

The Nature of Econometrics and Economic Data

Chapter 1 discusses the scope of econometrics and raises general issues that arise in the application of econometric methods. Section 1-1 provides a brief discussion about the purpose and scope of econometrics and how it fits into economic analysis. Section 1-2 provides examples of how one can start with an economic theory and build a model that can be estimated using data. Section 1-3 examines the kinds of data sets that are used in business, economics, and other social sciences. Section 1-4 provides an intuitive discussion of the difficulties associated with inferring causality in the social sciences.

1-1 What Is Econometrics?

Imagine that you are hired by your state government to evaluate the effectiveness of a publicly funded job training program. Suppose this program teaches workers various ways to use computers in the manufacturing process. The 20-week program offers courses during nonworking hours. Any hourly manufacturing worker may participate, and enrollment in all or part of the program is voluntary. You are to determine what, if any, effect the training program has on each worker's subsequent hourly wage.

Now, suppose you work for an investment bank. You are to study the returns on different investment strategies involving short-term U.S. treasury bills to decide whether they comply with implied economic theories.

The task of answering such questions may seem daunting at first. At this point, you may only have a vague idea of the kind of data you would need to collect. By the end of this introductory econometrics course, you should know how to use econometric methods to formally evaluate a job training program or to test a simple economic theory.

Econometrics is based upon the development of statistical methods for estimating economic relationships, testing economic theories, and evaluating and implementing government and business policy. A common application of econometrics is the forecasting of such important macroeconomic variables as interest rates, inflation rates, and gross domestic product (GDP). Whereas forecasts of economic indicators are highly visible and often widely published, econometric methods can be used in economic areas that have nothing to do with macroeconomic forecasting. For example, we will study the effects of political campaign expenditures on voting outcomes. We will consider the effect of school spending on student performance in the field of education. In addition, we will learn how to use econometric methods for forecasting economic time series.

Econometrics has evolved as a separate discipline from mathematical statistics because the former focuses on the problems inherent in collecting and analyzing nonexperimental economic data. **Nonexperimental data** are not accumulated through controlled experiments on individuals, firms, or segments of the economy. (Nonexperimental data are sometimes called **observational data**, or **retrospective data**, to emphasize the fact that the researcher is a passive collector of the data.) **Experimental data** are often collected in laboratory environments in the natural sciences, but they are more difficult to obtain in the social sciences. Although some social experiments can be devised, it is often impossible, prohibitively expensive, or morally repugnant to conduct the kinds of controlled experiments that would be needed to address economic issues. We give some specific examples of the differences between experimental and nonexperimental data in Section 1-4.

Naturally, econometricians have borrowed from mathematical statisticians whenever possible. The method of multiple regression analysis is the mainstay in both fields, but its focus and interpretation can differ markedly. In addition, economists have devised new techniques to deal with the complexities of economic data and to test the predictions of economic theories.

1-2 Steps in Empirical Economic Analysis

Econometric methods are relevant in virtually every branch of applied economics. They come into play either when we have an economic theory to test or when we have a relationship in mind that has some importance for business decisions or policy analysis. An **empirical analysis** uses data to test a theory or to estimate a relationship.

How does one go about structuring an empirical economic analysis? It may seem obvious, but it is worth emphasizing that the first step in any empirical analysis is the careful formulation of the question of interest. The question might deal with testing a certain aspect of an economic theory, or it might pertain to testing the effects of a government policy. In principle, econometric methods can be used to answer a wide range of questions.

In some cases, especially those that involve the testing of economic theories, a formal **economic model** is constructed. An economic model consists of mathematical equations that describe various relationships. Economists are well known for their building of models to describe a vast array of behaviors. For example, in intermediate microeconomics, individual consumption decisions, subject to a budget constraint, are described by mathematical models. The basic premise underlying these models is *utility maximization*. The assumption that individuals make choices to maximize their well-being, subject to resource constraints, gives us a very powerful framework for creating tractable economic models and making clear predictions. In the context of consumption decisions, utility maximization leads to a set of *demand equations*. In a demand equation, the quantity demanded of each commodity depends on the price of the goods, the price of substitute and complementary goods, the consumer's income, and the individual's characteristics that affect taste. These equations can form the basis of an econometric analysis of consumer demand.

Economists have used basic economic tools, such as the utility maximization framework, to explain behaviors that at first glance may appear to be noneconomic in nature. A classic example is Becker's (1968) economic model of criminal behavior.

EXAMPLE 1.1 Economic Model of Crime

In a seminal article, Nobel Prize winner Gary Becker postulated a utility maximization framework to describe an individual's participation in crime. Certain crimes have clear economic rewards, but most criminal behaviors have costs. The opportunity costs of crime prevent the criminal from participating in other activities such as legal employment. In addition, there are costs associated with the possibility of being caught and then, if convicted, the costs associated with incarceration. From Becker's perspective, the decision to undertake illegal activity is one of resource allocation, with the benefits and costs of competing activities taken into account.

Under general assumptions, we can derive an equation describing the amount of time spent in criminal activity as a function of various factors. We might represent such a function as

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7), \quad [1.1]$$

where

- y = hours spent in criminal activities,
- x_1 = "wage" for an hour spent in criminal activity,
- x_2 = hourly wage in legal employment,
- x_3 = income other than from crime or employment,
- x_4 = probability of getting caught,
- x_5 = probability of being convicted if caught,
- x_6 = expected sentence if convicted, and
- x_7 = age.

Other factors generally affect a person's decision to participate in crime, but the list above is representative of what might result from a formal economic analysis. As is common in economic theory, we have not been specific about the function $f(\cdot)$ in (1.1). This function depends on an underlying utility function, which is rarely known. Nevertheless, we can use economic theory—or introspection—to predict the effect that each variable would have on criminal activity. This is the basis for an econometric analysis of individual criminal activity.

Formal economic modeling is sometimes the starting point for empirical analysis, but it is more common to use economic theory less formally, or even to rely entirely on intuition. You may agree that the determinants of criminal behavior appearing in equation (1.1) are reasonable based on common sense; we might arrive at such an equation directly, without starting from utility maximization. This view has some merit, although there are cases in which formal derivations provide insights that intuition can overlook.

Next is an example of an equation that we can derive through somewhat informal reasoning.

EXAMPLE 1.2 Job Training and Worker Productivity

Consider the problem posed at the beginning of Section 1-1. A labor economist would like to examine the effects of job training on worker productivity. In this case, there is little need for formal economic theory. Basic economic understanding is sufficient for realizing that factors such as education, experience, and training affect worker productivity. Also, economists are well aware that workers are paid commensurate with their productivity. This simple reasoning leads to a model such as

$$wage = f(educ, exper, training), \quad [1.2]$$

where

- $wage$ = hourly wage,
- $educ$ = years of formal education,
- $exper$ = years of workforce experience, and
- $training$ = weeks spent in job training.

Again, other factors generally affect the wage rate, but equation (1.2) captures the essence of the problem.

After we specify an economic model, we need to turn it into what we call an **econometric model**. Because we will deal with econometric models throughout this text, it is important to know how an econometric model relates to an economic model. Take equation (1.1) as an example. The form of the function $f(\cdot)$ must be specified before we can undertake an econometric analysis. A second issue concerning (1.1) is how to deal with variables that cannot reasonably be observed. For example, consider the wage that a person can earn in criminal activity. In principle, such a quantity is well defined, but it would be difficult if not impossible to observe this wage for a given individual. Even variables such as the probability of being arrested cannot realistically be obtained for a given individual, but at least we can observe relevant arrest statistics and derive a variable that approximates the probability of arrest. Many other factors affect criminal behavior that we cannot even list, let alone observe, but we must somehow account for them.

The ambiguities inherent in the economic model of crime are resolved by specifying a particular econometric model:

$$\begin{aligned} \text{crime} = & \beta_0 + \beta_1 \text{wage} + \beta_2 \text{othinc} + \beta_3 \text{freqarr} + \beta_4 \text{freqconv} \\ & + \beta_5 \text{avgsen} + \beta_6 \text{age} + u, \end{aligned} \quad [1.3]$$

where

- crime* = some measure of the frequency of criminal activity,
- wage* = the wage that can be earned in legal employment,
- othinc* = the income from other sources (assets, inheritance, and so on),
- freqarr* = the frequency of arrests for prior infractions (to approximate the probability of arrest),
- freqconv* = the frequency of conviction, and
- avgsen* = the average sentence length after conviction.

The choice of these variables is determined by the economic theory as well as data considerations. The term u contains unobserved factors, such as the wage for criminal activity, moral character, family background, and errors in measuring things like criminal activity and the probability of arrest. We could add family background variables to the model, such as number of siblings, parents' education, and so on, but we can never eliminate u entirely. In fact, dealing with this *error term* or *disturbance term* is perhaps the most important component of any econometric analysis.

The constants $\beta_0, \beta_1, \dots, \beta_6$ are the *parameters* of the econometric model, and they describe the directions and strengths of the relationship between *crime* and the factors used to determine *crime* in the model.

A complete econometric model for Example 1.2 might be

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{training} + u, \quad [1.4]$$

where the term u contains factors such as “innate ability,” quality of education, family background, and the myriad other factors that can influence a person's wage. If we are specifically concerned about the effects of job training, then β_3 is the parameter of interest.

For the most part, econometric analysis begins by specifying an econometric model, without consideration of the details of the model's creation. We generally follow this approach, largely because careful derivation of something like the economic model of crime is time consuming and can take us into some specialized and often difficult areas of economic theory. Economic reasoning will play a role in our examples, and we will merge any underlying economic theory into the econometric model specification. In the economic model of crime example, we would start with an econometric model such as (1.3) and use economic reasoning and common sense as guides for choosing the variables. Although this approach loses some of the richness of economic analysis, it is commonly and effectively applied by careful researchers.

Once an econometric model such as (1.3) or (1.4) has been specified, various *hypotheses* of interest can be stated in terms of the unknown parameters. For example, in equation (1.3), we might hypothesize that *wage*, the wage that can be earned in legal employment, has no effect on criminal behavior. In the context of this particular econometric model, the hypothesis is equivalent to $\beta_1 = 0$.

An empirical analysis, by definition, requires data. After data on the relevant variables have been collected, econometric methods are used to estimate the parameters in the econometric model and to formally test hypotheses of interest. In some cases, the econometric model is used to make predictions in either the testing of a theory or the study of a policy's impact.

Because data collection is so important in empirical work, Section 1-3 will describe the kinds of data that we are likely to encounter.

1-3 The Structure of Economic Data

Economic data sets come in a variety of types. Whereas some econometric methods can be applied with little or no modification to many different kinds of data sets, the special features of some data sets must be accounted for or should be exploited. We next describe the most important data structures encountered in applied work.

1-3a Cross-Sectional Data

A **cross-sectional data set** consists of a sample of individuals, households, firms, cities, states, countries, or a variety of other units, taken at a given point in time. Sometimes, the data on all units do not correspond to precisely the same time period. For example, several families may be surveyed during different weeks within a year. In a pure cross-sectional analysis, we would ignore any minor timing differences in collecting the data. If a set of families was surveyed during different weeks of the same year, we would still view this as a cross-sectional data set.

An important feature of cross-sectional data is that we can often assume that they have been obtained by **random sampling** from the underlying population. For example, if we obtain information on wages, education, experience, and other characteristics by randomly drawing 500 people from the working population, then we have a random sample from the population of all working people. Random sampling is the sampling scheme covered in introductory statistics courses, and it simplifies the analysis of cross-sectional data. A review of random sampling is contained in Math Refresher C.

Sometimes, random sampling is not appropriate as an assumption for analyzing cross-sectional data. For example, suppose we are interested in studying factors that influence the accumulation of family wealth. We could survey a random sample of families, but some families might refuse to report their wealth. If, for example, wealthier families are less likely to disclose their wealth, then the resulting sample on wealth is not a random sample from the population of all families. This is an illustration of a sample selection problem.

Another violation of random sampling occurs when we sample from units that are large relative to the population, particularly geographical units. The potential problem in such cases is that the population is not large enough to reasonably assume the observations are independent draws. For example, if we want to explain new business activity across states as a function of wage rates, energy prices, corporate and property tax rates, services provided, quality of the workforce, and other state characteristics, it is unlikely that business activities in states near one another are independent. It turns out that the econometric methods that we discuss do work in such situations, but they sometimes need to be refined. For the most part, we will ignore the intricacies that arise in analyzing such situations and treat these problems in a random sampling framework, even when it is not technically correct to do so.

Cross-sectional data are widely used in economics and other social sciences. In economics, the analysis of cross-sectional data is closely aligned with the applied microeconomics fields, such as labor economics, state and local public finance, industrial organization, urban economics, demography, and health economics. Data on individuals, households, firms, and cities at a given point in time are important for testing microeconomic hypotheses and evaluating economic policies.

The cross-sectional data used for econometric analysis can be represented and stored in computers. Table 1.1 contains, in abbreviated form, a cross-sectional data set on 526 working individuals

TABLE 1.1 A Cross-Sectional Data Set on Wages and Other Individual Characteristics					
obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.
.
.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

for the year 1976. (This is a subset of the data in the file WAGE1.) The variables include *wage* (in dollars per hour), *educ* (years of education), *exper* (years of potential labor force experience), *female* (an indicator for gender), and *married* (marital status). These last two variables are binary (zero-one) in nature and serve to indicate qualitative features of the individual (the person is female or not; the person is married or not). We will have much to say about binary variables in Chapter 7 and beyond.

The variable *obsno* in Table 1.1 is the observation number assigned to each person in the sample. Unlike the other variables, it is not a characteristic of the individual. All econometrics and statistics software packages assign an observation number to each data unit. Intuition should tell you that, for data such as that in Table 1.1, it does not matter which person is labeled as observation 1, which person is called observation 2, and so on. The fact that the ordering of the data does not matter for econometric analysis is a key feature of cross-sectional data sets obtained from random sampling.

Different variables sometimes correspond to different time periods in cross-sectional data sets. For example, to determine the effects of government policies on long-term economic growth, economists have studied the relationship between growth in real per capita GDP over a certain period (say, 1960 to 1985) and variables determined in part by government policy in 1960 (government consumption as a percentage of GDP and adult secondary education rates). Such a data set might be represented as in Table 1.2, which constitutes part of the data set used in the study of cross-country growth rates by De Long and Summers (1991).

The variable *gpcrgdp* represents average growth in real per capita GDP over the period 1960 to 1985. The fact that *govcons60* (government consumption as a percentage of GDP) and *second60*

TABLE 1.2 A Data Set on Economic Growth Rates and Country Characteristics				
obsno	country	gpcrgdp	govcons60	second60
1	Argentina	0.89	9	32
2	Austria	3.32	16	50
3	Belgium	2.56	13	69
4	Bolivia	1.24	18	12
.
.
.
61	Zimbabwe	2.30	17	6

(percentage of adult population with a secondary education) correspond to the year 1960, while *gpcrgdp* is the average growth over the period from 1960 to 1985, does not lead to any special problems in treating this information as a cross-sectional data set. The observations are listed alphabetically by country, but nothing about this ordering affects any subsequent analysis.

1-3b Time Series Data

A **time series data** set consists of observations on a variable or several variables over time. Examples of time series data include stock prices, money supply, consumer price index, GDP, annual homicide rates, and automobile sales figures. Because past events can influence future events and lags in behavior are prevalent in the social sciences, time is an important dimension in a time series data set. Unlike the arrangement of cross-sectional data, the chronological ordering of observations in a time series conveys potentially important information.

A key feature of time series data that makes them more difficult to analyze than cross-sectional data is that economic observations can rarely, if ever, be assumed to be independent across time. Most economic and other time series are related, often strongly related, to their recent histories. For example, knowing something about the GDP from last quarter tells us quite a bit about the likely range of the GDP during this quarter, because GDP tends to remain fairly stable from one quarter to the next. Although most econometric procedures can be used with both cross-sectional and time series data, more needs to be done in specifying econometric models for time series data before standard econometric methods can be justified. In addition, modifications and embellishments to standard econometric techniques have been developed to account for and exploit the dependent nature of economic time series and to address other issues, such as the fact that some economic variables tend to display clear trends over time.

Another feature of time series data that can require special attention is the **data frequency** at which the data are collected. In economics, the most common frequencies are daily, weekly, monthly, quarterly, and annually. Stock prices are recorded at daily intervals (excluding Saturday and Sunday). The money supply in the U.S. economy is reported weekly. Many macroeconomic series are tabulated monthly, including inflation and unemployment rates. Other macro series are recorded less frequently, such as every three months (every quarter). GDP is an important example of a quarterly series. Other time series, such as infant mortality rates for states in the United States, are available only on an annual basis.

Many weekly, monthly, and quarterly economic time series display a strong seasonal pattern, which can be an important factor in a time series analysis. For example, monthly data on housing starts differ across the months simply due to changing weather conditions. We will learn how to deal with seasonal time series in Chapter 10.

Table 1.3 contains a time series data set obtained from an article by Castillo-Freeman and Freeman (1992) on minimum wage effects in Puerto Rico. The earliest year in the data set is the first

TABLE 1.3 Minimum Wage, Unemployment, and Related Data for Puerto Rico

obsno	year	avgmin	avgcov	prunemp	prgnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
.
.
.
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

observation, and the most recent year available is the last observation. When econometric methods are used to analyze time series data, the data should be stored in chronological order.

The variable *avgmin* refers to the average minimum wage for the year, *avgcov* is the average coverage rate (the percentage of workers covered by the minimum wage law), *prunemp* is the unemployment rate, and *prgnp* is the gross national product, in millions of 1954 dollars. We will use these data later in a time series analysis of the effect of the minimum wage on employment.

1-3C Pooled Cross Sections

Some data sets have both cross-sectional and time series features. For example, suppose that two cross-sectional household surveys are taken in the United States, one in 1985 and one in 1990. In 1985, a random sample of households is surveyed for variables such as income, savings, family size, and so on. In 1990, a *new* random sample of households is taken using the same survey questions. To increase our sample size, we can form a **pooled cross section** by combining the two years.

Pooling cross sections from different years is often an effective way of analyzing the effects of a new government policy. The idea is to collect data from the years before and after a key policy change. As an example, consider the following data set on housing prices taken in 1993 and 1995, before and after a reduction in property taxes in 1994. Suppose we have data on 250 houses for 1993 and on 270 houses for 1995. One way to store such a data set is given in Table 1.4.

Observations 1 through 250 correspond to the houses sold in 1993, and observations 251 through 520 correspond to the 270 houses sold in 1995. Although the order in which we store the data turns out not to be crucial, keeping track of the year for each observation is usually very important. This is why we enter *year* as a separate variable.

A pooled cross section is analyzed much like a standard cross section, except that we often need to account for secular differences in the variables across the time. In fact, in addition to increasing the sample size, the point of a pooled cross-sectional analysis is often to see how a key relationship has changed over time.

TABLE 1.4 Pooled Cross Sections: Two Years of Housing Prices

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85,500	42	1600	3	2.0
2	1993	67,300	36	1440	3	2.5
3	1993	134,000	38	2000	4	2.5
.
.
.
250	1993	243,600	41	2600	4	3.0
251	1995	65,000	16	1250	2	1.0
252	1995	182,400	20	2200	4	2.0
253	1995	97,500	15	1540	3	2.0
.
.
.
520	1995	57,200	16	1100	2	1.5

1-3d Panel or Longitudinal Data

A **panel data** (or *longitudinal data*) set consists of a time series for *each* cross-sectional member in the data set. As an example, suppose we have wage, education, and employment history for a set of individuals followed over a 10-year period. Or we might collect information, such as investment and financial data, about the same set of firms over a five-year time period. Panel data can also be collected on geographical units. For example, we can collect data for the same set of counties in the United States on immigration flows, tax rates, wage rates, government expenditures, and so on, for the years 1980, 1985, and 1990.

The key feature of panel data that distinguishes them from a pooled cross section is that the *same* cross-sectional units (individuals, firms, or counties in the preceding examples) are followed over a given time period. The data in Table 1.4 are not considered a panel data set because the houses sold are likely to be different in 1993 and 1995; if there are any duplicates, the number is likely to be so small as to be unimportant. In contrast, Table 1.5 contains a two-year panel data set on crime and related statistics for 150 cities in the United States.

There are several interesting features in Table 1.5. First, each city has been given a number from 1 through 150. Which city we decide to call city 1, city 2, and so on, is irrelevant. As with a pure cross section, the ordering in the cross section of a panel data set does not matter. We could use the city name in place of a number, but it is often useful to have both.

A second point is that the two years of data for city 1 fill the first two rows or observations, observations 3 and 4 correspond to city 2, and so on. Because each of the 150 cities has two rows of data, any econometrics package will view this as 300 observations. This data set can be treated as a pooled cross section, where the same cities happen to show up in each year. But, as we will see in Chapter 13, we can also use the panel structure to analyze questions that cannot be answered by simply viewing this as a pooled cross section.

In organizing the observations in Table 1.5, we place the two years of data for each city adjacent to one another, with the first year coming before the second in all cases. For just about every practical purpose, this is the preferred way for ordering panel data sets. Contrast this organization with the way the pooled cross sections are stored in Table 1.4. In short, the reason for ordering panel data as in Table 1.5 is that we will need to perform data transformations for each city across the two years.

Because panel data require replication of the same units over time, panel data sets, especially those on individuals, households, and firms, are more difficult to obtain than pooled cross sections. Not surprisingly, observing the same units over time leads to several advantages over cross-sectional data or even pooled cross-sectional data. The benefit that we will focus on in this text is that having

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350,000	8.7	440
2	1	1990	8	359,200	7.2	471
3	2	1986	2	64,300	5.4	75
4	2	1990	1	65,100	5.5	75
.
.
.
297	149	1986	10	260,700	9.6	286
298	149	1990	6	245,000	9.8	334
299	150	1986	25	543,000	4.3	520
300	150	1990	32	546,200	5.2	493

multiple observations on the same units allows us to control for certain unobserved characteristics of individuals, firms, and so on. As we will see, the use of more than one observation can facilitate causal inference in situations where inferring causality would be very difficult if only a single cross section were available. A second advantage of panel data is that they often allow us to study the importance of lags in behavior or the result of decision making. This information can be significant because many economic policies can be expected to have an impact only after some time has passed.

Most books at the undergraduate level do not contain a discussion of econometric methods for panel data. However, economists now recognize that some questions are difficult, if not impossible, to answer satisfactorily without panel data. As you will see, we can make considerable progress with simple panel data analysis, a method that is not much more difficult than dealing with a standard cross-sectional data set.

1-3e A Comment on Data Structures

Part 1 of this text is concerned with the analysis of cross-sectional data, because this poses the fewest conceptual and technical difficulties. At the same time, it illustrates most of the key themes of econometric analysis. We will use the methods and insights from cross-sectional analysis in the remainder of the text.

Although the econometric analysis of time series uses many of the same tools as cross-sectional analysis, it is more complicated because of the trending, highly persistent nature of many economic time series. Examples that have been traditionally used to illustrate the manner in which econometric methods can be applied to time series data are now widely believed to be flawed. It makes little sense to use such examples initially, because this practice will only reinforce poor econometric practice. Therefore, we will postpone the treatment of time series econometrics until Part 2, when the important issues concerning trends, persistence, dynamics, and seasonality will be introduced.

In Part 3, we will treat pooled cross sections and panel data explicitly. The analysis of independently pooled cross sections and simple panel data analysis are fairly straightforward extensions of pure cross-sectional analysis. Nevertheless, we will wait until Chapter 13 to deal with these topics.

1-4 Causality, *Ceteris Paribus*, and Counterfactual Reasoning

In most tests of economic theory, and certainly for evaluating public policy, the economist's goal is to infer that one variable (such as education) has a **causal effect** on another variable (such as worker productivity). Simply finding an association between two or more variables might be suggestive, but unless causality can be established, it is rarely compelling.

The notion of **ceteris paribus**—which means “other (relevant) factors being equal”—plays an important role in causal analysis. This idea has been implicit in some of our earlier discussion, particularly Examples 1.1 and 1.2, but thus far we have not explicitly mentioned it.

You probably remember from introductory economics that most economic questions are *ceteris paribus* by nature. For example, in analyzing consumer demand, we are interested in knowing the effect of changing the price of a good on its quantity demanded, while holding all other factors—such as income, prices of other goods, and individual tastes—fixed. If other factors are not held fixed, then we cannot know the causal effect of a price change on quantity demanded.

Holding other factors fixed is critical for policy analysis as well. In the job training example (Example 1.2), we might be interested in the effect of another week of job training on wages, with all other components being equal (in particular, education and experience). If we succeed in holding all other relevant factors fixed and then find a link between job training and wages, we can conclude that job training has a causal effect on worker productivity. Although this may seem pretty simple, even at this early stage it should be clear that, except in very special cases, it will not be possible to literally hold all else equal. The key question in most empirical studies is: Have enough other factors been held fixed to make a case for causality? Rarely is an econometric study evaluated without raising this issue.

In most serious applications, the number of factors that can affect the variable of interest—such as criminal activity or wages—is immense, and the isolation of any particular variable may seem like a hopeless effort. However, we will eventually see that, when carefully applied, econometric methods can simulate a *ceteris paribus* experiment.

The notion of *ceteris paribus* also can be described through **counterfactual reasoning**, which has become an organizing theme in analyzing various interventions, such as policy changes. The idea is to imagine an economic unit, such as an individual or a firm, in two or more different states of the world. For example, consider studying the impact of a job training program on workers' earnings. For each worker in the relevant population, we can imagine what his or her subsequent earnings would be under two states of the world: having participated in the job training program and having not participated. By considering these **counterfactual outcomes** (also called *potential outcomes*), we easily “hold other factors fixed” because the counterfactual thought experiment applies to each individual separately. We can then think of causality as meaning that the outcome—in this case, labor earnings—in the two states of the world differs for at least some individuals. The fact that we will eventually observe each worker in only one state of the world raises important problems of estimation, but that is a separate issue from the issue of what we mean by causality. We formally introduce an apparatus for discussing counterfactual outcomes in Chapter 2.

At this point, we cannot yet explain how econometric methods can be used to estimate *ceteris paribus* effects, so we will consider some problems that can arise in trying to infer causality in economics. We do not use any equations in this discussion. Instead, in each example, we will discuss what other factors we would like to hold fixed, and sprinkle in some counterfactual reasoning. For each example, inferring causality becomes relatively easy if we could conduct an appropriate experiment. Thus, it is useful to describe how such an experiment might be structured, and to observe that, in most cases, obtaining experimental data is impractical. It is also helpful to think about why the available data fail to have the important features of an experimental data set.

We rely, for now, on your intuitive understanding of such terms as *random*, *independence*, and *correlation*, all of which should be familiar from an introductory probability and statistics course. (These concepts are reviewed in Math Refresher B.) We begin with an example that illustrates some of these important issues.

EXAMPLE 1.3

Effects of Fertilizer on Crop Yield

Some early econometric studies [for example, Griliches (1957)] considered the effects of new fertilizers on crop yields. Suppose the crop under consideration is soybeans. Because fertilizer amount is only one factor affecting yields—some others include rainfall, quality of land, and presence of parasites—this issue must be posed as a *ceteris paribus* question. One way to determine the causal effect of fertilizer amount on soybean yield is to conduct an experiment, which might include the following steps. Choose several one-acre plots of land. Apply different amounts of fertilizer to each plot and subsequently measure the yields; this gives us a cross-sectional data set. Then, use statistical methods (to be introduced in Chapter 2) to measure the association between yields and fertilizer amounts.

As described earlier, this may not seem like a very good experiment because we have said nothing about choosing plots of land that are identical in all respects except for the amount of fertilizer. In fact, choosing plots of land with this feature is not feasible: some of the factors, such as land quality, cannot even be fully observed. How do we know the results of this experiment can be used to measure the *ceteris paribus* effect of fertilizer? The answer depends on the specifics of how fertilizer amounts are chosen. If the levels of fertilizer are assigned to plots independently of other plot features that affect yield—that is, other characteristics of plots are completely ignored when deciding on fertilizer amounts—then we are in business. We will justify this statement in Chapter 2.

The next example is more representative of the difficulties that arise when inferring causality in applied economics.

EXAMPLE 1.4 **Measuring the Return to Education**

Labor economists and policy makers have long been interested in the “return to education.” Somewhat informally, the question is posed as follows: If a person is chosen from the population and given another year of education, by how much will his or her wage increase? As with the previous examples, this is a *ceteris paribus* question, which implies that all other factors are held fixed while another year of education is given to the person. Notice the element of counterfactual reasoning here: we can imagine the wage of each individual varying with different levels of education, that is, in different states of the world. Eventually, we obtain data on each worker in only one state of the world: the education level they actually wound up with, through perhaps a complicated process of intellectual ability, motivation for learning, parental input, and societal influences.

We can imagine a social planner designing an experiment to get at this issue, much as the agricultural researcher can design an experiment to estimate fertilizer effects. Assume, for the moment, that the social planner has the ability to assign any level of education to any person. How would this planner emulate the fertilizer experiment in Example 1.3? The planner would choose a group of people and randomly assign each person an amount of education; some people are given an eighth-grade education, some are given a high school education, some are given two years of college, and so on. Subsequently, the planner measures wages for this group of people (where we assume that each person then works in a job). The people here are like the plots in the fertilizer example, where education plays the role of fertilizer and wage rate plays the role of soybean yield. As with Example 1.3, if levels of education are assigned independently of other characteristics that affect productivity (such as experience and innate ability), then an analysis that ignores these other factors will yield useful results. Again, it will take some effort in Chapter 2 to justify this claim; for now, we state it without support.

Unlike the fertilizer-yield example, the experiment described in Example 1.4 is unfeasible. The ethical issues, not to mention the economic costs, associated with randomly determining education levels for a group of individuals are obvious. As a logistical matter, we could not give someone only an eighth-grade education if he or she already has a college degree.

Even though experimental data cannot be obtained for measuring the return to education, we can certainly collect nonexperimental data on education levels and wages for a large group by sampling randomly from the population of working people. Such data are available from a variety of surveys used in labor economics, but these data sets have a feature that makes it difficult to estimate the *ceteris paribus* return to education. People *choose* their own levels of education; therefore, education levels are probably not determined independently of all other factors affecting wage. This problem is a feature shared by most nonexperimental data sets.

One factor that affects wage is experience in the workforce. Because pursuing more education generally requires postponing entering the workforce, those with more education usually have less experience. Thus, in a nonexperimental data set on wages and education, education is likely to be negatively associated with a key variable that also affects wage. It is also believed that people with more innate ability often choose higher levels of education. Because higher ability leads to higher wages, we again have a correlation between education and a critical factor that affects wage.

The omitted factors of experience and ability in the wage example have analogs in the fertilizer example. Experience is generally easy to measure and therefore is similar to a variable such as rainfall. Ability, on the other hand, is nebulous and difficult to quantify; it is similar to land quality in the fertilizer example. As we will see throughout this text, accounting for other observed factors, such as experience, when estimating the *ceteris paribus* effect of another variable, such as education, is relatively straightforward. We will also find that accounting for inherently unobservable factors, such as ability, is much more problematic. It is fair to say that many of the advances in econometric methods have tried to deal with unobserved factors in econometric models.

One final parallel can be drawn between Examples 1.3 and 1.4. Suppose that in the fertilizer example, the fertilizer amounts were not entirely determined at random. Instead, the assistant who

chose the fertilizer levels thought it would be better to put more fertilizer on the higher-quality plots of land. (Agricultural researchers should have a rough idea about which plots of land are of better quality, even though they may not be able to fully quantify the differences.) This situation is completely analogous to the level of schooling being related to unobserved ability in Example 1.4. Because better land leads to higher yields, and more fertilizer was used on the better plots, any observed relationship between yield and fertilizer might be spurious.

Difficulty in inferring causality can also arise when studying data at fairly high levels of aggregation, as the next example on city crime rates shows.

EXAMPLE 1.5 The Effect of Law Enforcement on City Crime Levels

The issue of how best to prevent crime has been, and will probably continue to be, with us for some time. One especially important question in this regard is: Does the presence of more police officers on the street deter crime?

The *ceteris paribus* question is easy to state: If a city is randomly chosen and given, say, ten additional police officers, by how much would its crime rates fall? Closely related to this thought experiment is explicitly setting up counterfactual outcomes: For a given city, what would its crime rate be under varying sizes of the police force? Another way to state the question is: If two cities are the same in all respects, except that city A has ten more police officers than city B, by how much would the two cities' crime rates differ?

It would be virtually impossible to find pairs of communities identical in all respects except for the size of their police force. Fortunately, econometric analysis does not require this. What we do need to know is whether the data we can collect on community crime levels and the size of the police force can be viewed as experimental. We can certainly imagine a true experiment involving a large collection of cities where we dictate how many police officers each city will use for the upcoming year.

Although policies can be used to affect the size of police forces, we clearly cannot tell each city how many police officers it can hire. If, as is likely, a city's decision on how many police officers to hire is correlated with other city factors that affect crime, then the data must be viewed as nonexperimental. In fact, one way to view this problem is to see that a city's choice of police force size and the amount of crime are *simultaneously determined*. We will explicitly address such problems in Chapter 16.

The first three examples we have discussed have dealt with cross-sectional data at various levels of aggregation (for example, at the individual or city levels). The same hurdles arise when inferring causality in time series problems.

EXAMPLE 1.6 The Effect of the Minimum Wage on Unemployment

An important, and perhaps contentious, policy issue concerns the effect of the minimum wage on unemployment rates for various groups of workers. Although this problem can be studied in a variety of data settings (cross-sectional, time series, or panel data), time series data are often used to look at aggregate effects. An example of a time series data set on unemployment rates and minimum wages was given in Table 1.3.

Standard supply and demand analysis implies that, as the minimum wage is increased above the market clearing wage, we slide up the demand curve for labor and total employment decreases. (Labor supply exceeds labor demand.) To quantify this effect, we can study the relationship between employment and the minimum wage over time. In addition to some special difficulties that can arise in dealing with time series data, there are possible problems with inferring causality. The minimum wage in the United States is not determined in a vacuum. Various economic and political forces impinge on the final minimum wage for any given year. (The minimum wage, once determined, is usually in place for several years, unless it is indexed for inflation.) Thus, it is probable that the amount of the minimum wage is related to other factors that have an effect on employment levels.

We can imagine the U.S. government conducting an experiment to determine the employment effects of the minimum wage (as opposed to worrying about the welfare of low-wage workers). The minimum wage could be randomly set by the government each year, and then the employment outcomes could be tabulated. The resulting experimental time series data could then be analyzed using fairly simple econometric methods. But this scenario hardly describes how minimum wages are set.

If we can control enough other factors relating to employment, then we can still hope to estimate the *ceteris paribus* effect of the minimum wage on employment. In this sense, the problem is very similar to the previous cross-sectional examples.

Even when economic theories are not most naturally described in terms of causality, they often have predictions that can be tested using econometric methods. The following example demonstrates this approach.

EXAMPLE 1.7 The Expectations Hypothesis

The *expectations hypothesis* from financial economics states that, given all information available to investors at the time of investing, the *expected* return on any two investments is the same. For example, consider two possible investments with a three-month investment horizon, purchased at the same time: (1) Buy a three-month T-bill with a face value of \$10,000, for a price below \$10,000; in three months, you receive \$10,000. (2) Buy a six-month T-bill (at a price below \$10,000) and, in three months, sell it as a three-month T-bill. Each investment requires roughly the same amount of initial capital, but there is an important difference. For the first investment, you know exactly what the return is at the time of purchase because you know the initial price of the three-month T-bill, along with its face value. This is not true for the second investment: although you know the price of a six-month T-bill when you purchase it, you do not know the price you can sell it for in three months. Therefore, there is uncertainty in this investment for someone who has a three-month investment horizon.

The actual returns on these two investments will usually be different. According to the expectations hypothesis, the expected return from the second investment, given all information at the time of investment, should equal the return from purchasing a three-month T-bill. This theory turns out to be fairly easy to test, as we will see in Chapter 11.

Summary

In this introductory chapter, we have discussed the purpose and scope of econometric analysis. Econometrics is used in all applied economics fields to test economic theories, to inform government and private policy makers, and to predict economic time series. Sometimes, an econometric model is derived from a formal economic model, but in other cases, econometric models are based on informal economic reasoning and intuition. The goals of any econometric analysis are to estimate the parameters in the model and to test hypotheses about these parameters; the values and signs of the parameters determine the validity of an economic theory and the effects of certain policies.

Cross-sectional, time series, pooled cross-sectional, and panel data are the most common types of data structures that are used in applied econometrics. Data sets involving a time dimension, such as time series and panel data, require special treatment because of the correlation across time of most economic time series. Other issues, such as trends and seasonality, arise in the analysis of time series data but not cross-sectional data.

In Section 1-4, we discussed the notions of causality, *ceteris paribus*, and counterfactuals. In most cases, hypotheses in the social sciences are *ceteris paribus* in nature: all other relevant factors must be fixed when studying the relationship between two variables. As we discussed, one way to think of the *ceteris paribus* requirement is to undertake a thought experiment where the same economic unit operates in different states of the world, such as different policy regimes. Because of the nonexperimental nature of most data collected in the social sciences, uncovering causal relationships is very challenging.

Key Terms

Causal Effect	Econometric Model	Panel Data
Ceteris Paribus	Economic Model	Pooled Cross Section
Counterfactual Outcomes	Empirical Analysis	Random Sampling
Counterfactual Reasoning	Experimental Data	Retrospective Data
Cross-Sectional Data Set	Nonexperimental Data	Time Series Data
Data Frequency	Observational Data	

Problems

- 1 Suppose that you are asked to conduct a study to determine whether smaller class sizes lead to improved student performance of fourth graders.
 - (i) If you could conduct any experiment you want, what would you do? Be specific.
 - (ii) More realistically, suppose you can collect observational data on several thousand fourth graders in a given state. You can obtain the size of their fourth-grade class and a standardized test score taken at the end of fourth grade. Why might you expect a negative correlation between class size and test score?
 - (iii) Would a negative correlation necessarily show that smaller class sizes cause better performance? Explain.
- 2 A justification for job training programs is that they improve worker productivity. Suppose that you are asked to evaluate whether more job training makes workers more productive. However, rather than having data on individual workers, you have access to data on manufacturing firms in Ohio. In particular, for each firm, you have information on hours of job training per worker (*training*) and number of nondefective items produced per worker hour (*output*).
 - (i) Carefully state the ceteris paribus thought experiment underlying this policy question.
 - (ii) Does it seem likely that a firm's decision to train its workers will be independent of worker characteristics? What are some of those measurable and unmeasurable worker characteristics?
 - (iii) Name a factor other than worker characteristics that can affect worker productivity.
 - (iv) If you find a positive correlation between *output* and *training*, would you have convincingly established that job training makes workers more productive? Explain.
- 3 Suppose at your university you are asked to find the relationship between weekly hours spent studying (*study*) and weekly hours spent working (*work*). Does it make sense to characterize the problem as inferring whether *study* "causes" *work* or *work* "causes" *study*? Explain.
- 4 States (and provinces) that have control over taxation sometimes reduce taxes in an attempt to spur economic growth. Suppose that you are hired by a state to estimate the effect of corporate tax rates on, say, the growth in per capita gross state product (GSP).
 - (i) What kind of data would you need to collect to undertake a statistical analysis?
 - (ii) Is it feasible to do a controlled experiment? What would be required?
 - (iii) Is a correlation analysis between GSP growth and tax rates likely to be convincing? Explain.

Computer Exercises

- C1 Use the data in WAGE1 for this exercise.
 - (i) Find the average education level in the sample. What are the lowest and highest years of education?
 - (ii) Find the average hourly wage in the sample. Does it seem high or low?
 - (iii) The wage data are reported in 1976 dollars. Using the Internet or a printed source, find the Consumer Price Index (CPI) for the years 1976 and 2013.

- (iv) Use the CPI values from part (iii) to find the average hourly wage in 2013 dollars. Now does the average hourly wage seem reasonable?
 - (v) How many women are in the sample? How many men?
- C2** Use the data in BWGHT to answer this question.
- (i) How many women are in the sample, and how many report smoking during pregnancy?
 - (ii) What is the average number of cigarettes smoked per day? Is the average a good measure of the “typical” woman in this case? Explain.
 - (iii) Among women who smoked during pregnancy, what is the average number of cigarettes smoked per day? How does this compare with your answer from part (ii), and why?
 - (iv) Find the average of *fatheduc* in the sample. Why are only 1,192 observations used to compute this average?
 - (v) Report the average family income and its standard deviation in dollars.
- C3** The data in MEAP01 are for the state of Michigan in the year 2001. Use these data to answer the following questions.
- (i) Find the largest and smallest values of *math4*. Does the range make sense? Explain.
 - (ii) How many schools have a perfect pass rate on the math test? What percentage is this of the total sample?
 - (iii) How many schools have math pass rates of exactly 50%?
 - (iv) Compare the average pass rates for the math and reading scores. Which test is harder to pass?
 - (v) Find the correlation between *math4* and *read4*. What do you conclude?
 - (vi) The variable *exppp* is expenditure per pupil. Find the average of *exppp* along with its standard deviation. Would you say there is wide variation in per pupil spending?
 - (vii) Suppose School A spends \$6,000 per student and School B spends \$5,500 per student. By what percentage does School A’s spending exceed School B’s? Compare this to $100 \cdot [\log(6,000) - \log(5,500)]$, which is the approximation percentage difference based on the difference in the natural logs. (See Section A.4 in Math Refresher A.)
- C4** The data in JTRAIN2 come from a job training experiment conducted for low-income men during 1976–1977; see Lalonde (1986).
- (i) Use the indicator variable *train* to determine the fraction of men receiving job training.
 - (ii) The variable *re78* is earnings from 1978, measured in thousands of 1982 dollars. Find the averages of *re78* for the sample of men receiving job training and the sample not receiving job training. Is the difference economically large?
 - (iii) The variable *unem78* is an indicator of whether a man is unemployed or not in 1978. What fraction of the men who received job training are unemployed? What about for men who did not receive job training? Comment on the difference.
 - (iv) From parts (ii) and (iii), does it appear that the job training program was effective? What would make our conclusions more convincing?
- C5** The data in FERTIL2 were collected on women living in the Republic of Botswana in 1988. The variable *children* refers to the number of living children. The variable *electric* is a binary indicator equal to one if the woman’s home has electricity, and zero if not.
- (i) Find the smallest and largest values of *children* in the sample. What is the average of *children*?
 - (ii) What percentage of women have electricity in the home?
 - (iii) Compute the average of *children* for those without electricity and do the same for those with electricity. Comment on what you find.
 - (iv) From part (iii), can you infer that having electricity “causes” women to have fewer children? Explain.