estimator (as opposed to the best *linear* unbiased estimator), this is not the case if the *specific* distribution is known. In this instance a lower bound, called the *Cramer–Rao lower bound*, for the variance (or variance–covariance matrix) of unbiased estimators can be calculated. Furthermore, if this lower bound is attained (which is not always the case), it is attained by a transformation of the maximum likelihood estimator (see section 2.9) that creates an unbiased estimator. See Kane (1968) pp. 187–8 for a discussion of the univariate case, and Kmenta (1971) pp. 159–60 for a discussion of the multivariate case. As an example, consider the sample mean statistic $\bar{X}$. Its variance, σ^2/T, is equal to the Cramer–Rao lower bound if the parent population is normal. Thus $\bar{X}$ is the best unbiased estimator of the mean of a normal population.

2.7

- Preference for the mean square error criterion over the unbiasedness criterion often hinges on the use to which the estimate is to be put. As an example of this, consider a man betting on horse races. If he is buying 'win' tickets, he will want an unbiased estimate of the winning horse, but if he is buying 'show' tickets it is not important that his horse wins the race (only that his horse finishes among the first three), so he will be willing to use a slightly biased estimator of the winning horse if it has a smaller variance.
- The difference between the variance of an estimator and its MSE is that the variance measures the dispersion of the estimator around its mean whereas the MSE measures its dispersion around the true value of the parameter being estimated. For unbiased estimators they are identical.
- Biased estimators with smaller variances than unbiased estimators are easy to find. For example, if $\hat{\beta}$ is an unbiased estimator with variance $V(\hat{\beta})$, then $0.9\hat{\beta}$ is a biased estimator with variance $0.81V(\hat{\beta})$. As a more relevant example, consider the fact that although $SSE/(T-K)$ is the best quadratic unbiased estimator of σ^2, as noted in section 2.6, it can be shown that among quadratic estimators the MSE estimator of σ^2 is $SSE/(T-K+2)$. See the technical notes to section 2.9.
- The MSE estimator has not been as popular as the best unbiased estimator because of the mathematical difficulties in its derivation. Furthermore, when it can be derived its formula often involves unknown coefficients (the value of β), making its application impossible. Monte Carlo studies have shown that approximating the estimator by using OLS estimates of the unknown parameters can sometimes circumvent this problem.
- Both the best unbiased and the MSE criteria rest heavily on the use of the quadratic loss function. Although this type of loss function has the disadvantage of being symmetrical, it has the advantage of being mathematically convenient and a reasonable approximation to more complicated loss functions.
- A set of non-linear estimators, called Stein-rule estimators, have been shown to be MSE estimators in certain estimating circumstances. Yancey and Judge (1976) discuss their applicability to econometrics. Efron and Morris (1977) have a good exposition.

2.8

- How large does the sample size have to be for estimators to display their asymptotic properties? The answer to this crucial question depends on the characteristics of the problem at hand. Goldfeld and Quandt (1972, p. 277) report an example in which a sample size of 30 is sufficiently large and an example in which a sample of 200 is required. They also note that large sample sizes are needed if interest focuses on estimation of estimator variances rather than on estimation of coefficients.
- An extremely appealing feature of probability limits is that the probability

limit of a non-linear function of an estimator is the non-linear function of the probability limit of the estimator, a property that does not hold true for expectations. Thus plim $g(\hat{\beta}) = g(\text{plim}\ \hat{\beta})$ but $Eg(\hat{\beta}) \neq g(E\hat{\beta})$, when g is a non-linear function. As a specific example, consider the problem of estimating $1/\beta$. Even if $\hat{\beta}$ is an unbiased estimator of β, $1/\hat{\beta}$ is *not* an unbiased estimator of $1/\beta$; but if $\hat{\beta}$ is a consistent estimator of β, $1/\hat{\beta}$ will be a consistent estimator of $1/\beta$.

It is this property that makes possible the algebraic derivation of asymptotic properties of many estimators in contexts in which the small-sample properties of these estimators cannot be deduced. Econometricians should not allow this attractive feature of asymptotic criteria to steer them away from examining the small-sample properties of alternative estimators however.

2.9

- Note that β^{MLE} is *not*, as is sometimes carelessly stated, the most probable value of β; the most probable value of β is β itself. (Only in a Bayesian interpretation, discussed later in this book, would the former statement be meaningful.) β^{MLE} is simply the value of β that maximizes the probability of drawing the sample actually obtained.
- Despite the fact that β^{MLE} is sometimes a biased estimator of β (although asymptotically unbiased), often a simple adjustment can be found that creates an unbiased estimator, and this unbiased estimator can be shown to be best unbiased (with no linearity requirement) through the relationship between the maximum likelihood estimator and the Cramer–Rao lower bound. For example, the maximum likelihood estimator of the variance of a random variable x is given by the formula

$$\sum_{i=1}^{T} (x_i - \bar{x})^2/T$$

which is a biased (but asymptotically unbiased) estimator of the true variance. By multiplying this expression by $T/(T-1)$, this estimator can be transformed into a best unbiased estimator.

- Maximum likelihood estimators have an invariance property similar to that of consistent estimators. The maximum likelihood estimator of a non-linear function of a parameter is the non-linear function of the maximum likelihood estimator of that parameter: $[g(\beta)]^{MLE} = g(\beta^{MLE})$ where g is a non-linear function. This greatly simplifies the algebraic derivations of maximum likelihood estimators, making adoption of this criterion more attractive.
- Goldfeld and Quandt (1972) conclude that the maximum likelihood technique performs well in a wide variety of applications and for relatively small sample sizes. It is particularly evident, from reading their book, that the maximum likelihood technique is well-suited to estimation involving non-linearities and other estimation problems. They do not feel that the computational costs of using this technique are prohibitive.
- Application of the maximum likelihood technique requires that the

econometrician assume a specific distribution for the error term. The normal distribution is invariably chosen for this purpose, usually on the grounds that the error term consists of the sum of a large number of random shocks and thus by the Central Limit Theorem can be considered to be approximately normally distributed. (See Bartels, 1977, for a warning on the use of this argument.) A more compelling reason is that the normal distribution is relatively easy to work with.

- Kmenta (1971) pp. 174–82 has a clear discussion of maximum likelihood estimation. A good brief exposition is in Kane (1968) pp. 177–80. Valavanis (1959) pp. 23–6, an econometrics text subtitled 'An Introduction to Maximum Likelihood Methods', has an interesting account of the meaning of the maximum likelihood technique.

2.10

- The criteria presented in this chapter are the major criteria used by econometricians to choose estimators. (An exception is the decision-theoretic criterion used in Bayesian estimation; this is discussed in chapter 10.) Other criteria do exist, but are less popular. An example is the *analogy principle of estimation*: estimate parameters by sample statistics that have the same property in the sample as the parameters do in the population. See chapter 2 of Goldberger (1968a) for an interpretation of the OLS estimator in these terms. This approach is sometimes called the *method of moments* because it implies that a moment of the population distribution will be estimated by the corresponding moment of the sample.
- Two good general references covering in more detail the material presented in this chapter are Kmenta (1971) pp. 8–15, 154–86, and Kane (1968) chapter 8.

Technical Notes

2.5

- The expected value of a variable x is defined formally as $Ex = \int xf(x)dx$ where f is the probability density function of x.

2.6

- In our discussion of unbiasedness, no confusion could arise from β being multidimensional: an estimator's expected value is either equal to β (in every dimension) or it is not. But in the case of the variance of an estimator, confusion could arise. An estimator β^* that is k-dimensional really consists of k different estimators, one for each dimension of β. These k different estimators all have their own variances. If all k of the variances associated with the estimator β^* are smaller than their respective counterparts of the estimator $\hat{\beta}$, then it is clear that the variance of β^* can be considered smaller than the variance of $\hat{\beta}$. For example, if β is two-dimensional, consisting of two separate parameters β_1 and β_2

$$\left(\text{i.e., } \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}\right),$$

an estimator β^* would consist of two estimators β_1^* and β_2^*. If β^* were an unbiased estimator of β, β_1^* would be an unbiased estimator of β_1 and β_2^* would be an unbiased estimator of β_2. The estimators β_1^* and β_2^* would each have variances. Suppose their variances were 3.1 and 7.4, respectively. Now suppose $\hat{\beta}$, consisting of $\hat{\beta}_1$ and $\hat{\beta}_2$, is another unbiased estimator, where $\hat{\beta}_1$ and $\hat{\beta}_2$ have variances 5.6 and 8.3, respectively. In this example, since the variance of β_1^* is less than the variance of $\hat{\beta}_1$ and the variance of β_2^* is less than the variance of $\hat{\beta}_2$, it is clear that the 'variance' of β^* is less than the variance of $\hat{\beta}$. But what if the variance of $\hat{\beta}_2$ were 6.3 instead of 8.3? Then it is *not* clear which 'variance' is smallest.

- An additional complication exists in comparing the variances of estimators of a multidimensional β. There may exist a non-zero covariance between the estimators of the separate components of β. For example, a positive covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$ implies that whenever $\hat{\beta}_1$ overestimates β_1, there is a tendency for $\hat{\beta}_2$ to overestimate β_2, making the complete estimate of β worse than would be the case were this covariance zero. Comparison of the 'variances' of multidimensional estimators should therefore somehow account for this covariance phenomenon.
- The 'variance' of a multidimensional estimator is called a variance–covariance matrix. If β^* is an estimator of a k-dimensional β, then the variance–covariance matrix of β^*, denoted by $V(\beta^*)$, is defined as a $k \times k$ matrix (a table with k entries in each direction) containing the variances of the k elements of β^* along the diagonal and the covariances in the off-diagonal positions. Thus

$$V(\beta^*) = \begin{bmatrix} V(\beta_1^*), & C(\beta_1^*, \beta_2^*), & \cdots, & C(\beta_1^*, \beta_k^*) \\ & V(\beta_2^*) & \ddots & \\ & & & V(\beta_k^*) \end{bmatrix}$$

where $V(\beta_k^*)$ is the variance of the kth element of β^* and $C(\beta_1^*, \beta_2^*)$ is the covariance between β_1^* and β_2^*. All this variance–covariance matrix does is array the relevant variances and covariances in a table. Once this is done, the econometrician can draw on mathematicians' knowledge of matrix algebra to suggest ways in which the variance–covariance matrix of one unbiased estimator could be considered 'smaller' than the variance–covariance matrix of another unbiased estimator.

- Consider three alternative ways of measuring smallness among variance–covariance matrices, all accomplished by transforming the matrices into single numbers and then comparing those numbers.

  (1) Choose the unbiased estimator whose variance–covariance matrix has the smallest *trace* (sum of diagonal elements).
  (2) Choose the unbiased estimator whose variance–covariance matrix has the smallest *determinant*.
  (3) Choose the unbiased estimator whose variance–covariance matrix minimizes a loss function consisting of a weighted sum of the individual variances and covariances.

  This last criterion seems sensible: a researcher can weight the variances and covariances according to the importance he or she subjectively feels their minimization should be given in choosing an estimator. It happens that in the context of an unbiased estimator this loss function can be expressed in an alternative form, as the expected value of a quadratic function of the difference between the estimate and the true parameter value, i.e., $E(\hat{\beta} - \beta)(\hat{\beta} - \beta)$. This alternative interpretation also makes good intuitive sense as a choice criterion for use in the estimating context.

  If the weights in the loss function described above are chosen so as to make it

impossible for the loss function to become negative (a reasonable request since it is supposed to be a *loss* function), then a very fortunate thing occurs. Under these circumstances it turns out that the variance–covariance matrix that minimizes the loss function also has the smallest trace and the smallest determinant. What is more, this result does *not* depend on the particular weights used in the loss function.

Although these three ways of defining a smallest matrix are reasonably straightforward, econometricians have chosen, for mathematical reasons, to use as their definition an equivalent but conceptually more difficult idea. This fourth rule says choose the unbiased estimator whose variance–covariance matrix, when subtracted from the variance–covariance matrix of any other unbiased estimator, leaves a non-negative definite matrix. (A matrix A is non-negative definite if the quadratic function formed by using the elements of A as parameters ($x'Ax$) takes on only non-negative values for all non-zero vectors x.)

Proofs of the equivalence of these four selection rules can be constructed by consulting Rothenberg (1973) p. 8, Theil (1971) p. 121, and Goldberger (1964) p. 38.

- A special case of the loss function is revealing. Suppose we choose the weighting such that the variance of any one element of the estimator has a very heavy weight, with all other weights negligible. This implies that each of the elements of the estimator with the 'smallest' variance–covariance matrix has individual minimum variance. (Thus the example given earlier of one estimator with individual variances 3.1 and 7.4 and another with variances 5.6 and 6.3 is unfair; these two estimators could be combined into a new estimator with variances 3.1 and 6.3.) This special case also indicates that in general covariances play no role in determining the best estimator.

2.7

- In the multivariate context the MSE criterion can be interpreted in terms of the 'smallest' (as defined in the technical notes to section 2.6) MSE matrix. This matrix, given by the formula $E(\hat{\beta}-\beta)(\hat{\beta}-\beta)'$, is a natural matrix generalization of the MSE criterion. It can be broken into the sum of the variance–covariance matrix and a matrix with the squares of the individual bias terms on the diagonal and the products of pairs of individual bias terms in the off-diagonal positions, just as the univariate MSE can be broken into the sum of the variance and the square of the bias. Minimizing this matrix is equivalent to minimizing the expectation of a general quadratic loss function written in terms of deviations of $\hat{\beta}$ from β, in a direct analogy to the discussion in the technical notes to section 2.6.

2.8

- As the sample size increases the sampling distribution of an estimator can change its basic mathematical form, as well as its mean and variance. If an estimator's limiting distribution is normal, it is said to be *asymptotically normally distributed.* According to the *Central Limit Theorem,* the sample mean statistic for samples from *any* population is distributed asymptotically normally.
- plim $\hat{\beta}=k$ means that the sampling distribution of $\hat{\beta}$ collapses on the value k as the sample size T approaches infinity (i.e., the sampling distribution of $\hat{\beta}$ approaches a degenerate distribution with all the probability concentrated at the value k). This 'convergence in probability' of $\hat{\beta}$ to k is usually written more formally as

$$\text{plim}\ \hat{\beta}=k \text{ if } \lim_{T\to\infty} \text{prob}\,(|\beta-k|<\delta)=1,$$

where δ is any arbitrarily small positive number. Roughly translated, this means

that by increasing the sample size indefinitely we can be almost certain of making β lie as close as we wish to k.

- A formal definition of consistency would be as follows: an estimator $\hat{\beta}$ of β is consistent if the probability that $\hat{\beta}$ differs in absolute value from β by less than some pre-assigned positive number δ (however small) can be made as close to 1 as desired by choosing a suitably large sample size.
- The probability limit and the asymptotic expectation of an estimator are not necessarily the same. Consider the example in which Prob $(\hat{\beta}=\beta)=1-1/T$ and Prob $(\hat{\beta}=T)=1/T$ where T is the sample size. It is easily seen that plim $\hat{\beta}=\beta$ but that

$$\lim_{T\to\infty} E(\hat{\beta})=\beta+1.$$

(Furthermore,

$$\lim_{T\to\infty} V(\hat{\beta})=\infty.)$$

- It is not unusual for the mean of the limiting distribution of $\hat{\beta}$ to exist without the expected value (and thus the asymptotic expectation) of $\hat{\beta}$ existing. For example, consider the statistic $1/\bar{X}$. Because of the (remote) possibility that $\bar{X}=0$, the expected value and variance of this statistic do not exist; its limiting distribution does exist (with mean $1/\mu$ and variance $\sigma^2/T\mu^4$) however, and its probability limit can be calculated as $1/\mu$. (Here $\bar{X}$ is the mean of a sample drawn randomly from a population with mean μ and variance σ^2.)
- Sometimes an estimator is defined to be consistent if (1) it is asymptotically unbiased, and (2) its asymptotic variance goes to zero as the sample size approaches infinity. (This definition is at times stated in terms of the limit of the mean square error of $\hat{\beta}$ being zero.) Although this is a *sufficient* condition for $\hat{\beta}$ to be consistent, it is *not* a *necessary* condition and thus should not be used as a definition of consistency. This is illustrated by the two preceding examples.
- The asymptotic variance cannot be calculated by taking the limit of the estimator's variance as the sample size goes to infinity, because this limit is often zero. Although strictly speaking this asymptotic variance should be computed by calculating the variance of the asymptotic distribution, operationally what is usually done is to perform the following calculation:

$$\text{Asy. Var}\ \hat{\beta}=1/T \lim_{T\to\infty} T\,V(\hat{\beta}).$$

- There do exist estimators that are asymptotically unbiased but not consistent. Consider the estimator of the population mean μ defined by

$$\hat{\mu}=\tfrac{1}{2}x_1+\frac{1}{2T}\sum_{i=2}^{T} x_i$$

Since

$$\lim_{T\to\infty} E\hat{\mu}=\mu,$$

then $\hat{\mu}$ is asymptotically unbiased. But plim $\hat{\mu}=\frac{1}{2}x_1+\frac{1}{2}\mu$. Note that the asymptotic variance of $\hat{\mu}$ is $\sigma^2/4+\sigma^2/4T$, which does not go to zero as the sample size T goes to infinity. (Here σ^2 is the variance of x.)

- The comments and examples given above note definite differences between probability limits and asymptotic expectations. Because probability limits are much easier to evaluate and work with, econometricians often ignore these differences and equate asymptotic expectations to probability limits. This is legitimate only if both the asymptotic expectation and the probability limit exist, and if

$$\lim_{T\to\infty} V(\hat{\beta}) = 0.$$

- Kmenta (1971) pp. 162–71 has a clear discussion of asymptotic properties. Other good discussions can be found in Maddala (1977) pp. 148–51 and in Johnston (1972) pp. 268–73.

2.9

- The mechanics of finding a maximum likelihood estimator are explained in most econometrics texts. Only an overview of this process is presented here. Consider a typical econometric problem of trying to find the maximum likelihood estimator of the vector

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

in the relationship $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ where T observations on y, x_2 and x_3 are available.

(1) The first step is to specify the nature of the distribution of the disturbance term ε. Suppose the disturbances are identically and independently distributed with probability density function $f(\varepsilon)$. For example, it could be postulated that ε is distributed normally with mean zero and variance σ^2, so that

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

(2) The second step is to rewrite the given relationship as $\varepsilon = y - \beta_1 - \beta_2 x_2 - \beta_3 x_3$ so that for each of the T observations on y, x_2 and x_3 we can calculate a value for ε as a function of β_1, β_2 and β_3.

(3) The third step is to form the likelihood function, the formula for the joint probability distribution of the sample, i.e., a formula proportional to the probability of drawing the particular error terms inherent in this sample. If the error terms are independent of each other this is given by the product of all the $f(\varepsilon)$s, one for each of the T sample observations. Replacing the εs by their expressions from step 2 above creates a complicated function of the sample data and the unknown parameters β_1, β_2 and β_3, plus any unknown parameters inherent in the probability density function f – in this case σ^2.

(4) The fourth step is to use mathematical maximization techniques to find the set of values of the unknown parameters (β_1, β_2, β_3 and σ^2), as functions of the sample data, that maximize this likelihood function.

- If the error terms are not identically distributed, or if they are not distributed independently of one another, the formula for the likelihood function is more complicated, and involves more unknown parameters. These are examples of special estimating problems for which this technique can become computationally difficult.

- The likelihood function is identical to the joint probability density function of the given sample. It is given a different name (i.e., the name 'likelihood') to denote the fact that in this context it is to be *interpreted* as a function of the parameter values (since it is to be maximized with respect to those parameter values) rather than, as is usually the case, being interpreted as a function of the sample data.

- If the disturbances were distributed normally, the MLE estimator of σ^2 is SSE/T. Drawing on similar examples reported in preceding sections, we see that estimation of the variance of a normally distributed population can be computed as $SSE/(T-1)$ (best unbiased), SSE/T (MLE), or $SSE/(T+1)$ (MSE), where in this context SSE is simply $\Sigma(x - \bar{x})^2$.

3. The Classical Linear Regression Model

3.1 Textbooks as Catalogues

In chapter 2 we learned that many of the estimating criteria held in high regard by econometricians (such as best unbiasedness and minimum mean square error) are characteristics of an estimator's sampling distribution. These characteristics cannot be determined unless a set of repeated samples can be taken or hypothesized; to take or hypothesize these repeated samples, knowledge of the way in which the observations are generated is necessary. Unfortunately, an estimator does not have the same characteristics for all ways in which the observations can be generated. This means that in some estimating situations a particular estimator has desirable properties but in other estimating situations it does *not* have desirable properties. Because there is no 'superestimator' having desirable properties in all situations, for each estimating problem (i.e., for each different way in which the observations can be generated) the econometrician must determine anew which estimator is preferred. An econometrics textbook can be characterized as a catalogue of which estimators are most desirable in what estimating situations. Thus a researcher facing a particular estimating problem simply turns to the catalogue to determine which estimator is most appropriate for him to employ in that situation. The purpose of this chapter is to explain how this catalogue is structured.

The cataloguing process described above is centred around a standard estimating situation referred to as the *classical linear regression model* (CLR model). It happens that in this standard situation the OLS estimator is considered the optimal estimator. This model consists of five assumptions concerning the way in which the data are generated. By changing these assumptions in one way or another, different estimating situations are created, in many of which the OLS estimator is no longer considered to be the optimal estimator. Most econometric problems can be characterized as situations in which one (or more) of these five assumptions is violated in a particular way. The catalogue works in a straightforward way: the estimating situation is modelled in the general mould of the CLR model and then the

researcher pinpoints the way in which this situation differs from the standard situation as described by the CLR model (i.e., he finds out which assumption of the CLR model is violated in this problem); he then turns to the textbook (catalogue) which tells him whether the OLS estimator retains its desirable properties, and if not what alternative estimator should be used. Because the econometrician is often not certain of whether or not the estimating situation he faces is one in which an assumption of the CLR model is violated, the catalogue also includes a listing of techniques useful in testing whether or not the CLR model assumptions are violated.

3.2 The Five Assumptions

The CLR model consists of five basic assumptions about the way in which the observations are generated.

(1) The *first assumption* of the CLR model is that the dependent variable can be calculated as a linear function of a specific set of independent variables, plus a disturbance term. The unknown coefficients of this linear function form the vector β and are assumed to be constants. Several violations of this assumption, called specification errors, are discussed in chapter 5.

(a) *wrong regressors* – the omission of relevant independent variables or the inclusion of irrelevant independent variables;
(b) *non-linearity* – the relationship between the dependent and independent variables is not linear;
(c) *changing parameters* – the parameters (β) do not remain constant during the period in which data was collected.

(2) The *second assumption* of the CLR model is that the expected value of the disturbance term is zero, i.e., the mean of the distribution from which the disturbance term is drawn is zero. Violation of this assumption leads to the *biased intercept* problem, discussed in chapter 6.

(3) The *third assumption* of the CLR model is that the disturbance terms all have the same variance and are not correlated with one another. Two major econometric problems, as discussed in chapter 7, are associated with violations of this assumption:

(a) *heteroskedasticity* – the disturbances do not all have the same variance;

(b) *autocorrelated errors* – the disturbances are correlated with one another.

(4) The *fourth assumption* of the CLR model is that the observations on the independent variable can be considered fixed in repeated samples, i.e., it is possible to repeat the sample with the same independent variables. Three important econometric problems, discussed in chapters 8A and 8B correspond to violations of this assumption.

(a) *errors in variables* – errors in measuring the independent variables;
(b) *autoregression* – using a lagged value of the dependent variable as an independent variable;
(c) *simultaneous equation estimation* – situations in which the dependent variables are determined by the simultaneous interaction of several relationships.

(5) The *fifth assumption* of the CLR model is that the number of observations is greater than the number of independent variables and that there are no linear relationships between the independent variables. Although this is viewed as an assumption for the general case, for a specific case it can easily be checked, so that it need not be assumed. The problem of multicollinearity (two or more independent variables being approximately linearly related in the sample data) is associated with this assumption. This is discussed in chapter 9.

All this is summarized in Table 3.1, which presents these five assumptions of the CLR model, shows the appearance they take when dressed in mathematical notation and lists the econometric problems most closely associated with violations of these assumptions. Later chapters in this book comment on the meaning and significance of these assumptions, note implications of their violation for the OLS estimator, discuss ways of determining whether or not they are violated and suggest new estimators appropriate to situations in which one of these assumptions must be replaced by an alternative assumption. Before moving on to this, however, more must be said about the character of the OLS estimator in the context of the CLR model, because of the central role it plays in the econometrician's 'catalogue'.

3.3 The OLS Estimator in the CLR Model

The central role of the OLS estimator in the econometrician's

catalogue is that of a standard against which all other estimators are compared. The reason for this is that the OLS estimator is extraordinarily popular. This popularity stems from the fact that in the context of the CLR model the OLS estimator has a large number of desirable properties, making it the overwhelming choice for the 'optimal' estimator when the estimating problem is accurately characterized by the CLR model. This is best illustrated by looking at the eight criteria listed in chapter 2 and determining how the OLS estimator rates on these criteria in the context of the CLR model.

(1) *Computational cost.* Because of the popularity of the OLS estimator, many packaged computer routines exist, as do standard short-cut means of hand computation. Whenever the functional form being estimated is linear, as it is in the CLR model, the OLS estimator involves very little computational cost.

(2) *Least squares.* Because the OLS estimator is designed to minimize the sum of squared residuals, it is automatically 'optimal' on this criterion.

(3) *Highest R^2.* Because the OLS estimator is optimal on the least squares criterion, it will automatically be optimal on the highest R^2 criterion.

(4) *Unbiasedness.* The assumptions of the CLR model can be used to show that the OLS estimator β^{OLS} is an unbiased estimator of β.

(5) *Best unbiasedness.* In the CLR model β^{OLS} is a linear estimator, i.e., it can be written as a linear function of the observations on the dependent variable. As noted earlier, it is unbiased. Among all linear unbiased estimators of β, it can be shown (in the context of the CLR model) to have the 'smallest' variance–covariance matrix. Thus the OLS estimator is the BLUE in the CLR model. If we add the additional assumption that the disturbances are distributed normally (creating the CNLR model – the *classical normal linear regression model*) it can be shown that the OLS estimator is the best unbiased estimator (i.e., best among *all* unbiased estimators, not just linear unbiased estimators).

(6) *Mean square error.* It is not the case that the OLS estimator is the minimum mean square error estimator in the CLR model. Even among linear estimators it is possible that a substantial reduction in variance can be obtained by adopting a slightly biased estimator. Unfortunately, econometricians have been unable to discover an estimator having minimum mean square error in the CLR model, due mainly to the fact that the MSE estimator depends on the unknown parameter β.

(7) *Asymptotic criteria.* Because the OLS estimator in the CLR

TABLE 3.1

Assumption	*Mathematical expression*		*Violations*	*Chapter in which discussed*
	Univariate	*Multivariate*		
1. Dependent variable a linear function of a specific set of independent variables, plus a disturbance	$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$, $t = 1, \ldots, T$	$Y = X\beta + \varepsilon$	Wrong regressors Non-linearity Changing parameters	5
2. Expected value of disturbance term is zero	$E\varepsilon_t = 0$, for all t	$E\varepsilon = 0$	Biased intercept	6
3. Disturbances have uniform variance and are uncorrelated	$E\varepsilon_t^2 = \sigma^2$, for all t	$E\varepsilon\varepsilon' = \sigma^2 I$	Heteroskedasticity Autocorrelated errors	7
4. Observations on independent variables can be considered fixed in repeated samples	x_t fixed in repeated samples	X fixed in repeated samples	Errors in variables Autoregression Simultaneous equations	8A, 8B
5. No exact linear relationships between independent variables and more observations than independent variables	$\sum_{t=1}^{T} (x_t - \bar{x})^2 = 0$	Rank of $X = K \leqslant T$	Multicollinearity	9

Explanatory Note: The mathematical terminology is explained in the technical notes to this section. The notation is as follows: Y is a vector of observations on the dependent variables; X is a matrix of observations on the independent variables; ε is a vector of disturbances; σ^2 is the variance of the disturbances; I is the identity matrix; K is the number of independent variables; T is the number of observations.

model is unbiased, it is also unbiased in samples of infinite size and thus is asymptotically unbiased. It can also be shown that the variance–covariance matrix of β^{OLS} goes to zero as the sample size goes to infinity, so that β^{OLS} is also a consistent estimator of β. Further, in the CNLR model it is asymptotically efficient.

(8) *Maximum likelihood.* It is impossible to calculate the maximum likelihood estimator given the assumptions of the CLR model, because these assumptions do not specify the functional form of the distribution of the disturbance terms. However, if the disturbances are assumed to be distributed normally (the CNLR model), it turns out that β^{MLE} is identical to β^{OLS}.

Thus whenever the estimating situation can be characterized by the CLR model, the OLS estimator meets practically all of the criteria econometricians consider relevant. It is no wonder, then, that this estimator has become so popular. It is in fact *too* popular; it is often used, without justification, in estimating situations that are not accurately represented by the CLR model. If some of the CLR model assumptions do not hold, many of the desirable properties of the OLS estimator no longer hold. If the OLS estimator does not have the properties that are thought to be of most importance, an alternative estimator must be found. Before moving to this aspect of our examination of econometrics, however, we spend a chapter discussing some concepts of and problems in inference, to provide a foundation for later chapters.

General Notes

3.1

- If more than one of the CLR model assumptions is violated at the same time, econometricians often find themselves in trouble because their catalogues usually tell them what to do if only *one* of the CLR model assumptions is violated. Much recent econometric research examines situations in which two assumptions of the CLR model are violated simultaneously. These situations will be discussed when it is appropriate to do so.

3.2

- Some additional econometric problems could be classified as violations of the first assumption of the CLR model, but are discussed separately as special topics in chapter 10:
  (a) *dummy variables* – independent variables that are qualitative and thus do not possess a numerical measure;

  (b) *qualitative or limited dependent variables* – situations in which the dependent variable is a dummy variable or has a limited range of possible values;
  (c) *extraneous information* – the existence of constraints on the parameter values or knowledge of estimates of parameters from previous studies.

3.3

- The process whereby the OLS estimator is applied to the data at hand is usually referred to by the terminology 'running a regression'. The dependent variable (the 'regressand') is said to be 'regressed' on the independent variables ('the regressors') to produce the OLS estimates. This terminology comes from a pioneering empirical study in which it was found that the mean height of children born of parents of a given height tends to 'regress' or move toward the population average height. See Maddala (1977) pp. 97–101 for further comment on this and for discussion of the meaning and interpretation of regression analysis.

- The result that the OLS estimator in the CLR model is the BLUE is often referred to as the *Gauss–Markov* theorem.

- The formula for the OLS estimator of a specific element of the β vector usually involves observations on *all* the independent variables (as well as observations on the dependent variable), not just observations on the independent variable corresponding to that particular element of β. This is because to obtain an accurate estimate of the influence of one independent variable on the dependent variable, the simultaneous influence of other independent variables on the dependent variable must be taken into account. Doing this ensures that the jth element of β^{OLS} reflects the influence of the jth independent variable on the dependent variable, holding all the other independent variables constant. Similarly, the formula for the variance of an element of β^{OLS} also usually involves observations on all the independent variables. These mechanical aspects of the OLS estimator are discussed in more detail in the technical notes to this section.

Technical Notes

3.2

- Econometricians are often careless in specifying the relationship between the dependent variable y and the independent variables $x_1, x_2, \ldots, x_k$. The regression model $y = g(x_1, \ldots, x_k) + \varepsilon$ is really a specification of how the conditional means $E(y|x_1, \ldots, x_k)$ are related to each other through the xs. The population regression function is written as $E(y|x_1, \ldots, x_k) = g(x)$; it describes how the average or expected value of y varies with the xs. Suppose g is a linear function so that the regression function is $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$. Each element of β^{OLS} (β_4^{OLS}, for example) is an estimate of the effect on the conditional expectation of y of a unit change in x_4, with all other xs held constant.

- In the CLR model the regression model is specified as $y = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ + disturbance, a formula that can be written down T times, once for each set of observations on the dependent and independent variables. This gives a large stack of equations, which can be consolidated via matrix notation as $Y = X\beta + \varepsilon$. Here Y is a vector containing the T observations on the dependent variable y; X is a matrix consisting of K columns, each column being a vector of T observations on one of the independent variables; and ε is a vector containing the T unknown disturbances.

3.3

- The general formula for the OLS estimator β^{OLS} is $(X'X)^{-1}X'Y$. For the simple regression function $y = \beta_2 x_2$ this produces the formula $\Sigma x_2 y/\Sigma x_2^2$ for β_2^{OLS} where the summation is over the T observations. For the simple regression function $y = \beta_1 + \beta_2 x_2$ it produces the formula

$$\Sigma(x_2 - \bar{x}_2)y/\Sigma(x_2 - \bar{x}_2)^2 \qquad \text{or} \qquad \Sigma(x_2 - \bar{x}_2)(y - \bar{y})/\Sigma(x_2 - \bar{x}_2)^2$$

for β_2^{OLS}. As more regressors are added these formulae become more complicated.

- Except for the intercept term, the OLS estimates remain unchanged when calculated with observations expressed as deviations about their respective means. This fact is often used to simplify algebraic manipulations.

- The formula for the variance–covariance matrix β^{OLS} is $\sigma^2(X'X)^{-1}$ where σ^2 is the variance of the disturbance term. For the simple case in which the regression function is $y = \beta_1 + \beta_2 x_2$ this gives the formula $\sigma^2/\Sigma(x_2 - \bar{x}_2)^2$ for the variance of β_2^{OLS}. Note that if the variation in the regressor values is substantial the denominator of this expression will be large, tending to make the variance of β^{OLS} small.

- The variance–covariance matrix of β^{OLS} is usually unknown because σ^2 is usually unknown. It is estimated by $s^2(X'X)^{-1}$ where s^2 is an estimator of σ^2. The estimator s^2 is usually given by the formula $\hat{\varepsilon}'\hat{\varepsilon}/(T-K) = \Sigma\hat{\varepsilon}_i^2/(T-K)$ where $\hat{\varepsilon}$ is the estimate of the disturbance vector, calculated as $(Y - \hat{Y})$ where $\hat{Y}$ is $X\beta^{OLS}$. In the CLR model s^2 is the best quadratic unbiased estimator of σ^2; in the CNLR model it is best unbiased.

- The OLS estimating formula is such that the estimate of the influence of one independent variable (say, x_4) on the dependent variable is calculated while controlling for the simultaneous influence of the other independent variables on the dependent variable. To accomplish this, the OLS estimation procedure uses only variation in x_4 that is 'unique' to x_4 in calculating the OLS estimate of β_4. This can be explained by noting that the OLS estimation formula can be obtained by taking the residuals x_4^* of a regression of x_4 on all other independent variables and then regressing the dependent variable y on x_4^* (using the formula $\Sigma x_4^* y/\Sigma(x_4^*)^2$ to get β_4^{OLS}.

- Similarly the variance of β_4^{OLS} can be calculated by $\sigma^2/\Sigma(x_4^*)^2$. If the variation in x_4 is almost totally explained by variation in the other independent variables (i.e., if x_4 is highly collinear with these variables), the variation in x_4^* will be quite small, making the denominator of this expression small, tending to make the variance of β_4^{OLS} large. This phenomenon is discussed at greater length in chapter 9.

- Although in general the formula for β_4^{OLS} involves observations on all the independent variables, there is an exception. Whenever x_4 is *orthogonal* to all the other xs (i.e., whenever x_4 is uncorrelated in the sample with all the other xs) the formula for β_4^{OLS} depends only on observations on x_4 (as well as, of course, observations on y).