INTRODUCTION AND THE LINEAR REGRESSION MODEL

I

This part consists of four chapters.

Chapter 1 discusses the changes in the methodology that have taken place since the 1950s and 1960s, and gives an outline of the book.

Chapter 2 reviews some basic results in statistics. Most students will have covered this introductory material in courses in statistics. This chapter also provides an introduction to matrix algebra.

Chapter 3 covers the simple regression model and Chapter 4 the multiple regression model in detail. This forms the basics of linear regression under the assumption of independent and identically distributed normal errors. These assumptions are relaxed in Part II.

1

W H A T   I S   E C O N O M E T R I C S ?

What is in this Chapter?

This chapter explains the scope and methodology of econometrics. Econometrics deals with the application of statistical tools to economic data. The first task an econometrician faces is that of formulating an economic relationship, which is necessarily a simplified model of the real-world process. Estimation and testing of these models with observed data and the use of the estimated models for prediction and policy analysis are the other two major goals of econometrics.

This chapter also contains a schematic depiction of the various methodological steps involved in an econometric analysis.

1.1   What is Econometrics?

Literally speaking, the word "econometrics" means "measurement in economics." This is too broad a definition to be of any use, because most of economics is concerned with measurement. We measure our gross national product, employment, money supply, exports, imports, price indexes, and so on. What we mean by *econometrics* is:

> The application of statistical and mathematical methods to the analysis of economic data, with the purpose of giving empirical content to economic theories and verifying them or refuting them.

In this respect, econometrics is distinguished from mathematical economics, which consists of the application of mathematics only, and the theories derived need not necessarily have an empirical content.

The application of statistical tools to economic data has a very long history. Stigler (1954) notes that the first "empirical" demand schedule was published in 1699 by Charles Davenant, and that the first modern statistical demand studies were made by Rodulfo Enini, an Italian statistician, in 1907. The main impetus to the development of econometrics, however, came with the establishment of the Econometric Society in 1930 and the publication of the journal *Econometrica* in January 1933.

Before any statistical analysis with economic data can be done, one needs a clear mathematical formulation of the relevant economic theory. To take a very simple example, saying that the demand curve is downward sloping is not enough. We have to write the statement in mathematical form. This can be done in several ways. For instance, defining q as the quantity demanded and p as price, we can write

$$q = \alpha + \beta p \quad \beta < 0$$

or

$$q = Ap^{\beta} \quad \beta < 0$$

As we will see later in the book, one of the major problems we face is the fact that economic theory is rarely informative about functional forms. We have to use statistical methods to choose the functional form, as well.

1.2 Economic and Econometric Models

The first task an econometrician faces is that of formulating an econometric model. What is a model?

A *model* is a simplified representation of a real-world process. For instance, saying that the quantity demanded of oranges depends on the price of oranges is a simplified representation, because there are a host of other variables that one can think of that determine the demand for oranges. For instance, income of consumers, an increase in diet consciousness ("drinking coffee causes cancer, so you better switch to orange juice," etc.), an increase or decrease in the price of apples, and so on. However, there is no end to this stream of other variables. In a remote sense, even the price of gasoline can affect the demand for oranges.

Many scientists have argued in favor of simplicity, because simple models are easier to understand, communicate, and test empirically with data. This is the position of Karl Popper (1959, p. 142) and Milton Friedman (1953, p. 14). The choice of a simple model to explain complex real-world phenomena leads to two criticisms:

1. The model is oversimplified.
2. The assumptions are unrealistic.

For instance, in our example of the demand for oranges, to say that it depends only on the price of oranges is an oversimplification and also an unrealistic assumption. To the criticism of oversimplification, one can argue that it is better to start with a simplified model and progressively construct more complicated models. This is the idea expressed by Koopmans (1957, pp. 142–143). On the other hand, there are some who argue in favor of starting with a very general model and simplifying it progressively based on the data available.[1] The famous statistician L. J. (Jimmy) Savage used to say that "a model should be as big as an elephant." Whatever the relative merits of this alternative approach are, we will start with simple models and progressively build more complicated models.

---

[1] This is the approach suggested by J. D. Sargan and notably David F. Hendry.

The other criticism we have mentioned is that of "unrealistic assumptions." To this criticism Friedman (1953, pp. 14–15) argues that the assumptions of a theory are never descriptively realistic. He says:

> The relevant question to ask about the "assumptions" of a theory is not whether they are descriptively "realistic" for they never are, but whether they are sufficiently good approximations for the purpose at hand. And this question can be answered by only seeing whether the theory works, which means whether it yields sufficiently accurate predictions.

Returning to our example of demand for oranges, to say that it depends only on the price of oranges is a descriptively unrealistic assumption. However, the inclusion of other variables, such as income and price of apples in the model, does not render the model more descriptively realistic. Even this model can be considered to be based on unrealistic assumptions because it leaves out many other variables (such as health consciousness, etc.). But the issue is which model is more useful for predicting the demand for oranges. This issue can be decided only from the data we have and the data we can get.

In practice, we include in our model all the variables that we think are relevant for our purpose and dump the rest of the variables in a basket called "disturbance." This brings us to the distinction between an economic model and an econometric model.

An *economic model* is a set of assumptions that approximately describes the behavior of an economy (or a sector of an economy). An *econometric model* consists of the following:

1. A set of behavioral equations derived from the economic model. These equations involve some observed variables and some "disturbances" (which are a catch-all for all the variables considered as irrelevant for the purpose of this model as well as all unforeseen events).
2. A statement of whether there are errors of observation in the observed variables.
3. A specification of the probability distribution of the "disturbances" (and errors of measurement).

With these specifications we can proceed to test the empirical validity of the economic model and use it to make forecasts or use it in policy analysis.

Taking the simplest example of a demand model, the econometric model usually consists of:

1. The behavioral equation

$$q = \alpha + \beta p + u$$

where q is the quantity demanded and p the price. Here p and q are the observed variables and u is a disturbance term.
2. A specification of the probability distribution of u, which says that $E(u|p) = 0$ and that the values of u for the different observations are independently and normally distributed with mean zero and variance σ^2.

With these specifications one proceeds to test empirically the law of demand or the hypothesis that $\beta < 0$. One can also use the estimated demand function for prediction and policy purposes.

1.3   The Aims and Methodology of Econometrics

The aims of econometrics are:

1. Formulation of econometric models, that is, formulation of economic models in an empirically testable form. Usually, there are several ways of formulating the econometric model from an economic model, because we have to choose the functional form, the specification of the stochastic structure of the variables, and so on. This part constitutes the *specification aspect* of the econometric work.
2. Estimation and testing of these models with observed data. This part constitutes the *inference aspect* of the econometric work.
3. Use of these models for prediction and policy purposes.

During the 1950s and 1960s the inference aspect received a lot of attention and the specification aspect very little. The major preoccupation of econometricians had been the statistical estimation of correctly specified econometric models. During the late 1940s the Cowles Foundation provided a major breakthrough in this respect, but the statistical analysis presented formidable computational problems. Thus the 1950s and 1960s were spent mostly in devising alternative estimation methods and alternative computer algorithms. Not much attention was paid to errors in the specification or to errors in observations.[2] With the advent of high-speed computers, all this has, however, changed. The estimation problems are no longer formidable and econometricians have turned attention to other aspects of econometric analysis.

We can schematically depict the various steps involved in an econometric analysis, as was done before the emphasis on specification analysis. This is shown in Figure 1.1. Since the entries in the boxes are self-explanatory, we will not elaborate on them. The only box that needs an explanation is box 4, "prior information." This refers to any information that we might have on the unknown parameters in the model. This information can come from economic theory or from previous empirical studies.

There has, however, been considerable dissatisfaction with the scheme shown in Figure 1.1. Although one can find instances of dissatisfaction earlier, it was primarily during the 1970s that arguments were levied against the one-way traffic shown in Figure 1.1. We will discuss three of these arguments.

1. In Figure 1.1 there is no feedback from the econometric testing of economic theories to the formulation of economic theories (i.e., from box 6 to box 1). It has been argued that econometricians are not just handmaidens of economic theorists. It is not true that they just take the theories they are given and test them, learning nothing from the tests. So we need an arrow from box 6 to box 1.
2. The same is true regarding the data collection agencies. It is not true that they just gather whatever data they can and the econometricians use whatever data are given them. (The word *data* comes from the Latin word *datum*, which means *given*.) There ought to be feedback from boxes 2 and 5 to box 3.
3. Regarding box 6 itself, it has been argued that the hypothesis testing refers only to the hypotheses suggested by the original economic model. This depends on the assumption that the specification

---

[2] There was some work on specification errors in the early 1960s by Theil and Griliches, but this referred to omitted-variable bias (see Chapter 4). Griliches' lecture notes (unpublished) at the University of Chicago were titled "Specification Errors in Econometrics."
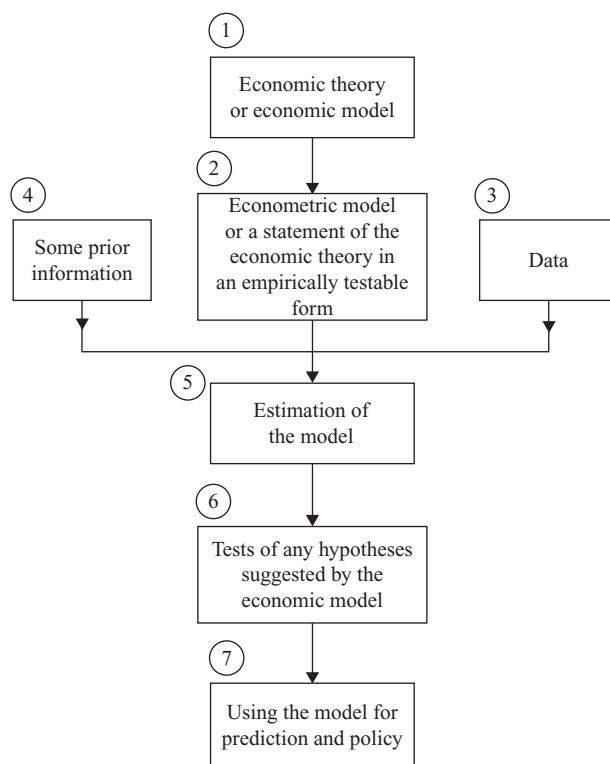
Figure 1.1   A schematic description of the steps involved in an econometric analysis of economic models.

adopted in box 2 is correct. However, what we should be doing is testing the adequacy of the original specification as well. Thus we need an extra box of specification testing and diagnostic checking. There will also be feedback from this box to box 2, that is, the specification tests will result in a new specification for the econometric models. These problems, which have been one of the most important major developments in econometrics in the 1970s, are treated in Chapter 10.

The new developments that have been suggested are shown figuratively in Figure 1.2. Some of the original boxes have been deleted or condensed. The schematic description in Figure 1.2 is illustrative only and should not be interpreted literally. The important things to note are the feedback:

1. From econometric results to economic theory.
2. From specification testing and diagnostic checking to revised specification of the economic model.
3. From the econometric model to data.

In the foregoing scheme we have talked of only one theory, but often there are many competing theories, and one of the main purposes of econometrics is to help in the choice among competing theories. This is the problem of model selection, which is discussed in Chapter 10.
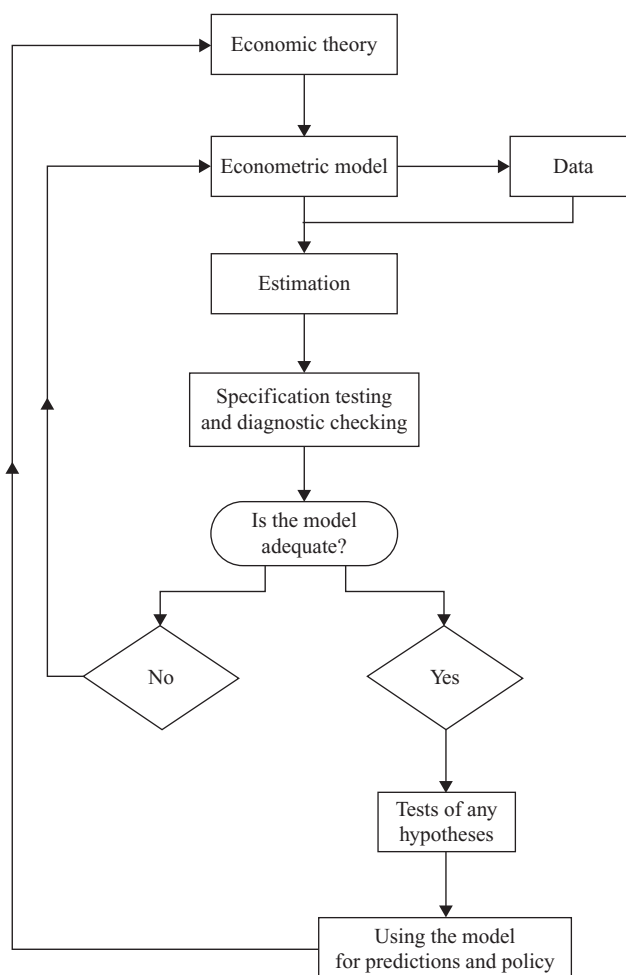
Figure 1.2   A revised schematic description of the steps involved in an econometric analysis of economic models.

1.4   What Constitutes a Test of an Economic Theory?

Earlier, we stated that one of the aims of econometrics was that of testing economic theories. An important question that arises is: What constitutes a test? As evidence of a successful test of economic theory, it is customary to report that the signs of the estimated coefficients in an econometric model are correct. This approach may be termed the approach of *confirming* economic theories. The problem with this approach is that, as Mark Blaug (1980, p. 261) points out:

> In many areas of economics, different econometric studies reach conflicting conclusions and, given the available data, there are frequently no effective methods for deciding which conclusion is correct. In consequence, contradictory hypotheses continue to co-exist sometimes for decades or more.

A more valid test of an economic theory is if it can give predictions that are better than those of alternative theories suggested earlier. Thus one needs to compare a given model with earlier models. This approach of evaluating alternative theories has received greater attention in recent years. The problems associated with model comparison and model selection are dealt with in Chapter 10.

Summary and an Outline of the Book

The preceding discussion suggests some changes that have taken place during the last decade in the development of econometric methods. The earlier emphasis (during the 1950s and 1960s) was on efficient estimation of a given model. The emphasis has now shifted to specification testing, diagnostic checking, model comparison, and an adequate formulation of the expectational variables, given the pervasive role of expectations in almost all economic theories.

A large part of the book, however, will be devoted to developments that took place in the 1960s and 1970s, because these developments form the basis for recent work. However, many later chapters are devoted to recent advances in the areas of specification testing and model selection in time-series and panel data models. Chapter 16 will deal with some current developments in the area of inference in small samples, including Monte Carlo and bootstrap methods. In addition, the other chapters also describe recent developments. For instance, rational expectations models are discussed in Chapter 13, recent developments on tests for serial correlation are reviewed in Chapter 6, and vector autoregressions, unit roots, and cointegration are discussed in Chapter 14.

The book is divided into three parts:

Part I: Introduction and the Linear Regression Model consists of four chapters. Chapter 2 is a review of basic results in statistics. It also has an introduction to matrix algebra. Further results in matrix algebra on characteristic roots and vectors are in the appendix to Chapter 7. Chapter 3 is on the simple regression model and Chapter 4 is on the multiple regression model. In three specialized appendices to Chapter 4, we introduce nonlinear estimation and large-sample theory, including ML, GIVE, and GMM methods. These chapters cover the basics of regression based on the assumption that the errors are $IN(0, \sigma^2)$.

Part II: Violation of the Assumptions of the Basic Model has seven chapters. Chapter 5 is on heteroskedasticity and Chapter 6 is on autocorrelation. Chapter 7 is on multicollinearity. Chapter 8 is on dummy variables and truncated variables. Chapter 9 is on simultaneous equations. Chapter 10 discusses diagnostic checking of the basic linear regression model, specification testing, model selection on nonlinear models, and tests for nonnormality. Chapter 11 is on errors in variables.

These chapters relax the several assumptions made regarding the basic regression model discussed in Part I.

Part III: Special Topics consists of five chapters. Chapter 12 introduces the basic time-series analysis, including ARMA models and the Box–Jenkins approach. Chapter 13 analyzes distributed lags and models of expectations. Chapter 14 is on vector autoregressions, unit roots, and cointegration. Chapter 15 is on the estimation of various panel data models, and Chapter 16 is on small-sample inference and bootstrap methods.

References

Blaug, M.,1980, *The Methodology of Economics*, Cambridge: Cambridge University Press.

Friedman, M., 1953, "The Methodology of Positive Economics," in *Essays in Positive Economics*, Chicago: University of Chicago Press.

Koopmans, T. C., 1957, *Three Essays on the State of Economics Science*, New York: McGraw-Hill.

Popper, K. F., 1959, *The Logic of Scientific Discovery*, London: Hutchinson.

Stigler, G. J., 1954, "The Early History of Empirical Studies of Consumer Behavior," *The Journal of Political Economy* (reprinted in G. J. Stigler, 1965, *Essays in the History of Economics*, Chicago: University of Chicago Press).