1

THE NATURE AND SCOPE OF ECONOMETRICS

> Econometrics may be defined as the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observations, related by appropriate methods of inference.
>
> Paul Samuelson

> Econometrics may be defined as the social science in which tools of economic theory, mathematics, and statistical inference are applied to the analysis of economic phenomena.
>
> Arthur S. Goldberger

Research in economics, finance, management, marketing, and related disciplines is becoming increasingly quantitative. Beginning students in these fields are encouraged, if not required, to take a course or two in econometrics—a field of study that has become quite popular. This chapter gives the beginner an overview of what econometrics is all about.

1.1 WHAT IS ECONOMETRICS?

Simply stated, **econometrics** means economic measurement. Although quantitative measurement of economic concepts such as the gross domestic product (GDP), unemployment, inflation, imports, and exports is very important, the scope of econometrics is much broader, as can be seen from the following definitions:

> Econometrics may be defined as the social science in which the tools of economic theory, mathematics, and statistical inference are applied to the analysis of economic phenomena.[1]

[1] Arthur S. Goldberger, *Econometric Theory,* Wiley, New York, 1964, p. 1.

> Econometrics, the result of a certain outlook on the role of economics, consists of the application of mathematical statistics to economic data to lend empirical support to the models constructed by mathematical economics and to obtain numerical results.[2]

1.2 WHY STUDY ECONOMETRICS?

As the preceding definitions suggest, econometrics makes use of economic theory, mathematical economics, economic statistics (i.e., economic data), and mathematical statistics. Yet, it is a subject that deserves to be studied in its own right for the following reasons.

Economic theory makes statements or hypotheses that are mostly qualitative in nature. For example, microeconomic theory states that, other things remaining the same (the famous *ceteris paribus* clause of economics), an increase in the price of a commodity is expected to decrease the quantity demanded of that commodity. Thus, economic theory postulates a negative or inverse relationship between the price and quantity demanded of a commodity—this is the widely known law of downward-sloping demand or simply *the law of demand*. But the theory itself does not provide any numerical measure of the strength of the relationship between the two; that is, it does not tell by how much the quantity demanded will go up or down as a result of a certain change in the price of the commodity. It is the econometrician's job to provide such numerical estimates. Econometrics gives empirical (i.e., based on observation or experiment) content to most economic theory. If we find in a study or experiment that when the price of a unit increases by a dollar the quantity demanded goes down by, say, 100 units, we have not only confirmed the law of demand, but in the process, we have also provided a numerical estimate of the relationship between the two variables—price and quantity.

The main concern of **mathematical economics** is to express economic theory in mathematical form or equations (or models) without regard to measurability or empirical verification of the theory. Econometrics, as noted earlier, is primarily interested in the empirical verification of economic theory. As we will show shortly, the econometrician often uses mathematical models proposed by the mathematical economist but puts these models in forms that lend themselves to empirical testing.

[2] P. A. Samuelson, T. C. Koopmans, and J. R. N. Stone, "Report of the Evaluative Committee for *Econometrica,*" *Econometrica,* vol. 22, no. 2, April 1954, pp. 141–146.

Economic statistics is mainly concerned with collecting, processing, and presenting economic data in the form of charts, diagrams, and tables. This is the economic statistician's job. He or she collects data on the GDP, employment, unemployment, prices, and so on. These data constitute the raw data for econometric work. But the economic statistician does not go any further because he or she is not primarily concerned with using the collected data to test economic theories.

Although *mathematical statistics* provides many of the tools employed in the trade, the econometrician often needs special methods because of the unique nature of most economic data, namely, that the data are not usually generated as the result of a controlled experiment. The econometrician, like the meteorologist, generally depends on data that cannot be controlled directly. Thus, data on consumption, income, investments, savings, prices, and so on, which are collected by public and private agencies, are nonexperimental in nature. The econometrician takes these data as given. This creates special problems not normally dealt with in mathematical statistics. Moreover, such data are likely to contain errors of measurement, of either omission or commission, and the econometrician may be called upon to develop special methods of analysis to deal with such errors of measurement.

For students majoring in economics and business, there is a pragmatic reason for studying econometrics. After graduation, in their employment, they may be called upon to forecast sales, interest rates, and money supply or to estimate demand and supply functions or price elasticities for products. Quite often, economists appear as expert witnesses before federal and state regulatory agencies on behalf of their clients or the public at large. Thus, an economist appearing before a state regulatory commission that controls prices of gas and electricity may be required to assess the impact of a proposed price increase on the quantity demanded of electricity before the commission will approve the price increase. In situations like this, the economist may need to develop a demand function for electricity for this purpose. Such a demand function may enable the economist to estimate the price elasticity of demand, that is, the percentage change in the quantity demanded for a percentage change in the price. Knowledge of econometrics is very helpful in estimating such demand functions.

It is fair to say that econometrics has become an integral part of training in economics and business.

It may be added the technics and methods developed in econometrics have found uses in several other areas of social sciences, in politics and international relations, in agricultural and medical sciences, as some of the examples discussed in this book will reveal as we progress through the book.

1.3 THE METHODOLOGY OF ECONOMETRICS

How does one actually do an econometric study? Broadly speaking, econometric analysis proceeds along the following lines.

1. The object of research
2. Collecting data
3. Specifying the mathematical model of theory
4. Specifying the statistical, or econometric, model of theory
5. Estimating the parameters of the chosen econometric model
6. Checking for model adequacy: model specification testing
7. Testing hypotheses derived from the model
8. Using the model for prediction or forecasting

To illustrate the methodology, consider this question: Do economic conditions affect people's decisions to enter the labor force, that is, their willingness to work? As a measure of economic conditions, suppose we use the unemployment rate (UNR), and as a measure of labor force participation, we use the labor force participation rate (LFPR). Data on UNR and LFPR are regularly published by the government. So to answer the question, we proceed as follows.

1. The Object of Research

The starting point is to find out what economic theory has to say on the subject you want to study. In labor economics, there are two rival hypotheses about the effect of economic conditions on people's willingness to work. The **discouraged-worker hypothesis (effect)** states that when economic conditions worsen, as reflected in a higher unemployment rate, many unemployed workers give up hope of finding a job and drop out of the labor force. On the other hand, the **added-worker hypothesis (effect)** maintains that when economic conditions worsen, many secondary workers who are not currently in the labor market (e.g., mothers with children) may decide to join the labor force if the main breadwinner in the family loses his or her job. Even if the jobs these secondary workers get are low paying, the earnings will make up some of the loss in income suffered by the primary breadwinner.

Whether, on balance, the labor force participation rate will increase or decrease will depend on the relative strengths of the added-worker and discouraged-worker effects. If the added-worker effect dominates, LFPR will increase even when the unemployment rate is high. Contrarily, if the discouraged-worker effect dominates, LFPR will decrease. How do we find this out? This now becomes our empirical question.

2. Collecting Data

For empirical purposes, therefore, we need quantitative information on the two variables. There are three types of data that are generally available for empirical analysis.

1. Time series
2. Cross-sectional
3. Pooled (a combination of time series and cross-sectional)

Times-series data are collected over a period of time, such as the data on GDP, employment, unemployment, money supply, or government deficits. Such data may be collected at regular intervals—daily (e.g., stock prices), weekly (e.g., money supply), monthly (e.g., the unemployment rate), quarterly (e.g., GDP), or annually (e.g., government budget). So-called **high-frequency data** are collected over an extremely short-period time, such as seconds and minutes. In **flash trading** in stock and foreign exchange markets, such high-frequency data have now become common. These data may be **quantitative** in nature (e.g., prices, income, money supply) or **qualitative** (e.g., male or female, employed or unemployed, married or unmarried, White or Black). As we will show, qualitative variables, also called *dummy* or *categorical* variables, can be every bit as important as quantitative variables.

Since successive observations in time-series data may be correlated, they pose special problems for regressions involving time-series data, particularly the problem of **autocorrelation,** a topic we discuss at length in Chapter 10 with appropriate examples.

Time-series data pose another problem, namely, that they may not be **stationary.** Loosely speaking, *a time series is stationary if its mean and variance do not vary systematically over time.* In Chapter 11 on time-series econometrics, we examine the nature of stationary and nonstationary time series and show the special statistical problems created by the latter. If we are dealing with time-series data, we will denote the observations subscript by t (e.g., Y_t, X_t).

Cross-sectional data are data on one or more variables collected at one point in time, such as the census of population conducted by the U.S. Census Bureau every 10 years (the most recent was on April 1, 2010; the results of the 2020 census are not yet available at the time of writing); the surveys of consumer expenditures conducted by the University of Michigan; and the opinion polls such as those conducted by Gallup, Harris, and other polling organizations. Like time-series data, cross-sectional data have their particular problems, particularly the problem of **heterogeneity.** For example, if you collect data on executive salaries in a given industry at the same point in time, heterogeneity arises because the data may contain small-, medium-, and large-size companies with their own management style and policies. In Chapter 5, we show how the **size or scale effect** of heterogeneous companies can be taken into account.

In **pooled data,** we have elements of both time-series and cross-sectional data. For example, if we collect data on the unemployment rate for 10 countries for a period of 20 years, the data will constitute an example of pooled data—data on the unemployment rate for each country for the 20-year period will form time-series data, whereas data on the unemployment rate for the 10 countries for any single year will be cross-sectional data. In pooled data, we will have 200 observations—20 annual observations for each of the 10 countries.

There is a special type of pooled data called **panel data,** also called **longitudinal** or **micropanel data,** in which the same cross-sectional unit, say, a family or firm, is surveyed over time. For example, the U.S. Department of Commerce conducts a census of housing at periodic intervals. At each periodic survey, the same household (or the people living at the same address) is interviewed to find out if there has been any change in the housing and financial conditions of that household since the last survey. The panel data that result from repeatedly interviewing the same household at periodic intervals provide very useful information on the dynamics of household behavior.

We denote panel data by the double subscript *it*. Thus, Y_{it} will denote the (cross-sectional) observation for the *i*th unit at time *t*.

Quality of the Data. The researcher must check carefully the reputation of the agency that collects the data, for very often the data contain errors of measurement, errors of omission of some observations, or errors of systematic rounding and the like. Data collected in public polls or in marketing surveys may be biased because of nonresponse or incomplete response from the participants. Sometimes the data are available only at a highly aggregated, or macro, level, which may not tell us much about the individual entities included in the aggregate. *It should always be kept in mind that the results of research are only as good as the quality of the data.*

Since an individual researcher does not have the luxury of collecting data on their own, very often they have to depend on secondary sources. But every effort must be made to check the quality of the data used in empirical analysis.

Data Revisions. Macro data on variables such as GDP, consumer price index (CPI), and other economic variables are often revised upward or downward as initially published data may be tentative. It behooves the researcher to keep track of the revised data.

Not only that, macro and micro economic data are often "jolted" by unusual events, such as the great recession of 2008 and the following several years, which was triggered by the collapse of the housing market boon that was set in motion by the subpar loans that were given by real estate brokers and banks. This collapse spilled over into the stock market. The severe recession that started in the United States very quickly spread across the globe, so such unusual events should be taken into account in analyzing economic data.

A startling example is the coronavirus disease 2019 (COVID-19) pandemic that started in one country in March 2019 and quickly spread to other countries, with devastating effects on their economies. In the United States, according to the U.S. Centers for Disease Control and Prevention, as of March 29, 2021, the total number of COVID-19 cases was 30,085,827 and the total number of deaths was 546,704. The long-term consequences of COVID-19 have yet to be assessed. So doing econometric analysis in such situations such as this is very challenging, to say the least.

Sources of the Data. A word is in order regarding data sources. The success of any econometric study hinges on the quality as well as the quantity of data. Fortunately, the Internet has opened up a veritable wealth of data. In Appendix 1A, we give addresses of several websites that have all kinds of microeconomic and macroeconomic data. Students should be familiar with such sources of data, as well as how to access or download them. Of course, these data are continually updated so the reader should find the latest available data.

Data From Statistical Packages. Statistical packages, such as EViews, Stata, Minitab, and SAS, have data sets for expository purposes. The Federal Reserve Bank of St. Louis has extensive data on several macroeconomic variables in Excel format that can be directly imported into Eviews (http://research.stlouisfed.org/fred-addin), and FRED economic data are extremely useful for empirical research. Stata can also import FRED data in Stata format by issuing the command *findit Freduse* while you use Stata.

For our analysis, we obtained the time-series data shown in Table 1-1 of the book's website. This table gives data on the civilian labor force participation rate (CLFPR) and the civilian unemployment rate (CUNR), defined as the number of civilians unemployed as a percentage of the civilian labor force, for the United States for the period 1980–2007.[3] The data beyond this period are given in Problem 1.10 (see Table 1-2 found on the book's website).

[3] We consider here only the aggregate CLFPR and CUNR, but data are available by age, sex, and ethnic composition.

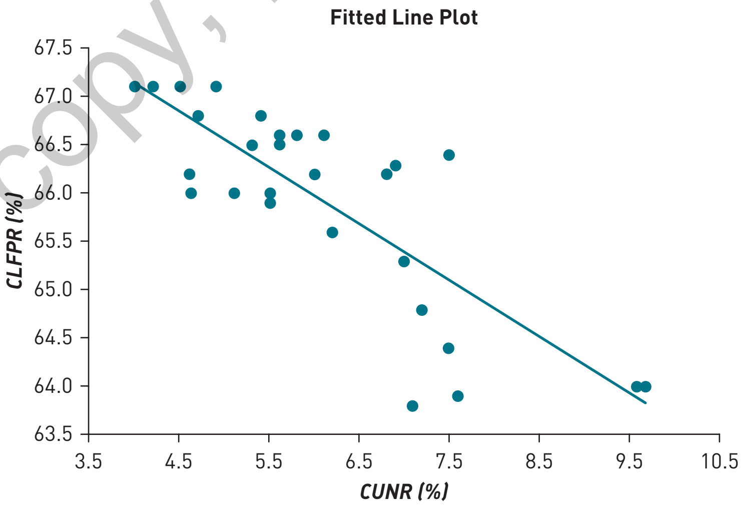
Unlike physical sciences, most data collected in economics (e.g., GDP, money supply, Dow Jones index, car sales) are nonexperimental in that the data-collecting agency (e.g., government) may not have any direct control over the data. Thus, the data on labor force participation and unemployment are based on the information provided to the government by participants in the labor market. In a sense, the government is a passive collector of these data and may not be aware of the added- or discouraged-worker hypotheses, or any other hypothesis, for that matter. Therefore, the collected data may be the result of several factors affecting the labor force participation decision made by the individual person. That is, the same data may be compatible with more than one theory.

3. Specifying the Mathematical Model of Labor Force Participation

To see how *CLFPR* behaves in relation to *CUNR*, the first thing we should do is plot the data for these variables in a **scatter diagram,** or **scattergram,** as shown in Figure 1-1.

The scattergram shows that *CLFPR* and *CUNR* are inversely related, perhaps suggesting that, on balance, the discouraged-worker effect is stronger than the added-worker effect.[4] As a *first approximation,* we can draw a straight line through the scatter

FIGURE 1-1 ⬢ Regression plot for civilian labor force participation rate (%) and civilian unemployment rate (%)



[4] On this, see Shelly Lundberg, "The Added Worker Effect," *Journal of Labor Economics,* vol. 3, January 1985, pp. 11–37.

points and write the relationship between *CLFPR* and *CUNR* by the following simple mathematical model:

$$CLFPR = B_1 + B_2\ CUNR \qquad \textbf{(1.1)}$$

Equation (1.1) states that *CLFPR* is *linearly* related to *CUNR*. B_1 and B_2 are known as the **parameters** of the linear function.[5] B_1 is also known as the **intercept;** it gives the value of *CLFPR* when *CUNR* is zero.[6] B_2 is known as the **slope.** The *slope measures the rate of change in CLFPR for a unit change in CUNR* or, more generally, the rate of change in the value of the variable on the left-hand side of the equation for a unit change in the value of the variable on the right-hand side. The slope coefficient B_2 can be positive (if the added-worker effect dominates the discouraged-worker effect) or negative (if the discouraged-worker effect dominates the added-worker effect). Figure 1-1 suggests that in the present case, it is negative.

4. Specifying the Statistical, or Econometric, Model of Labor Force Participation

The purely mathematical model of the relationship between *CLFPR* and *CUNR* given in Equation (1.1), although of prime interest to the mathematical economist, is of limited appeal to the econometrician, for such a model assumes an *exact,* or *deterministic, relationship* between the two variables; that is, for a given *CUNR,* there is a unique value of *CLFPR.* In reality, one rarely finds such neat relationships between economic variables. Most often, the relationships are *inexact,* or *statistical,* in nature.

This is seen clearly from the scattergram given in Figure 1-1. Although the two variables are inversely related, the relationship between them is not perfectly or exactly linear, for if we draw a straight line through the 28 data points, not all the data points will lie exactly on that straight line. Recall that to draw a straight line, we need only two points.[7] Why don't the 28 data points lie exactly on the straight line specified by the mathematical model, Equation (1.1)? Remember that our data on labor force and unemployment are nonexperimentally collected. Therefore, as noted earlier, besides the added- and discouraged-worker hypotheses, there may be other forces affecting labor force participation decisions. As a result, the observed relationship between *CLFPR* and *CUNR* is likely to be imprecise.

[5] Broadly speaking, a parameter is an unknown quantity that may vary over a certain set of values. In statistics, a probability distribution function (PDF) of a random variable is often characterized by its parameters, such as its mean and variance. This topic is discussed in greater detail in Appendixes A and B.

[6] In Chapter 2, we give a more precise interpretation of the intercept in the context of regression analysis.

[7] We even tried to fit a parabola to the scatter points given in Figure 1-1, but the results were not materially different from the linear specification.

Let us allow for the influence of all other variables affecting *CLFPR* in a catchall variable *u* and write Equation (1.2) as follows:

$$CLFPR = B_1 + B_2 CUNR + u \tag{1.2}$$

where *u* represents the **random error term,** or simply the **error term.**[8] We let *u* represent all those forces (besides *CUNR*) that affect *CLFPR* but are not explicitly introduced in the model, as well as purely random forces. As we will see in Part II, the error term distinguishes econometrics from purely mathematical economics.

Equation (1.2) is an example of a *statistical,* or *empirical* or *econometric, model.* More precisely, it is an example of what is known as a **linear regression model,** which is a prime subject of this book. In such a model, the variable appearing on the left-hand side of the equation is called the **dependent variable,** and the variable on the right-hand side is called the **independent,** or **explanatory, variable.** In linear regression analysis, our primary objective is to explain the behavior of one variable (the dependent variable) in relation to the behavior of one or more other variables (the explanatory variables), allowing for the fact that the relationship between them is inexact.

Notice that the econometric model, Equation (1.2), is derived from the mathematical model, Equation (1.1), which shows that mathematical economics and econometrics are mutually complementary disciplines. This is clearly reflected in the definition of econometrics given at the outset.

Before proceeding further, a warning regarding **causation** is in order. In the regression model, Equation (1.2), we have stated that *CLFPR* is the dependent variable and *CUNR* is the independent, or explanatory, variable. Does that mean that the two variables are *causally* related; that is, is *CUNR* the cause and *CLFPR* the effect? In other words, does regression imply causation? Not necessarily. As Kendall and Stuart note, "A statistical relationship, however strong and however suggestive, can never establish causal connection: our ideas of causation must come from outside statistics, ultimately from some theory or other."[9] In our example, it is up to economic theory (e.g., the discouraged-worker hypothesis) to establish the cause-and-effect relationship, if any, between the dependent and explanatory variables. If causality cannot be established, it is better to call the relationship, Equation (1.2), a *predictive relationship:* Given *CUNR,* can we predict *CLFPR?*

[8] In statistical lingo, the random error term is known as the stochastic error term.

[9] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics,* Charles Griffin, New York, 1961, vol. 2, chap. 26, p. 279.

5. Estimating the Parameters of the Chosen Econometric Model

Given the data on *CLFPR* and *CUNR,* such as that in Table 1-1, how do we estimate the parameters of the model, Equation (1.2), namely, B_1 and B_2? That is, how do we find the numerical values (i.e., **estimates**) of these **parameters?** This will be the focus of our attention in Part II, where we develop the appropriate methods of computation, especially the method of *ordinary least squares (OLS).* Using OLS and the data given in Table 1-1, we obtained the following results:

$$\widehat{CLFPR} = 69.4620 - 0.5814CUNR \qquad \textbf{(1.3)}$$

Note that we have put the symbol Λ on *CLFPR* (read as "CLFPR hat") to remind us that Equation (1.3) is an *estimate* of Equation (1.2). The estimated regression line is shown in Figure 1-1, along with the actual data points.

As Equation (1.3) shows, the estimated value of B_1 is ≈ 69.5 and that of B_2 is ≈ −0.58, where the symbol ≈ means approximately. Thus, if the unemployment rate goes up by one unit (i.e., one percentage point), *ceteris paribus, CLFPR* is expected to decrease *on the average* by about 0.58 percentage points; that is, as economic conditions worsen, on average, there is a net decrease in the labor force participation rate of about 0.58 percentage points, perhaps suggesting that the discouraged-worker effect dominates. We say "on the average" because the presence of the error term *u,* as noted earlier, is likely to make the relationship somewhat imprecise. This is vividly seen in Figure 1-1, where the points not on the estimated regression line are the actual participation rates and the (vertical) distance between them and the points on the regression line are the estimated *u*s. As we will see in Chapter 2, the estimated *u*s are called *residuals.* In short, the estimated regression line, Equation (1.3), gives the relationship between *average CLFPR* and *CUNR*, that is, on average, how *CLFPR* responds to a unit change in *CUNR.* The value of about 69.5 suggests that the average value of *CLFPR* will be about 69.5% if the *CUNR* is zero; that is, about 69.5% of the civilian working-age population will participate in the labor force if there is full employment (i.e., zero unemployment).[10]

6. Checking for Model Adequacy: Model Specification Testing

How adequate is our model, Equation (1.3)? It is true that a person will take into account labor market conditions as measured by, say, the unemployment rate before entering the labor market. For example, in 1982 (a recession year), the civilian unemployment rate was about 9.7%. Compared to that, in 2001, it was only 4.7%. A person

[10] This is, however, a mechanical interpretation of the intercept. We will see in Chapter 2 how to interpret the intercept term meaningfully in a given context.

is more likely to be discouraged from entering the labor market when the unemployment rate is more than 9% than when it is 5%. But other factors also enter into labor force participation decisions. For example, hourly wages, or earnings, prevailing in the labor market also will be an important decision variable. In the short run at least, a higher wage may attract more workers to the labor market, other things remaining the same (*ceteris paribus*). To see its importance, in Table 1-1, we have also given data on real average hourly earnings (AHE82), where real earnings are measured in 1982 dollars. To take into account the influence of AHE82, we now consider the following model:

$$\widehat{CLFPR} = B_1 + B_2 CUNR + B_3\, AHE82 + u \qquad \textbf{(1.4)}$$

Equation (1.4) is an example of a *multiple linear regression model,* in contrast to Equation (1.2), which is an example of a *simple* (*two-variable* or *bivariate*) *linear regression model.* In the two-variable model, there is a single explanatory variable, whereas in a multiple regression, there are several, or multiple, explanatory variables. Notice that in the multiple regression, Equation (1.4), we also have included the error term, *u,* for no matter how many explanatory variables one introduces in the model, one cannot fully explain the behavior of the dependent variable. How many variables one introduces in the multiple regression is a decision that the researcher will have to make in a given situation. Of course, the underlying economic theory will often tell what these variables might be. However, keep in mind the warning given earlier that regression does not mean causation; the relevant theory must determine whether one or more explanatory variables are related to the dependent variable.

How do we estimate the parameters of the multiple regression, Equation (1.4)? We cover this topic in Chapter 4, after we discuss the two-variable model in Chapters 2 and 3. We consider the two-variable case first because it is the building block of the multiple regression model. As we shall see in Chapter 4, the multiple regression model is in many ways a straightforward extension of the two-variable model.

For our illustrative example, the empirical counterpart of Equation (1.4) is as follows (these results are based on OLS):

$$CLFPR = 81.2267 - 0.6384 CUNR - 1.4449 AHE82 \qquad \textbf{(1.5)}$$

These results are interesting because both the slope coefficients are negative. The negative coefficient of *CUNR* suggests that, *ceteris paribus* (i.e., holding the influence of *AHE82* constant), a one-percentage-point increase in the unemployment rate leads, on average, to about a 0.64-percentage-point decrease in *CLFPR,* perhaps once again supporting the discouraged-worker hypothesis. On the other hand, holding the influence of *CUNR* constant, an increase in real average hourly earnings of one dollar, on average, leads to about a 1.44-percentage-point decline in *CLFPR*.[11] Does the negative coefficient for *AHE82* make economic sense? Would one not expect a positive coefficient—the higher the hourly earnings, the higher the attraction of the labor market? However, one could justify the negative coefficient by recalling the twin concepts of microeconomics, namely, the *income effect* and the *substitution effect*.[12]

Which model do we choose, Equation (1.3) or Equation (1.5)? Since Equation (1.5) *encompasses* Equation (1.3) and adds an additional dimension (earnings) to the analysis, we may choose Equation (1.5). After all, Equation (1.2) was based implicitly on the assumption that variables other than the unemployment rate were held constant. But where do we stop? For example, labor force participation may also depend on family wealth, number of children under age 6 (this is especially critical for married women thinking of joining the labor market), availability of daycare centers for young children, religious beliefs, availability of welfare benefits, unemployment insurance, and so on. Even if data on these variables are available, we may not want to introduce them all in the model because the purpose of developing an econometric model is not to capture total reality but just its salient features. If we decide to include every conceivable variable in the regression model, the model will be so unwieldy that it will be of little practical use. The model ultimately chosen should be a reasonably good replica of the underlying reality, but keeping in mind the **principle of parsimony** or **Ockham's razor.** William Ockham (1285–1349), an English philosopher, held that complicated explanation should not be accepted without good reason, or as he put it, "It is vain to do with more what can be done with less." In Chapter 7, we will discuss this question further and find out how one can go about developing a model.

7. Testing Hypotheses Derived From the Model

Having finally settled on a model, we may want to perform **hypothesis testing.** That is, we may want to find out whether the estimated model makes economic sense and whether the results obtained conform with the underlying economic theory. For example, the discouraged-worker hypothesis postulates a negative relationship between labor force participation and the unemployment rate. Is this hypothesis borne out by our results? Our statistical results seem to be in conformity with this hypothesis because the estimated coefficient of *CUNR* is negative.

[11] As we will discuss in Chapter 4, the coefficients of CUNR and AHE82 given in Equation (1.5) are known as *partial regression coefficients*. In that chapter, we will discuss the precise meaning of partial regression coefficients.

[12] Consult any standard textbook on microeconomics. One intuitive justification of this result is as follows. Suppose both spouses are in the labor force and the earnings of one spouse rise substantially. This may prompt the other spouse to withdraw from the labor force without substantially affecting the family income.

However, hypothesis testing can be complicated. In our illustrative example, suppose someone told us that in a prior study, the coefficient of *CUNR* was found to be about –1. Are our results in agreement? If we rely on the model, Equation (1.3), we might get one answer, but if we rely on Equation (1.5), we might get another answer. How do we resolve this question? Although we will develop the necessary tools to answer such questions, we should keep in mind that the answer to a particular hypothesis may depend on the model we finally choose.

The point worth remembering is that in regression analysis, we may be interested not only in estimating the parameters of the regression model but also in testing certain hypotheses suggested by economic theory and/or prior empirical experience.

Although the basic principles of hypothesis testing are covered in a basic course in statistics, Appendix D discusses this topic at some length for the benefit of the reader as a refresher course.

8. Using the Model for Prediction or Forecasting

Having gone through this multistage procedure, you can legitimately ask the following question: What do we do with the estimated model, such as Equation (1.5)? Quite naturally, we would like to use it for **prediction,** or **forecasting.** For instance, suppose we have 2008 data on the *CUNR* and *AHE82.* Assume these values are 6.0 and 10, respectively. If we put these values in Equation (1.5), we obtain 62.9473% as the predicted value of *CLFPR* for 2008. That is, if the unemployment rate in 2008 were 6.0% and the real hourly earnings were $10, the civilian labor force participation rate for 2008 would be about 63%. Of course, when data on *CLFPR* for 2008 actually become

TABLE 1-3 Summary of the Steps Involved in Econometric Analysis

Step	Example
1. Statement of theory	The added/discouraged-worker hypothesis
2. Collection of data	Table 1-1
3. Mathematical model of theory	$CLFPR = B_1 + B_2 CUNR$
4. Econometric model of theory	$CLFPR = B_1 + B_2 CUNR + u$
5. Parameter estimation	$CLFPR = 69.462 - 0.5814 CUNR$
6. Model adequacy check	$CLFPR = 81.3 - 0.638 CUNR - 1.445 AHE82$
7. Hypothesis test	$B_2 < 0$ or $B_2 > 0$
8. Prediction	What is *CLFPR*, given values of *CUNR* and *AHE82?*

available, we can compare the predicted value with the actual value (see Problem 1.10). The discrepancy between the two will represent the *prediction error*. Naturally, we would like to keep the prediction error as small as possible.

Although we examined econometric methodology using an example from labor economics, we should point out that a similar procedure can be employed to analyze quantitative relationships between variables in any field of knowledge. As a matter of fact, regression analysis has been used in politics, international relations, psychology, sociology, meteorology, and many other disciplines. As an example, see Problem 1.9.

1.4 THE ROAD AHEAD

Now that we have provided a glimpse of the nature and scope of econometrics, let us see what lies ahead. The book is divided into four parts.

Part I introduces the reader to the bread-and-butter tool of econometrics, namely, the *classical linear regression model* (*CLRM*). A thorough understanding of CLRM is a must in order to follow research in the general areas of economics and business.

Part II considers the practical aspects of regression analysis and discusses a variety of problems that the practitioner will have to tackle when one or more assumptions of the CLRM do not hold.

Part III discusses two comparatively advanced topics, time-series econometrics and panel data regression models.

Part IV, consisting of Appendixes A, B, C, and D, reviews the basics of probability and statistics for the benefit of those readers whose knowledge of statistics has become rusty. The reader should have some previous background in introductory statistics.

This book keeps the needs of the beginner in mind. The discussion of most topics is straightforward and unencumbered with mathematical proofs, derivations, and so on.[13] I firmly believe that the apparently forbidding subject of econometrics can be taught to beginners in such a way that they can see the value of the subject without getting bogged down in mathematical and statistical minutiae. The student should keep in mind that an introductory econometrics course is just like the introductory statistics course he or she has already taken. As in statistics, econometrics is primarily about estimation and hypothesis testing. What is different, and generally much more interesting and useful, is that the parameters being estimated or tested are not

[13] Some of the proofs and derivations are presented in our *Basic Econometrics,* 5th ed., McGraw-Hill, New York, 2009. A more mathematical treatment is given in Damodar N. Gujarati, *Linear Regression: A Mathematical Introduction,* SAGE, Los Angeles, 2018.

just means and variances but relationships between variables, which is what much of economics and other social sciences is all about.

A final word: The availability of comparatively inexpensive computer software packages has now made econometrics readily accessible to beginners. In this book, we will largely use four software packages: EViews, Excel, STATA, and MINITAB. These packages are readily available and widely used. Once students get used to such packages, they will soon realize that learning econometrics is really great fun, and they will have a better appreciation of the much maligned "dismal" science of economics.

KEY TERMS AND CONCEPTS

The key terms and concepts introduced in this chapter, and page numbers where they are referenced, are as follows:

Econometrics 1
Mathematical economics 2
Discouraged-worker hypothesis (effect) 4
Added-worker hypothesis (effect) 4
Time-series data: Quantitative and qualitative 5
High-frequency data 5
Flash trading 5
Autocorrelation 5
Stationary 5
Cross-sectional data 5
Heterogeneity 6
Size or scale effect 6
Pooled data 6
Panel (or longitudinal or micropanel data) 6
Scatter diagram (scattergram) 8
Parameters: Intercept and slopes 9
Random error term (error term) 10
Linear regression model: Dependent variable, independent (or explanatory) variable 10
Causation 10
Parameter estimates 11
Principle of parsimony or Ockham's razor 13
Hypothesis testing 16
Prediction (forecasting) 16

QUESTIONS

1.1. Suppose a local government decides to increase the tax rate on residential properties under its jurisdiction. What will be the effect of this on the prices of residential houses? Follow the eight-step procedure discussed in the text to answer this question.

1.2. How do you perceive the role of econometrics in decision making in business and government?

1.3. Suppose you are an economic adviser to the chairman of the Federal Reserve Board (the Fed), and he asks you whether it is advisable to increase the money supply to bolster the economy. What factors would you take into account in your advice? How would you use econometrics in your advice?

1.4. To reduce the dependence on foreign oil supplies, the government is thinking of increasing the federal taxes on gasoline. Suppose the Ford Motor Company has hired you to assess the impact of the tax increase on the demand for its cars. How would you go about advising the company?

1.5. President Joe Biden plans to propose to the U.S. Congress an infrastructure investment plan (highways, bridges, tunnels, etc.) at a cost of about $2 trillion. To pay for this, he also plans to increase the tax rate on high-income earners as well as private corporations, although the details are yet to be worked out. How would you design an econometric study to assess the economic consequences, both short term and long term, of his proposal?

PROBLEMS

1.6. **Table 1-4** on the book's website gives monthly data on the closing prices of the Dow Jones Industrial Average and the Standard & Poor's 500 stock market indexes. The data are from Yahoo Finance's historical stock quotations page.

- **a.** Plot these data with time on the horizontal axis and the two variables on the vertical axis. If you prefer, you may use a separate figure for each variable.
- **b.** What relationships do you expect to find between the two indexes? Why?
- **c.** For each variable, "eyeball" a regression line from the scattergram.
- **d.** Obtain monthly data for the two variables for the period from January 2012 to December 2020 and repeat questions **a, b,** and **c** and find out if there are any changes in the results. If so, what might account for the change?

1.7. **Table 1-5** on the book's website gives data on the exchange rate between the U.K. pound and the U.S. dollar (number of U.K. pounds per U.S. dollar), as well as the consumer price indexes in the two countries for the period 1985–2007.

- **a.** Plot the exchange rate (ER) and the two consumer price indexes against time, measured in years.
- **b.** Divide the U.S. CPI by the U.K. CPI and call it the relative price ratio (RPR).
- **c.** Plot ER against RPR.
- **d.** Visually sketch a regression line through the scatter points.
- **e.** Update the data in Table 1-5 to year 2020. Repeat questions **a, b, c,** and **d** and find out if there is any changes in the results. What accounts for the change, if any, in the results?

1.8. **Table 1-6** on the textbook website contains data on 1,247 cars for 2008.[14] To find out if there is there a relationship between a car's MPG (miles per gallon) and the number of cylinders it has:

- **a.** Create a scatterplot of the combined MPG for the vehicles based on the number of cylinders.
- **b.** Sketch a line that seems to fit the data.
- **c.** What type of relationship is indicated by the plot?

[14] Data were collected from the U.S. Department of Energy website at http://www.fueleconomy.gov/.

1.9. **Table 1-7** on the book's website gives data on Corruption Perception Index and GDP per worker.

a. Plot Corruption Perception Index against GDP per worker.

b. A priori, what kind of relationship do you expect between the two variables?

c. Does the scattergram suggest that the relationship between the two variables is linear (i.e., a straight line)? If so, sketch the regression line.

1.10. **Table 1-2** on the website updates the data given in Table 1-1 for the years 2001–2016. For the years 2001–2007, the CLFR and CUNR figures are the same as those shown in Table 1-1. However, the AHE82 figures differ in the two periods. As pointed out in the text, the differences are usually due to data revisions.

a. Plot CLFR and CUNR as in Figure 1-1. What difference due you see in the two scattergrams?

b. Is the relationship between the variables linear as in Figure 1-1? If so, visually sketch a regression line through the scatterplot.

c. Is there a "break" in the data in the sense that after a certain date, the relationship between the two variables has changed? Can you spot that break point?

d. Based on the data in Table 1-2, the regression results corresponding to Equation (1.3) are as follows:

$$CLFR\ = 66.5245 - 0.2465CUNR$$

How does t his regression differ from the one shown in Equation (1.3)? What may be the reason for the difference?

e. The regression results corresponding to Equation (1.5) using the data in Table 1-2 are as follows:

$$CL\hat{F}PR = 1121761 + 0.0150CUNR - 5.4385AHE82$$

How does this regression differ from the one shown in Equation (1.5)? What might explain the difference between the two regression results?

Note: The full results of the preceding two regressions will be discussed in Chapter 3 after we discuss the theory behind regression analysis.

1.11. **Table 1-8** on the book's website gives quarterly data on real personal consumption expenditure (RPCE) and real personal disposable (after-tax) income (RPDI) for the years 2014–2019.

a. Plot RPCE and RPDI on the same graph. What is your impression about the two time series?

b. Graph RPCE against RPDI. What does the scattergram show?

c. Visually sketch a regression line through the scatter points. What does it show?

d. Save the data for further analysis in subsequent chapters.

1.12. Based on the data for 1947–2002, Kellsted and Whitten obtained the following regression:[15]

$$M_t = 74.00 - 2.71GDP_t$$

where M = percentage of households in which a married couple is present and GDP = gross domestic product.

a. Does this result make sense?

b. How would you interpret the regression?

c. Is there a cause-and-effect relationship between the two variables?

d. The regression results give above may be an example of what is called spurious or nonsense regression. We may have more to say about it in a later chapter.

1.13. **Table 1.9** on the book's website gives data on the following variables for 99 countries

[15] Paul M. Kellstedt and Guy D. Whitten, *The Fundamentals of Political Science Research,* Cambridge University Press, 2nd ed., New York, 2013, p. 262.

obtained from the *Human Development Report* for 1994.

LifeExp = 1992 life expectancy at birth

TV = Televisions per 100 people

PopDoc = Population per doctor (1990)

GDP = real GDP per person adjusted for PPP (purchasing power parity)

a. Plot life expectancy against each of the other variables in separate graphs.

b. A priori, what do you expect the relationship is between LifeExp and each of the other variables: positive, negative, or no relationship?

SUGGESTIONS FOR FURTHER READING

"The Usefulness of Applied Econometrics to the Policy Maker," Address by R. Frances, President, Federal Bank of St. Louis, at the National Association of Business Economist Seminar, Chicago, Illinois, April 4, 1973, Federal Bank of St. Louis, May 1973.

"What Is Econometrics?" International Monetary Fund, Finance and Development, December 2011, vol. 48, No. 4 (https://www.imf.org/extenal/pubs/ft/famd/2011/12/basics.htm).

On corruption, read https://ourworldindata.org/corruption.

APPENDIX 1A: Economic Data on the World Wide Web[16]

Economic Statistics Briefing Room: An excellent source of data on output, income, employment, unemployment, earnings, production and business activity, prices and money, credits and security markets, and international statistics.

http://www.whitehouse.gov/fsbr/esbr.htm

Federal Reserve System Beige Book: Gives a summary of current economic conditions by the Federal Reserve District. There are 12 Federal Reserve Districts.

www.federalreserve.gov/FOMC/BeigeBook/2008

National Bureau of Economic Research (NBER) Home Page: This highly regarded private economic research institute has extensive data on asset prices, labor, productivity, money supply, business cycle indicators, and so on. NBER has many links to other websites.

http://www.nber.org

Panel Study: Provides data on longitudinal survey of representative sample of U.S. individuals and families. These data have been collected annually since 1968.

http://www.umich.edu/-psid

The Federal Web Locator: Provides information on almost every sector of the federal government; has international links.

www.lib.auburn.edu/madd/docs/fedloc.html

WebEC: Resources in Economics: A most comprehensive library of economic facts and figures.

www.helsinki.fi/WebEc

American Stock Exchange: Information on some 700 companies listed on the second largest stock market.

[16] It should be noted that this list is by no means exhaustive. The sources listed here are updated continually.

http://www.amex.com/

Bureau of Economic Analysis (BEA) Home Page: This agency of the U.S. Department of Commerce, which publishes the *Survey of Current Business,* is an excellent source of data on all kinds of economic activities.

www.bea.gov

Business Cycle Indicators: You will find data on about 256 economic time series.

http://www.globalexposure.com/bci.html

CIA Publication: You will find the *World Fact Book* (annual).

www.cia.gov/library/publications

Energy Information Administration (Department of Energy [DOE]): Economic information and data on each fuel category.

http://www.eia.doe.gov/

FRED Database: Federal Reserve Bank of St. Louis publishes historical economic and social data, which include interest rates, monetary and business indicators, exchange rates, and so on.

http://www.stls.frb.org/fred/

International Trade Administration: Offers many web links to trade statistics, cross-country programs, and so on.

http://www.ita.doc.gov/

STAT-USA Databases: The National Trade Data Bank provides the most comprehensive source of international trade data and export promotion information. It also contains extensive data on demographic, political, and socioeconomic conditions for several countries.

http://www.stat-usa.gov/

Bureau of Labor Statistics: The home page contains data related to various aspects of employment, unemployment, and earnings and provides links to other statistical websites.

http://stats.bls.gov

U.S. Census Bureau Home Page: Prime source of social, demographic, and economic data on income, employment, income distribution, and poverty.

http://www.census.gov/

General Social Survey: Annual personal interview survey data on U.S. households that began in 1972. More than 35,000 have responded to some 2,500 different questions covering a variety of data.

www.norc.org/GCS+Website

Institute for Research on Poverty: Data collected by nonpartisan and nonprofit university-based research center on a variety of questions relating to poverty and social inequality.

http://www.ssc.wisc.edu/irp/

Social Security Administration: The official website of the Social Security Administration with a variety of data.

http://www.ssa.gov

Federal Deposit Insurance Corporation, Bank Data and Statistics

http://www.fdic.gov/bank/statistical/

Federal Reserve Board, Economic Research and Data

http://www.federalreserve.gov/econresdata

U.S. Census Bureau, Home Page

http://www.census.gov

U.S. Department of Energy, Energy Information Administration

www.eia.doe.gov/overview_hd.html

U.S. Department of Health and Human Services, National Center for Health Statistics

http://www.cdc.gov/nchs

U.S. Department of Housing and Urban Development, Data Sets

http://www.huduser.org/datasets/pdrdatas.html

U.S. Department of Labor, Bureau of Labor Statistics

http://www.bls.gov

U.S. Department of Transportation, TranStats

http://www.transtats.bts.gov

U.S. Department of the Treasury, Internal Revenue Service, Tax Statistics

http://www.irs.gov/taxstats

Rockefeller Institute of Government, State and Local Fiscal Data

www.rockinst.org/research/sl_finance

American Economic Association, Resources for Economists

http://www.rfe.org

American Statistical Association, Business and Economic Statistics

www.amstat.org/publications/jbes

American Statistical Association, Statistics in Sports

http://www.amstat.org/sections/sis/

European Central Bank, Statistics

http://www.ecb.int/stats

World Bank, Data and Statistics

http://www.worldbank.org/data

International Monetary Fund, Statistical Topics

http://www.imf.org/external/np/sta/

Penn World Tables

http://pwt.econ.upenn.edu

Current Population Survey

http://www.bls.census.gov/cps/

Consumer Expenditure Survey

http://www.bls.gov/cex/

Survey of Consumer Finances

http://www.federalreserve.gov/pubs/oss/

City and County Data Book

http://www.census.gov/statab/www/ccdb.html

Panel Study of Income Dynamics

http://psidonline.isr.umich.edu

National Longitudinal Surveys

http://www.bls.gov/nls/

National Association of Home Builders, Economic and Housing Data

http://www.nahb.org/page.aspx/category/sectionID=113

National Science Foundation, Division of Science Resources Statistics

http://www.nsf.gov/sbe/srs/

Economic Report of the President

http://www.gpoaccess.gov/eop/

Various Economic Data Sets

http://www.economy.com/freelunch/

The Economist Market Indicators

http://www.economist.com/markets/indicators

Statistical Resources on the Military

http://www.lib.umich.edu/govdocs/stmil.html

World Economic Indicators

http://devdata.worldbank.org/

Economic Time Series Data

http://www.economagic.com/

United Nations Population Division's Annual Estimates and Projections

http://unstats.un.org/unsd/default.htm

United Nations Statistics Division-UNdata

http://data.un.org/Default.aspx

World Bank Data

http://databank.worldbank.org/