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."

Chapter 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. First, and foremost, they are *economists*, capable of utilizing economic theory to improve their empirical analyses of the problems they address. 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; owing 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 die for curry, ping-pong balls for turtle's eggs, and, for Chalifougnac vintage 1883, a can of turpentine. (Valavanis, 1959, p. 83)

How has this state of affairs come about? One reason is that prestige in the econometrics profession hinges on technical expertise rather than on the hard work required to collect good data:

> It is the preparation skill of the econometric chef that catches the professional eye, not the quality of the raw materials in the meal, or the effort that went into procuring them. (Griliches, 1994, p. 14)

Criticisms of econometrics along these lines are not uncommon. Rebuttals cite improvements in data collection, 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. (Malinvaud, 1966, p. 514)

Modern econometrics texts try to infuse this art into students by providing a large number of detailed examples of empirical application. This important dimension of econometrics texts lies beyond the scope of this book, although Chapter 22 on applied econometrics provides some perspective on this. 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 or 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. Errors in measuring the explaining variable(s) (as opposed to the variable being explained) create a serious econometric problem, discussed in chapter 10. The terminology "errors in variables" is also used to refer to measurement errors.
3. *Human indeterminacy.* Some people believe that human behavior 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 behavior.

Associated with any explanatory relationship are unknown constants, 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 usually employed 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 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., if the dimension of β is 3), 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 estimate by showing, for example, that the estimating formula used to produce that estimate, the 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, in particular the meaning of "usually" and of "quite close," 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. Many textbooks use the letter b to denote the OLS estimator.

1.4 Good and Preferred Estimators

Any fool can produce an estimator of β, since literally an infinite number of them exists; that is, 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 could be used to generate the "preferred" estimate of β. What further distinguishes an econometrician is the 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 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 What is Econometrics?

- 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 in econometrics. Geweke, Horowitz, and Pesaran (2007) is a concise history and overview of recent advances in econometrics. Hendry and Morgan (1995) is a collection of papers of historical importance in the development of econometrics, with excellent commentary. Epstein (1987), Morgan (1990), and Qin (1993) are extended histories; see also Morgan (1990a). Shorter histories, complementing one another, are Farebrother (2006) and Gilbert and Qin (2006). Hendry (1980) notes that the word "econometrics" should not be confused with "eco-nomystics," "economic-tricks," or "icon-ometrics." Econometrics actually comes in several different flavors, reflecting different methodological approaches to research; Hoover (2006) is a good summary.
- Just as the study of economics has split into two halves, microeconomics and macroeconomics, econometrics has divided into two halves, microeconometrics and time-series analysis. Data for microeconometrics tend to be disaggregated, so that heterogeneity of individuals and firms plays a much more prominent role than in time-series data for which data tend to be aggregated. Aggregation averages away heterogeneity, leading to data and relationships that have continuity and smoothness features. Disaggregated data, on the other hand, frequently reflect discrete, nonlinear behavior, presenting special estimating/inference problems. But time-series data have their own special estimating/inference problems, such as unit roots. Panel data, containing observations on microeconomic decision makers over time, blend microeconometric and time-series data, creating yet more special estimating/inference problems. The later chapters of this book address these special problems.
- Before and during the 1960s econometric estimation techniques were based on analytical expressions derived via mathematics. During the 1970s and 1980s the range of econometrics was extended by utilizing numerical optimization algorithms (see chapter 23) to produce estimates for situations in which analytical solutions were not available. More recently, a new generation of econometric techniques has arisen, based on simulation methods (again, see chapter 23) that enable estimation in circumstances in which the criterion functions to be optimized do not have tractable expressions, or in applications of Bayesian methods. The computer has played a prominent role in making progress possible on these technical fronts. One purpose of this book is to make these and other technical dimensions of econometrics more understandable and so alleviate two dangers this progress has produced, articulated below by two of the more respected members of the econometric profession.

  Think first why you are doing what you are doing before attacking the problem with all of the technical arsenal you have and churning out a paper that may be mathematically imposing but of limited practical use. (G. S. Maddala, as quoted by Hsiao, 2003, p. vii)

  The cost of computing has dropped exponentially, but the cost of thinking is what it always was. That is why we see so many articles with so many regressions and so little thought. (Zvi Griliches, as quoted by Mairesse, 2003, p. xiv)

- The discipline of econometrics has grown so rapidly, and in so many different directions, that disagreement regarding the definition of econometrics has grown rather than diminished over the past decade. Reflecting this, at least one prominent econometrician, Goldberger (1989, p. 151), has concluded that "nowadays my definition would be that econometrics is what econometricians do." One thing that econometricians do that is not discussed in this book is serve as expert witnesses in court cases. Fisher (1986) has an interesting account of this dimension of econometric

work; volume 113 of the *Journal of Econometrics* (2003) has several very informative papers on econometrics in the courts. Judge *et al.* (1988, p. 81) remind readers that "econometrics is *fun*!"

- Granger (2001) discusses the differences between econometricians and statisticians. One major distinguishing feature of econometrics is that it focuses on ways of dealing with data that are awkward/dirty because they were not produced by controlled experiments. In recent years, however, controlled experimentation in economics has become more common. Burtless (1995) summarizes the nature of such experimentation and argues for its continued use. Heckman and Smith (1995) is a strong defense of using traditional data sources. Much of this argument is associated with the selection bias phenomenon (discussed in chapter 17) – people in an experimental program inevitably are not a random selection of all people, particularly with respect to their unmeasured attributes, and so results from the experiment are compromised. Friedman and Sunder (1994) is a primer on conducting economic experiments. Meyer (1995) discusses the attributes of "natural" experiments in economics.
- Keynes (1939) described econometrics as "statistical alchemy," an attempt to turn the base metal of imprecise data into the pure gold of a true parameter estimate. He stressed that in economics there is no such thing as a real parameter because all parameters associated with economic behavior are local approximations applying to a specific time and place. Mayer (1993, chapter 10), Summers (1991), Brunner (1973), Rubner (1970), Streissler (1970), and Swann (2006, chapters 5 and 6) are good sources of cynical views of econometrics, summed up dramatically by McCloskey (1994, p. 359): "most allegedly empirical research in economics is unbelievable, uninteresting or both." More critical comments on econometrics appear in this book in section 10.3 on errors in variables and chapter 20 on prediction. Fair (1973) and Fromm and Schink (1973) are examples of studies defending the use of sophisticated econometric techniques. The use of econometrics in the policy context has been hampered by the (inexplicable?) operation of "Goodhart's Law" (1978), namely that all econometric models break down when used for policy. The finding of Dewald, Thursby, and Anderson (1986) that there is a remarkably high incidence of inability to replicate empirical studies in economics, does not promote a favorable view of econometricians.
- In a book provocatively titled *Putting Econometrics in its Place*, Swann (2006) complains that econometrics has come to play a too-dominant role in applied economics; it is viewed as a universal solvent when in fact it is no such thing. He argues that a range of alternative methods, despite their many shortcomings, should be used to supplement econometrics. In this regard he discusses at length the possible contributions of experimental economics, surveys and questionnaires, simulation, engineering economics, economic history and the history of economic thought, case studies, interviews, common sense and intuition, and metaphors. Each of these, including econometrics, has strengths and weaknesses. Because they complement one another, however, a wise strategy would be to seek information from as many of these techniques as is feasible. He summarizes this approach by appealing to a need in economics to respect and assimilate "vernacular knowledge" of the economy, namely information gathered by laypeople from their everyday interaction with markets. In support of this view, Bergmann (2007) complains that empirical work in economics ignores information that could be obtained by interviewing economic decision makers; Bartel, Ichniowski, and Shaw (2004) advocate "insider econometrics," in which information obtained by interviewing/surveying knowledgeable insiders (decision makers) is used to guide traditional econometric analyses. Along these same lines, feminist economists have complained that traditional econometrics contains a male bias. They urge econometricians to broaden their teaching and research methodology to encompass the collection of primary data of different types, such as survey or interview data, and the use of qualitative studies which are not based on the exclusive use of "objective" data. See MacDonald (1995), Nelson (1995), and Bechtold (1999). King, Keohane, and Verba (1994) discuss how research

using qualitative studies can meet traditional scientific standards. See also Helper (2000).

- What has been the contribution of econometrics to the development of economic science? Some would argue that empirical work frequently uncovers empirical regularities that inspire theoretical advances. For example, the difference between time-series and cross-sectional estimates of the MPC prompted development of the relative, permanent, and life-cycle consumption theories. But many others view econometrics with scorn, as evidenced by the following quotes:

> We don't genuinely take empirical work seriously in economics. It's not the source by which economists accumulate their opinions, by and large. (Leamer in Hendry, Leamer, and Poirier, 1990, p. 182)
>
> The history of empirical work that has been persuasive–that has changed people's understanding of the facts in the data and which economic models understand those facts – looks a lot more different than the statistical theory preached in econometrics textbooks. (Cochrane, 2001, p. 302)
>
> Very little of what economists will tell you they know, and almost none of the content of the elementary text, has been discovered by running regressions. Regressions on government-collected data have been used mainly to bolster one theoretical argument over another. But the bolstering they provide is weak, inconclusive, and easily countered by someone else's regressions. (Bergmann, 1987, p. 192)
>
> No economic theory was ever abandoned because it was rejected by some empirical econometric test, nor was a clear cut decision between competing theories made in light of the evidence of such a test. (Spanos, 1986, p. 660)
>
> I invite the reader to try ... to identify a meaningful hypothesis about economic behavior that has fallen into disrepute because of a formal statistical test. (Summers, 1991, p. 130)

This reflects the belief that economic data are not powerful enough to test and choose among theories, and that as a result econometrics has shifted from being a tool for testing theories to being a tool for exhibiting/displaying theories. Because economics is a nonexperimental science, often the data are weak, and, because of this, empirical evidence provided by econometrics is frequently inconclusive; in such cases, it should be qualified as such. Griliches (1986) comments at length on the role of data in econometrics, and notes that they are improving; Aigner (1988) stresses the potential role of improved data. This is summed up nicely by Samuelson (as quoted in Card and Krueger, 1995, p. 355): "In economics it takes a theory to kill a theory, facts can only dent a theorist's hide."

- The criticisms above paint a discouraging view of econometrics, but as cogently expressed by Masten (2002, p. 428), econometricians do have a crucial role to play in economics:

> In the main, empirical research is regarded as subordinate to theory. Theorists perform the difficult and innovative work of conceiving new and sometimes ingenious explanations for the world around us, leaving empiricists the relatively mundane task of gathering data and applying tools (supplied by theoretical econometricians) to support or reject hypotheses that emanate from the theory.
>
> To be sure, facts by themselves are worthless, "a mass of descriptive material waiting for a theory, or a fire," as Coase (1984, p. 230), in characteristic form, dismissed the contribution of the old-school institutionalists. But without diminishing in any way the creativity inherent in good theoretical work, it is worth remembering that theory without evidence is, in the end, just speculation. Two questions that theory alone can never answer are (1) which of the logically possible explanations for observed phenomena is the most probable?; and (2) are the phenomena that constitute the object of our speculations important?

- 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!

1.2 The Disturbance Term

- The error term associated with a relationship need not necessarily be additive, as it is in the example cited. For some nonlinear functions it is often convenient to specify the error term in a multiplicative form. 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.)
- Some econometricians prefer to define the relationship between C and Y discussed earlier as "the mean of C conditional on Y is $f(Y)$," written as $E(C \mid Y) = f(Y)$. This spells out more explicitly what econometricians have in mind when using this specification. The conditional expectation interpretation can cause some confusion. Suppose wages are viewed as a function of education, gender, and marriage status. Consider an unmarried male with 12 years of education. The conditional expectation of such a person's income is the value of y averaged over all unmarried males with 12 years of education. This says nothing about what would happen to a particular individual's income if he were to get married. The coefficient on marriage status tells us what the average difference is between married and unmarried people, much of which may be due to unmeasured characteristics that differ between married and unmarried people. A positive coefficient on marriage status tells us that married people have different unmeasured characteristics that tend to cause higher earnings; it does not mean that getting married will increase one's income. On the other hand, it could be argued that getting married creates economies in organizing one's nonwork life, which enhances earning capacity. This would suggest that getting married would lead to some increase in earnings, but in light of earlier comments, the coefficient on marriage status would surely be an overestimate of this effect.
- 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 Estimates and Estimators

- 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 with alternative estimators. For example, an estimator may be described as "unbiased" or "efficient." Frequently, estimates are described using the same terminology, so that reference might be made to an "unbiased" estimate. Technically this is incorrect because estimates are single numbers – it is the estimating formula, the estimator, that is unbiased, not the estimate. This technical error has become so commonplace that it is now generally understood that when one refers to an "unbiased" estimate one merely means that it has been produced by an estimator that is unbiased.

1.4 Good and Preferred Estimators

- 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.
- 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.

Technical Notes

1.1 What is Econometrics?

- In the macroeconomic context, in particular in research on real business cycles, a computational simulation procedure called *calibration* is often employed as an alternative to traditional econometric analysis. In this procedure, economic theory plays a much more prominent role than usual. Indeed, Pagan (1998, p. 611) claims that "it is this belief in the pre-eminence of theory that distinguishes a calibrator from a non-calibrator." This theory supplies ingredients to a general equilibrium model designed to address a specific economic question. This model is then "calibrated" by setting parameter values equal to average values of economic ratios known not to have changed much over time or equal to empirical estimates from microeconomic studies. A computer simulation produces output from the model, with adjustments to model and parameters made until the output from these simulations has qualitative characteristics (such as correlations between variables of interest) matching those of the real world. Once this qualitative matching is achieved, the model is simulated to address the primary question of interest. Kydland and Prescott (1996) is a good exposition of this approach. Note that in contrast to traditional econometrics, no real estimation is involved, and no measures of uncertainty, such as confidence intervals, are produced.

  Econometricians have not viewed this technique with favor, primarily because there is so little emphasis on evaluating the quality of the output using traditional testing/assessment procedures. Hansen and Heckman (1996), a cogent critique, note (p. 90) that "Such models are often elegant, and the discussions produced from using them are frequently stimulating and provocative, but their empirical foundations are not secure. What credibility should we attach to numbers produced from their 'computational experiments,' and why should we use their 'calibrated models' as a basis for serious quantitative policy evaluation?" Pagan (1998, p. 612) is more direct: "The idea that a model should be used just because the 'theory is strong', without a demonstration that it provides a fit to an actual economy, is mind-boggling."

  Dawkins, Srinivasan, and Whalley (2001) is an excellent summary of calibration and the debates that surround it. Despite all this controversy, calibration exercises are useful supplements to traditional econometric analyses because they widen the range of empirical information used to study a problem.

Chapter 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. Efficiency;
6. Mean square error (MSE);
7. Asymptotic properties;
8. Maximum likelihood.

Discussion of one major criterion, robustness (insensitivity to violations of the assumptions under which the estimator has desirable properties as measured by the criteria above), is postponed to chapter 21. Since econometrics can be characterized as a search for estimators satisfying one or more of these criteria, care is taken in the discussion of the 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 with 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; this could occur, for example, when using an iterative routine to find parameter estimates for a problem involving several nonlinearities.

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 Figure 2.1. The line labeled $\hat{y}$ is the estimated 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

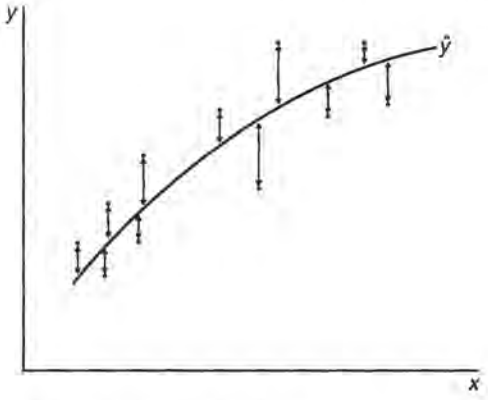

Figure 2.1 Minimizing the sum of squared residuals.

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* (OLS) 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, which 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 nonlinear 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 the 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 5.5 and 6.2).

2.5 Unbiasedness

Suppose we perform the conceptual experiment of taking what is called a *repeated* sample: by 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 Figure 2.2, where the sampling distribution of the estimator is labeled $f(\beta^*)$. 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 Figure 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. Appendix A at the end of this book discusses sampling distributions at greater length. 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.

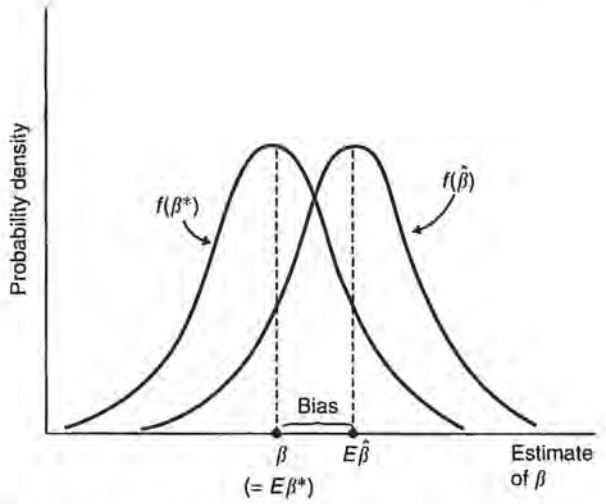

Figure 2.2 Using the sampling distribution to illustrate bias.

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, that is, 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 Figure 2.2, β^* is seen to be unbiased, whereas $\hat{\beta}$ has a bias 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." In one respect this is without import because in reality we only have one sample. A better way to interpret the desirability of the unbiasedness property is to view one sample as producing a single random draw out of an estimator's sampling distribution, and then ask, "If I have one random draw out of a sampling distribution would I prefer to draw out of a sampling distribution centered over the unknown parameter or out of a distribution centered over some other value?"

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 section 2.10.

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 centered 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 Efficiency

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 *efficient* estimator among all unbiased estimators. Why it is considered the most desirable of all unbiased estimators is easy to visualize. In Figure 2.3 the sampling distributions of 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 or her estimate was close to β than would a researcher using β^*. Would you prefer to obtain your estimate by making a single random draw out of an unbiased sampling distribution with a small variance or out of an unbiased sampling distribution with a large variance?

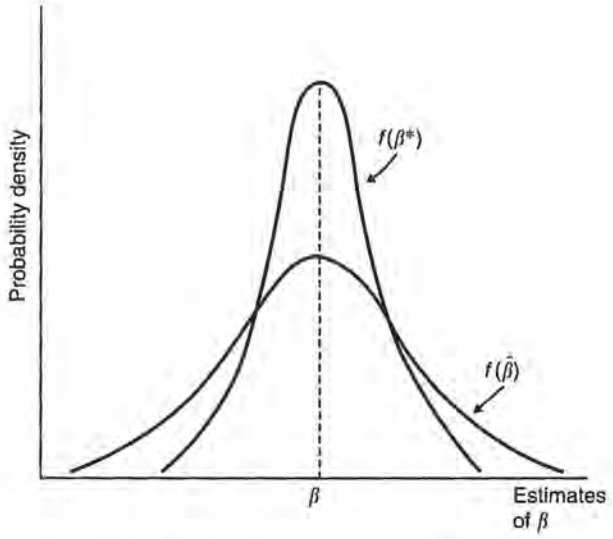

Figure 2.3 Using the sampling distribution to illustrate efficiency.

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, at least implicitly, 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 a further restriction that the estimator 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 and unbiased and that 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, that is, 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

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 Figure 2.4. The sampling distribution of $\hat{\beta}$, the best unbiased estimator, is labeled $f(\hat{\beta})$. β^* is a biased estimator with sampling distribution $f(\beta^*)$. It is apparent from Figure 2.4 that, although $f(\beta^*)$ is not centered 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 $\hat{\beta}$. Would you prefer to obtain your estimate of β by making a single random draw out of $f(\beta^*)$ or out of $f(\hat{\beta})$?

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., choosing 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

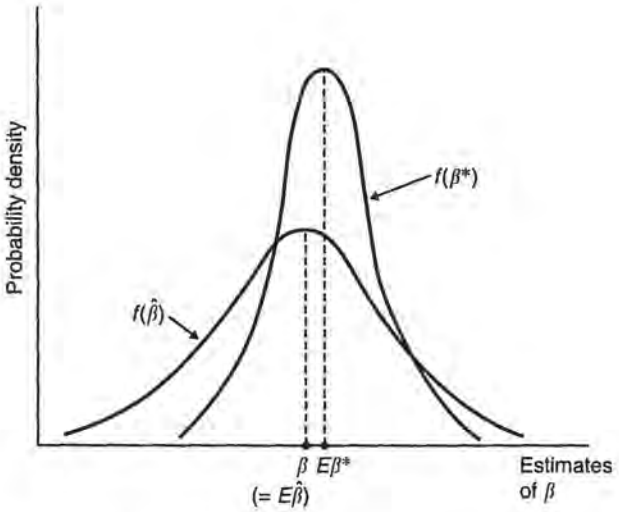

Figure 2.4 MES trades off bias and variance.

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 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 MSE criterion. The popularity of the MSE 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 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 12, 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 centered 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 may 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 centered 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 an *asymptotic distribution* and defining desirable asymptotic or "large-sample properties" of an estimator in terms of the character of its asymptotic distribution. The discussion below of this concept and how it is used is heuristic (and not technically correct); a more formal exposition appears in appendix C at the end of this book.

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 asymptotic distribution of $\hat{\beta}$. Two basic estimator properties are defined in terms of the asymptotic distribution.

1. If the asymptotic distribution of $\hat{\beta}$ becomes 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 $\text{plim}\ \hat{\beta} = k$; if $\text{plim}\ \hat{\beta} = \beta$, then $\hat{\beta}$ is said to be *consistent*.
2. The variance of the asymptotic 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*.

At considerable risk of oversimplification, the plim can be thought of as the large-sample equivalent of the expected value, and so $\text{plim}\ \hat{\beta} = \beta$ is the large-sample equivalent of unbiasedness. Consistency can be crudely conceptualized as the large-sample equivalent of the minimum MSE 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.

Figure 2.5 illustrates the basic appeal of asymptotic properties. For sample size 20, the sampling distribution of β^* is shown as $f(\beta^*)_{20}$. Since this sampling distribution is not centered over β, the estimator β^* is biased. As shown in Figure 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 centered over β (i.e., it becomes less biased), and it becomes less spread out (i.e., its variance becomes smaller). If β^* was 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

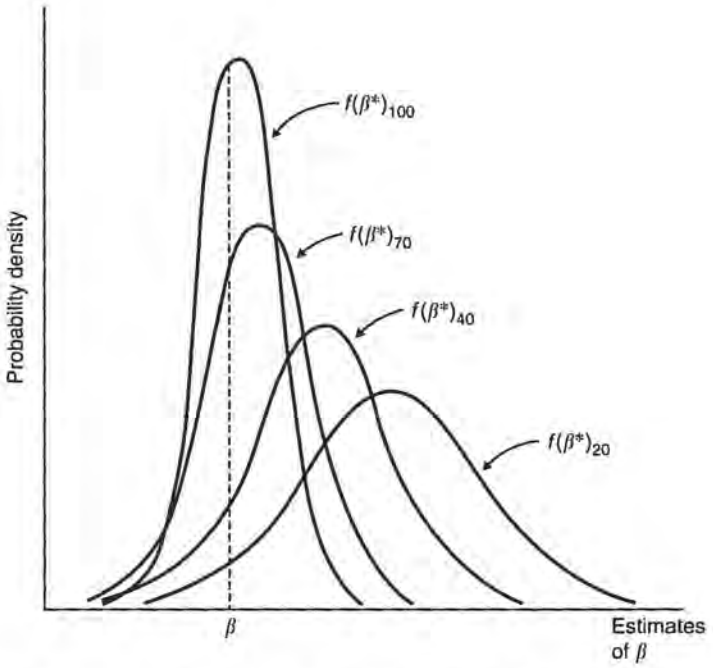

Figure 2.5 How sampling distribution can change as the sample size grows.

unbiasedness, best unbiasedness, and minimum MSE, 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 (see section 2.10) have shown that in general this supposition is warranted.

The message of the discussion above is that when estimators with attractive small-sample properties cannot be found, one may wish to choose an estimator on the basis of its large-sample properties. There is an additional reason for interest in asymptotic properties, however, of equal importance. Often the derivation of small-sample properties of an estimator is algebraically intractable, whereas derivation of large-sample properties is not. This is because, as explained in the technical notes, the expected value of a nonlinear function of a statistic is not the nonlinear function of the expected value of that statistic, whereas the plim of a nonlinear function of a statistic is equal to the nonlinear function of the plim of that statistic.

These two features of asymptotics give rise to the following four reasons for why asymptotic theory has come to play such a prominent role in econometrics.

1. When no estimator with desirable small-sample properties can be found, as is often the case, econometricians are forced to choose estimators on the basis of their asymptotic properties. An example is the choice of the OLS estimator when a lagged value of the dependent variable serves as a regressor. See chapter 10.

2. Small-sample properties of some estimators are extraordinarily difficult to calculate, in which case using asymptotic algebra can provide an indication of what the small-sample properties of this estimator are likely to be. An example is the plim of the OLS estimator in the simultaneous equations context. See chapter 11.
3. Formulas based on asymptotic derivations are useful approximations to formulas that otherwise would be very difficult to derive and estimate. An example is the formula in the technical notes used to estimate the variance of a nonlinear function of an estimator.
4. Many useful estimators and test statistics might never have been found had it not been for algebraic simplifications made possible by asymptotic algebra. An example is the development of LR, W, and LM test statistics for testing nonlinear restrictions. See chapter 4.

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} that gives the greatest probability of obtaining the observed data.

This idea is illustrated in Figure 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 Figure 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 with 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

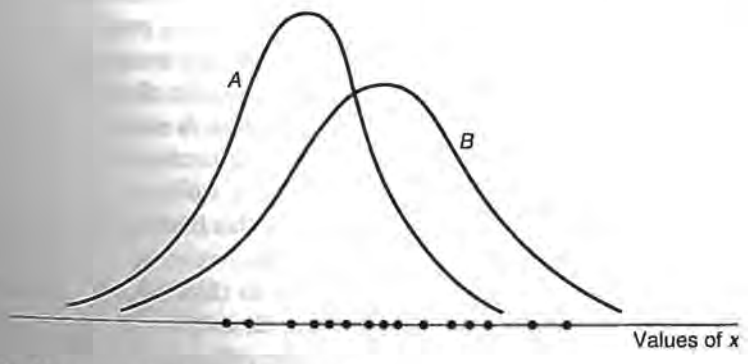

Figure 2.6 Maximum likelihood estimation.

μ and σ^2. The MLE is the particular pair of values μ^{MLE} and $(\sigma^2)^{MLE}$ that creates the greatest probability of having obtained the sample in question; that is, 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 MLE is discussed 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, it is distributed asymptotically normally, and its asymptotic variance can be found via a standard formula (the Cramer–Rao lower bound – see the technical notes to this section). Its only major theoretical drawback is that in order to calculate the MLE, the econometrician must assume a *specific* (e.g., normal) distribution for the error term. Most econometricians seem willing to do this.

These properties make maximum likelihood estimation very appealing for situations in which it is impossible to find estimators with desirable small-sample properties, a situation that arises all too often in practice. In spite of this, however, until recently maximum likelihood estimation has not been popular, mainly because of high computational cost. Considerable algebraic manipulation is required before estimation, and most types of MLE problems require substantial input preparation for available computer packages. But econometricians' attitudes to MLEs have changed recently, for several reasons. Advances in computers and related software have dramatically reduced the computational burden. Many interesting estimation problems have been solved through the use of MLE techniques, rendering this approach more useful (and in the process advertising its properties more widely). And instructors have been teaching students the theoretical aspects of MLE techniques, enabling them to be more comfortable with the algebraic manipulations they require.

2.10 Monte Carlo Studies

A Monte Carlo study is a computer simulation exercise designed to shed light on the small-sample properties of competing estimators for a given estimating problem. They are called upon whenever, for that particular problem, there exist potentially attractive estimators whose small-sample properties cannot be derived theoretically. Estimators with unknown small-sample properties are continually being proposed in the econometric literature, so Monte Carlo studies have become quite common, especially now that computer technology has made their undertaking quite cheap. This is one good reason for having a good understanding of this technique. A more important reason is that a thorough understanding of Monte Carlo studies guarantees an understanding of the repeated sample and sampling distribution concepts, which are crucial to an understanding of econometrics. Appendix A at the end of this book has more on sampling distributions and their relation to Monte Carlo studies.

The general idea behind a Monte Carlo study is to (1) model the data-generating process, (2) generate several sets of artificial data, (3) employ these data and an estimator to create several estimates, and (4) use these estimates to gauge the sampling distribution properties of that estimator for the particular data-generating process

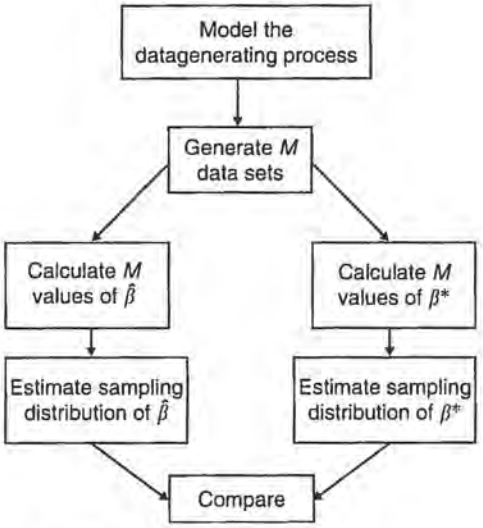

Figure 2.7 Structure of a Monte Carlo study.

under study. This is illustrated in Figure 2.7 for a context in which we wish to compare the properties of two competing estimators. These four steps are described below:

1. *Model the data-generating process* Simulation of the process thought to be generating the real-world data for the problem at hand requires building a model for the computer to mimic the data-generating process, including its stochastic component(s). For example, it could be specified that N (the sample size) values of X, Z, and an error term generate N values of Y according to $Y = \beta_1 + \beta_2 X + \beta_3 Z + \varepsilon$, where the β_i are specific, known numbers, the N values of X and Z are given, exogenous, observations on explanatory variables, and the N values of ε are drawn randomly from a normal distribution with mean zero and known variance σ^2. (Computers are capable of generating such random error terms.) Any special features thought to characterize the problem at hand must be built into this model. For example, if $\beta_2 = \beta_3^{-1}$ then the values of β_2 and β_3 must be chosen such that this is the case. Or if the variance σ^2 varies from observation to observation, depending on the value of Z, then the error terms must be adjusted accordingly. An important feature of the study is that all of the (usually unknown) parameter values are *known* to the person conducting the study (because this person chooses these values).
2. *Create sets of data* With a model of the data-generating process built into the computer, artificial data can be created. The key to doing this is the stochastic element of the data-generating process. A sample of size N is created by obtaining N values of the stochastic variable ε and then using these values, in conjunction with the rest of the model, to generate N values of Y. This yields one complete sample of size N, namely N observations on each of Y, X, and Z, corresponding to the particular set of N error terms drawn. Note that this artificially generated set

of sample data could be viewed as an *example* of real-world data that a researcher would be faced with when dealing with the kind of estimation problem this model represents. Note especially that the set of data obtained depends crucially on the particular set of error terms drawn. A different set of error terms would create a different data set (because the Y values are different) *for the same problem*. Several of these examples of data sets could be created by drawing different sets of N error terms. Suppose this is done, say, 2000 times, generating 2000 sets of sample data, each of sample size N. These are called repeated samples.

3. *Calculate estimates* Each of the 2000 repeated samples can be used as data for an estimator $\hat{\beta}$, say, creating 2000 estimated $\hat{\beta}_{3i}$ ($i = 1, 2, \ldots, 2000$) of the parameter β_3. These 2000 estimates can be viewed as random "drawings" from the sampling distribution of $\hat{\beta}_3$.
4. *Estimate sampling distribution properties* These 2000 drawings from the sampling distribution of $\hat{\beta}_3$ can be used as data to estimate the properties of this sampling distribution. The properties of most interest are its expected value and variance, estimates of which can be used to estimate bias and MSE.
   (a) The *expected value* of the sampling distribution of $\hat{\beta}_3$ is estimated by the average of the 2000 estimates:

$$\text{Estimated expected value} = \bar{\hat{\beta}}_3 = \sum_{i=1}^{2000} \hat{\beta}_{3i} \Big/ 2000$$

   (b) The *bias* of $\hat{\beta}_3$ is estimated by subtracting the known true value of β_3 from the average:

$$\text{Estimated bias} = \bar{\hat{\beta}}_3 - \beta_3$$

   (c) The *variance* of the sampling distribution of $\hat{\beta}_3$ is estimated by using the traditional formula for estimating variance:

$$\text{Estimated variance} = \sum_{i=1}^{2000} (\hat{\beta}_{3i} - \bar{\hat{\beta}}_3)^2 \Big/ 1999$$

   (d) The MSE of $\hat{\beta}_3$ is estimated by the average of the squared differences between $\hat{\beta}$ and the true value of β_3:

$$\text{Estimated MSE} \sum_{i=1}^{2000} (\hat{\beta}_{3i} - \beta_3)^2 \Big/ 2000$$

At stage 3 above an alternative estimator $\hat{\beta}_3^*$ could also have been used to calculate 2000 estimates, as suggested in Figure 2.7. If so, the properties of the samplin

distribution of $\hat{\beta}_3^*$ could also be estimated and then compared with those of the sampling distribution of $\hat{\beta}_3$. (Here $\hat{\beta}_3$ could be, for example, the OLS estimator and $\hat{\beta}_3^*$ any competing estimator such as an instrumental variable estimator, the least absolute error estimator or a generalized least squares estimator. These estimators are discussed in later chapters.) On the basis of this comparison, the person conducting the Monte Carlo study may be in a position to recommend one estimator in preference to another for the sample size N. By repeating such a study for progressively greater values of N, it is possible to investigate how quickly an estimator attains its asymptotic properties.

2.11 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 MSE 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 checks 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 whether the circumstances under which the OLS estimator is desirable are met, and, if not, suggest appropriate alternative estimators. (Unfortunately, in practice this is too often not done, with the OLS estimator being adopted without justification.) The next chapter explains how the econometrician orders this investigation.

General Notes

2.2 Computational Cost

- Computational cost has been reduced significantly by the development of extensive computer software for econometricians. The more prominent of these are EVIEWS, GAUSS, LIMDEP, PC-GIVE, RATS, SAS, SHAZAM, SPSS, STATA, and TSP. For those wanting to code special estimation procedures themselves, this can be done using features of these software packages, or specialized software such as GAUSS, MATLAB, and OX. The *Journal of Applied Econometrics* and the *Journal of Economic Surveys* both publish software reviews regularly. All these packages are very comprehensive, encompassing most of the econometric techniques discussed in textbooks. For applications that they do not cover, in most cases, specialized programs exist. 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 by feeding data to a computer and saying SHAZAM.
- The rapid drop in the cost of computer-intensive analysis has markedly changed econometrics. Now there is much more analysis using graphics, nonparametrics, simulation, bootstrapping, Monte Carlo, Bayesian statistics, and data exploration/mining, all discussed in later chapters.
- Problems with the accuracy of computer calculations are ignored in practice, but can be considerable, as discussed at length by McCullough and Vinod (1999). See also Aigner (1971, pp. 99–101) and Rhodes (1975).

2.3 Least Squares

- Experiments have shown that OLS estimates tend to correspond to the average of laymen's "freehand" attempts to fit a line to a scatter of data. See Mosteller *et al.* (1981).
- In Figure 2.1 the residuals were measured as the vertical distances from the observations to the estimated line. A natural alternative to this vertical measure is the orthogonal measure – the distance from the observation to the estimating line along a line perpendicular to the estimating line. This infrequently seen alternative is discussed in Malinvaud (1966, pp. 7–11); it is sometimes used when measurement errors plague the data, as discussed in section 10.2.

2.4 Highest R^2

- R^2 is called the coefficient of determination. It is the square of the correlation coefficient between y and its OLS estimate $\hat{y}$.
- The total variation of the dependent variable y about its mean, $\sum (y-\bar{y})^2$, is called SST (the total sum of squares); the "explained" variation, the sum of squared deviations of the estimated values of the dependent variable about their mean, $\sum (\hat{y}-\bar{y})^2$ is called SSR (the regression sum of squares); and the "unexplained" variation, the sum of squared residuals, is called SSE (the error sum of squares). R^2 is then given by SSR/SST or by 1 – (SSE/SST).
- What is a high R^2? There is no generally accepted answer to this question. In dealing with time series data, very high R^2s are not unusual, because of common trends. Ames and Reiter (1961) found, for example, that on average the R^2 of a relationship between a randomly chosen variable and its own value lagged one period is about 0.7, and that an R^2 in excess of 0.5 could be obtained by selecting an economic time series and regressing it against two to six other randomly selected economic time series. For cross-sectional data, typical R^2s are not nearly so high. A more meaningful R^2 for time series data can be calculated by first removing the time trend by getting the residuals from regressing y on a time trend, and then regressing these residuals on the explanatory variables and a time trend. See Wooldridge (1991).
- The OLS estimator maximizes R^2. Since the R^2 measure is used as an index of how well an

estimator "fits" the sample data, the OLS estimator is often called the "best-fitting" estimator. A high R^2 is often called a "good fit."

- Because the R^2 and OLS criteria are formally identical, objections to the latter apply to the former. The most frequently voiced of these is that searching for a good fit is likely to generate parameter estimates tailored to the particular sample at hand rather than to the underlying "real world." Further, a high R^2 is not necessary for "good" estimates; R^2 could be low because of a high variance of the disturbance terms, and our estimate of β could be "good" on other criteria, such as those discussed in later sections of this chapter.
- The neat breakdown of the total variation into the "explained" and "unexplained" variations that allows meaningful interpretation of the R^2 statistic is valid only under three conditions. First, the estimator in question must be the OLS estimator. Second, the relationship being estimated must be linear. Thus the R^2 statistic only gives the percentage of the variation in the dependent variable explained *linearly* by variation in the independent variables. And third, the linear relationship being estimated must include a constant, or intercept, term. The formulas for R^2 can still be used to calculate an R^2 for estimators other than the OLS estimator, for nonlinear cases, and for cases in which the intercept term is omitted; it 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 (1971, 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.
- Running a regression without an intercept is the most common way of obtaining an R^2 outside the 0–1 range. To see how this could happen, draw a scatter of points in (x, y) space with an estimated OLS line such that there is a substantial intercept. Now draw in the OLS line that would be estimated if it were forced to go through the origin. In both cases SST is identical (because the same y observations are used). But in the second case the SSE and the SSR could be gigantic, because the $\hat{\varepsilon}$s and the $(\hat{y} - \bar{\hat{y}})$s could be huge. Thus if R^2 is calculated as 1 − SSE/SST, a negative number could result; if it is calculated as SSR/SST, a number greater than one could result.
- 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 favorite 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 savings, 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 savings 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.
- R^2 is also sensitive to the range of variation of the independent variable, basically because a wider range of the independent variables will cause a wider range of the dependent variable and so affect R^2 as described above. A consequence of this is that it makes no sense to compare R^2 across different samples – do not compare the R^2 for data from one country with the R^2 for data from another country, for example. Comparing estimates of the variance of the error term would make more sense.
- In general, econometricians are interested in obtaining "good" parameter estimates where "good" is not defined in terms of R^2. Consequently the measure R^2 is not of much importance in econometrics. Unfortunately, however, many practitioners act as though it is important, for reasons that are not entirely clear, as noted by Cramer (1987, p. 253):

> These measures of goodness of fit have a fatal attraction. Although it is generally conceded among insiders that they do not mean a thing, high values are still a source of pride and satisfaction to their authors, however hard they may try to conceal these feelings.

- Because of this, the meaning and role of R^2 are discussed at some length throughout this book. Section 5.5 and its general notes extend the discussion of this section. Comments are offered in the general notes of other sections when appropriate. For example, one should be aware that R^2s from two equations with different dependent variables should not be compared, and that adding dummy variables (to capture seasonal influences, for example) can inflate R^2, and that regressing on group means overstates R^2 because the error terms have been averaged.

2.5 Unbiasedness

- In contrast to the OLS and R^2 criteria, the unbiasedness criterion (and the other criteria related to the sampling distribution) says 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 MSE 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. None the less, unbiasedness has enjoyed remarkable popularity among practitioners. Part of the reason for this may be due to the emotive content of the terminology: who can stand up in public and state that they prefer *biased* estimators?
- The main objection to the unbiasedness criterion is summarized nicely by the story of the three econometricians who go duck hunting. The first shoots about a foot in front of the duck, the second about a foot behind; the third yells, "We got him!"

2.6 Efficiency

- Cochrane (2001, p. 303) has a sobering view of efficiency: "I can think of no case in which the application of a clever statistical model to wring the last ounce of efficiency out of a data set, changing t statistics from 1.5 to 2.5, substantially changed the way people think about an issue."
- We have seen that efficiency has a trade-off with unbiasedness. It also has a trade-off with robustness. To produce efficiency, extra information about the data-generating process is incorporated into estimation, causing the estimator to be sensitive to the veracity of this extra information. By definition, robust estimators, discussed in chapter 21, are not affected much by violation of the assumptions under which they have been derived.
- Often econometricians forget that although the BLUE property is attractive, its requirement that the estimator be linear can sometimes be restrictive. If the errors have been generated from a "fat-tailed" distribution, for example, so that relatively high errors occur frequently, linear unbiased estimators are inferior to several popular nonlinear unbiased estimators, called robust estimators.See chapter 21.
- 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/$(N-K)$, where N is the number of observations and K is the number of explanatory variables (including a constant), is under general conditions the best quadratic unbiased estimator of σ^2. When K does not include the constant (intercept) term, this formula is written as SSE$/(N-K-1)$.
- 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 of the error 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) creating an unbiased estimator. As an example, consider the sample mean statistic $\bar{x}$. Its variance, σ^2/N, is equal to the Cramer–Rao lower bound if the parent population is normal. Thus, $\bar{x}$ is the best unbiased estimator (whether linear or not) of the mean of a normal population.

2.7 Mean Square Error

- Preference for the MSE criterion over the unbiasedness criterion often hinges on the use to which the estimate is 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 $\text{SSE}/(N-K)$ 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 $\text{SSE}/(N-K+2)$.
- 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.

2.8 Asymptotic Properties

- 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 observant reader of the discussion in the body of this chapter might wonder why the large-sample equivalent of the expected value is defined as the plim rather than being called the "asymptotic expectation." In practice most people use the two terms synonymously, as is done in this book, but technically the latter refers to the limit of the expected value, which is usually, but not always, the same as the plim. Consistency, which is the criterion of relevance in the asymptotic context, relates to plim, not asymptotic expectation; asymptotic specialists get upset when reference is made to asymptotic expectation. For discussion see the technical notes to appendix C.

2.9 Maximum Likelihood

- 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.
- The asymptotic variance of the MLE is usually equal to the Cramer–Rao lower bound, the lowest asymptotic variance that a consistent estimator can have. This is why the MLE is asymptotically efficient. Consequently, the variance (not just the asymptotic variance) of the MLE is estimated by an estimate of the Cramer–Rao lower bound.

The formula for the Cramer–Rao lower bound is given in the technical notes to this section.

- 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

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

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

- Maximum likelihood estimators have an invariance property similar to that of consistent estimators. The maximum likelihood estimator of a nonlinear function of a parameter is the nonlinear function of the maximum likelihood estimator of that parameter: $[g(\beta)]^{MLE} = g(\beta^{MLE})$ where g is a nonlinear 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 nonlinearities and unusual estimation problems. Even in 1972 they did not feel that the computational costs of MLE were prohibitive.
- Application of the maximum likelihood estimation technique requires that a specific distribution for the error term be chosen. In the context of regression, 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. See the general notes to chapter 4 for further discussion. In later chapters we encounter situations (such as count data and logit models) in which a distribution other than the normal is employed. It must be noted, though, that maximum likelihood estimation is usually applied in contexts in which estimation is based on the distribution of the dependent variable rather than the distribution of the error term, as evidenced in applications discussed in later chapters. The distribution of an error term is usually involved, however; the *change-of-variable theorem*, discussed in the technical notes to section 2.9, is used to move from the error density to the dependent variable density.
- Kmenta (1986, pp. 175–83) 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 Monte Carlo Studies

- In this author's opinion, understanding Monte Carlo studies is one of the most important elements of studying econometrics, not because a student may need actually to do a Monte Carlo study, but because an understanding of Monte Carlo studies guarantees an understanding of the concept of a sampling distribution and the uses to which it is put. For examples and advice on Monte Carlo methods see Smith (1973) and Kmenta (1986, chapter 2). Hendry (1984) is a more advanced reference. Barreto and Howland (2006) is a text emphasizing Monte Carlo studies. Appendix A at the end of this book provides further discussion of sampling distributions and Monte Carlo studies. Several exercises in appendix D illustrate Monte Carlo studies.
- If a researcher is worried that the specific parameter values used in the Monte Carlo study may

influence the results, it is wise to choose the parameter values equal to the estimated parameter values using the data at hand, so that these parameter values are reasonably close to the true parameter values. Furthermore, the Monte Carlo study should be repeated using nearby parameter values to check for sensitivity of the results. Bootstrapping is a special Monte Carlo method designed to reduce the influence of assumptions made about the parameter values and the error distribution. Section 4.6 of chapter 4 has an extended discussion.

- The Monte Carlo technique can be used to examine test statistics as well as parameter estimators. For example, a test statistic could be examined to see how closely its sampling distribution matches, say, a chi-square. In this context, interest would undoubtedly focus on determining its size (type I error for a given critical value) and power, particularly as compared with alternative test statistics.
- By repeating a Monte Carlo study for several different values of the factors that affect the outcome of the study, such as sample size or nuisance parameters, one obtains several estimates of, say, the critical values of a test statistic. These estimated critical values can be used as observations with which to estimate a functional relationship between the critical values and the factors affecting these critical values. This relationship is called a *response surface*. McDonald (1998) has a good exposition in the context of finding critical values for unit root and cointegration test statistics. See also Davidson and MacKinnon (1993, pp. 755–63). MacKinnon (1991) is a good example. He specifies the response surface for critical values for cointegration tests (see chapter 19) as $\beta_\infty + \beta_1 N^{-1} + \beta_2 N^{-2}$ for sample size N, and provides values for the β_is for different combinations of significance levels, number of variables in the cointegrating relationship, the presence of an intercept, and the presence of a trend. Notice that the subscript on the intercept reminds us that it is the asymptotic critical value.
- It is common to hold the values of the explanatory variables fixed during repeated sampling when conducting a Monte Carlo study. Whenever the values of the explanatory variables are affected by the error term, such as in the cases of simultaneous equations, measurement error, or the lagged value of a dependent variable serving as a regressor, this is illegitimate and must not be done – the process generating the data must be properly mimicked. But in other cases it is not obvious if the explanatory variables should be fixed. If the sample exhausts the population, such as would be the case for observations on all cities in Washington state with population greater than 30,000, it would not make sense to allow the explanatory variable values to change during repeated sampling. On the other hand, if a sample of wage-earners is drawn from a very large potential sample of wage-earners, one could visualize the repeated sample as encompassing the selection of wage-earners as well as the error term, and so one could allow the values of the explanatory variables to vary in some representative way during repeated samples. Doing this allows the Monte Carlo study to produce an estimated sampling distribution which is not sensitive to the characteristics of the particular wage-earners in the sample; fixing the wage-earners in repeated samples produces an estimated sampling distribution conditional on the observed sample of wage-earners, which may be what one wants if decisions are to be based on that sample.

2.11 Adding Up

- Other, less prominent, criteria exist for selecting point estimates, some examples of which follow.
  - (a) *Admissibility* An estimator is said to be admissible (with respect to some criterion) if, for at least one value of the unknown b, it cannot be beaten on that criterion by any other estimator.
  - (b) *Minimax* A minimax estimator is one that minimizes the maximum expected loss, usually measured as MSE, generated by competing estimators as the unknown β varies through its possible values.
  - (c) *Robustness* An estimator is said to be robust if its desirable properties are not sensitive to violations of the conditions under which it is optimal. In general, a robust estimator is

applicable to a wide variety of situations, and is relatively unaffected by a small number of bad data values. See chapter 21.

(d) *MELO* In the Bayesian approach to statistics (see chapter 14), a decision-theoretic approach is taken to estimation; an estimate is chosen such that it minimizes an expected loss function and is called the MELO (minimum expected loss) estimator. Under general conditions, if a quadratic loss function is adopted, the mean of the posterior distribution of β is chosen as the point estimate of β and this has been interpreted in the non-Bayesian approach as corresponding to minimization of average risk. (Risk is the sum of the MSEs of the individual elements of the estimator of the vector β.) See Zellner (1978).

(e) *Analogy principle* Parameters are estimated by sample statistics that have the same property in the sample as the parameters do in the population. See chapter 2 of Goldberger (1968b) for an interpretation of the OLS estimator in these terms. Manski (1988) gives a more complete treatment. This approach is sometimes called the *method of moments* because it implies that a moment of the population distribution should be estimated by the corresponding moment of the sample. See the technical notes.

(f) *Indirect inference* Sometimes model estimation is extremely difficult, but it may be possible easily to simulate from this model (given parameter values β^*), and easily estimate an approximate model with parameter values δ. Find the β^* values that cause the simulated data to produce δ estimates that are closest to the δ estimates obtained using the actual data. A more detailed discussion appears in chapter 23.

(g) *Nearness/concentration* Some estimators have infinite variances and for that reason are often dismissed. With this in mind, Fiebig (1985) suggests using as a criterion the *probability of nearness* (prefer $\hat{\beta}$ to β^* if prob $(|\hat{\beta} - \beta| < |\beta^* - \beta|) \geq 0.5$) or the *probability of concentration* (prefer $\hat{\beta}$ to β^* if prob $(|\hat{\beta} - \beta| < \delta) >$ prob $(|\beta^* - \beta| < \delta)$.

- Two good introductory references for the material of this chapter are Kmenta (1986, pp. 9–16, 97–108, 156–72) and Kane (1968, chapter 8).

Technical Notes

2.5 Unbiasedness

- The expected value of a variable x is defined formally as $Ex = \int x f(x) dx$ where f is the probability density function (sampling distribution) of x. Thus $E(\hat{\beta})$ could be viewed as a weighted average of all possible values of $\hat{\beta}$ where the weights are proportional to the heights of the density function (i.e., the sampling distribution) of $\hat{\beta}$.

2.6 Efficiency

- In this author's experience, student assessment of sampling distributions is hindered, more than anything else, by confusion about how to calculate an estimator's variance. This confusion arises for several reasons.
  1. There is a crucial difference between a variance and an estimate of that variance, something that often is not well understood.
  2. Many instructors assume that some variance formulas are "common knowledge," retained from previous courses.
  3. It is frequently not apparent that the derivations of variance formulas all follow a generic form.
  4. Students are expected to recognize that some formulas are special cases of more general formulas.
  5. Discussions of variance, and appropriate formulas, are seldom gathered together in one place for easy reference.

  Appendix B has been included at the end of this book to alleviate this confusion, supplementing the material in these technical notes.
- 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 nonzero 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 if this covariance was 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 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,

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

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 four 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 for which any given linear combination of its elements has the smallest variance.
  4. Choose the unbiased estimator whose variance–covariance matrix minimizes a *risk* function consisting of a weighted sum of the individual variances and covariances. (A risk function is the expected value of a traditional loss function, such as the square of the difference between an estimate and what it is estimating.)

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 risk 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; that is, $E(\hat{\beta} - \beta)'Q(\hat{\beta} - \beta)$. This alternative interpretation also makes good intuitive sense as a choice criterion for use in the estimating context.

- If the weights in the risk function described above, the elements of Q, are chosen so as to make it impossible for this risk function to be negative (a reasonable request, since if it were negative it would be a gain, not a loss), then a very fortunate thing occurs. Under these circumstances all four of these criteria lead to the same choice of estimator. What is more, this result does *not* depend on the particular weights used in the risk function.
- Although these four 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 fifth 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. Thus to ensure a non-negative risk function as described above, the weighting matrix Q must be non-negative definite.)

  Proofs of the equivalence of these five 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 risk 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 Mean Square Error

- 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. In practice, however, this generalization is shunned in favor of the sum of the MSEs of all the individual components of $\hat{\beta}$, a definition of *risk* that has come to be the usual meaning of the term.

2.8 Asymptotic Properties

- The econometric literature has become full of asymptotics, so much so that at least one prominent econometrician, Leamer (1988), has complained that there is too much of it. Appendix C of this book provides an introduction to the technical dimension of this important area of econometrics, supplementing the items that follow.
- The reason for the important result that $Eg(x) \neq g(Ex)$ for g nonlinear is illustrated in Figure 2.8. On the horizontal axis are measured values of $\hat{\beta}$, the sampling distribution of which is portrayed by pdf($\hat{\beta}$), with values of $g(\hat{\beta})$ measured on the vertical axis. Values A and B of $\hat{\beta}$, equidistant from $E\hat{\beta}$, are traced to give $g(A)$ and $g(B)$. Note that $g(B)$ is much farther from $g(E\hat{\beta})$ than is $g(A)$: high values of $\hat{\beta}$ lead to values of $g(\hat{\beta})$ considerably above $g(E\hat{\beta})$, but low values of $\hat{\beta}$ lead to values of $g(\hat{\beta})$ only slightly below $g(E\hat{\beta})$. Consequently, the sampling distribution of $g(\hat{\beta})$ is asymmetric, as shown by pdf[$g(\hat{\beta})$], and in this example the expected value of $g(\hat{\beta})$ lies above $g(E\hat{\beta})$.

  If g were a linear function, the asymmetry portrayed in Figure 2.8 would not arise and thus we would have $Eg(\hat{\beta}) = g(E\hat{\beta})$. For g nonlinear, however, this result does not hold.

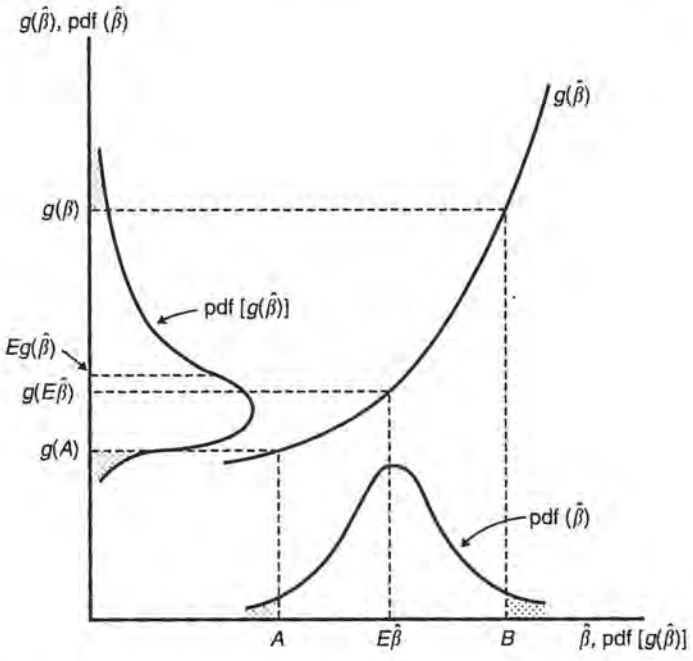

Figure 2.8 Why the expected value of a nonlinear function is not the nonlinear function of the expected value.

Suppose now that we allow the sample size to become very large, and suppose that plim $\hat{\beta}$ exists and is equal to $E\hat{\beta}$ in Figure 2.8. As the sample size becomes very large, the sampling distribution pdf($\hat{\beta}$) begins to collapse on plim $\hat{\beta}$; that is, its variance becomes very, very small. The points A and B are no longer relevant since values near them now occur with negligible probability. Only values of $\hat{\beta}$ very, very close to plim $\hat{\beta}$ are relevant; such values when traced through $g(\hat{\beta})$ are very, very close to g(plim $\hat{\beta}$). Clearly, the distribution of $g(\hat{\beta})$ collapses on g plim $\hat{\beta}$) as the distribution of $\hat{\beta}$ collapses on plim $\hat{\beta}$. Thus plim $g(\hat{\beta}) = g$(plim $\hat{\beta}$), for g a continuous function.

For a simple example of this phenomenon, let g be the square function, so that $g(\hat{\beta}) = \hat{\beta}^2$. From the well-known result that $V(x) = E(x^2) - (Ex)^2$, we can deduce that $E(\hat{\beta}^2)$ $(E\hat{\beta})^2 + V(\hat{\beta})$. Clearly, $E(\hat{\beta}^2) \neq (E\hat{\beta})^2$, but if the variance of $\hat{\beta}$ goes to zero as the sample size goes to infinity, then plim$(\hat{\beta}^2) =$ (plim $\hat{\beta})^2$. The case of $\hat{\beta}$ equal to the sample mean statistic provides an easy example of this.

Note that in Figure 2.8 the modes, as well as the expected values, of the two densities do not correspond. An explanation of this can be constructed with the help of the *change-of-variable theorem* discussed in the technical notes to section 2.9.

- An approximate correction factor can be estimated to reduce the small-sample bias discussed here. For example, suppose an estimate $\hat{\beta}$ of β is distributed normally with mean β and variance $V(\hat{\beta})$. Then $\exp(\hat{\beta})$ is distributed lognormally with mean $\exp[\beta + \frac{1}{2}V(\hat{\beta})]$, suggesting that $\exp(\beta)$ could be estimated by $\exp[\hat{\beta} - \frac{1}{2}\hat{V}(\hat{\beta})]$ which, although biased, should have less bias than $\exp(\hat{\beta})$. If in this same example, the original error was not distributed normally, so that $\hat{\beta}$ was not distributed normally, a Taylor series expansion could be used to deduce an appropriate correction factor. Expand $\exp(\hat{\beta})$ around $E\hat{\beta} = \beta$ to get

$$\exp(\hat{\beta}) = \exp(\beta) + (\hat{\beta} - \beta)\exp(\beta) + \tfrac{1}{2}(\hat{\beta} - \beta)^2 \exp(\beta)$$

plus higher-order terms that are neglected. Taking the expected value of both sides produces

$$E\exp(\hat{\beta}) = \exp\beta[1 + \tfrac{1}{2}V(\hat{\beta})]$$

suggesting that $\exp\beta$ could be estimated by

$$\exp(\hat{\beta})[1 + \tfrac{1}{2}\hat{V}(\hat{\beta})]^{-1}.$$

For discussion and examples of these kinds of adjustments, see Miller (1984), Kennedy (1981a, 1983), and Goldberger (1968a). An alternative way of producing an estimate of a nonlinear function $g(\beta)$ is to calculate many values of $g(\hat{\beta} + \varepsilon)$, where ε is an error with mean zero and variance equal to the estimated variance of $\hat{\beta}$, and average them. For more on this "smearing" estimate see Duan (1983).

- An application of the adjustment discussed above, frequently exposited incorrectly in textbooks, is to cases in which a regression has produced an unbiased estimate $\widehat{\ln y}$ of $\ln y$ and a forecast of y is desired. Think of $\widehat{\ln y}$ as being equal to $\ln y$ plus a forecast error (fe). If the errors in the regression are distributed normally, then $\widehat{\ln y}$ is distributed normally, with mean $\ln y$ and variance $V(\text{fe})$. From above, the expected value of $\exp(\widehat{\ln y})$ is $\exp\{\ln y + \tfrac{1}{2}V(\text{fe})\}$ which is clearly biased as a forecast of y. A reasonable correction is to forecast using $\exp\{\widehat{\ln y} - \tfrac{1}{2}\hat{V}(\text{fe})\}$ where $\hat{V}(\text{fe})$ is an estimate of the variance of the forecast error. See Kennedy (1983). The formula for this variance can be found in example (d) of section 5 of appendix B; its magnitude depends on the regressor values associated with the value to be forecast. Estimation of this variance can most easily be done by using an observation-specific dummy as described in chapter 15.
- When g is a linear function, the variance of $g(\hat{\beta})$ is given by the square of the slope of g times the variance of $\hat{\beta}$; that is, $V(a + bx) = b^2V(x)$. When g is a continuous nonlinear function its variance is difficult to calculate; econometricians deal with this problem by using an estimate of the asymptotic variance of $g(\hat{\beta})$. As noted above in the context of Figure 2.8, when the sample size becomes very large only values of $\hat{\beta}$ very, very close to plim $\hat{\beta}$ are relevant, and in this range a linear approximation to $g(\hat{\beta})$ is adequate. The slope of such a linear approximation is given by the first derivative of g with respect to $\hat{\beta}$. Thus the asymptotic variance of $g(\hat{\beta})$ is calculated as the square of this first derivative times the asymptotic variance of $\hat{\beta}$, with this derivative evaluated at $\hat{\beta} = \text{plim}\,\hat{\beta}$ for the theoretical variance, and evaluated at $\hat{\beta}$ for the estimated variance. See appendix B for what is done when $g(\hat{\beta})$ or $\hat{\beta}$ is a vector.

2.9 Maximum Likelihood

- The likelihood of a sample is often identified with the "probability" of obtaining that sample, something which is, strictly speaking, not correct. The use of this terminology is accepted, however, because of an implicit understanding, articulated by Press *et al.* (1992, p. 652): "If the y_i's take on continuous values, the probability will always be zero unless we add the phrase, 'plus or minus some fixed Δy on each data point.' So let's always take this phrase as understood."
- 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.
- The mechanics of finding a maximum likelihood estimator are explained in most econometrics texts. Because of the importance of maximum likelihood estimation in the econometric literature, an example 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 $y = \beta_1 + \beta_2 x + \beta_3 z + \varepsilon$ where N observations on y, x, and z 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

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

2. The second step is to rewrite the given relationship as $\varepsilon = y - \beta_1 - \beta_2 x - \beta_3 z$ so that for the ith value of ε we have

$$f(\varepsilon_i) = (2\pi\sigma^2)^{-1/2} \times \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_1 - \beta_2 x_i - \beta_3 z_i)^2\right\}.$$

3. The third step is to form the *likelihood function*, the formula for the joint probability distribution of the sample, that is, 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 N sample observations. For the example at hand, this creates the likelihood function

$$L = (2\pi\sigma^2)^{-N/2} \times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i - \beta_3 z_i)^2\right\},$$

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 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. Since the parameter values that maximize L also maximize $\ln L$, and the latter task is easier, attention usually focuses on the log-likelihood function. In this example,

$$\ln L = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i - \beta_3 z_i)^2$$

In some simple cases, such as this one, the maximizing values of this function (i.e., the MLEs) can be found using standard algebraic maximizing techniques. In most cases, however, a numerical search technique (described in chapter 23) must be employed to find the MLE.

- There are two circumstances in which the technique presented above must be modified.

1. *Density of y not equal to density of* ε We have observations on y, not ε. Thus, the likelihood function should be structured from the density of y, not the density of ε. The technique described above implicitly assumes that the density of y, $f(y)$, is identical to $f(\varepsilon)$, the density of ε, so that we can replace ε in this formula by $y - X\beta$. But this is not necessarily the case. The probability of obtaining a value of ε in the small range $d\varepsilon$ is given by $f(\varepsilon)d\varepsilon$; this implies an equivalent probability for y of $f(y)|dy|$ where $f(y)$ is the density function of y and $|dy|$ is the absolute value of the range of y values corresponding to $d\varepsilon$. Thus, because of $f(\varepsilon)d\varepsilon = f(y)|dy|$, we can calculate $f(y)$ as $f(\varepsilon)|d\varepsilon/dy|$.

In the example given above $f(y)$ and $f(\varepsilon)$ are identical since $|d\varepsilon/dy|$ is one. But suppose our example were as above except that we had

$$(y^\lambda - 1)/\lambda = \beta_1 + \beta_2 x + \beta_3 z + \varepsilon$$

where λ is an extra parameter. (This is known as the *Box–Cox transformation*, discussed in chapter 6). In this case, $d\varepsilon/dy = y^{\lambda-1}$ so that

$$f(y_i) = y_i^{\lambda-1} f(\varepsilon_i) = y_i^{\lambda-1} (2\pi\sigma^2)^{-1/2} \times \exp\{-[(y^\lambda - 1)/\lambda - \beta_1 - \beta_2 x - \beta_3 z]^2/2\sigma^2\}$$

This method of finding the density of y when y is a function of another variable ε whose density is known, is referred to as the *change-of-variable theorem*. The multivariate analogue of $|d\varepsilon/dy|$ is the absolute value of the *Jacobian* of the transformation – the determinant of the matrix of first derivatives of the vector ε with respect to the vector y. Judge *et al.* (1988, pp. 30–6) have a good exposition.

2. *Observations not independent* In the examples above, the observations were independent

of one another so that the density values for each observation could simply be multiplied together to obtain the likelihood function. When the observations are not independent, for example, if a lagged value of the regressand appears as a regressor, or if the errors are autocorrelated, an alternative means of finding the likelihood function must be employed. There are two ways of handling this problem.

(a) *Using a multivariate density* A multivariate density function gives the density of an entire vector of ε rather than of just one element of that vector (i.e., it gives the "probability" of obtaining the entire set of ε_i). For example, the multivariate normal density function for the vector ε is given (in matrix terminology) by the formula

$$f(\varepsilon) = (2\pi\sigma^2)^{-N/2}\left|\det\Omega\right|^{-1/2} \times \exp\left\{\frac{1}{-2\sigma^2}\varepsilon'\Omega^{-1}\varepsilon\right\}$$

where $\sigma^2\Omega$ is the variance–covariance matrix of the vector ε. This formula itself can serve as the likelihood function (i.e., there is no need to multiply a set of densities together since this formula has implicitly already done that, as well as taking account of interdependencies among the data). Note that this formula gives the density of the vector ε, not the vector y. Since what is required is the density of y, a multivariate adjustment factor equivalent to the univariate $|d\varepsilon/dy|$ used earlier is necessary. This adjustment factor is $|\det d\varepsilon/dy|$ where $d\varepsilon/dy$ is a matrix containing in its ijth position the derivative of the ith observation of ε with respect to the jth observation of y. It is called the *Jacobian* of the transformation from ε to y. Watts (1973) has a good explanation of the Jacobian.

(b) *Using a transformation* It may be possible to transform the variables of the problem so as to be able to work with errors that are independent. For example, suppose we have

$$y = \beta_1 + \beta_2 x + \beta_3 z + \varepsilon$$

but ε is such that $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$ where u_t is a normally distributed error with mean zero and variance σ^2_u. The εs are not independent of one another, so the density for the vector ε cannot be formed by multiplying together all the individual densities; the multivariate density formula given earlier must be used, where Ω is a function of ρ and σ^2 is a function of ρ and σ^2_u. But the u errors are distributed independently, so the density of the u vector can be formed by multiplying together all the individual u_t densities. Some algebraic manipulation allows u_t to be expressed as

$$u_t = (y_t - \rho y_{t-1}) - \beta_1(1-\rho) - \beta_2(x_t - \rho x_{t-1}) - \beta_3(z_t - \rho z_{t-1}).$$

(There is a special transformation for u_1; see the technical notes to section 8.4 where autocorrelated errors are discussed.) The density of the y vector, and thus the required likelihood function, is then calculated as the density of the u vector times the Jacobian of the transformation from u to y. In the example at hand, this second method turns out to be easier, since the first method (using a multivariate density function) requires that the determinant of Ω be calculated, a difficult task.

- Working through examples in the literature of the application of these techniques is the best way to become comfortable with them and to become aware of the uses to which MLEs can be put. To this end see Beach and MacKinnon (1978a), Savin and White (1978), Lahiri and Egy (1981), Spitzer (1982), Seaks and Layson (1983), and Layson and Seaks (1984).
- The Cramer–Rao lower bound is a matrix given by the formula

$$-\left[E\frac{\partial^2 \ln L}{\partial\theta^2}\right]^{-1}$$

where θ is the vector of unknown parameters (including σ^2) for the MLE estimates of which the Cramer–Rao lower bound is the asymptotic variance–covariance matrix. Its estimation is accomplished by inserting the MLE estimates of the unknown parameters. The inverse of the Cramer–Rao lower bound is called the *information matrix*.

- If a random variable x is distributed normally with variance σ^2, the MLE estimator of σ^2 is $\Sigma(x - \bar{x})^2/N$. From results reported earlier in this chapter, competing estimators are $\Sigma(x - \bar{x})^2/(N - 1)$, the best unbiased estimator, and $\Sigma(x - \bar{x})^2/(N + 1)$, the minimum MSE estimator. They are identical asymptotically, but not in small samples.

2.11 Adding Up

- The analogy principle of estimation is often called the *method of moments* because typically moment conditions (such as that $EX'\varepsilon = 0$, the covariance between the explanatory variables and the error is zero) are utilized to derive estimators using this technique. For example, consider a variable x with unknown mean μ. The mean μ of x is the first moment, so we estimate μ by the first moment (the average) of the data, $\bar{x}$. This procedure is not always so easy. Suppose, for example, that the density of x is given by $f(x) = \lambda x^{\lambda - 1}$ for $0 \leq x \leq 1$ and zero elsewhere. The expected value of x is $\lambda/(\lambda + 1)$ so the method of moments estimator λ^* of λ is found by setting $\bar{x} = \lambda^*/(\lambda^* + 1)$ and solving to obtain $\lambda^* = \bar{x}/(1 - \bar{x})$. In general, we are usually interested in estimating several parameters and so will require as many of these moment conditions as there are parameters to be estimated, in which case finding estimates involves solving these equations simultaneously.
- Consider, for example, estimating α and β in $y = \alpha + \beta x + \varepsilon$. Because ε is specified to be an independent error, the expected value of the product of x and ε is zero, an "orthogonality" or "moment" condition. This suggests that estimation could be based on setting the product of x and the residual $\varepsilon^* = y - \alpha^* - \beta^* x$ equal to zero, where α^* and β^* are the desired estimates of α and β. Similarly, the expected value of ε (its first moment) is specified to be zero, suggesting that estimation could be based on setting the average of the ε^* equal to zero. This gives rise to two equations in two unknowns:

$$\Sigma(y - \alpha^* - \beta^* x)x = 0$$

$$\Sigma(y - \alpha^* - \beta^* x) = 0$$

which a reader might recognize as the normal equations of the OLS estimator. It is not unusual for a method of moments estimator to turn out to be a familiar estimator, a result which gives it some appeal. Greene (2008, pp. 429–36) has a good textbook exposition.

- This approach to estimation is straightforward so long as the number of moment conditions is equal to the number of parameters to be estimated. But what if there are more moment conditions than parameters? In this case there will be more equations than unknowns and it is not obvious how to proceed. The *generalized method of moments* (GMM) procedure, exposited in section 8.5, deals with this case.
- Bera and Bilias (2002) have an advanced but very interesting discussion of relationships among a wide variety of different approaches to estimation.

Chapter 3

The Classical Linear Regression Model

3.1 Textbooks as Catalogs

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 catalog of which estimators are most desirable in what estimating situations. Thus, a researcher facing a particular estimating problem simply turns to the catalog to determine which estimator is most appropriate for him or her to employ in that situation. The purpose of this chapter is to explain how this catalog is structured.

The cataloging process described above is centered around a standard estimating situation referred to as the *classical linear regression model* (CLR model). It happens that in this standard situation the ordinary least squares (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 catalog works in a straightforward way: the estimating

situation is modeled in the general mold of the CLR model and the researcher pinpoints the way in which this situation differs from the standard situation as described by the CLR model (i.e., finds out which assumption of the CLR model is violated in this problem); he or she then turns to the textbook (catalog) to see whether the OLS estimator retains its desirable properties, and if not what alternative estimator should be used. Because econometricians often are not certain of whether the estimating situation they face is one in which an assumption of the CLR model is violated, the catalog 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 6:
   (a) *Wrong regressors* – the omission of relevant independent variables or the inclusion of irrelevant independent variables.
   (b) *Nonlinearity* – when the relationship between the dependent and independent variables is not linear.
   (c) *Changing parameters* – when the parameters (β) do not remain constant during the period in which data were collected.
2. The *second assumption* of the CLR model is that the expected value of the disturbance term is zero; that is, 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 7.
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, discussed in chapter 8, are associated with violations of this assumption:
   (a) *Heteroskedasticity* – when the disturbances do not all have the same variance.
   (b) *Autocorrelated errors* – when 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; that is, it is possible to redraw the sample with the same independent variable values. Three important econometric problems, discussed in chapters 10 and 11, 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 exact 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 12.

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 we move 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 "catalog."

Table 3.1 The assumptions of the CLR model.

Assumption	*Mathematical expression*		*Violations*	*Chapter in which discussed*
	Bivariate	*Multivariate*		
1. Dependent variable a linear function of a specific set of independent variables, plus a disturbance	$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$, $t = 1, \ldots, N$	$Y = X\beta + \varepsilon$	Wrong regressors Nonlinearity Changing parameters	6
2. Expected value of disturbance term is zero	$E\varepsilon_t = 0$, for all t	$E\varepsilon = 0$	Biased intercept	7
3. Disturbances have uniform variance and are uncorrelated	$E\varepsilon_t\varepsilon_r = 0, t \neq r$ $= \sigma^2, t = r$	$E\varepsilon\varepsilon' = \sigma^2 I$	Heteroskedasticity Autocorrelated errors	8
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	10 11
5. No exact linear relationships between independent variables and more observations than independent variables	$\sum_{t=1}^{N}(x_t - \bar{x})^2 \neq 0$	Rank of $X = K \leq N$	Perfect multicollinearity	12

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

3.3 The OLS Estimator in the CLR Model

The central role of the OLS estimator in the econometrician's catalog 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.* All econometric software packages estimate OLS in a flash, and many popular nonstatistical software packages, such as Excel, do so as well.
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; that is, it can be written as a linear function of the errors. 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 best linear unbiased estimator (BLUE) in the CLR model. If we add the additional assumption that the disturbances are distributed normally (creating the *classical normal linear regression model* [CNLR 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. This is the OLS estimator's weakest point; chapters 12 and 13 discuss several estimators whose appeal lies in the possibility that they may beat OLS on the mean square error (MSE) criterion.
7. *Asymptotic criteria.* Because the OLS estimator in the CLR 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), then 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 may 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 will discuss in the next chapter some concepts of and problems in inference, to provide a foundation for later chapters.

General Notes

3.1 Textbooks as Catalogs

- The econometricians' catalog is not viewed favorably by all. Consider the opinion of Worswick (1972, p. 79): "[Econometricians] are not, it seems to me, engaged in forging tools to arrange and measure actual facts so much as making a marvelous array of pretend-tools which would perform wonders if ever a set of facts should turn up in the right form."
- Bibby and Toutenburg (1977, pp. 72–3) note that the CLR model, what they call the general linear model (GLM), can be a trap, a snare, and a delusion. They quote Whitehead as saying: "Seek simplicity ... and distrust it," and go on to explain how use of the linear model can change in undesirable ways the nature of the debate on the phenomenon being examined in the study in question. For example, casting the problem in the mold of the CLR model narrows the question by restricting its terms of reference to a particular model based on a particular set of data; it trivializes the question by focusing attention on apparently meaningful yet potentially trivial questions concerning the values of unknown regression coefficients; and it "technicalizes" the debate, obscuring the real questions at hand, by turning attention to technical statistical matters capable of being understood only by experts.

  They warn users of the GLM by noting that, "it certainly eliminates the complexities of hardheaded thought, especially since so many computer programs exist. For the soft-headed analyst who doesn't want to think too much, an off-the-peg computer package is simplicity itself, especially if it cuts through a mass of complicated data and provides a few easily reportable coefficients. Occam's razor has been used to justify worse barbarities: but razors are dangerous things and should be used carefully."
- If more than one of the CLR model assumptions is violated at the same time, econometricians often find themselves in trouble because their catalogs 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 appropriate.

3.3 The OLS Estimator in the CLR Model

- 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. The 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 tended to "regress" or move towards 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. Regression analysis is the heart and soul of econometrics, as noted by Fiedler (1977, p. 63): "Most economists think of God as working great multiple regressions in the sky." Critics note that the *New Standard Dictionary* defines regression as "The diversion of psychic energy ... into channels of fantasy."

- 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.
- Because the OLS estimator is so popular, and because it so often plays a role in the formulation of alternative estimators, it is important that its mechanical properties be well understood. The most effective way of expositing these characteristics is through the use of a Venn diagram called the Ballentine. Suppose the CLR model applies, with Y determined by X and an error term. In Figure 3.1 the circle Y represents variation in the dependent variable Y and the circle X represents variation in the independent variable X. The overlap of X with Y, the blue area, represents variation that Y and X have in common in the sense that this variation in Y can be explained by X via an OLS regression. The blue area reflects information employed by the estimating procedure in estimating the slope coefficient β_x; the larger this area, the more information is used to form the estimate and thus the smaller is its variance.

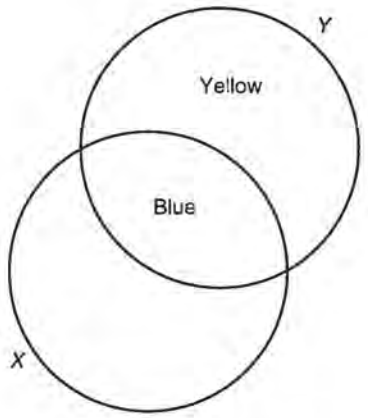

Figure 3.1 Defining the Ballentine Venn diagram.

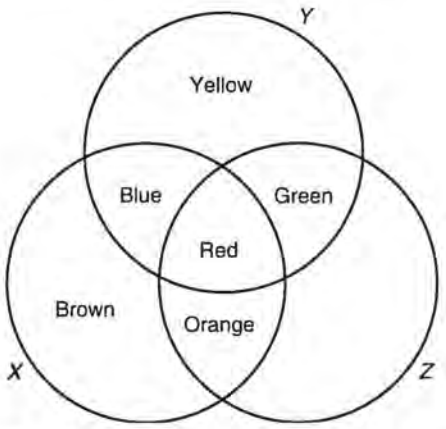

Figure 3.2 Interpreting multiple regression with the Ballentine.

Now consider Figure 3.2, in which a Ballentine for a case of two explanatory variables, X and Z, is portrayed (i.e., now Y is determined by both X and Z). In general, the X and Z circles will overlap, reflecting some collinearity between the two; this is shown in Figure 3.2 by the red-plus-orange area. If Y were regressed on X alone, information in the blue-plus-red area would be used to estimate β_x, and if Y were regressed on Z alone, information in the green-plus-red area would be used to estimate β_z. What happens, though, if Y is regressed on X and Z together?

In the multiple regression of Y on X and Z together, the OLS estimator uses the information in the blue area to estimate β_x and the information in the green area to estimate β_z, *discarding the information in the red area*. The information in the blue area corresponds to variation in Y that matches up uniquely with variation in X; using this information should therefore produce an unbiased estimate of β_x. Similarly, information in the green area corresponds to variation in Y that matches up uniquely with variation in Z; using this information should produce an unbiased estimate of β_z. The information in the red area is not used because it reflects variation in Y that is determined by variation in *both* X and Z, the relative contributions of which are not *a priori* known. In the blue area, for example, variation in Y is all due to variation in X, so matching up this variation in Y with variation in X should allow accurate estimation of β_x. But in the red area, matching up these variations will be misleading because not all variation in Y is due to variation in X.

- Notice that regression Y on X and Z together creates unbiased estimates of β_x and β_z, whereas regressing Y on X and Z separately creates biased estimates of β_x and β_z because this latter method uses the red area. But notice also that, because the former method discards the red area, it uses less information to produce its slope coefficient estimates and thus these estimates will have larger variances. As is invariably the case in econometrics, the price of obtaining unbiased estimates is higher variances.
- Whenever X and Z are orthogonal to one another (have zero collinearity) they do not overlap as in Figure 3.2 and the red area disappears. Because there is no red area in this case, regressing Y on X alone or on Z alone produces the same estimates of β_x and β_z as if Y were regressed on X and Z together. Thus, although in general the OLS estimate of a specific element of the β vector involves observations on *all* the regressors, in the case of orthogonal regressors it involves observations on only one regressor (the one for which it is the slope coefficient estimate).
- Whenever X and Z are highly collinear and therefore overlap a lot, the blue and green areas become very small, implying that when Y is regressed on X and Z together very little information is used to estimate β_x and β_z. This causes the variances of these estimates to be very large. Thus, the impact of multicollinearity is to raise the variances of the OLS estimates. Perfect collinearity causes the X and Z circles to overlap completely; the blue and green areas disappear and estimation is impossible. Multicollinearity is discussed at length in chapter 12.
- In Figure 3.1 the blue area represents the variation in Y explained by X. Thus, R^2 is given as the ratio of the blue area to the entire Y circle. In Figure 3.2 the blue-plus-red-plus-green area represents the variation in Y explained by X and Z together. (Note that the red area is discarded only for the purpose of estimating the coefficients, not for predicting Y; once the coefficients are estimated, all variation in X and Z is used to predict Y.) Thus, the R^2 resulting from the multiple regression is given by the ratio of the blue-plus-red-plus-green area to the entire Y circle. Notice that there is no way of allocating portions of the total R^2 to X and Z because the red area variation is explained by *both*, in a way that cannot be disentangled. Only if X and Z are orthogonal, and the red area disappears, can the total R^2 be allocated unequivocally to X and Z separately.
- The yellow area represents variation in Y attributable to the error term, and thus the magnitude of the yellow area represents the magnitude of σ^2, the variance of the error term. This implies, for example, that if, in the context of Figure 3.2, Y had been regressed on only X, omitting Z, σ^2 would be estimated by the yellow-plus-green area, an overestimate.
- The Ballentine was named by its originators Cohen and Cohen (1975) after a brand of US beer whose logo resembles Figure 3.2. Their use of the Ballentine was confined to the exposition of various concepts related to R^2. Kennedy (1981b) extended its use to the exposition of other aspects of regression. It turns out that the Ballentine can mislead on occasion, particularly when used to exposit R^2 concepts. A limitation of the Ballentine is that it is necessary in certain cases for the red area to represent a negative quantity

(Suppose the two explanatory variables X and Z each have positive coefficients, but in the data X and Z are negatively correlated: X alone could do a poor job of explaining variation in Y because, for example, the impact of a high value of X is offset by a low value of Z. The red area would have to be negative!) This problem notwithstanding, the use of the Ballentine to exposit bias and variance magnitudes for regression is retained in this book, on the grounds that the benefits of its illustrative power outweigh the danger that it will lead to error. The Ballentine is used here as a metaphoric device illustrating some regression results; it should not be given meaning beyond that.

- An alternative geometric analysis of OLS, using vector geometry, is sometimes used. Davidson and MacKinnon (1993, chapter 1) have a good exposition.

Technical Notes

3.2 The Five Assumptions

- The regression model $y = g(x_1, \ldots, x_k) + \varepsilon$ is really a specification of how the conditional means $E(y \mid x_1, \ldots, x_k)$ are related to each other through x. The population regression function is written as $E(y \mid x_1, \ldots, x_k) = g(x)$; it describes how the average or expected value of y varies with x. Suppose g is a linear function so that the regression function is $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon$. 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_i, with all other x held constant.
- The fourth assumption of the CLR model is that the observations on the explanatory variables can be considered fixed in repeated samples; that is, it is possible to redraw the sample with the same explanatory variable values. This is often weakened to read that the explanatory variables are random but independent of the error term. The examples of violations of this assumption given earlier (errors in variables, autoregression, and simultaneous equations) were all instances in which the explanatory variables were random and *not* independent of the error term.

In many instances the explanatory variables are such that they can be considered fixed in repeated samples, for example, when there is one observation on each of the 50 states so that the sample exhausts the population. But in many instances the observations do not exhaust the population. A sample of a thousand individuals from the Current Population Survey (CPS) is an example. In this latter instance we could ask how would the parameter estimates vary when we draw a set of observations on a new set of a thousand individuals along with a new set of error terms: the nature of the conceptual repeated sample is different! There is no reason to believe that a new draw of a thousand observations from the CPS is related to a new draw of error terms, so the weaker version of the fourth assumption is satisfied. Consequently, the OLS estimator continues to be BLUE (although one might complain that in a sense it is no longer linear). It is straightforward to show that it is unbiased, but a difficulty arises when finding the formula for its variance–covariance matrix. The usual formula is $\sigma^2(X'X)^{-1}$ but when X is stochastic rather than fixed this formula becomes $\sigma^2 E[(X'X)^{-1}]$. The difficulty occurs because $E[(X'X)^{-1}]$ is the expected value of a complicated nonlinear function of a stochastic variable. As seen in the technical notes to section 2.8, the expected value of a nonlinear function is not equal to the nonlinear function of the expected value. Because of this $(X'X)^{-1}$ is a biased estimate of $E[(X'X)^{-1}]$. Econometricians wishing to avoid assuming that the explanatory variables are fixed in repeated samples use two means of dealing with this problem, neither of which is fully satisfactory. First, they may talk in terms of $\sigma^2(X'X)^{-1}$ being the variance of OLS *conditional* on X and so use this traditional formula. But this is just another way of saying that we are holding X constant in repeated samples! Second, they may revert to asymptotic criteria so that although biased, $\sigma^2(X'X)^{-1}$ is a consistent estimate of $\sigma^2 E[(X'X)^{-1}]$, and so continue to use this traditional formula. This is a bit questionable

because in small samples it means that estimation of the variance is biased downward because it does not account for variability coming from the change in explanatory variable observations over repeated samples. Stock and Watson (2007) is a textbook adopting the weaker version of assumption 4, employing the asymptotic approach. They argue that the asymptotic approach is necessary in any event because it is unlikely that errors are distributed normally. (In large samples, the OLS estimator is distributed normally, regardless of how the errors are distributed.)

- In the CLR model, the regression model is specified as $y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k +$ disturbance, a formula that can be written down N 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 N observations on the dependent variable y; X is a matrix consisting of K columns, each column being a vector of N observations on one of the independent variables; and ε is a vector containing the N unknown disturbances.

3.3 The OLS Estimator in the CLR Model

- The formula for β^{OLS} is $(X'X)^{-1}X'Y$. A proper derivation of this is accomplished by minimizing the sum of squared errors. An easy way of remembering this formula is to premultiply $Y = X\beta + \varepsilon$ by X' to get $X'Y = X'X\beta + X'\varepsilon$, drop the $X'\varepsilon$, and then solve for β.
- 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 $y = \beta_1 + \beta_2 x$ this gives the formula $\sigma^2/\Sigma(x - \bar{x})^2$ 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 $\hat{\varepsilon}'\hat{\varepsilon}/(N - K) = \Sigma\hat{\varepsilon}_i^2/(N - K)$ 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.
- By discarding the red area in Figure 3.2, the OLS formula ensures that its estimates of the influence of one independent variable are calculated while controlling for the simultaneous influence of the other independent variables, that is, the interpretation of, say, the jth element of β^{OLS} is as an estimate of the influence of the jth explanatory variable, holding all other explanatory variables constant. That the red area is discarded can be emphasized by noting that the OLS estimate of, say, β_x can be calculated from either the regression of Y on X and Z together or the regression of Y on X "residualized" with respect to Z (i.e., with the influence of Z removed). In Figure 3.2, if we were to regress X on Z we would be able to explain the red-plus-orange area; the residuals from this regression, the blue-plus-brown area, are called X residualized for Z. Now suppose that Y is regressed on X residualized for Z. The overlap of the Y circle with the blue-plus-brown area is the blue area, so exactly the same information is used to estimate β_x in this method as is used when Y is regressed on X and Z together, resulting in an identical estimate of β_x.
  Notice further that, if Y were also residualized for Z, producing the yellow-plus-blue area, regressing the residualized Y on the residualized X would also produce the same estimate of β_x since their overlap is the blue area. An important implication of this result is that, for example, running a regression on data from which a linear time trend has been removed will produce exactly the same coefficient estimates as when a linear time trend is included among the regressors in a regression run on raw data. As another example, consider the removal of a linear seasonal influence; running a regression on linearly deseasonalized data will produce exactly the same coefficient estimates as if the linear seasonal influence were included as an extra regressor in a regression run on raw data.

- A variant of OLS called *stepwise regression* is to be avoided. It consists of regressing Y on each explanatory variable separately and keeping the regression with the highest R^2. (A variant looks for the regressor with the highest t statistic.) This determines the estimate of the slope coefficient of that regression's explanatory variable. Then the residuals from this regression are used as the dependent variable in a new search using the remaining explanatory variables and the procedure is repeated. Suppose that, for the example of Figure 3.2, the regression of Y on X produced a higher R^2 than the regression of Y on Z. Then the estimate of β_x would be formed using the information in the blue-plus-red area. Note that this estimate is biased. Econometricians often denigrate statisticians on the grounds that they espouse such algorithmic searches. Leamer (2007, p.101) expresses this cogently:

  > We don't rely on stepwise regression or any other automated statistical pattern recognition to pull understanding from our data sets because there is currently no way of providing the critical contextual inputs into these algorithms and because an understanding of the context is absolutely critical to making sense of our noisy non-experimental data. The last person you want to analyze an economic data set is a statistician, which is what you get when you run stepwise regression.

- The Ballentine can be used to illustrate several variants of R^2. Consider, for example, the simple R^2 between Y and Z in Figure 3.2. If the area of the Y circle is normalized to be unity, this simple R^2, denoted R^2_{yz}, is given by the red-plus-green area. The *partial* R^2 between Y and Z is defined as reflecting the influence of Z on Y *after* accounting for the influence of X. It is measured by obtaining the R^2 from the regression of Y corrected for X on Z corrected for X, and is denoted $R^2_{yz.x}$. Our earlier use of the Ballentine makes it easy to deduce that in Figure 3.2 it is given as the green area divided by the yellow-plus-green area. The reader might like to verify that it is given by the formula

$$R^2_{yz.x} = (R^2 - R^2_{yx})/(1 - R^2_{yx}).$$

- The OLS estimator has several well-known mechanical properties with which students should become intimately familiar – instructors tend to assume this knowledge after the first lecture or two on OLS. Listed below are the more important of these properties; proofs can be found in most textbooks. The context is $y = \alpha + \beta x + \varepsilon$.
  1. If $\beta = 0$ so that the only regressor is the intercept, y is regressed on a column of ones, producing $\alpha^{OLS} = \bar{y}$, the average of the y observations.
  2. If $\alpha = 0$ so there is no intercept and one explanatory variable, y is regressed on a column of x values, producing $\beta^{OLS} = \Sigma xy/\Sigma x^2$.
  3. If there is an intercept and one explanatory variable x

$$\begin{aligned}\beta^{OLS} &= \Sigma(x - \bar{x})(y - \bar{y})/\Sigma(x - \bar{x})^2 \\ &= \Sigma(x - \bar{x})y/\Sigma(x - \bar{x})^2.\end{aligned}$$

  4. If observations are expressed as deviations from their means, $y^* = y - \bar{y}$ and $x^* = x - \bar{x}$, then $\beta^{OLS} = \Sigma x^* y^*/\Sigma x^{*2}$. This follows from (3) above. Lower case letters are sometimes reserved to denote deviations from sample means.
  5. The intercept can be estimated as $\bar{y} - \beta^{OLS}\bar{x}$ or, if there are more explanatory variables, as $\bar{y} - \Sigma\beta_i^{OLS}\bar{x}_i$. This comes from the first normal equation, the equation that results from setting the partial derivative of SSE (the sum of squared errors) with respect to α equal to zero (to minimize the SSE).
  6. An implication of (5) is that the sum of the OLS residuals equals zero; in effect the intercept is estimated by the value that causes the sum of the OLS residuals to equal zero.
  7. The predicted, or estimated, y values are calculated as $\hat{y}_i = \alpha^{OLS} + \beta^{OLS}x_i$. An implication of (6) is that the mean of the $\hat{y}$ values equals the mean of the actual y values: $\bar{\hat{y}} = \bar{y}$.
  8. An implication of (5), (6), and (7) above is that the OLS regression line passes through the overall mean of the data points.

9. Adding a constant to a variable, or scaling a variable, has a predictable impact on the OLS estimates. For example, multiplying the x observations by 10 will multiply β^{OLS} by one-tenth, and adding 6 to the y observations will increase α^{OLS} by 6.
10. A linear restriction on the parameters can be incorporated into a regression by eliminating one coefficient from that equation and running the resulting regression using transformed variables. For an example see the general notes to section 4.3.
11. The "variation" in the dependent variable is the "total sum of squares" $SST = \Sigma(y - \bar{y})^2 = y'y - N\bar{y}^2$ where $y'y$ is matrix notation for Σy^2, and N is the sample size.
12. The "variation" explained linearly by the independent variables is the "regression sum of squares," $SSR = \Sigma(\hat{y} - \bar{\hat{y}})^2 = \hat{y}'\hat{y} - N\bar{y}^2$.
13 The sum of squared errors from a regression is $SSE = (y - \hat{y})'(y - \hat{y}) = y'y - \hat{y}'\hat{y} = SST - SSR$. (Note that textbook notation varies. Some authors use SSE for "explained sum of squares" and *SSR* for "sum of squared residuals," creating results that look to be the opposite of those given here.)
14 SSE is often calculated by $\Sigma y^2 - \alpha^{OLS}\Sigma y - \beta^{OLS}\Sigma xy$, or in the more general matrix notation $y'y - \beta^{OLS\prime}X'y$.
15 The coefficient of determination, $R^2 = SSR/SST = 1 - SSE/SST$ is maximized by OLS because OLS minimizes SSE. R^2 is the squared correlation coefficient between y and $\hat{y}$; it is the fraction of the "variation" in y that is explained linearly by the explanatory variables.
16 When no intercept is included, it is possible for R^2 to lie outside the zero to one range. See the general notes to section 2.4.
17 Minimizing with some extra help cannot make the minimization less successful. Thus SSE decreases (or in unusual cases remains unchanged) when an additional explanatory variable is added; R^2 must therefore rise (or remain unchanged).
18 Because the explanatory variable(s) is (are) given as much credit as possible for explaining changes in y, and the error as little credit as possible, ε^{OLS} is uncorrelated with the explanatory variable(s) and thus with $\hat{y}$ (because $\hat{y}$ is a linear function of the explanatory variable(s)).
19 The estimated coefficient of the ith regressor can be obtained by regressing y on this regressor "residualized" for the other regressors (the residuals from a regression of the ith regressor on all the other regressors). The same result is obtained if the "residualized" y is used as the regressand instead of y. These results were explained earlier in these technical notes with the help of the Ballentine.

Chapter 4

Interval Estimation and Hypothesis Testing

4.1 Introduction

In addition to estimating parameters, econometricians often wish to construct confidence intervals for their estimates and test hypotheses concerning parameters. To strengthen the perspective from which violations of the classical linear regression (CLR) model are viewed in the following chapters, this chapter provides a brief discussion of these principles of inference in the context of traditional applications found in econometrics.

Under the null hypothesis most test statistics have a distribution that is tabulated in appendices at the back of statistics books, the most common of which are the standard normal, the t, the chi-square, and the F distributions. In small samples the applicability of all these distributions depends on the errors in the CLR model being normally distributed, something that is not one of the CLR model assumptions. For situations in which the errors are not distributed normally, it turns out that in most cases a traditional test statistic has an asymptotic distribution equivalent to one of these tabulated distributions; with this as justification, testing/interval estimation proceeds in the usual way, ignoring the small-sample bias. For expository purposes, this chapter's discussion of inference is couched in terms of the classical normal linear regression (CNLR) model, in which the assumptions of the CLR model are augmented by assuming that the errors are distributed normally.

4.2 Testing a Single Hypothesis: The t Test

Hypothesis tests on, and interval estimates of, single parameters are straightforward applications of techniques familiar to all students of elementary statistics. In the CNLR model, the ordinary least squares (OLS) estimator β^{OLS} generates estimates that are

distributed joint-normally in repeated samples. This means that $\beta_1^{OLS}, \beta_2^{OLS}, \ldots, \beta_k^{OLS}$ are all connected to one another (through their covariances). In particular, this means that β_3^{OLS}, say, is distributed normally with mean β_3 (since the OLS estimator is unbiased) and variance $V(\beta_3^{OLS})$ is equal to the third diagonal element of the variance–covariance matrix of β^{OLS}. The square root of $V(\beta_3^{OLS})$ is the standard deviation of β_3^{OLS}. Using the normal table and this standard deviation, interval estimates can be constructed and hypotheses can be tested.

A major drawback to this procedure is that the variance–covariance matrix of β^{OLS} is not usually known (because σ^2, the variance of the disturbances, which appears in the formula for this variance–covariance matrix, is not usually known). Estimating σ^2 by s^2, as discussed in the technical notes to section 3.3, allows an estimate of this matrix to be created. The square root of the third diagonal element of this matrix is the standard error of $V(\beta_3{}^{OLS})$, an estimate of the standard deviation of $V(\beta_3{}^{OLS})$. With this estimate the t table can be used in place of the normal table to test hypotheses or construct interval estimates.

The use of such t tests, as they are called, is so common that econometric software packages have included in their estimation output a number called the t statistic for each parameter estimate. This gives the value of the parameter estimate divided by its estimated standard deviation (the standard error). This value can be compared directly to critical values in the t table to test the hypothesis that that parameter is equal to zero. In some research reports, this t statistic is printed in parentheses underneath the parameter estimates, creating some confusion because sometimes the standard errors appear in this position. (A negative number in parentheses would have to be a t value, so that this would indicate that these numbers were t values rather than standard errors.)

4.3 Testing a Joint Hypothesis: the *F* Test

Suppose that a researcher wants to test the joint hypothesis that, say, the fourth and fifth elements of β are equal to 1.0 and 2.0, respectively. That is, he or she wishes to test the hypothesis that the sub-vector

$$\begin{bmatrix} \beta_4 \\ \beta_5 \end{bmatrix}$$

is equal to the vector

$$\begin{bmatrix} 1.0 \\ 2.0 \end{bmatrix}$$

This is a different question from the two separate questions of whether β_4 is equal to 1.0 and whether β_5 is equal to 2.0. It is possible, for example, to accept the hypothesis

that β_4 is equal to 1.0 and also to accept the hypothesis that β_5 is equal to 2.0, but to *reject* the joint hypothesis that

$$\begin{bmatrix} \beta_4 \\ \beta_5 \end{bmatrix}$$

is equal to

$$\begin{bmatrix} 1.0 \\ 2.0 \end{bmatrix}$$

The purpose of this section is to explain how the F test is used to test such joint hypotheses. The following section explains how a difference between results based on separate tests and joint tests could arise.

The F statistic for testing a set of J linear constraints in a regression with K parameters (including the intercept) and N observations takes the generic form

$$\frac{\left[\text{SSE}(\text{constrained}) - \text{SSE}(\text{unconstrained})\right]/J}{\text{SSE}(\text{unconstrained})/(N-K)}$$

where the degrees of freedom for this F statistic are J and $N-K$. This generic form is worth memorizing – it is extremely useful for structuring F tests for a wide variety of special cases, such as Chow tests (chapter 6) and tests involving dummy variables (chapter 15).

When the constraints are true, because of the error term they will not be satisfied exactly by the data; so the SSE (error sum of squares) will increase when the constraints are imposed – minimization subject to constraints will not be as successful as minimization without constraints. But if the constraints are true, the per-constraint increase in SSE should not be large relative to the influence of the error term. The numerator has the "per-constraint" change in SSE due to imposing the constraints and the denominator has the "per-error" contribution to SSE. (The minus K in this expression corrects for degrees of freedom, explained in the general notes.) If their ratio is "too big" we would be reluctant to believe that it happened by chance, concluding that it must have happened because the constraints are false. High values of this F statistic thus lead us to reject the null hypothesis that the constraints are true.

How does one find the constrained SSE? A constrained regression is run to obtain the constrained SSE. The easiest example is the case of constraining a coefficient to be equal to zero – just run the regression omitting that coefficient's variable. To run a regression constraining β_4^{OLS} to be 1.0 and β_5^{OLS} to be 2.0, subtract 1.0 times the fourth regressor and 2.0 times the fifth regressor from the dependent variable and regress this new, constructed dependent variable on the remaining regressors. In general, to incorporate a linear restriction into a regression, use the restriction to solve out one of the parameters, and rearrange the resulting equation to form a new regression involving constructed variables. An explicit example is given in the general notes.

4.4 Interval Estimation for a Parameter Vector

Interval estimation in the multidimensional case is best illustrated by a two-dimensional example. Suppose that the sub-vector

$$\begin{bmatrix} \beta_4 \\ \beta_5 \end{bmatrix}$$

is of interest. The OLS estimate of this sub-vector is shown as the point in the center of the rectangle in Figure 4.1. Using the t table and the square root of the fourth diagonal term in the estimated variance–covariance matrix of β^{OLS}, a 95% confidence interval can be constructed for β_4. This is shown in Figure 4.1 as the interval from A to B; β_4^{OLS} lies halfway between A and B. Similarly, a 95% confidence interval can be constructed for β_5; it is shown in Figure 4.1 as the interval from C to D and is drawn larger than the interval AB to reflect an assumed larger standard error for β_5^{OLS}.

An interval estimate for the sub-vector

$$\begin{bmatrix} \beta_4 \\ \beta_5 \end{bmatrix}$$

is a *region* or area that, when constructed in repeated samples, covers the true valu (β_4, β_5) in, say, 95% of the samples. Furthermore, this region should for an efficier

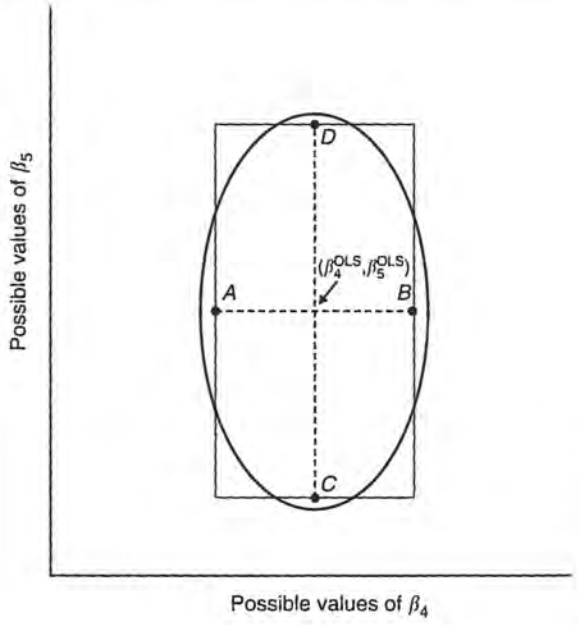

Figure 4.1 A confidence region with zero covariance.

estimate be the smallest such region possible. A natural region to choose for this purpose is the rectangle formed by the individual interval estimates, as shown in Figure 4.1. If β_4^{OLS} and β_5^{OLS} have zero covariance, then in repeated sampling rectangles calculated in this fashion will cover the unknown point (β_4, β_5) in $0.95 \times 0.95 = 90.25\%$ of the samples. (In repeated samples the probability is 0.95 that the β_4 confidence interval covers β_4, as is the probability that the β_5 confidence interval covers β_5; thus the probability for both β_4 and β_5 to be covered simultaneously is 0.95×0.95.)

Evidently, this rectangle is not "big" enough to serve as a 95% joint confidence interval. Where should it be enlarged? Because the region must be kept as small as possible, the enlargement must come in those parts that have the greatest chance of covering (β_4, β_5) in repeated samples. The corners of the rectangle will cover (β_4, β_5) in a repeated sample whenever β_4^{OLS} and β_5^{OLS} are simultaneously a long way from the mean values of β_4 and β_5. The probability in repeated samples of having these two unlikely events occur simultaneously is very small. Thus the areas just outside the rectangle near the points *A*, *B*, *C*, and *D* are more likely to cover (β_4, β_5) in repeated samples than are the areas just outside the corners of the rectangle: the rectangle should be made bigger near the points *A*, *B*, *C*, and *D*. Further thought suggests that the areas just outside the points *A*, *B*, *C*, and *D* are more likely, in repeated samples, to cover (β_4, β_5) than the areas just *inside* the corners of the rectangle. Thus the total region should be adjusted by chopping a lot of area off the corners and extending slightly the areas near the points *A*, *B*, *C*, and *D*. In fact, the *F* statistic described earlier allows the econometrician to derive the confidence region as an ellipse, as shown in Figure 4.1.

The ellipse in Figure 4.1 represents the case of zero covariance between β_4^{OLS} and β_5^{OLS}. If β_4^{OLS} and β_5^{OLS} have a positive covariance (an estimate of this covariance is found in either the fourth column and fifth row or the fifth column and fourth row of the estimate of the variance–covariance matrix of β^{OLS}), whenever β_4^{OLS} is an overestimate of β_4, β_5^{OLS} is likely to be an overestimate of β_5, and whenever β_4^{OLS} is an underestimate of β_4, β_5^{OLS} is likely to be an underestimate of β_5. This means that the area near the top right-hand corner of the rectangle and the area near the bottom left-hand corner are no longer as unlikely to cover (β_4, β_5) in repeated samples; it also means that the areas near the top left-hand corner and bottom right-hand corner are even less likely to cover (β_4, β_5). In this case the ellipse representing the confidence region is tilted to the right, as shown in Figure 4.2. In the case of negative covariance between β_4^{OLS} and β_5^{OLS}, the ellipse is tilted to the left. In all cases, the ellipse remains centered on the point $(\beta_4^{OLS}, \beta_5^{OLS})$.

This two-dimensional example illustrates the possibility, mentioned earlier, of accepting two individual hypotheses but rejecting the corresponding joint hypothesis. Suppose the hypothesis is that $\beta_4 = 0$ and $\beta_5 = 0$, and suppose the point (0, 0) lies inside a corner of the rectangle in Figure 4.1, but outside the ellipse. Testing the hypothesis $\beta_4 = 0$ using a *t* test concludes that β_4 is insignificantly different from zero (because the interval *AB* contains zero), and testing the hypothesis $\beta_5 = 0$ concludes that β_5 is insignificantly different from zero (because the interval *CD* contains zero). But testing the joint hypothesis

$$\begin{bmatrix} \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

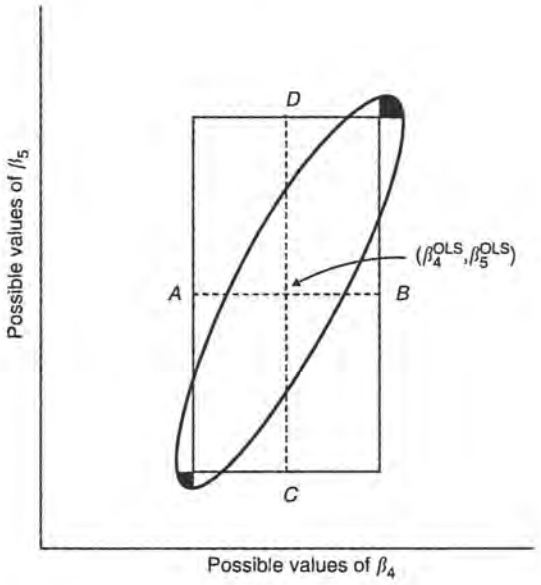

Figure 4.2 A confidence region with positive covariance.

using an F test concludes that

$$\begin{bmatrix} \beta_4 \\ \beta_5 \end{bmatrix}$$

is significantly different from the zero vector because (0, 0) lies outside the ellips
In this example, one can confidently say that *at least one* of the two variables has
significant influence on the dependent variable, but one cannot with confidence assi
that influence to either of the variables individually. The typical circumstance in whi
this comes about is in the case of multicollinearity (see chapter 12), in which indepe
dent variables are related so that it is difficult to tell which of the variables deserv
credit for explaining variation in the dependent variable. Figure 4.2 is representative
the multicollinearity case.

In three dimensions the confidence region becomes a confidence volume and
represented diagrammatically by an ellipsoid. In higher dimensions diagrammatic r
resentation is impossible, but the hypersurface corresponding to a critical value of
F statistic can be called a multidimensional ellipsoid.

4.5 LR, W, and LM Statistics

The F test discussed above is applicable whenever we are testing linear restrict
in the context of the CNLR model. Whenever the problem cannot be cast into

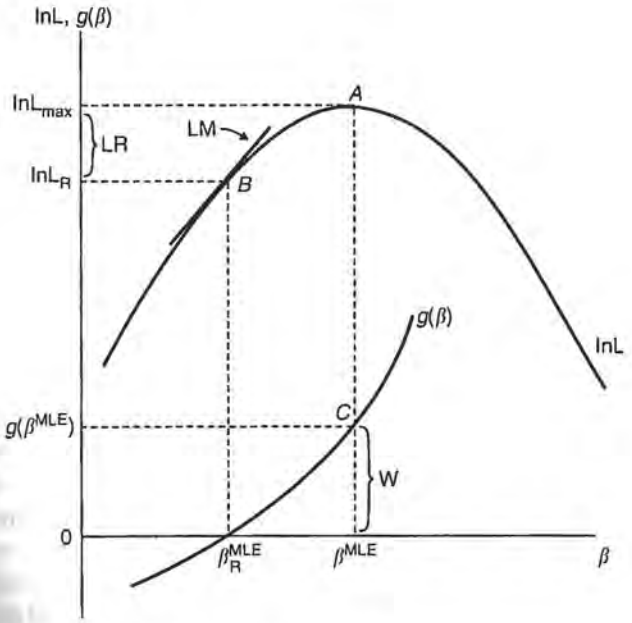

Figure 4.3 Explaining the LR, W, and LM statistics.

mold – for example, if the restrictions are nonlinear, the model is nonlinear in the parameters, or the errors are distributed non-normally – this procedure is inappropriate and is usually replaced by one of three asymptotically equivalent tests. These are the *likelihood ratio* (LR) test, the *Wald* (W) test, and the *Lagrange multiplier* (LM) test. The test statistics associated with these tests have unknown small-sample distributions, but are each distributed asymptotically as a chi-square (χ^2) with degrees of freedom equal to the number of restrictions being tested.

These three test statistics are based on three different rationales. Consider Figure 4.3, in which the log-likelihood ($\ln L$) function is graphed as a function of β, the parameter being estimated. β^{MLE} is, by definition, the value of β at which $\ln L$ attains its maximum. Suppose the restriction being tested is written as $g(\beta) = 0$, satisfied at the value β_R^{MLE} where the function $g(\beta)$ cuts the horizontal axis:

1. *The LR test* If the restriction is true, then $\ln L_R$, the maximized value of lnL imposing the restriction, should not be *significantly* less than $\ln L_{max}$, the unrestricted maximum value of $\ln L$. The LR test tests whether $(\ln L_{max} - \ln L_R)$ is significantly different from zero.
2. *The W test* If the restriction $g(\beta) = 0$ is true, then $g(\beta^{MLE})$ should not be *significantly* different from zero. The W test tests whether β^{MLE} (the unrestricted estimate of β) violates the restriction by a significant amount.
3. *The LM test* The log-likelihood function $\ln L$ is maximized at point A where the slope of $\ln L$ with respect to β is zero. If the restriction is true, then the slope of $\ln L$ at point B should not be *significantly* different from zero. The LM test tests

whether the slope of $\ln L$, evaluated at the restricted estimate, is significantly different from zero.

When faced with three statistics with identical asymptotic properties, econometricians would usually choose among them on the basis of their small-sample properties, as determined by Monte Carlo studies. In this case, however, it happens that computational cost plays a dominant role in this respect. To calculate the LR statistic, both the restricted and the unrestricted estimates of β must be calculated. If neither is difficult to compute, then the LR test is computationally the most attractive of the three tests. To calculate the W statistic only the unrestricted estimate is required; if the restricted but not the unrestricted estimate is difficult to compute, owing to a nonlinear restriction, for example, the W test is computationally the most attractive. To calculate the LM statistic, only the restricted estimate is required; if the unrestricted but not the restricted estimate is difficult to compute – for example, when imposing the restriction transforms a nonlinear functional form into a linear functional form – the LM test is the most attractive. In cases in which computational considerations are not of consequence, the LR test is the test of choice.

4.6 Bootstrapping

Testing hypotheses exploits knowledge of the sampling distributions of test statistics when the null is true, and constructing confidence intervals requires knowledge of estimators' sampling distributions. Unfortunately, this "knowledge" is often questionable, or unavailable, for a variety of reasons:

1. Assumptions made concerning the distribution of the error term may be false. For example, the error may not be distributed normally, or even approximately normally, as is often assumed.
2. Algebraic difficulties in calculating the characteristics of a sampling distribution often cause econometricians to undertake such derivations assuming that the sample size is very large. The resulting "asymptotic" results may not be close approximations for the sample size of the problem at hand.
3. For some estimating techniques, such as minimizing the median squared error, even asymptotic algebra cannot produce formulas for estimator variances.
4. A researcher may obtain an estimate by undertaking a series of tests, the results of which lead eventually to adoption of a final estimation formula. This search process makes it impossible to derive algebraically the character of the sampling distribution.

One way of dealing with these problems is to perform a Monte Carlo study: data are simulated to mimic the process thought to be generating the data, the estimate or test statistic is calculated and this process is repeated several thousand times to allow computation of the character of the sampling distribution of the estimator or test statistic. To tailor the Monte Carlo study to the problem at hand, initial parameter estimates are

used as the "true" parameter values, and the actual values of the explanatory variables are employed as the "fixed in repeated sample" values of the explanatory variables. But this tailoring is incomplete because in the Monte Carlo study the errors must be drawn from a known distribution such as the normal. This is a major drawback of the traditional Monte Carlo methodology in this context.

The bootstrap is a special Monte Carlo procedure that circumvents this problem. It does so by assuming that the unknown distribution of the error term can be adequately approximated by a discrete distribution that gives equal weight to each of the residuals from the original estimation. With a reasonable sample size, in typical cases most of the residuals should be small in absolute value, so that although each residual is given equal weight (and thus is equally likely to be chosen in random draws from this distribution), small residuals predominate, causing random draws from this distribution to produce small values much more frequently than large values. This procedure, which estimates sampling distributions by using only the original data (and so "pulls itself up by its own bootstraps"), has proved to be remarkably successful. In effect, it substitutes computing power, the price of which has dramatically decreased, for theorem-proving, whose price has held constant or even increased as we have adopted more complicated estimation procedures.

The bootstrap begins by estimating the model in question and saving the residuals. It performs a Monte Carlo study, using the estimated parameter values as the "true" parameter values and the actual values of the explanatory variables as the fixed explanatory variable values. During this Monte Carlo study errors are drawn, with replacement, from the set of original residuals. In this way account is taken of the unknown distribution of the true errors. This "residual-based" technique is only appropriate whenever each error is equally likely to be drawn for each observation. If this is not the case, an alternative bootstrapping method is employed. See the general notes for further discussion.

General Notes

4.1 Introduction

- It is extremely convenient to assume that errors are distributed normally, but there exists little justification for this assumption. Tiao and Box (1973, p. 13) speculate that "Belief in universal near-Normality of disturbances may be traced, perhaps, to early feeding on a diet of asymptotic Normality of maximum likelihood and other estimators." Poincaré is said to have claimed that "everyone believes in the [Gaussian] law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an empirical fact." Several tests for normality exist; for a textbook exposition see Maddala (1977, pp. 305–8). See also Judge *et al.* (1985, pp. 882–7). The consequences of non-normality of the fat-tailed kind, implying infinite variance, are quite serious, since hypothesis testing and interval estimation cannot be undertaken meaningfully. Faced with such non-normality, two options exist. First, one can employ robust estimators, as described in chapter 21. And second, one can transform the data to create transformed errors that are closer to being normally distributed. For discussion see Maddala (1977, pp. 314–17).
- Testing hypotheses is viewed by some with scorn. Consider, for example, the remark of Johnson (1971, p. 2): "The 'testing of hypotheses' is

frequently merely a euphemism for obtaining plausible numbers to provide ceremonial adequacy for a theory chosen and defended on *a priori* grounds." For a completely opposite cynical view, Blaug (1980, p. 257) feels that econometricians "express a hypothesis in terms of an equation, estimate a variety of forms for that equation, select the best fit, discard the rest, and then adjust the theoretical argument to rationalize the hypothesis that is being tested."

- It should be borne in mind that despite the power, or lack thereof, of hypothesis tests, often conclusions are convincing to a researcher only if supported by personal experience. Nelson (1995, p. 141) captures this subjective element of empirical research by noting that "what often really seems to matter in convincing a male colleague of the existence of sex discrimination is not studies with 10000 'objective' observations, but rather a particular single direct observation: the experience of his own daughter."
- Hypothesis tests are usually conducted using a type I error rate (probability of rejecting a true null) of 5%, but there is no good reason why 5% should be preferred to some other percentage. The father of statistics, R. A. Fisher, suggested it in an obscure 1923 paper, and it has been blindly followed ever since. Rosnow and Rosenthal (1989, p. 1277) recognize that "surely, God loves the .06 as much as the .05." By increasing the type I error rate, the type II error rate (the probability of accepting the null when it is false) is lowered, so the choice of type I error rate should be determined by the relative costs of the two types of error, but this issue is usually ignored by all but Bayesians (see chapter 14). The .05 is chosen so often that it has become a tradition, prompting Kempthorne and Doerfler (1969, p. 231) to opine that "statisticians are people whose aim in life is to be wrong 5% of the time!"
- Most hypothesis tests fall into one of two categories. Suppose we are testing the null that the slope of x in a regression is zero. One reason we are doing this could be that we are genuinely interested in whether this slope is zero, perhaps because it has some substantive policy implication or is crucial to supporting some economic theory. This is the category for which the traditional choice of a 5% type I error rate is thought to be applicable. But it may be that we have no real interest in this parameter and that some other parameter in this regression is of primary interest. In this case, the reason we are testing this null hypothesis is because if we fail to reject it we can drop this explanatory variable from the estimating equation, thereby improving estimation of this other parameter. In this context, the traditional choice of 5% for the type I error is no longer an obvious choice, something that is not well recognized by practitioners. As explained in chapter 6, omitting a relevant explanatory variable in general causes bias. Because most econometricians fear bias, they need to be very careful that they do not drop an explanatory variable that belongs in the regression. Because of this they want the power of their test (the probability of rejecting the null when it is false) to be high, to ensure that they do not drop a relevant explanatory variable. But choosing a low type I error, such as 5%, means that power will be much lower than if a type I error of, say, 30% was chosen. Somehow the type I error needs to be chosen so as to maximize the quality of the estimate of the parameter of primary interest. Maddala and Kim (1998, p. 140) suggest a type I error of about 25%. Further discussion of this important practical issue occurs in the general notes to section 5.2, in section 6.2 and its general notes, and in the technical notes to section 13.5.
- For a number of reasons, tests of significance can sometimes be misleading. A good discussion can be found in Bakan (1966). One of the more interesting problems in this respect is the fact that almost any parameter can be found to be significantly different from zero if the sample size is sufficiently large. (Almost every relevant independent variable will have *some* influence, however small, on a dependent variable; increasing the sample size will reduce the variance and eventually make this influence "statistically significant.") Thus, although a researcher wants a large sample size to generate more accurate estimates, too large a sample size might cause difficulties in interpreting the usual tests of significance.

McCloskey and Ziliak (1996) look carefully at a large number of empirical studies in economics and conclude that researchers seem not to appreciate that statistical significance does not imply economic significance. One must ask if the magnitude of the coefficient in question is large enough for its explanatory variable to have a meaningful (as opposed to "significant") influence on the dependent variable. This is called the *too-large sample size problem*. One suggestion for dealing with this problem is to report *beta coefficient* estimates – scale the usual coefficient estimates so that they measure the number of standard deviation changes in the dependent variable due to a standard deviation change in the explanatory variable. A second suggestion is to adjust the significance level downward as the sample size grows; for a formalization see Leamer (1978, pp. 88–9, 104–5). See also Attfield (1982). Leamer would also argue (1988, p. 331) that this problem would be resolved if researchers recognized that genuinely interesting hypotheses are neighborhoods, not points. Another interesting dimension of this problem is the question of what significance level should be employed when replicating a study with new data; conclusions must be drawn by considering both sets of data as a unit, not just the new set of data. For discussion see Busche and Kennedy (1984). Another interesting example in this context is the propensity for published studies to contain a disproportionately large number of type I errors; studies with statistically significant results tend to get published, whereas those with insignificant results do not. For comment see Feige (1975). Yet another example that should be mentioned here is pretest bias, discussed in chapter 13.

- In psychometrics these problems with significance testing have given rise to a book entitled "What if there were no significance tests?" (Harlow, Mulaik, and Steiger, 1997) and journal policies not to publish papers that do not report effect size (the magnitude of a treatment's impact, usually measured in terms of standard deviations of the phenomenon in question). Loftus's (1993, p. 250) opinion that "hypothesis testing is overstated, overused and practically useless as a means of illuminating what the data in some experiment are trying to tell us," is shared by many. Nester (1996) has a collection of similar quotes berating significance testing. One way of alleviating this problem is to report confidence intervals rather than hypothesis test results; this allows a reader to see directly the magnitude of the parameter estimate along with its uncertainty.

  In econometrics, McCloskey (1998, chapter 8) summarizes her several papers on the subject, chastising the profession for its tendency to pay undue homage to significance testing. McCloskey and Ziliak (1996, p. 112) cogently sum up this view as follows:

  > No economist has achieved scientific success as a result of a statistically significant coefficient. Massed observations, clever common sense, elegant theorems, new policies, sagacious economic reasoning, historical perspective, relevant accounting: these have all led to scientific success. Statistical significance has not.

  Ziliak and McCloskey (2004) is an update of their earlier study, finding that researchers continue to abuse significance tests; this paper is followed by a set of interesting commentaries.
- Tukey (1969) views significance testing as "sanctification" of a theory, with a resulting unfortunate tendency for researchers to stop looking for further insights. Sanctification via significance testing should be replaced by searches for additional evidence, both corroborating evidence, and, especially, disconfirming evidence. If your theory is correct, are there testable implications? Can you explain a range of interconnected findings? Can you find a bundle of evidence consistent with your hypothesis but inconsistent with alternative hypotheses? Abelson (1995, p. 186) offers some examples. A related concept is encompassing: Can your theory encompass its rivals in the sense that it can explain other models' results? See Hendry (1988).
- Inferences from a model may be sensitive to the model specification, the validity of which may be in doubt. A *fragility analysis* is recommended to deal with this; it examines the range of inferences resulting from the range of believable model

specifications. See Leamer and Leonard (1983) and Leamer (1983a).

- Armstrong (1978, pp. 406–7) advocates the use of the method of multiple hypotheses, in which research is designed to compare two or more reasonable hypotheses, in contrast to the usual advocacy strategy in which a researcher tries to find confirming evidence for a favorite hypothesis. (Econometricians, like artists, tend to fall in love with their models!) It is claimed that the latter procedure biases the way scientists perceive the world, and that scientists employing the former strategy progress more rapidly.
- Keuzenkamp and Magnus (1995) have an interesting and informative discussion of the different purposes served by hypothesis testing and of the meaning of "significance."

4.2 Testing a Single Hypothesis: The *t* Test

- A *t* test can be used to test any single linear constraint. Suppose $y = \alpha + \beta x + \delta w + \varepsilon$ and we wish to test $\beta + \delta = 1$. A *t* test is formulated by rewriting the constraint so that it is equal to zero, in this case as $\beta + \delta - 1 = 0$, estimating the left-hand side as $\beta^{OLS} + \delta^{OLS} - 1$ and dividing this by the square root of its estimated variance to form a *t* statistic with degrees of freedom equal to the sample size minus the number of parameters estimated in the regression. Estimation of the variance of $(\beta^{OLS} + \delta^{OLS} - 1)$ is a bit messy, but can be done using the elements in the estimated variance–covariance matrix from the OLS regression. This messiness can be avoided by using an *F* test, as explained in the general notes to the following section.
- Another way of avoiding this messiness for a single linear hypothesis is by twisting the specification to produce an artificial regression in which one of the "coefficients" is the linear restriction under test. Consider the example above in which we wish to test $\beta + \delta = 1$, rewritten as $\beta + \delta - 1 = 0$. Set $\theta = \beta + \delta - 1$, solve for $\beta = \theta - \delta + 1$, substitute into the original specification to get $y = \alpha + (\theta - \delta + 1)x + \delta w + \varepsilon$ and rearrange to get $y - x = \alpha + \theta x + \delta(w - x) + \varepsilon$. Regressing $y - x$ on an intercept, x and $w - x$ will produce estimates of θ and its variance. Its *t* statistic can be used to test the null that $\theta = 0$.
- Nonlinear constraints are usually tested by using a W, LR, or LM test, but sometimes an "asymptotic" *t* test is encountered: the nonlinear constraint is written with its right-hand side equal to zero, the left-hand side is estimated and then divided by the square root of an estimate of its asymptotic variance to produce the asymptotic *t* statistic. It is the square root of the corresponding *W* test statistic. The asymptotic variance of a nonlinear function was discussed in chapter 2.

4.3 Testing a Joint Hypothesis: The *F* Test

- If there are only two observations, a linear function with one independent variable (i.e., two parameters) will fit the data perfectly, *regardless* of what independent variable is used. Adding a third observation will destroy the perfect fit, but the fit will remain quite good, simply because there is effectively only one observation to explain. It is to correct this phenomenon that statistics are adjusted for *degrees of freedom* – the number of "free" or linearly independent observations used in the calculation of the statistic. For all of the *F* tests cited in this section, the degrees of freedom appropriate for the numerator is the number of restrictions being tested. The degrees of freedom for the denominator is $N - K$, the number of observations less the number of parameters being estimated. $N - K$ is also the degrees of freedom for the *t* statistic mentioned in section 4.2.
- The degrees of freedom of a statistic is the number of quantities that enter into the calculation of the statistic minus the number of constraints connecting these quantities. For example, the formula used to compute the sample variance involves the sample mean statistic. This places a constraint on the data – given the sample mean any one data point can be determined by the other $(N - 1)$ data points. Consequently, there are in effect only $(N - 1)$ unconstrained observations available to estimate the sample variance

the degrees of freedom of the sample variance statistic is $(N-1)$.

- A special case of the F statistic is automatically reported by most regression packages – the F statistic for the "overall significance of the regression." This F statistic tests the hypothesis that all the slope coefficients are zero. The constrained regression in this case would have only an intercept.
- To clarify further how one runs a constrained regression, suppose, for example, that $y = \alpha + \beta x + \delta w + \varepsilon$ and we wish to impose the constraint that $\beta + \delta = 1$. Substitute $\beta = 1 - \delta$ and rearrange to get $y - x = \alpha + \delta(w - x) + \varepsilon$. The restricted SSE is obtained from regressing the constructed variable $(y - x)$ on a constant and the constructed variable $(w - x)$. Note that because the dependent variable has changed it will not be meaningful to compare the R^2 of this regression with that of the original regression.
- In the preceding example it should be clear that it is easy to construct an F test of the hypothesis that $\beta + \delta = 1$. The resulting F statistic will be the square of the t statistic that could be used to test this same hypothesis (described in the preceding section, involving a messy computation of the required standard error). This reflects the general result that the square of a t statistic is an F statistic (with numerator degrees of freedom equal to one and denominator degrees of freedom equal to the t test degrees of freedom). With the exception of testing a single coefficient equal to a specific value, it is usually easier to perform an F test than a t test. Note that the square root of an F statistic is not equal to a t statistic unless the degrees of freedom of the numerator is one.
- By dividing the numerator and denominator of the F statistic by SST (total sum of squares), the total variation in the dependent variable F can be written in terms of R^2 and ΔR^2. This method is not recommended, however, because often the restricted SSE is obtained by running a regression with a different dependent variable than that used by the regression run to obtain the unrestricted SSE (as in the example above), implying different SSTs and incompatible R^2s.
- An F statistic with p and n degrees of freedom is the ratio of two independent chi-square statistics, each divided by its degrees of freedom, p for the numerator and n for the denominator. For the standard F statistic that we have been discussing, the chi-square on the denominator is SSE, the sum of squared OLS residuals, with degrees of freedom $T-K$, divided by σ^2. Asymptotically, $SSE/(T-K)$ equals σ^2, so the denominator becomes unity, leaving F equal to the numerator chi-square divided by its degrees of freedom p. Thus, asymptotically pF is distributed as a chi-square with degrees of freedom p. This explains why test statistics derived on asymptotic arguments are invariably expressed as chi-square statistics rather than as F statistics. In small samples it cannot be said that this approach, calculating the chi-square statistic and using critical values from the chi-square distribution, is definitely preferred to calculating the F statistic and using critical values from the F distribution: the choice of chi-square statistic here is an econometric ritual.
- One application of the F test is in testing for causality. It is usually assumed that movements in the dependent variable are caused by movements in the independent variable(s), but the existence of a relationship between these variables proves neither the existence of causality nor its direction. Using the dictionary meaning of causality, it is impossible to test for causality. Granger developed a special definition of causality which econometricians use in place of the dictionary definition; strictly speaking, econometricians should say "Granger-cause" in place of "cause," but usually they do not. A variable x is said to Granger-cause y if prediction of the current value of y is enhanced by using past values of x. This definition is implemented for empirical testing by regressing y on past, current, and future values of x; if causality runs one way, from x to y, the set of coefficients of the future values of x should test insignificantly different from the zero vector (via an F test), and the set of coefficients of the past values of x should test significantly different from zero. Before running this regression both data sets are transformed (using the same transformation), so as to eliminate any autocorrelation

in the error attached to this regression. (This is required to permit use of the F test; chapter 8 examines the problem of autocorrelated errors.) Great controversy exists over the appropriate way of conducting this transformation and the extent to which the results are sensitive to the transformation chosen. Other criticisms focus on the possibility of expected future values of x affecting the current value of y, and, in general, the lack of full correspondence between Granger-causality and causality. (Consider, for example, the fact that Christmas card sales Granger-cause Christmas!) In essence, Granger-causality just means precedence. Bishop (1979) has a concise review and references to the major studies on this topic. Darnell (1994, pp. 41–3) has a concise textbook exposition.

4.4 Interval Estimation for a Parameter Vector

- Figure 4.2 can be used to illustrate another curiosity – the possibility of accepting the hypothesis that

$$\begin{bmatrix}\beta_4\\ \beta_5\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix}$$

on the basis of an F test while rejecting the hypothesis that $\beta_4 = 0$, and the hypothesis that $\beta_5 = 0$ on the basis of individual t tests. This would be the case if, for the sample at hand, the point (0, 0) fell in either of the small shaded areas (in the upper right or lower left) of the ellipse in Figure 4.2. For a summary discussion of the possible cases that could arise here, along with an example of this seldom encountered curiosity, see Geary and Leser (1968).

4.5 LR, W, and LM Statistics

- Figure 4.3 is taken from Buse (1982) who uses it to conduct a more extensive discussion of the W, LR, and LM statistics, noting, for example, that the geometric distances being tested depend on the second derivatives of the log-likelihood function, which enter into the test statistics through variances (recall that these second derivatives appeared in the Cramer–Rao lower bound). Engle (1984) has an extensive discussion of the W, LR, and LM test statistics. Greene (2008, pp. 498–507) is a good textbook exposition.
- An alternative derivation of the LM statistic gives rise to its name. The Lagrange multiplier technique is used to maximize subject to restrictions; if the restrictions are not binding, the vector of Lagrange multipliers is zero. When maximizing the log-likelihood subject to restrictions, true restrictions shoud be close to being satisfied by the data and so the value of the Lagrange multiplier vector should be close to zero. Consequently, we can test the restrictions by testing the vector of Lagrange multipliers against the zero vector. This produces the LM test.
- Critical values from the χ^2 distribution are used for the LR, W, and LM tests, in spite of the fact that in small samples they are not distributed as χ^2. This is a weakness of all three of these tests. Furthermore, it has been shown by Berndt and Savin (1977) that in linear models in small samples, the values of these test statistics are such that $W \geq LR \geq LM$ for the same data, testing for the same restrictions. Consequently, it is possible for conflict among these tests to arise in the sense that in small samples a restriction could be accepted on the basis of one test but rejected on the basis of another. Zaman (1996, pp. 411–12) argues that the third-order terms in the asymptotic expansions of the W, LR, and LM tests are different and upon examination the LR test is to be favored in small samples. Dagenais and Dufour (1991, 199[illegible] conclude that W tests and some forms of LM tests are not invariant to changes in the measurement units, the representation of the null hypothesis and reparameterizations, and so recommend the LR test.
- For the special case of testing linear restrictions in the CNLR model with σ^2 known, the LR, W, and LM tests are equivalent to the F test (which in this circumstance, because σ^2 is known, becomes a χ^2 test). When σ^2 is unknown, see Vandaele (1981) for the relationships among these tests. In general, the W and LM test statistics are [illegible]

on a quadratic approximation to the log-likelihood (and so are equivalent if the log-likelihood is quadratic); for this reason, Meeker and Escobar (1995) claim that confidence regions based on the LR statistic are superior.

- In many cases it turns out that the parameters characterizing several misspecifications are functionally independent of each other, so that the information matrix is block-diagonal. In this case the LM statistic for testing all the misspecifications jointly is the sum of the LM statistics for testing each of the misspecifications separately. The same is true for the W and LR statistics.
- A nonlinear restriction can be written in different ways. For example, the restriction $\alpha\beta - 1 = 0$ could be written as $\alpha - 1/\beta = 0$, or the restriction $\theta = 1$ could be written as $\ln\theta = 0$. Gregory and Veall (1985) find that the Wald test statistic is sensitive to which way the restriction is written. It would be wise to formulate the restriction in the simplest possible way, avoiding quotients. The former versions in the two examples above would be recommended.
- In chapter 8 much will be made of the fact that the OLS variance estimates are biased whenever the variance–covariance matrix of the error term is nonspherical. As explained in chapter 8 a very popular (and recommended) way of dealing with this is to employ a "robust" estimate of the OLS variance–covariance matrix, which avoids this bias in large samples. A great advantage of the Wald test is that it easily incorporates this adjustment; the LR and LM tests do not. This is one reason why the Wald test is the most popular of the W, LR, and LM tests; another reason is that it is the test most familiar to practitioners, with t values (the square root of a W test) reported automatically in software output.

4.6 Bootstrapping

- Jeong and Maddala (1993) is a good survey of bootstrapping in an econometric context. Li and Maddala (1996) extend this survey, concentrating on time series data. Ruiz and Pascual (2002) survey bootstrapping for financial time series data. Veall (1987, 1992) are good examples of econometric applications, and Veall (1989, 1998) are concise surveys of such applications. Kennedy (2001) is a good elementary exposition. Efron and Tibshirani (1993) is a detailed exposition. Brownstone and Valletta (2001) is a concise exposition. MacKinnon (2006) is a good survey, emphasizing that bootstrapping does not work well in all contexts and often needs to be undertaken in special ways.
- Davidson and MacKinnon (2000) suggest a means of determining how many bootstraps are required to calculate for testing purposes. Efron (1987) suggests that estimation of bias and variance requires only about 200, but estimation of confidence intervals, and thus use for hypothesis testing, requires about 2000. Booth and Sarkar (1998) find that about 800 bootstrap resamples are required to estimate variance properly.
- An implicit assumption of bootstrapping is that the errors are exchangeable, meaning that each error, which in this case is one of the N residuals (sample size N), is equally likely to occur with each observation. This may not be true. For example, larger error variances might be associated with larger values of one of the explanatory variables (i.e., a form of heteroskedasticity – see chapter 8), in which case large errors are more likely to occur whenever there are large values of this explanatory variable. A variant of the bootstrap called the complete, or paired, bootstrap is employed to deal with this problem. Each of the N observations in the original sample is written as a vector of values containing an observation on the dependent variable and an associated observation for each of the explanatory variables. Observations for a Monte Carlo repeated sample are drawn with replacement from the set of these vectors.

  This technique introduces three innovations. First, it implicitly employs the true, unknown errors because they are part of the dependent variable values, and keeps these unknown errors paired with the original explanatory variable values with which they were associated. Second, it does not employ estimates of the unknown parameters, implicitly using the true parameter values (and the true functional form).

And third, it no longer views the explanatory variable values as fixed in repeated samples, assuming instead that these values were drawn from a distribution adequately approximated by a discrete distribution giving equal weight to each observed vector of values on the explanatory variables. A larger sample size is needed for this to be representative of the population of explanatory variable values. This makes sense in a context in which the observations are a small subset of a large population of similar observations. Unfortunately, it does not make sense if the original observations exhaust the population, as would be the case, for example, if they were observations on all large Canadian cities. This would especially be the case if there was one city that was markedly different than the others (very large, for example); the bootstrapped samples could contain this city more than once and so can be misleading. It would also not make sense in a context in which a researcher selected the values of the explanatory variables to suit the study rather than via some random process. It also would not be suitable for a problem in which the errors are autocorrelated in that the error for one observation is related to the error for another; in this case a bootstrapping residuals technique would have to be used with an appropriate modification to create the desired error correlation in each bootstrapped sample. The message here is that the bootstrapping procedure must be carefully thought out for each application.

- To find the sampling distribution of a test statistic on the null hypothesis, the null hypothesis parameter values should be used when creating Monte Carlo repeated samples. In general, as with all Monte Carlo studies, every effort should be made to create the bootstrap samples in a way that incorporates all known facets of the data-generating process. As an example, consider the residuals from estimating a nonlinear functional form. Unlike when estimating a linear function, the average of these residuals may not be zero; before bootstrapping the residuals should be recentered (by subtracting their average from each residual).
- The most common use of the bootstrap by practitioners is to estimate standard errors in contexts in which standard errors are difficult to compute. Here are three examples.
  - **(a)** The estimating procedure may involve two steps, with the first step computing an estimated or expected value of a variable and the second step using this variable to estimate an unknown parameter. Calculation of the standard error in this context is difficult because of the extra stochastic ingredient due to the first step.
  - **(b)** The desired coefficient estimate may be a nonlinear function of two estimates, for example, $\hat{\theta} = \hat{\beta}/\hat{\delta}$. The delta method (see appendix B) could be used to estimate the variance of $\hat{\theta}$, but it has only asymptotic justification.
  - **(c)** An estimation procedure may have been used for which there does not exist a pushbutton in the software for robust variance estimates to guard against heteroskedasticity of unknown form (see chapter 8).

  In each of these examples a bootstrapping procedure would be used to produce B coefficient estimates and then the sample variance of these estimates would be used to estimate the variance of the estimator. Standard errors are estimated by the square root of the variance measure.
- The second-most common use of the bootstrap by practitioners is to adjust hypothesis tests for incorrect type I error rates. The non-nested J test, for example, has a type I error rate that exceeds its nominal rate (i.e., in repeated samples the J statistic exceeds the α% critical value from the *table* more than α% of the time). Similarly, LM tests that rely on the outer product of the gradient (OPG) estimate of the variance–covariance matrix have type I error rates that differ from what they are supposed to be. In these cases a bootstrapping procedure can be used to produce B values of the relevant test statistic and then we can see where the actual statistic value falls in the distribution of these B statistics. So, for example, if B is 1999 we have 2000 values of the test statistic (1999 bootstrapped values plus the original, actual value), and if 39 of these values exceed the actual statistic value the p va

(one-sided) of our test is 2%. In general, as illustrated in this example, for bootstrap tests B should be chosen such that $\alpha(B+1)$ is an integer, where α is the type I error rate.

- An observant reader may have noticed an inconsistency between the preceding two popular applications of bootstrapping. The main reason for estimating standard errors via the bootstrapping procedure is to undertake hypothesis testing or to produce confidence intervals. This would be done by utilizing a critical value from one of the statistical tables provided at the back of most textbooks. But these tables rely on errors being distributed normally, or rely on asymptotic justifications that are invalid in small samples. Part of the whole point of bootstrapping is to avoid having to rely on this assumption. The essence of the hypothesis-testing methodology described above is to calculate special critical values applicable to the specific problem at hand. This suggests that a standard error estimate calculated via bootstrapping should not be used for hypothesis testing except in circumstances in which one is confident that the critical values from the usual tables are applicable. A similar caveat applies when estimating confidence intervals. The estimated standard error (sterr) can be used for this purpose, but it should not in general be combined with the traditional critical values. Instead, we should bootstrap to find the critical values applicable to the problem at hand. Suppose in the example above we had 1000 values of the t statistic. If we order them from smallest to largest and pick out the 50th value (critlow) and the 950th value (crithigh) these values will be the critical values we seek for a two-sided 90% confidence interval. This confidence interval would be formed by taking our estimated coefficient and subtracting critlow*sterr and adding crithigh*sterr. This is our "bootstrapped" confidence interval; it could be asymmetric around the coefficient estimate, in contrast to the traditional confidence interval that is symmetric. This is an example of an *asymptotic refinement* that makes the bootstrap procedure perform better in small samples than formulas based on traditional asymptotic theory.

Technical Notes

4.1 Introduction

- A *type I error* is concluding the null hypothesis is false when it is true; a *type II error* is concluding the null hypothesis is true when it is false. Traditional testing methodologies set the probability of a type I error (called the *size*, usually denoted α, called the *significance level*) equal to an arbitrarily determined value (typically 5%) and then maximize the *power* (one minus the probability of a type II error) of the test. A test is called *uniformly most powerful* (UMP) if it has greater power than any other test of the same size for all degrees of falseness of the hypothesis. Econometric theorists work hard to develop fancy tests with high power, but, as noted by McAleer (1994, p. 334), a test that is never used has zero power, suggesting that tests must be simple to perform if they are to have power.
- A test is *consistent* if its power goes to one as the sample size grows to infinity, something that usually happens if the test is based on a consistent estimate. Many tests are developed and defended on the basis of asymptotics, with most such tests being consistent; this creates a dilemma – how can the power of such tests be compared when asymptotically they all have power one? This problem is solved through the concepts of a *local alternative* and *local power*. For the null hypothesis $\beta = \beta_0$, the alternative hypothesis is indexed to approach the null as the sample size N approaches infinity, so that, for example, the alternative $\beta \neq \beta_0$ becomes the local alternative $\beta_N = \beta_0 + \Delta\beta/\sqrt{N}$. Now an increase in N increases power, but this is balanced by a move of the alternative towards the null; the local alternative is in general constructed so as to make the power approach a well-defined limit as N approaches infinity. This limit is called the local power, and is what is used to compare consistent tests.
- Power varies with the degree of falseness of the null hypothesis. (It also varies, of course, with the precision of estimation, affected by things like sample size, error variance, and variation in regressors.) If the null is true, power is

equal to the probability of a type I error, the significance level of the test; if the null is grossly false, power should be close to 100%. Because the degree of falseness of the null is not known, the power of a test is not known. This creates the following unfortunate dilemma. Suppose a null is "accepted" (i.e., not rejected). We would like to conclude that this acceptance is because the null is true, but it may simply be because the power of our test is low. What should be done here (but, embarrassingly, is not) is report power for some meaningful alternative hypothesis; this would give readers of a report some sense of how seriously to take the results of an hypothesis test.

4.3 Testing a Joint Hypothesis: The *F* Test

- The ΔSSE that appears in the numerator of the F statistic sometimes appears in other guises in textbooks. If, for example, the test for β is equal to a specific vector β_0, then $\Delta SSE = (\beta^{OLS} - \beta_0)' X'X(\beta^{OLS} - \beta_0)$. This can be shown algebraically, but it is instructive to see why it makes sense. Assuming the CNLR model applies, under the null hypothesis β^{OLS} is distributed normally with mean β_0 and variance–covariance matrix $\sigma^2(X'X)^{-1}$. Thus $(\beta^{OLS} - \beta_0)$ is distributed normally with mean zero and variance $\sigma^2(X'X)^{-1}$, implying that $(\beta^{OLS} - \beta_0)'X'X(\beta^{OLS} - \beta_0)/\sigma^2$ is distributed as a chi-square. (This is explained in the technical notes to section 4.5.) This chi-square is the numerator chi-square of the F statistic (an F statistic is the ratio of two independent chi-squares, each divided by its degrees of freedom); the σ^2 gets canceled out by a σ^2 that appears in the denominator chi-square.

4.5 LR, W, and LM Statistics

- The LR test statistic is computed as $-2 \ln \lambda$ where λ is the *likelihood ratio*, the ratio of the constrained maximum of the likelihood (i.e., under the null hypothesis) to the unconstrained maximum of the likelihood. This is just $2(\ln L_{max} - \ln L_R)$, easily calculated by estimating MLE unrestricted, estimating again restricted, and picking out the maximized log-likelihood values reported by the software.
- The W statistic is computed using a generalized version of the χ^2 which is very useful to know. A sum of J independent, squared standard normal variables is distributed as χ^2 with J degrees of freedom. (This in effect defines a χ^2 distribution in most elementary statistics texts.) Thus, if the J elements θ_j of θ are distributed normally with mean zero, variance σ_j^2 and zero covariance, then $Q=\Sigma\theta_j^2/\sigma_j^2$ is distributed as a χ^2 with J degrees of freedom. This can be written in matrix terminology as $Q=\theta'V^{-1}\theta$ where V is a diagonal matrix with σ_j^2 as its diagonal elements. Generalizing in the obvious way, we obtain $\theta'V^{-1}\theta$ distributed as a χ^2 with J degrees of freedom, where the $J \times 1$ vector θ is distributed multivariate normally with mean zero and variance–covariance matrix V.

  For the W statistic, θ is a vector $\hat{g}$ of the J restrictions evaluated at β^{MLE}, and V, the variance–covariance matrix of $\hat{g}$, is given by $G'CG$ where G is the $(K \times J)$ matrix of derivatives of $\hat{g}$ with respect to β and C is the Cramer–Rao lower bound, representing the asymptotic variance of β^{MLE}. (The technical notes of section 2.8 and appendix B provide an explanation of why the asymptotic variance of $\hat{g}$ is given by $G'CG$.) Placing hats over G and C to indicate that they are evaluated at β^{MLE}, we obtain $W = \hat{g}'[\hat{G}'\hat{C}\hat{G}]^{-1}\hat{g}$.
- Calculation of the LM statistic can be undertaken by the formula $\hat{d}'\hat{C}\hat{d}$, sometimes referred to as the *score test*. $\hat{d}$ is a $K \times 1$ vector of the slopes (first derivatives) of lnL with respect to β, evaluated at β_R^{MLE}, the restricted estimate of β. It is called the *score vector*, or the *gradient vector*, or often just the *score*. $\hat{C}$ is an estimate of the Cramer–Rao lower bound. Different ways of estimating the Cramer–Rao lower bound give rise to a variety of LM statistics with identical asymptotic properties but slightly different small-sample properties. For discussion of the various different ways of computing the LM statistic, and an evaluation of their relative merits, see Davidson and MacKinnon (1983).
- If the model in question can be written as $Y = h(x;\beta) + \varepsilon$ where h is either a linear or nonlinear functional form and the ε are distributed independent normally with zero mean and constant variance, an auxiliary regression can be employed

to facilitate calculation of the LM statistic for a test of some portion of β equal to a specific vector. Consider H, the vector of the K derivatives of h with respect to β. Each element of this vector could be evaluated for each of the N observations, using β_R^{MLE}, the restricted estimate of β. This would give a set of N "observations" on each of the K derivatives. Consider also $\hat{\varepsilon}$, the vector of N residuals resulting from the calculation of β_R^{MLE}. Suppose $\hat{\varepsilon}$ is regressed on the K derivatives in H. Then the product of the resulting R^2 and the sample size N yields the LM statistic: LM = NR^2. For a derivation of this, and an instructive example illustrating its application, see Breusch and Pagan (1980, pp. 242–3). Additional examples of the derivation and use of the LM statistic can be found in Godfrey (1978), Breusch and Pagan (1979), Harvey (1981, pp. 167–74), and Tse (1984).

- Here is a very simple example of the NR^2 version of the LM test, often encountered in the literature. Suppose we have the CNLR model $y = \alpha + \beta x + \gamma z + \delta w + \varepsilon$ and we wish to test the joint null hypothesis that $\gamma = \delta = 0$. The restricted MLE residuals $\hat{\varepsilon}$ are obtained by regressing y on x. The derivative of y with respect to α is a column of ones, with respect to β is a column of x values, with respect to γ is a column of z values, and with respect to δ is a column of w values. The LM test is computed as NR^2 from regressing $\hat{\varepsilon}$ on an intercept (the column of ones), x, z, and w. In essence we are trying to see if the restricted residuals can be explained by z and w.
- Conditional moment tests, discussed in chapter 5, give rise to a different NR^2 version of the LM test in which a column of ones is regressed on the score vectors, where R^2 is the uncentered R^2 from this regression. (Uncentered means that the dependent variable, which in this case is always zero, does not have its mean subtracted from it when calculating the total sum of squares.) In this special case, this NR^2 is equal to the explained sum of squares from this regression, so this version of the LM test is sometimes described as the explained sum of squares from a regression of a column of ones on the scores. This test statistic is extremely easy to calculate, but unfortunately is not reliable because in small samples its type I error can be grossly inflated. This is because it is based on the OPG variant of the information matrix, as explained below. Nonetheless, some, such as Verbeek (2000), believe that its computational simplicity overcomes its unreliability. As with most such statistics, its problems can be greatly alleviated via bootstrapping.
- It is noted in appendix B that there are three different ways of estimating the information matrix. This implies that there are three different ways of estimating the variance–covariance matrix needed for calculating the W and LM tests. In general, the OPG variant is inferior to the alternatives and should be avoided; see, for example, Bera and McKenzie (1986). Unfortunately, however, some of the computationally attractive ways of calculating the LM statistic implicitly have built into them the OPG calculation for the variance–covariance matrix of the MLE, causing the size of the resulting LM statistic to be too large. In particular, versions of the LM test that are calculated as the explained sum of squares from regressing a column of ones on first derivatives are suspect. Davidson and MacKinnon (1983) suggest an alternative way of calculating the LM statistic for a wide variety of applications, through running what they call a *double-length regression* (DLR), which retains the computational attractiveness of the OPG variant of the LM test, but avoids its shortcomings. Godfrey (1988, pp. 82–4) has a good discussion. See also Davidson and MacKinnon (1988). Davidson and MacKinnon (1993, pp. 492–502) is a good textbook exposition. Again, bootstrapping can help.

4.6 Bootstrapping

- When drawing OLS residuals for bootstrapping they should be adjusted upwards by multiplying by the square root of $N/(N-K)$ to account for the fact that although the OLS residuals are unbiased estimates of the errors, they underestimate their absolute value.

- A lesson not immediately evident from the discussion in the general notes is that bootstrapping should investigate the sampling distribution of an "asymptotically pivotal" statistic, a statistic whose sampling distribution does not depend on the true values of the parameters (most test statistics are pivotal). For example, rather than bootstrapping the sampling distribution of a parameter estimate, the sampling distribution of the associated t statistic should be bootstrapped. The sampling distribution of the t statistic can be used indirectly to produce confidence intervals, as described earlier in the general notes, rather than calculating confidence intervals directly using the sampling distribution of the parameter estimate.
- The bootstrap can be used to estimate the bias of an estimate. Generate bootstrap samples using $\hat{\beta}$ and then see if the average of the bootstrap estimates is close to $\hat{\beta}$. If not, a bias is evident, and an obvious adjustment can be made to $\hat{\beta}$. This bias correction is seldom used, however, because the bootstrap estimate can be more variable than the $\hat{\beta}$, and any bias is often quite small relative to the standard error of $\hat{\beta}$.
- There are many variants of the bootstrap. One of the most peculiar, and most successful for dealing with heteroskedasticity (heteroskedasticity is discussed in chapter 8), is the *wild bootstrap*. In this procedure, when drawing bootstrapped residuals each residual $\hat{\varepsilon}$ is replaced with either $-0.618\ \hat{\varepsilon}$ or $1.618\ \hat{\varepsilon}$, with probability 0.7236 and 0.2764, respectively. This causes the new residual to have mean zero and variance $\hat{\varepsilon}^2$, forcing heteroskedasticity into the bootstrap draws. Although this is not a good way of estimating the actual heteroskedasticity, this bootstrapping procedure, more successful than the paired bootstrap, works because what is relevant happens when this heteroskedasticity is averaged over bootstrap draws. This is similar to why the heteroskedasticity-consistent variance–covariance matrix estimate (discussed in chapter 8) works. See question 17 in section HH of appendix D for how this peculiar distribution has come about.
- An alternative computer-based means of estimating a sampling distribution of a test statistic is that associated with a randomization/permutation test. The rationale behind this testing methodology is that if an explanatory variable has no influence on a dependent variable then it should make little difference to the outcome of the test statistic if the values of this explanatory variable are shuffled and matched up with different dependent variable values. By performing this shuffling thousands of times, each time calculating the test statistic, the hypothesis can be tested by seeing if the original test statistic value is unusual relative to the thousands of test statistic values created by the shufflings. Notice how different is the meaning of the sampling distribution – it no longer corresponds to "what would happen if we drew different bundles of errors"; now it corresponds to "what would happen if the independent variable values were paired with different dependent variable values." Hypothesis testing is based on viewing the test statistic as having resulted from playing a game of chance; the randomization view of testing claims that there is more than one way to play a game of chance with one's data! For further discussion of this testing methodology in the econometrics context see Kennedy (1995) and Kennedy and Cade (1996). Noreen (1989) is a good elementary reference.

Chapter 5

Specification

5.1 Introduction

At one time, econometricians tended to assume that the model provided by economic theory represented accurately the real-world mechanism generating the data, and viewed their role as one of providing "good" estimates for the key parameters of that model. If any uncertainty was expressed about the model specification, there was a tendency to think in terms of using econometrics to "find" the real-world data-generating mechanism. Both these views of econometrics are obsolete. It is now generally acknowledged that econometric models are "false" and that there is no hope, or pretense, that through them "truth" will be found. Feldstein's (1982, p. 829) remarks are typical of this view: "in practice all econometric specifications are necessarily 'false' models. ... The applied econometrician, like the theorist, soon discovers from experience that a useful model is not one that is 'true' or 'realistic' but one that is parsimonious, plausible and informative." This is echoed by an oft-quoted remark attributed to George Box, "All models are wrong, but some are useful," and another from Theil (1971, p. vi): "Models are to be used, but not to be believed." In Leamer's (2004, p. 555) view "The goal of an empirical economist should be not to determine the truthfulness of a model but rather the domain of usefulness."

In light of this recognition, econometricians have been forced to articulate more clearly what econometric models are, one view being that they "are simply rough guides to understanding" (Quah, 1995, p. 1596). There is some consensus that models are metaphors, or windows, through which researchers view the observable world, and that their adoption depends not upon whether they can be deemed "true" but rather upon whether they can be said to (1) correspond to the facts and (2) be useful. Econometric specification analysis therefore is a means of formalizing what is meant by "corresponding to the facts" and "being useful," thereby defining what is meant by a "correctly specified model." From this perspective, econometric analysis becomes much more than estimation and inference in the context of a given model; in conjunction