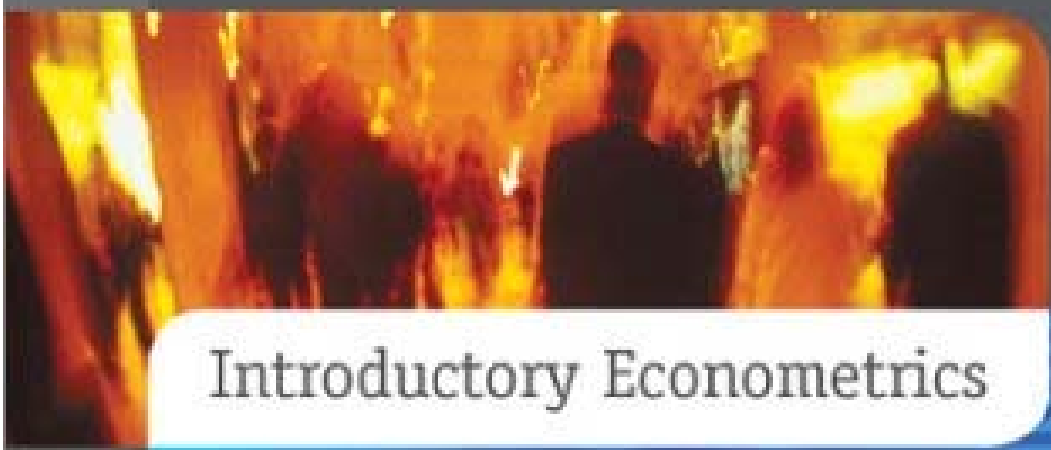
JEFFREY M. WOOLDRIDGE



Introductory Econometrics

A MODERN APPROACH • 2E

Chapter One

The Nature of Econometrics and Economic Data

Chapter 1 discusses the scope of econometrics and raises general issues that result from the application of econometric methods. 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 the inference of 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 twenty-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. The most common application of econometrics is the forecasting of such important macroeconomic variables as interest rates, inflation rates, and gross domestic product. While forecasts of economic indicators are highly visible and are 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** 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 much more difficult to obtain in the social sciences. While 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

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7), \tag{1.1}$$

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

$x_7 =$ 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 where formal derivations provide insights that intuition can overlook.

Here is an example of an equation that was derived 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

$$wage = f(educ, exper, training) \tag{1.2}$$

where *wage* is hourly wage, *educ* is years of formal education, *exper* is years of workforce experience, and *training* is weeks spent in job training. Again, other factors generally affect the wage rate, but (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**. Since 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:

$$crime = \beta_0 + \beta_1 wage_m + \beta_2 othinc + \beta_3 freqarr + \beta_4 freqconv + \beta_5 avgsen + \beta_6 age + u, \tag{1.3}$$

where *crime* is some measure of the frequency of criminal activity, $wage_m$ is the wage that can be earned in legal employment, *othinc* is the income from other sources (assets, inheritance, etc.), *freqarr* is the frequency of arrests for prior infractions (to approximate the probability of arrest), *freqconv* is the frequency of conviction, and *avgsen* is 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 unob-

served 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 β_0, β_1, …, β_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

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training + u, \tag{1.4}$$

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. While 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 $wage_m$, 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. While 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.

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 section 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 Appendix 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, an advanced topic that we will discuss in Chapter 17.

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 for the year 1976. (This is a subset of the data in the file WAGE1.RAW.) 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

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

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 one, which person is called Observation Two, 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, in order to determine the effects of government policies on long-term economic growth, economists have studied the relationship between growth in real per capita gross domestic product (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).

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

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* (percent 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 order of the observations is listed alphabetically by country, but there is nothing about this ordering that affects any subsequent analysis.

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, gross domestic product, 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 it more difficult to analyze than cross-sectional data is the fact 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 gross domestic product from last quarter tells us quite a bit about the likely range of the GDP during this quarter, since GDP tends to remain fairly stable from one quarter to

the next. While 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 employment rates. Other macro series are recorded less frequently, such as every three months (every quarter). Gross domestic product 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 differs 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 observation, and the most recent year available is the last

Table 1.3

Minimum Wage, Unemployment, and Related Data for Puerto Rico

obsno	*year*	*avgmin*	*avgcov*	*unemp*	*gnp*
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. 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), *unemp* is the unemployment rate, and *gnp* is the gross national product. We will use these data later in a time series analysis of the effect of the minimum wage on employment.

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. In order to increase our sample size, we can form a **pooled cross section** by combining the two years. Because random samples are taken in each year, it would be a fluke if the same household appeared in the sample during both years. (The size of the sample is usually very small compared with the number of households in the United States.) This important factor distinguishes a pooled cross section from a panel data set.

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, when there was 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. While 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.

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 ten-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, etc., for the years 1980, 1985, and 1990.

The key feature of panel data that distinguishes it from a pooled cross section is the fact that the *same* cross-sectional units (individuals, firms, or counties in the above

Table 1.4

Pooled Cross Sections: Two Years of Housing Prices

obsno	*year*	*hprice*	*proptax*	*sqrft*	*bdrms*	*bthrms*
1	1993	85500	42	1600	3	2.0
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243600	41	2600	4	3.0
251	1995	65000	16	1250	2	1.0
252	1995	182400	20	2200	4	2.0
253	1995	97500	15	1540	3	2.0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57200	16	1100	2	1.5

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.

Table 1.5

A Two-Year Panel Data Set on City Crime Statistics

obsno	*city*	*year*	*murders*	*population*	*unem*	*police*
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

A second useful 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. Since 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 two pooled cross sections, where the same cities happen to show up in the same year. But, as we will see in Chapters 13 and 14, we can also use the panel structure to respond to 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 sev-

eral advantages over cross-sectional data or even pooled cross-sectional data. The benefit that we will focus on in this text is that having multiple observations on the same units allows us to control 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 it often allows us to study the importance of lags in behavior or the result of decision making. This information can be significant since 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 which is not much more difficult than dealing with a standard cross-sectional data set.

A Comment on Data Structures

Part 1 of this text is concerned with the analysis of cross-sectional data, as 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.

While the econometric analysis of time series uses many of the same tools as cross-sectional analysis, it is more complicated due to 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, since 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 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 AND THE NOTION OF CETERIS PARIBUS IN ECONOMETRIC ANALYSIS

In most tests of economic theory, and certainly for evaluating public policy, the economist's goal is to infer that one variable has a **causal effect** on another variable (such as crime rate or 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. While 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.

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. For each example, the problem of inferring causality disappears if an appropriate experiment can be carried out. 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 fails to have the important features of an experimental data set.

We rely for now on your intuitive understanding of terms such as *random*, *independence*, and *correlation*, all of which should be familiar from an introductory probability and statistics course. (These concepts are reviewed in Appendix 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. Since 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.

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. One approach is to emulate the fertilizer experiment in Example 1.3: Choose a group of people, randomly give each person an amount of education (some people have an eighth grade education, some are given a high school education, etc.), and then measure their wages (assuming that each 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 infeasible. The moral 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, and 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 work force. Since pursuing more education generally requires postponing entering the work force, 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. Since 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 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 problematical. 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 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.

EXAMPLE 1.5

(The Effect of Law Enforcement on City Crime Levels)

The issue of how best to prevent crime has, 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 10 additional police officers, by how much would its crime rates fall? Another way to state the question is: If two cities are the same in all respects, except that city A has 10 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.

While 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. While 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 is an example of 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: while 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 economic fields to test economic theories, 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 goal of any econometric analysis is 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 ceteris paribus and causal inference. 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. Because of the nonexperimental nature of most data collected in the social sciences, uncovering causal relationships is very challenging.

KEY TERMS

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

Chapter Two

The Simple Regression Model

The simple regression model can be used to study the relationship between two variables. For reasons we will see, the simple regression model has limitations as a general tool for empirical analysis. Nevertheless, it is sometimes appropriate as an empirical tool. Learning how to interpret the simple regression model is good practice for studying multiple regression, which we'll do in subsequent chapters.

2.1 DEFINITION OF THE SIMPLE REGRESSION MODEL

Much of applied econometric analysis begins with the following premise: y and x are two variables, representating some population, and we are interested in "explaining y in terms of x," or in "studying how y varies with changes in x." We discussed some examples in Chapter 1, including: y is soybean crop yield and x is amount of fertilizer; y is hourly wage and x is years of education; y is a community crime rate and x is number of police officers.

In writing down a model that will "explain y in terms of x," we must confront three issues. First, since there is never an exact relationship between two variables, how do we allow for other factors to affect y? Second, what is the functional relationship between y and x? And third, how can we be sure we are capturing a ceteris paribus relationship between y and x (if that is a desired goal)?

We can resolve these ambiguities by writing down an equation relating y to x. A simple equation is

$$y = \beta_0 + \beta_1 x + u. \tag{2.1}$$

Equation (2.1), which is assumed to hold in the population of interest, defines the **simple linear regression model**. It is also called the *two-variable linear regression model* or *bivariate linear regression model* because it relates the two variables x and y. We now discuss the meaning of each of the quantities in (2.1). (Incidentally, the term "regression" has origins that are not especially important for most modern econometric applications, so we will not explain it here. See Stigler [1986] for an engaging history of regression analysis.)

When related by (2.1), the variables y and x have several different names used interchangeably, as follows. y is called the **dependent variable**, the **explained variable**, the **response variable**, the **predicted variable**, or the **regressand**. x is called the **independent variable**, the **explanatory variable**, the **control variable**, the **predictor variable**, or the **regressor**. (The term **covariate** is also used for x.) The terms "dependent variable" and "independent variable" are frequently used in econometrics. But be aware that the label "independent" here does not refer to the statistical notion of independence between random variables (see Appendix B).

The terms "explained" and "explanatory" variables are probably the most descriptive. "Response" and "control" are used mostly in the experimental sciences, where the variable x is under the experimenter's control. We will not use the terms "predicted variable" and "predictor," although you sometimes see these. Our terminology for simple regression is summarized in Table 2.1.

Table 2.1

Terminology for Simple Regression

y	x
Dependent Variable	Independent Variable
Explained Variable	Explanatory Variable
Response Variable	Control Variable
Predicted Variable	Predictor Variable
Regressand	Regressor

The variable u, called the **error term** or **disturbance** in the relationship, represents factors other than x that affect y. A simple regression analysis effectively treats all factors affecting y other than x as being unobserved. You can usefully think of u as standing for "unobserved."

Equation (2.1) also addresses the issue of the functional relationship between y and x. If the other factors in u are held fixed, so that the change in u is zero, $\Delta u = 0$, then x has a *linear* effect on y:

$$\Delta y = \beta_1 \Delta x \text{ if } \Delta u = 0. \qquad \textbf{(2.2)}$$

Thus, the change in y is simply β_1 multiplied by the change in x. This means that β_1 is the **slope parameter** in the relationship between y and x holding the other factors in u fixed; it is of primary interest in applied economics. The **intercept parameter** β_0 also has its uses, although it is rarely central to an analysis.

EXAMPLE 2.1
(Soybean Yield and Fertilizer)

Suppose that soybean yield is determined by the model

$$yield = \beta_0 + \beta_1 fertilizer + u, \tag{2.3}$$

so that $y = yield$ and $x = fertilizer$. The agricultural researcher is interested in the effect of fertilizer on yield, holding other factors fixed. This effect is given by β_1. The error term u contains factors such as land quality, rainfall, and so on. The coefficient β_1 measures the effect of fertilizer on yield, holding other factors fixed: $\Delta yield = \beta_1 \Delta fertilizer$.

EXAMPLE 2.2
(A Simple Wage Equation)

A model relating a person's wage to observed education and other unobserved factors is

$$wage = \beta_0 + \beta_1 educ + u. \tag{2.4}$$

If *wage* is measured in dollars per hour and *educ* is years of education, then β_1 measures the change in hourly wage given another year of education, holding all other factors fixed. Some of those factors include labor force experience, innate ability, tenure with current employer, work ethics, and innumerable other things.

The linearity of (2.1) implies that a one-unit change in x has the *same* effect on y, regardless of the initial value of x. This is unrealistic for many economic applications. For example, in the wage-education example, we might want to allow for *increasing* returns: the next year of education has a *larger* effect on wages than did the previous year. We will see how to allow for such possibilities in Section 2.4.

The most difficult issue to address is whether model (2.1) really allows us to draw ceteris paribus conclusions about how x affects y. We just saw in equation (2.2) that β_1 *does* measure the effect of x on y, holding all other factors (in u) fixed. Is this the end of the causality issue? Unfortunately, no. How can we hope to learn in general about the ceteris paribus effect of x on y, holding other factors fixed, when we are ignoring all those other factors?

As we will see in Section 2.5, we are only able to get reliable estimators of β_0 and β_1 from a random sample of data when we make an assumption restricting how the unobservable u is related to the explanatory variable x. Without such a restriction, we will not be able to estimate the ceteris paribus effect, β_1. Because u and x are random variables, we need a concept grounded in probability.

Before we state the key assumption about how x and u are related, there is one assumption about u that we can always make. As long as the intercept β_0 is included in the equation, nothing is lost by assuming that the average value of u in the population is zero.

Mathematically,

$$E(u) = 0. \tag{2.5}$$

Importantly, assume (2.5) says nothing about the relationship between u and x but simply makes a statement about the distribution of the unobservables in the population. Using the previous examples for illustration, we can see that assumption (2.5) is not very restrictive. In Example 2.1, we lose nothing by normalizing the unobserved factors affecting soybean yield, such as land quality, to have an average of zero in the population of all cultivated plots. The same is true of the unobserved factors in Example 2.2. Without loss of generality, we can assume that things such as average ability are zero in the population of all working people. If you are not convinced, you can work through Problem 2.2 to see that we can always redefine the intercept in equation (2.1) to make (2.5) true.

We now turn to the crucial assumption regarding how u and x are related. A natural measure of the association between two random variables is the *correlation coefficient*. (See Appendix B for definition and properties.) If u and x are *uncorrelated*, then, as random variables, they are not *linearly* related. Assuming that u and x are uncorrelated goes a long way toward defining the sense in which u and x should be unrelated in equation (2.1). But it does not go far enough, because correlation measures only linear dependence between u and x. Correlation has a somewhat counterintuitive feature: it is possible for u to be uncorrelated with x while being correlated with functions of x, such as x^2. (See Section B.4 for further discussion.) This possibility is not acceptable for most regression purposes, as it causes problems for interpretating the model and for deriving statistical properties. A better assumption involves the *expected value of u given x*.

Because u and x are random variables, we can define the conditional distribution of u given any value of x. In particular, for any x, we can obtain the expected (or average) value of u for that slice of the population described by the value of x. The crucial assumption is that the average value of u does *not* depend on the value of x. We can write this as

$$E(u|x) = E(u) = 0, \tag{2.6}$$

where the second equality follows from (2.5). The first equality in equation (2.6) is the new assumption, called the **zero conditional mean assumption**. It says that, for any given value of x, the average of the unobservables is the same and therefore must equal the average value of u in the entire population.

Let us see what (2.6) entails in the wage example. To simplify the discussion, assume that u is the same as innate ability. Then (2.6) requires that the average level of ability is the same regardless of years of education. For example, if $E(abil|8)$ denotes the average ability for the group of all people with eight years of education, and $E(abil|16)$ denotes the average ability among people in the population with 16 years of education, then (2.6) implies that these must be the same. In fact, the average ability level must be the same for *all* education levels. If, for example, we think that average ability increases with years of education, then (2.6) is false. (This would happen if, on average, people with more ability choose to become more educated.) As we cannot observe innate ability, we have no way of knowing whether or not average ability is the

same for all education levels. But this is an issue that we must address before applying simple regression analysis.

In the fertilizer example, if fertilizer amounts are chosen independently of other features of the plots, then (2.6) will hold: the average land quality will not depend on the amount of fertilizer. However, if more fertilizer is put on the higher quality plots of land, then the expected value of u changes with the level of fertilizer, and (2.6) fails.

QUESTION 2.1

Suppose that a score on a final exam, *score*, depends on classes attended (*attend*) and unobserved factors that affect exam performance (such as student ability):

$$score = \beta_0 + \beta_1 attend + u \qquad \textbf{(2.7)}$$

When would you expect this model to satisfy (2.6)?

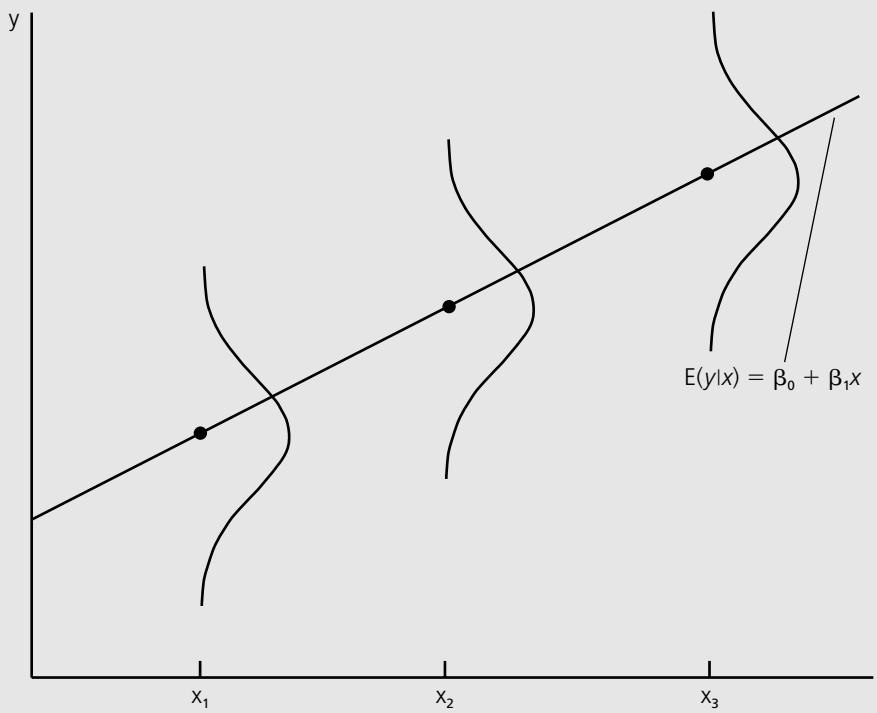
Assumption (2.6) gives β_1 another interpretation that is often useful. Taking the expected value of (2.1) conditional on x and using $E(u|x) = 0$ gives

$$E(y|x) = \beta_0 + \beta_1 x \qquad \textbf{(2.8)}$$

Equation (2.8) shows that the **population regression function** (**PRF**), $E(y|x)$, is a linear function of x. The linearity means that a one-unit increase in x changes the *expect-*

Figure 2.1

$E(y|x)$ as a linear function of x.



ed value of y by the amount β_1. For any given value of x, the distribution of y is centered about $E(y|x)$, as illustrated in Figure 2.1.

When (2.6) is true, it is useful to break y into two components. The piece $\beta_0 + \beta_1 x$ is sometimes called the *systematic part* of y—that is, the part of y explained by x—and u is called the *unsystematic part*, or the part of y not explained by x. We will use assumption (2.6) in the next section for motivating estimates of β_0 and β_1. This assumption is also crucial for the statistical analysis in Section 2.5.

2.2 DERIVING THE ORDINARY LEAST SQUARES ESTIMATES

Now that we have discussed the basic ingredients of the simple regression model, we will address the important issue of how to estimate the parameters β_0 and β_1 in equation (2.1). To do this, we need a sample from the population. Let $\{(x_i,y_i): i=1,\ldots,n\}$ denote a random sample of size n from the population. Since these data come from (2.1), we can write

$$y_i = \beta_0 + \beta_1 x_i + u_i \qquad \textbf{(2.9)}$$

for each i. Here, u_i is the error term for observation i since it contains all factors affecting y_i other than x_i.

As an example, x_i might be the annual income and y_i the annual savings for family i during a particular year. If we have collected data on 15 families, then $n = 15$. A scatter plot of such a data set is given in Figure 2.2, along with the (necessarily fictitious) population regression function.

We must decide how to use these data to obtain estimates of the intercept and slope in the population regression of savings on income.

There are several ways to motivate the following estimation procedure. We will use (2.5) and an important implication of assumption (2.6): in the population, u has a zero mean and is uncorrelated with x. Therefore, we see that u has zero expected value and that the *covariance* between x and u is zero:

$$E(u) = 0 \qquad \textbf{(2.10)}$$

$$\text{Cov}(x,u) = E(xu) = 0, \qquad \textbf{(2.11)}$$

where the first equality in (2.11) follows from (2.10). (See Section B.4 for the definition and properties of covariance.) In terms of the observable variables x and y and the unknown parameters β_0 and β_1, equations (2.10) and (2.11) can be written as

$$E(y - \beta_0 - \beta_1 x) = 0 \qquad \textbf{(2.12)}$$

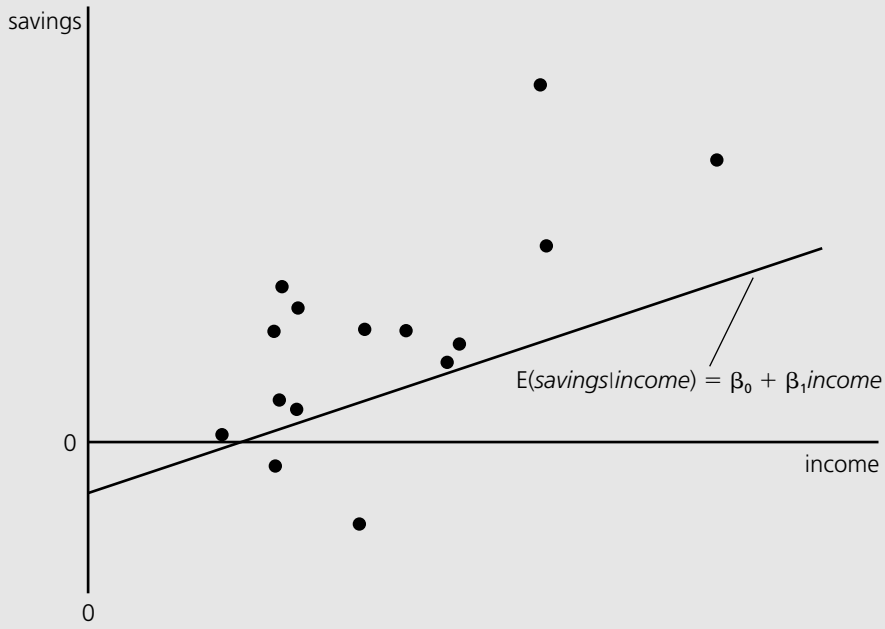
and

$$E[x(y - \beta_0 - \beta_1 x)] = 0, \qquad \textbf{(2.13)}$$

respectively. Equations (2.12) and (2.13) imply two restrictions on the joint probability distribution of (x,y) in the population. Since there are two unknown parameters to estimate, we might hope that equations (2.12) and (2.13) can be used to obtain good esti-

Figure 2.2

Scatterplot of savings and income for 15 families, and the population regression $E(savings|income) = \beta_0 + \beta_1 income$.



mators of β_0 and β_1. In fact, they can be. Given a sample of data, we choose estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ to solve the *sample* counterparts of (2.12) and (2.13):

$$n^{-1}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0. \tag{2.14}$$

$$n^{-1}\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0. \tag{2.15}$$

This is an example of the *method of moments* approach to estimation. (See Section C.4 for a discussion of different estimation approaches.) These equations can be solved for $\hat{\beta}_0$ and $\hat{\beta}_1$.

Using the basic properties of the summation operator from Appendix A, equation (2.14) can be rewritten as

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}, \tag{2.16}$$

where $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ is the sample average of the y_i and likewise for $\bar{x}$. This equation allows us to write $\hat{\beta}_0$ in terms of $\hat{\beta}_1$, $\bar{y}$, and $\bar{x}$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{2.17}$$

Therefore, once we have the slope estimate $\hat{\beta}_1$, it is straightforward to obtain the intercept estimate $\hat{\beta}_0$, given $\bar{y}$ and $\bar{x}$.

Dropping the n^{-1} in (2.15) (since it does not affect the solution) and plugging (2.17) into (2.15) yields

$$\sum_{i=1}^{n} x_i(y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i) = 0$$

which, upon rearrangement, gives

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i(x_i - \bar{x}).$$

From basic properties of the summation operator [see (A.7) and (A.8)],

$$\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \text{ and } \sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Therefore, provided that

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0, \tag{2.18}$$

the estimated slope is

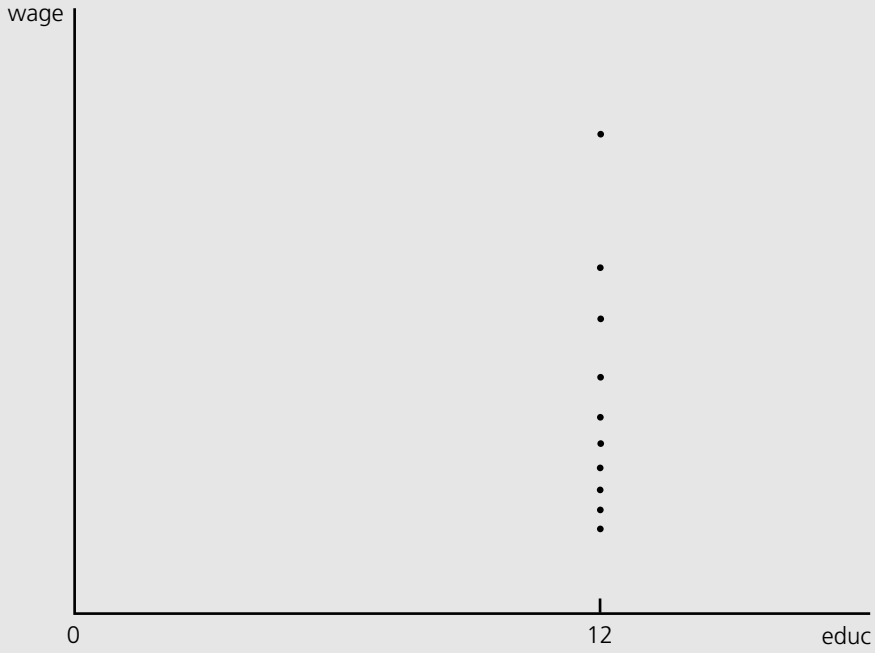
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{2.19}$$

Equation (2.19) is simply the sample covariance between x and y divided by the sample variance of x. (See Appendix C. Dividing both the numerator and the denominator by $n - 1$ changes nothing.) This makes sense because β_1 equals the population covariance divided by the variance of x when $E(u) = 0$ and $Cov(x,u) = 0$. An immediate implication is that if x and y are positively correlated in the sample, then $\hat{\beta}_1$ is positive; if x and y are negatively correlated, then $\hat{\beta}_1$ is negative.

Although the method for obtaining (2.17) and (2.19) is motivated by (2.6), the only assumption needed to compute the estimates for a particular sample is (2.18). This is hardly an assumption at all: (2.18) is true provided the x_i in the sample are not all equal to the same value. If (2.18) fails, then we have either been unlucky in obtaining our sample from the population or we have not specified an interesting problem (x does not vary in the population.). For example, if $y = wage$ and $x = educ$, then (2.18) fails only if everyone in the sample has the same amount of education. (For example, if everyone is a high school graduate. See Figure 2.3.) If just one person has a different amount of education, then (2.18) holds, and the OLS estimates can be computed.

Figure 2.3

A scatterplot of wage against education when $educ_i = 12$ for all i.



The estimates given in (2.17) and (2.19) are called the **ordinary least squares** (**OLS**) estimates of β_0 and β_1. To justify this name, for any $\hat{\beta}_0$ and $\hat{\beta}_1$, define a **fitted value** for y when $x = x_i$ such as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \tag{2.20}$$

for the given intercept and slope. This is the value we predict for y when $x = x_i$. There is a fitted value for each observation in the sample. The **residual** for observation i is the difference between the actual y_i and its fitted value:

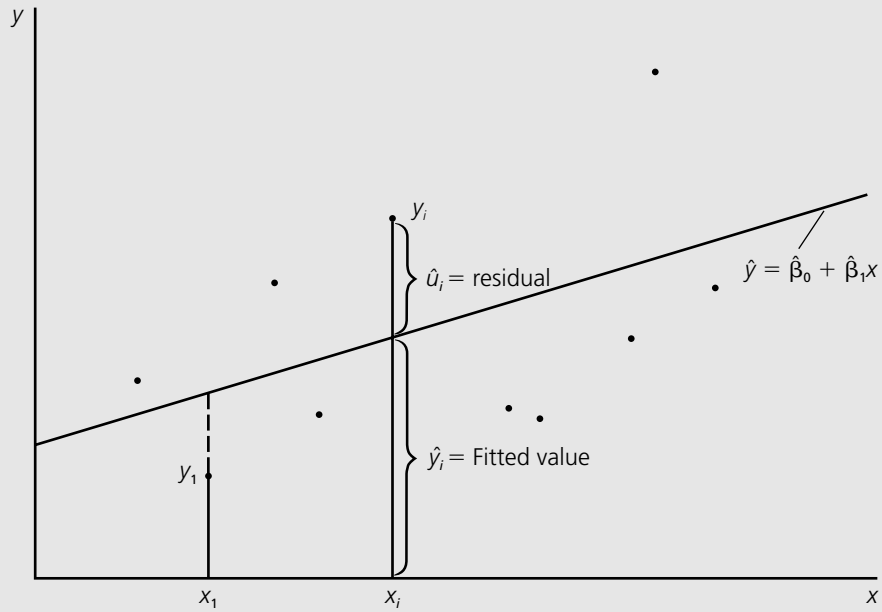
$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i. \tag{2.21}$$

Again, there are n such residuals. (These are *not* the same as the errors in (2.9), a point we return to in Section 2.5.) The fitted values and residuals are indicated in Figure 2.4.

Now, suppose we choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to make the **sum of squared residuals**,

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2, \tag{2.22}$$

Figure 2.4
Fitted values and residuals.



as small as possible. The appendix to this chapter shows that the conditions necessary for $(\hat{\beta}_0, \hat{\beta}_1)$ to minimize (2.22) are given exactly by equations (2.14) and (2.15), without n^{-1}. Equations (2.14) and (2.15) are often called the **first order conditions** for the OLS estimates, a term that comes from optimization using calculus (see Appendix A). From our previous calculations, we know that the solutions to the OLS first order conditions are given by (2.17) and (2.19). The name "ordinary least squares" comes from the fact that these estimates minimize the sum of squared residuals.

Once we have determined the OLS intercept and slope estimates, we form the **OLS regression line**:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \qquad \textbf{(2.23)}$$

where it is understood that $\hat{\beta}_0$ and $\hat{\beta}_1$ have been obtained using equations (2.17) and (2.19). The notation $\hat{y}$, read as "y hat," emphasizes that the predicted values from equation (2.23) are estimates. The intercept, $\hat{\beta}_0$, is the predicted value of y when $x = 0$, although in some cases it will not make sense to set $x = 0$. In those situations, $\hat{\beta}_0$ is not, in itself, very interesting. When using (2.23) to compute predicted values of y for various values of x, we must account for the intercept in the calculations. Equation (2.23) is also called the **sample regression function** (**SRF**) because it is the estimated version of the population regression function $E(y|x) = \beta_0 + \beta_1 x$. It is important to remember that the PRF is something fixed, but unknown, in the population. Since the SRF is

obtained for a given sample of data, a new sample will generate a different slope and intercept in equation (2.23).

In most cases the slope estimate, which we can write as

$$\hat{\beta}_1 = \Delta\hat{y}/\Delta x, \tag{2.24}$$

is of primary interest. It tells us the amount by which $\hat{y}$ changes when x increases by one unit. Equivalently,

$$\Delta\hat{y} = \hat{\beta}_1\Delta x, \tag{2.25}$$

so that given any change in x (whether positive or negative), we can compute the predicted change in y.

We now present several examples of simple regression obtained by using real data. In other words, we find the intercept and slope estimates with equations (2.17) and (2.19). Since these examples involve many observations, the calculations were done using an econometric software package. At this point, you should be careful not to read too much into these regressions; they are not necessarily uncovering a causal relationship. We have said nothing so far about the statistical properties of OLS. In Section 2.5, we consider statistical properties after we explicitly impose assumptions on the population model equation (2.1).

EXAMPLE 2.3
(CEO Salary and Return on Equity)

For the population of chief executive officers, let y be annual salary (*salary*) in thousands of dollars. Thus, $y = 856.3$ indicates an annual salary of \$856,300, and $y = 1452.6$ indicates a salary of \$1,452,600. Let x be the average return equity (*roe*) for the CEO's firm for the previous three years. (Return on equity is defined in terms of net income as a percentage of common equity.) For example, if $roe = 10$, then average return on equity is 10 percent.

To study the relationship between this measure of firm performance and CEO compensation, we postulate the simple model

$$salary = \beta_0 + \beta_1 roe + u.$$

The slope parameter β_1 measures the change in annual salary, in thousands of dollars, when return on equity increases by one percentage point. Because a higher *roe* is good for the company, we think $\beta_1 > 0$.

The data set CEOSAL1.RAW contains information on 209 CEOs for the year 1990; these data were obtained from *Business Week* (5/6/91). In this sample, the average annual salary is \$1,281,120, with the smallest and largest being \$223,000 and \$14,822,000, respectively. The average return on equity for the years 1988, 1989, and 1990 is 17.18 percent, with the smallest and largest values being 0.5 and 56.3 percent, respectively.

Using the data in CEOSAL1.RAW, the OLS regression line relating *salary* to *roe* is

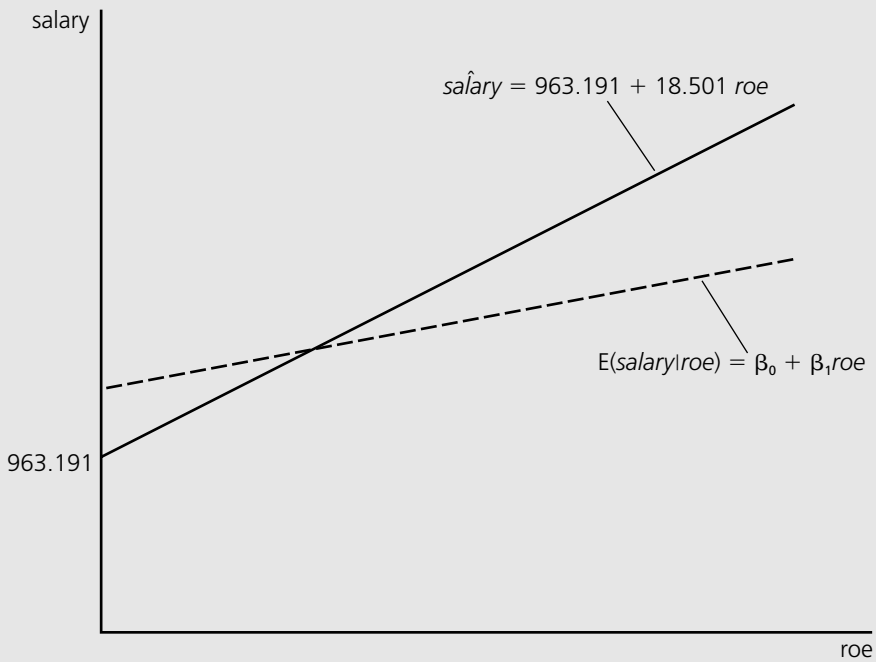
$$\widehat{salary} = 963.191 + 18.501\ roe, \tag{2.26}$$

where the intercept and slope estimates have been rounded to three decimal places; we use "*salary* hat" to indicate that this is an estimated equation. How do we interpret the equation? First, if the return on equity is zero, $roe = 0$, then the predicted *salary* is the intercept, 963.191, which equals \$963,191 since *salary* is measured in thousands. Next, we can write the predicted change in salary as a function of the change in *roe*: $\Delta\widehat{salary} = 18.501(\Delta roe)$. This means that if the return on equity increases by one percentage point, $\Delta roe = 1$, then *salary* is predicted to change by about 18.5, or \$18,500. Because (2.26) is a linear equation, this is the estimated change regardless of the initial salary.

We can easily use (2.26) to compare predicted salaries at different values of *roe*. Suppose $roe = 30$. Then $\widehat{salary} = 963.191 + 18.501(30) = 1518.221$, which is just over \$1.5 million. However, this does *not* mean that a particular CEO whose firm had an $roe = 30$ earns \$1,518,221. There are many other factors that affect salary. This is just our prediction from the OLS regression line (2.26). The estimated line is graphed in Figure 2.5, along with the population regression function E(*salary*|*roe*). We will never know the PRF, so we cannot tell how close the SRF is to the PRF. Another sample of data will give a different regression line, which may or may not be closer to the population regression line.

Figure 2.5

The OLS regression line $\widehat{salary} = 963.191 + 18.50\,roe$ and the (unknown) population regression function.



EXAMPLE 2.4
(Wage and Education)

For the population of people in the work force in 1976, let $y = wage$, where *wage* is measured in dollars per hour. Thus, for a particular person, if $wage = 6.75$, the hourly *wage* is \$6.75. Let $x = educ$ denote years of schooling; for example, $educ = 12$ corresponds to a complete high school education. Since the average wage in the sample is \$5.90, the consumer price index indicates that this amount is equivalent to \$16.64 in 1997 dollars.

Using the data in WAGE1.RAW where $n = 526$ individuals, we obtain the following OLS regression line (or sample regression function):

$$\widehat{wage} = -0.90 + 0.54\, educ. \qquad \textbf{(2.27)}$$

We must interpret this equation with caution. The intercept of -0.90 literally means that a person with no education has a predicted hourly wage of -90 cents an hour. This, of course, is silly. It turns out that no one in the sample has less than eight years of education, which helps to explain the crazy prediction for a zero education value. For a person with eight years of education, the predicted wage is $\widehat{wage} = -0.90 + 0.54(8) = 3.42$, or \$3.42 per hour (in 1976 dollars).

QUESTION 2.2

The estimated wage from (2.27), when $educ = 8$, is \$3.42 in 1976 dollars. What is this value in 1997 dollars? (*Hint:* You have enough information in Example 2.4 to answer this question.)

The slope estimate in (2.27) implies that one more year of education increases hourly wage by 54 cents an hour. Therefore, four more years of education increase the predicted wage by $4(0.54) = 2.16$ or \$2.16 per hour. These are fairly large effects. Because of the linear nature of (2.27), another year of education increases the wage by the same amount, regardless of the initial level of education. In Section 2.4, we discuss some methods that allow for nonconstant marginal effects of our explanatory variables.

EXAMPLE 2.5
(Voting Outcomes and Campaign Expenditures)

The file VOTE1.RAW contains data on election outcomes and campaign expenditures for 173 two-party races for the U.S. House of Representatives in 1988. There are two candidates in each race, A and B. Let *voteA* be the percentage of the vote received by Candidate A and *shareA* be the the percentage of total campaign expenditures accounted for by Candidate A. Many factors other than *shareA* affect the election outcome (including the quality of the candidates and possibly the dollar amounts spent by A and B). Nevertheless, we can estimate a simple regression model to find out whether spending more relative to one's challenger implies a higher percentage of the vote.

The estimated equation using the 173 observations is

$$\widehat{voteA} = 40.90 + 0.306\, shareA. \qquad \textbf{(2.28)}$$

This means that, if the share of Candidate A's expenditures increases by one percentage point, Candidate A receives almost one-third of a percentage point more of the

total vote. Whether or not this is a causal effect is unclear, but the result is what we might expect.

In some cases, regression analysis is not used to determine causality but to simply look at whether two variables are positively or negatively related, much like a standard correlation analysis. An example of this occurs in Problem 2.12, where you are asked to use data from Biddle and Hamermesh (1990) on time spent sleeping and working to investigate the tradeoff between these two factors.

QUESTION 2.3

In Example 2.5, what is the predicted vote for Candidate A if *shareA* = 60 (which means 60 percent)? Does this answer seem reasonable?

A Note on Terminolgy

In most cases, we will indicate the estimation of a relationship through OLS by writing an equation such as (2.26), (2.27), or (2.28). Sometimes, for the sake of brevity, it is useful to indicate that an OLS regression has been run without actually writing out the equation. We will often indicate that equation (2.23) has been obtained by OLS in saying that we *run the regression of*

$$y \text{ on } x, \tag{2.29}$$

or simply that we *regress* y on x. The positions of y and x in (2.29) indicate which is the dependent variable and which is the independent variable: we always regress the dependent variable on the independent variable. For specific applications, we replace y and x with their names. Thus, to obtain (2.26), we regress *salary* on *roe* or to obtain (2.28), we regress *voteA* on *shareA*.

When we use such terminology in (2.29), we will always mean that we plan to estimate the intercept, $\hat{\beta}_0$, along with the slope, $\hat{\beta}_1$. This case is appropriate for the vast majority of applications. Occasionally, we may want to estimate the relationship between y and x *assuming* that the intercept is zero (so that $x = 0$ implies that $\hat{y} = 0$); we cover this case briefly in Section 2.6. Unless explicitly stated otherwise, we always estimate an intercept along with a slope.

2.3 MECHANICS OF OLS

In this section, we cover some algebraic properties of the fitted OLS regression line. Perhaps the best way to think about these properties is to realize that they are features of OLS for a particular sample of data. They can be contrasted with the *statistical* properties of OLS, which requires deriving features of the sampling distributions of the estimators. We will discuss statistical properties in Section 2.5.

Several of the algebraic properties we are going to derive will appear mundane. Nevertheless, having a grasp of these properties helps us to figure out what happens to the OLS estimates and related statistics when the data are manipulated in certain ways, such as when the measurement units of the dependent and independent variables change.

Fitted Values and Residuals

We assume that the intercept and slope estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, have been obtained for the given sample of data. Given $\hat{\beta}_0$ and $\hat{\beta}_1$, we can obtain the fitted value $\hat{y}_i$ for each observation. [This is given by equation (2.20).] By definition, each fitted value of $\hat{y}_i$ is on the OLS regression line. The OLS residual associated with observation i, $\hat{u}_i$, is the difference between y_i and its fitted value, as given in equation (2.21). If $\hat{u}_i$ is positive, the line underpredicts y_i; if $\hat{u}_i$ is negative, the line overpredicts y_i. The ideal case for observation i is when $\hat{u}_i = 0$, but in most cases *every* residual is not equal to zero. In other words, none of the data points must actually lie on the OLS line.

EXAMPLE 2.6
(CEO Salary and Return on Equity)

Table 2.2 contains a listing of the first 15 observations in the CEO data set, along with the fitted values, called *salaryhat*, and the residuals, called *uhat*.

Table 2.2

Fitted Values and Residuals for the First 15 CEOs

obsno	*roe*	*salary*	*salaryhat*	*uhat*
1	14.1	1095	1224.058	−129.0581
2	10.9	1001	1164.854	−163.8542
3	23.5	1122	1397.969	−275.9692
4	5.9	578	1072.348	−494.3484
5	13.8	1368	1218.508	149.4923
6	20.0	1145	1333.215	−188.2151
7	16.4	1078	1266.611	−188.6108
8	16.3	1094	1264.761	−170.7606
9	10.5	1237	1157.454	79.54626
10	26.3	833	1449.773	−616.7726
11	25.9	567	1442.372	−875.3721
12	26.8	933	1459.023	−526.0231

continued

Table 2.2 (*concluded*)

obsno	*roe*	*salary*	*salaryhat*	*uhat*
13	14.8	1339	1237.009	101.9911
14	22.3	937	1375.768	−438.7678
15	56.3	2011	2004.808	6.191895

The first four CEOs have lower salaries than what we predicted from the OLS regression line (2.26); in other words, given only the firm's *roe*, these CEOs make less than what we predicted. As can be seen from the positive *uhat*, the fifth CEO makes more than predicted from the OLS regression line.

Algebraic Properties of OLS Statistics

There are several useful algebraic properties of OLS estimates and their associated statistics. We now cover the three most important of these.

(1) The sum, and therefore the sample average of the OLS residuals, is zero.

Mathematically,

$$\sum_{i=1}^{n} \hat{u}_i = 0. \tag{2.30}$$

This property needs no proof; it follows immediately from the OLS first order condition (2.14), when we remember that the residuals are defined by $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. In other words, the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are *chosen* to make the residuals add up to zero (for any data set). This says nothing about the residual for any particular observation i.

(2) The sample covariance between the regressors and the OLS residuals is zero. This follows from the first order condition (2.15), which can be written in terms of the residuals as

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0. \tag{2.31}$$

The sample average of the OLS residuals is zero, so the left hand side of (2.31) is proportional to the sample covariance between x_i and $\hat{u}_i$.

(3) The point $(\bar{x},\bar{y})$ is always on the OLS regression line. In other words, if we take equation (2.23) and plug in $\bar{x}$ for x, then the predicted value is $\bar{y}$. This is exactly what equation (2.16) shows us.

EXAMPLE 2.7
(Wage and Education)

For the data in WAGE1.RAW, the average hourly wage in the sample is 5.90, rounded to two decimal places, and the average education is 12.56. If we plug *educ* = 12.56 into the OLS regression line (2.27), we get $\widehat{wage} = -0.90 + 0.54(12.56) = 5.8824$, which equals 5.9 when rounded to the first decimal place. The reason these figures do not exactly agree is that we have rounded the average wage and education, as well as the intercept and slope estimates. If we did not initially round any of the values, we would get the answers to agree more closely, but this practice has little useful effect.

Writing each y_i as its fitted value, plus its residual, provides another way to intepret an OLS regression. For each i, write

$$y_i = \hat{y}_i + \hat{u}_i. \tag{2.32}$$

From property (1) above, the average of the residuals is zero; equivalently, the sample average of the fitted values, $\hat{y}_i$, is the same as the sample average of the y_i, or $\bar{\hat{y}} = \bar{y}$. Further, properties (1) and (2) can be used to show that the sample covariance between $\hat{y}_i$ and $\hat{u}_i$ is zero. Thus, we can view OLS as decomposing each y_i into two parts, a fitted value and a residual. The fitted values and residuals are uncorrelated in the sample.

Define the **total sum of squares (SST)**, the **explained sum of squares (SSE)**, and the **residual sum of squares (SSR)** (also known as the sum of squared residuals), as follows:

$$\text{SST} \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{2.33}$$

$$\text{SSE} \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \tag{2.34}$$

$$\text{SSR} \equiv \sum_{i=1}^{n} \hat{u}_i^2. \tag{2.35}$$

SST is a measure of the total sample variation in the y_i; that is, it measures how spread out the y_i are in the sample. If we divide SST by $n - 1$, we obtain the sample variance of y, as discussed in Appendix C. Similarly, SSE measures the sample variation in the $\hat{y}_i$ (where we use the fact that $\bar{\hat{y}} = \bar{y}$), and SSR measures the sample variation in the $\hat{u}_i$. The total variation in y can always be expressed as the sum of the explained variation and the unexplained variation SSR. Thus,

$$\text{SST} = \text{SSE} + \text{SSR}. \tag{2.36}$$

Proving (2.36) is not difficult, but it requires us to use all of the properties of the summation operator covered in Appendix A. Write

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= \sum_{i=1}^{n}[\hat{u}_i + (\hat{y}_i - \bar{y})]^2 \\ &= \sum_{i=1}^{n}\hat{u}_i^2 + 2\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ &= \text{SSR} + 2\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i - \bar{y}) + \text{SSE}.\end{aligned}$$

Now (2.36) holds if we show that

$$\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i - \bar{y}) = 0. \qquad \textbf{(2.37)}$$

But we have already claimed that the sample covariance between the residuals and the fitted values is zero, and this covariance is just (2.37) divided by $n-1$. Thus, we have established (2.36).

Some words of caution about SST, SSE, and SSR are in order. There is no uniform agreement on the names or abbreviations for the three quantities defined in equations (2.33), (2.34), and (2.35). The total sum of squares is called either SST or TSS, so there is little confusion here. Unfortunately, the explained sum of squares is sometimes called the "regression sum of squares." If this term is given its natural abbreviation, it can easily be confused with the term residual sum of squares. Some regression packages refer to the explained sum of squares as the "model sum of squares."

To make matters even worse, the residual sum of squares is often called the "error sum of squares." This is especially unfortunate because, as we will see in Section 2.5, the errors and the residuals are different quantities. Thus, we will always call (2.35) the residual sum of squares or the sum of squared residuals. We prefer to use the abbreviation SSR to denote the sum of squared residuals, because it is more common in econometric packages.

Goodness-of-Fit

So far, we have no way of measuring how well the explanatory or independent variable, x, explains the dependent variable, y. It is often useful to compute a number that summarizes how well the OLS regression line fits the data. In the following discussion, be sure to remember that we assume that an intercept is estimated along with the slope.

Assuming that the total sum of squares, SST, is not equal to zero—which is true except in the very unlikely event that all the y_i equal the same value—we can divide (2.36) by SST to get $1 = \text{SSE/SST} + \text{SSR/SST}$. The **R-squared** of the regression, sometimes called the **coefficient of determination**, is defined as

$$R^2 = \text{SSE/SST} = 1 - \text{SSR/SST}. \tag{2.38}$$

R^2 is the ratio of the explained variation compared to the total variation, and thus it is interpreted as the *fraction of the sample variation in y that is explained by x*. The second equality in (2.38) provides another way for computing R^2.

From (2.36), the value of R^2 is always between zero and one, since SSE can be no greater than SST. When interpreting R^2, we usually multiply it by 100 to change it into a percent: $100 \cdot R^2$ is the *percentage of the sample variation in y that is explained by x*.

If the data points all lie on the same line, OLS provides a perfect fit to the data. In this case, $R^2 = 1$. A value of R^2 that is nearly equal to zero indicates a poor fit of the OLS line: very little of the variation in the y_i is captured by the variation in the $\hat{y}_i$ (which all lie on the OLS regression line). In fact, it can be shown that R^2 is equal to the *square* of the sample correlation coefficient between y_i and $\hat{y}_i$. This is where the term "*R*-squared" came from. (The letter *R* was traditionally used to denote an estimate of a population correlation coefficient, and its usage has survived in regression analysis.)

EXAMPLE 2.8
(CEO Salary and Return on Equity)

In the CEO salary regression, we obtain the following:

$$\widehat{salary} = 963.191 + 18.501\, roe \tag{2.39}$$

$$n = 209,\ R^2 = 0.0132$$

We have reproduced the OLS regression line and the number of observations for clarity. Using the *R*-squared (rounded to four decimal places) reported for this equation, we can see how much of the variation in salary is actually explained by the return on equity. The answer is: not much. The firm's return on equity explains only about 1.3% of the variation in salaries for this sample of 209 CEOs. That means that 98.7% of the salary variations for these CEOs is left unexplained! This lack of explanatory power may not be too surprising since there are many other characteristics of both the firm and the individual CEO that should influence salary; these factors are necessarily included in the errors in a simple regression analysis.

In the social sciences, low *R*-squareds in regression equations are not uncommon, especially for cross-sectional analysis. We will discuss this issue more generally under multiple regression analysis, but it is worth emphasizing now that a seemingly low *R*-squared does not necessarily mean that an OLS regression equation is useless. It is still possible that (2.39) is a good estimate of the ceteris paribus relationship between *salary* and *roe*; whether or not this is true does *not* depend directly on the size of *R*-squared. Students who are first learning econometrics tend to put too much weight on the size of the *R*-squared in evaluating regression equations. For now, be aware that using *R*-squared as the main gauge of success for an econometric analysis can lead to trouble.

Sometimes the explanatory variable explains a substantial part of the sample variation in the dependent variable.

EXAMPLE 2.9
(Voting Outcomes and Campaign Expenditures)

In the voting outcome equation in (2.28), $R^2 = 0.505$. Thus, the share of campaign expenditures explains just over 50 percent of the variation in the election outcomes for this sample. This is a fairly sizable portion.

2.4 UNITS OF MEASUREMENT AND FUNCTIONAL FORM

Two important issues in applied economics are (1) understanding how changing the units of measurement of the dependent and/or independent variables affects OLS estimates and (2) knowing how to incorporate popular functional forms used in economics into regression analysis. The mathematics needed for a full understanding of functional form issues is reviewed in Appendix A.

The Effects of Changing Units of Measurement on OLS Statistics

In Example 2.3, we chose to measure annual salary in thousands of dollars, and the return on equity was measured as a percent (rather than as a decimal). It is crucial to know how *salary* and *roe* are measured in this example in order to make sense of the estimates in equation (2.39).

We must also know that OLS estimates change in entirely expected ways when the units of measurement of the dependent and independent variables change. In Example 2.3, suppose that, rather than measuring salary in thousands of dollars, we measure it in dollars. Let *salardol* be salary in dollars (*salardol* = 845,761 would be interpreted as \$845,761.). Of course, *salardol* has a simple relationship to the salary measured in thousands of dollars: $salardol = 1{,}000 \cdot salary$. We do not need to actually run the regression of *salardol* on *roe* to know that the estimated equation is:

$$\widehat{salardol} = 963{,}191 + 18{,}501\, roe. \qquad \textbf{(2.40)}$$

We obtain the intercept and slope in (2.40) simply by multiplying the intercept and the slope in (2.39) by 1,000. This gives equations (2.39) and (2.40) the *same* interpretation. Looking at (2.40), if *roe* = 0, then $\widehat{salardol} = 963{,}191$, so the predicted salary is \$963,191 [the same value we obtained from equation (2.39)]. Furthermore, if *roe* increases by one, then the predicted salary increases by \$18,501; again, this is what we concluded from our earlier analysis of equation (2.39).

Generally, it is easy to figure out what happens to the intercept and slope estimates when the dependent variable changes units of measurement. If the dependent variable is multiplied by the constant c—which means each value in the sample is multiplied by c—then the OLS intercept and slope estimates are also multiplied by c. (This assumes nothing has changed about the independent variable.) In the CEO salary example, $c = 1{,}000$ in moving from *salary* to *salardol*.

We can also use the CEO salary example to see what happens when we change the units of measurement of the independent variable. Define *roedec* = *roe*/100 to be the decimal equivalent of *roe*; thus, *roedec* = 0.23 means a return on equity of 23 percent. To focus on changing the units of measurement of the independent variable, we return to our original dependent variable, *salary*, which is measured in thousands of dollars. When we regress *salary* on *roedec*, we obtain

QUESTION 2.4

Suppose that salary is measured in hundreds of dollars, rather than in thousands of dollars, say *salarhun*. What will be the OLS intercept and slope estimates in the regression of *salarhun* on *roe*?

$$\widehat{salary} = 963.191 + 1850.1\ roedec. \tag{2.41}$$

The coefficient on *roedec* is 100 times the coefficient on *roe* in (2.39). This is as it should be. Changing *roe* by one percentage point is equivalent to $\Delta roedec = 0.01$. From (2.41), if $\Delta roedec = 0.01$, then $\Delta\widehat{salary} = 1850.1(0.01) = 18.501$, which is what is obtained by using (2.39). Note that, in moving from (2.39) to (2.41), the independent variable was divided by 100, and so the OLS slope estimate was multiplied by 100, preserving the interpretation of the equation. Generally, if the independent variable is divided or multiplied by some nonzero constant, c, then the OLS slope coefficient is also multiplied or divided by c respectively.

The intercept has not changed in (2.41) because *roedec* = 0 still corresponds to a zero return on equity. In general, changing the units of measurement of only the independent variable does not affect the intercept.

In the previous section, we defined R-squared as a goodness-of-fit measure for OLS regression. We can also ask what happens to R^2 when the unit of measurement of either the independent or the dependent variable changes. Without doing any algebra, we should know the result: the goodness-of-fit of the model should not depend on the units of measurement of our variables. For example, the amount of variation in salary, explained by the return on equity, should not depend on whether salary is measured in dollars or in thousands of dollars or on whether return on equity is a percent or a decimal. This intuition can be verified mathematically: using the definition of R^2, it can be shown that R^2 is, in fact, invariant to changes in the units of y or x.

Incorporating Nonlinearities in Simple Regression

So far we have focused on *linear* relationships between the dependent and independent variables. As we mentioned in Chapter 1, linear relationships are not nearly general enough for all economic applications. Fortunately, it is rather easy to incorporate many nonlinearities into simple regression analysis by appropriately defining the dependent and independent variables. Here we will cover two possibilities that often appear in applied work.

In reading applied work in the social sciences, you will often encounter regression equations where the dependent variable appears in logarithmic form. Why is this done? Recall the wage-education example, where we regressed hourly wage on years of education. We obtained a slope estimate of 0.54 [see equation (2.27)], which means that each additional year of education is predicted to increase hourly wage by 54 cents.

Because of the linear nature of (2.27), 54 cents is the increase for either the first year of education or the twentieth year; this may not be reasonable.

Suppose, instead, that the *percentage* increase in wage is the same given one more year of education. Model (2.27) does not imply a constant percentage increase: the percentage increases depends on the initial wage. A model that gives (approximately) a constant percentage effect is

$$\log(wage) = \beta_0 + \beta_1 educ + u, \tag{2.42}$$

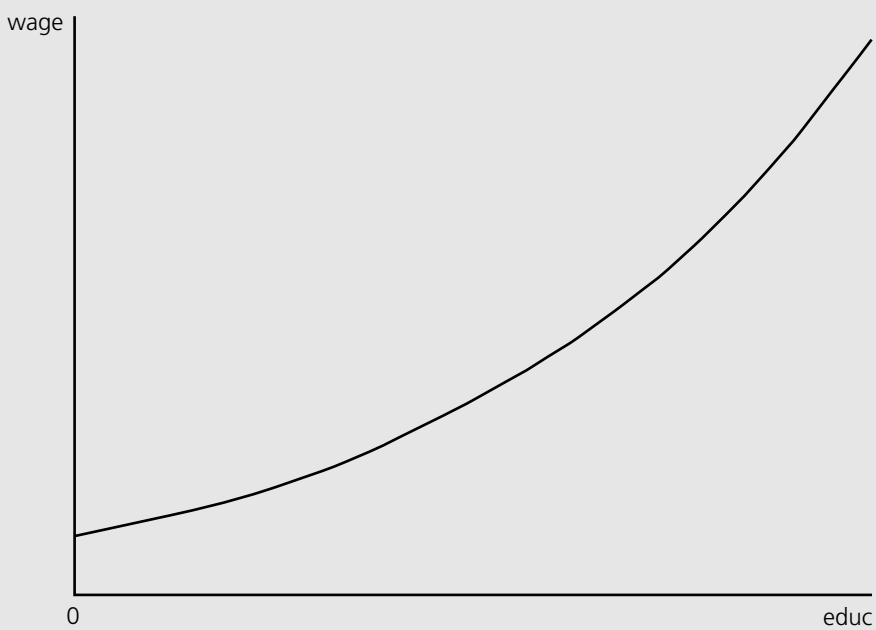
where $\log(\cdot)$ denotes the *natural* logarithm. (See Appendix A for a review of logarithms.) In particular, if $\Delta u = 0$, then

$$\%\Delta wage \approx (100 \cdot \beta_1)\Delta educ. \tag{2.43}$$

Notice how we multiply β_1 by 100 to get the percentage change in *wage* given one additional year of education. Since the percentage change in *wage* is the same for each additional year of education, the change in *wage* for an extra year of education *increases* as education increases; in other words, (2.42) implies an *increasing* return to education. By exponentiating (2.42), we can write $wage = \exp(\beta_0 + \beta_1 educ + u)$. This equation is graphed in Figure 2.6, with $u = 0$.

Figure 2.6

wage $= \exp(\beta_0 + \beta_1 educ)$, *with* $\beta_1 > 0$.



Estimating a model such as (2.42) is straightforward when using simple regression. Just define the dependent variable, y, to be $y = \log(wage)$. The independent variable is represented by $x = educ$. The mechanics of OLS are the same as before: the intercept and slope estimates are given by the formulas (2.17) and (2.19). In other words, we obtain $\hat{\beta}_0$ and $\hat{\beta}_1$ from the OLS regression of $\log(wage)$ on $educ$.

EXAMPLE 2.10
(A Log Wage Equation)

Using the same data as in Example 2.4, but using log(*wage*) as the dependent variable, we obtain the following relationship:

$$\widehat{\log(wage)} = 0.584 + 0.083\, educ \qquad \textbf{(2.44)}$$

$$n = 526,\ R^2 = 0.186.$$

The coefficient on *educ* has a percentage interpretation when it is multiplied by 100: *wage* increases by 8.3 percent for every additional year of education. This is what economists mean when they refer to the "return to another year of education."

It is important to remember that the main reason for using the log of *wage* in (2.42) is to impose a constant percentage effect of education on *wage*. Once equation (2.42) is obtained, the natural log of *wage* is rarely mentioned. In particular, it is *not* correct to say that another year of education increases log(*wage*) by 8.3%.

The intercept in (2.42) is not very meaningful, as it gives the predicted log(*wage*), when $educ = 0$. The *R*-squared shows that *educ* explains about 18.6 percent of the variation in log(*wage*) (*not wage*). Finally, equation (2.44) might not capture all of the nonlinearity in the relationship between wage and schooling. If there are "diploma effects," then the twelfth year of education—graduation from high school—could be worth much more than the eleventh year. We will learn how to allow for this kind of nonlinearity in Chapter 7.

Another important use of the natural log is in obtaining a **constant elasticity model**.

EXAMPLE 2.11
(CEO Salary and Firm Sales)

We can estimate a constant elasticity model relating CEO salary to firm sales. The data set is the same one used in Example 2.3, except we now relate *salary* to *sales*. Let *sales* be annual firm sales, measured in millions of dollars. A constant elasticity model is

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u, \qquad \textbf{(2.45)}$$

where β_1 is the elasticity of *salary* with respect to *sales*. This model falls under the simple regression model by defining the dependent variable to be $y = \log(salary)$ and the independent variable to be $x = \log(sales)$. Estimating this equation by OLS gives

$$\widehat{\log(salary)} = 4.822 + 0.257 \log(sales) \quad \textbf{(2.46)}$$

$$n = 209, R^2 = 0.211.$$

The coefficient of log(*sales*) is the estimated elasticity of *salary* with respect to *sales*. It implies that a 1 percent increase in firm sales increases CEO salary by about 0.257 percent—the usual interpretation of an elasticity.

The two functional forms covered in this section will often arise in the remainder of this text. We have covered models containing natural logarithms here because they appear so frequently in applied work. The interpretation of such models will not be much different in the multiple regression case.

It is also useful to note what happens to the intercept and slope estimates if we change the units of measurement of the dependent variable when it appears in logarithmic form. Because the change to logarithmic form approximates a proportionate change, it makes sense that *nothing* happens to the slope. We can see this by writing the rescaled variable as $c_1 y_i$ for each observation i. The original equation is $\log(y_i) = \beta_0 + \beta_1 x_i + u_i$. If we add $\log(c_1)$ to both sides, we get $\log(c_1) + \log(y_i) = [\log(c_1) + \beta_0] + \beta_1 x_i + u_i$, or $\log(c_1 y_i) = [\log(c_1) + \beta_0] + \beta_1 x_i + u_i$. (Remember that the sum of the logs is equal to the log of their product as shown in Appendix A.) Therefore, the slope is still β_1, but the intercept is now $\log(c_1) + \beta_0$. Similarly, if the independent variable is $\log(x)$, and we change the units of measurement of x before taking the log, the slope remains the same but the intercept does not change. You will be asked to verify these claims in Problem 2.9.

We end this subsection by summarizing four combinations of functional forms available from using either the original variable or its natural log. In Table 2.3, x and y stand for the variables in their original form. The model with y as the dependent variable and x as the independent variable is called the *level-level* model, because each variable appears in its level form. The model with $\log(y)$ as the dependent variable and x as the independent variable is called the *log-level* model. We will not explicitly discuss the *level-log* model here, because it arises less often in practice. In any case, we will see examples of this model in later chapters.

Table 2.3

Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
level-level	y	x	$\Delta y = \beta_1 \Delta x$
level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

The last column in Table 2.3 gives the interpretation of β_1. In the log-level model, $100 \cdot \beta_1$ is sometimes called the **semi-elasticity** of y with respect to x. As we mentioned in Example 2.11, in the log-log model, β_1 is the **elasticity** of y with respect to x. Table 2.3 warrants careful study, as we will refer to it often in the remainder of the text.

The Meaning of "Linear" Regression

The simple regression model that we have studied in this chapter is also called the simple *linear* regression model. Yet, as we have just seen, the general model also allows for certain *nonlinear* relationships. So what does "linear" mean here? You can see by looking at equation (2.1) that $y = \beta_0 + \beta_1 x + u$. The key is that this equation is linear in the *parameters*, β_0 and β_1. There are no restrictions on how y and x relate to the original explained and explanatory variables of interest. As we saw in Examples 2.7 and 2.8, y and x can be natural logs of variables, and this is quite common in applications. But we need not stop there. For example, nothing prevents us from using simple regression to estimate a model such as $cons = \beta_0 + \beta_1 \sqrt{inc} + u$, where *cons* is annual consumption and *inc* is annual income.

While the mechanics of simple regression do not depend on how y and x are defined, the interpretation of the coefficients does depend on their definitions. For successful empirical work, it is much more important to become proficient at interpreting coefficients than to become efficient at computing formulas such as (2.19). We will get much more practice with interpreting the estimates in OLS regression lines when we study multiple regression.

There are plenty of models that *cannot* be cast as a linear regression model because they are not linear in their parameters; an example is $cons = 1/(\beta_0 + \beta_1 inc) + u$. Estimation of such models takes us into the realm of the *nonlinear regression model*, which is beyond the scope of this text. For most applications, choosing a model that can be put into the linear regression framework is sufficient.

2.5 EXPECTED VALUES AND VARIANCES OF THE OLS ESTIMATORS

In Section 2.1, we defined the population model $y = \beta_0 + \beta_1 x + u$, and we claimed that the key assumption for simple regression analysis to be useful is that the expected value of u given any value of x is zero. In Sections 2.2, 2.3, and 2.4, we discussed the algebraic properties of OLS estimation. We now return to the population model and study the *statistical* properties of OLS. In other words, we now view $\hat{\beta}_0$ and $\hat{\beta}_1$ as *estimators* for the parameters β_0 and β_1 that appear in the population model. This means that we will study properties of the distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ over different random samples from the population. (Appendix C contains definitions of estimators and reviews some of their important properties.)

Unbiasedness of OLS

We begin by establishing the unbiasedness of OLS under a simple set of assumptions. For future reference, it is useful to number these assumptions using the prefix "SLR" for simple linear regression. The first assumption defines the population model.

ASSUMPTION SLR.1 (LINEAR IN PARAMETERS)
In the population model, the dependent variable y is related to the independent variable x and the error (or disturbance) u as

$$y = \beta_0 + \beta_1 x + u, \tag{2.47}$$

where β_0 and β_1 are the population intercept and slope parameters, respectively.

To be realistic, y, x, and u are all viewed as random variables in stating the population model. We discussed the interpretation of this model at some length in Section 2.1 and gave several examples. In the previous section, we learned that equation (2.47) is not as restrictive as it initially seems; by choosing y and x appropriately, we can obtain interesting nonlinear relationships (such as constant elasticity models).

We are interested in using data on y and x to estimate the parameters β_0 and, especially, β_1. We assume that our data were obtained as a random sample. (See Appendix C for a review of random sampling.)

ASSUMPTION SLR.2 (RANDOM SAMPLING)
We can use a random sample of size n, $\{(x_i,y_i): i = 1,2,\ldots,n\}$, from the population model.

We will have to address failure of the random sampling assumption in later chapters that deal with time series analysis and sample selection problems. Not all cross-sectional samples can be viewed as outcomes of random samples, but many can be.

We can write (2.47) in terms of the random sample as

$$y_i = \beta_0 + \beta_1 x_i + u_i, \; i = 1,2,\ldots,n, \tag{2.48}$$

where u_i is the error or disturbance for observation i (for example, person i, firm i, city i, etc.). Thus, u_i contains the unobservables for observation i which affect y_i. The u_i should not be confused with the residuals, $\hat{u}_i$, that we defined in Section 2.3. Later on, we will explore the relationship between the errors and the residuals. For interpreting β_0 and β_1 in a particular application, (2.47) is most informative, but (2.48) is also needed for some of the statistical derivations.

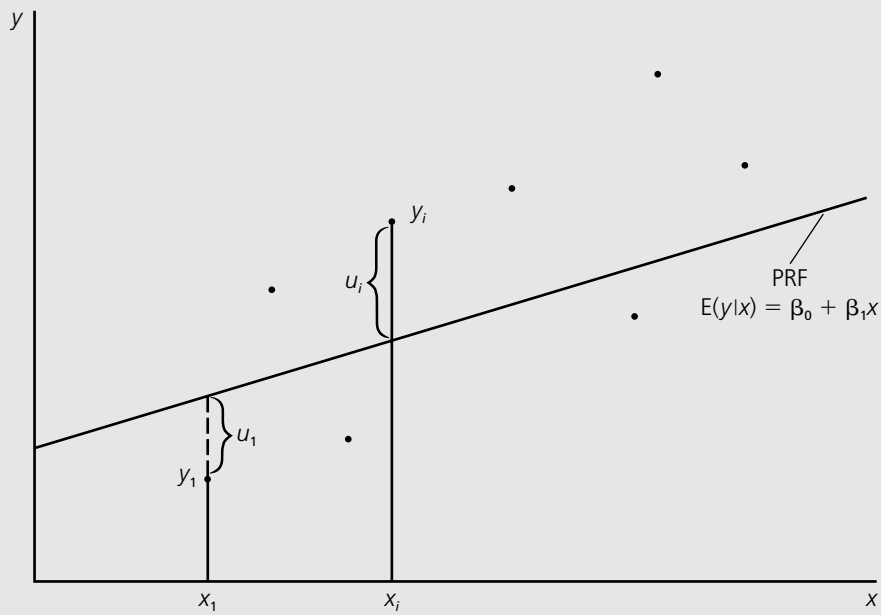
The relationship (2.48) can be plotted for a particular outcome of data as shown in Figure 2.7.

In order to obtain unbiased estimators of β_0 and β_1, we need to impose the zero conditional mean assumption that we discussed in some detail in Section 2.1. We now explicitly add it to our list of assumptions.

ASSUMPTION SLR.3 (ZERO CONDITIONAL MEAN)
$E(u|x) = 0$.

Figure 2.7

Graph of $y_i = \beta_0 + \beta_1 x_i + u_i$.



For a random sample, this assumption implies that $E(u_i|x_i) = 0$, for all $i = 1,2,\ldots,n$.

In addition to restricting the relationship between u and x in the population, the zero conditional mean assumption—coupled with the random sampling assumption—allows for a convenient technical simplification. In particular, we can derive the statistical properties of the OLS estimators as *conditional* on the values of the x_i in our sample. Technically, in statistical derivations, conditioning on the sample values of the independent variable is the same as treating the x_i as *fixed in repeated samples*. This process involves several steps. We first choose n sample values for $x_1, x_2, \ldots, x_n$ (These can be repeated.). Given these values, we then obtain a sample on y (effectively by obtaining a random sample of the u_i). Next another sample of y is obtained, using the *same* values for $x_1, \ldots, x_n$. Then another sample of y is obtained, again using the same x_i. And so on.

The fixed in repeated samples scenario is not very realistic in nonexperimental contexts. For instance, in sampling individuals for the wage-education example, it makes little sense to think of choosing the values of *educ* ahead of time and then sampling individuals with those particular levels of education. Random sampling, where individuals are chosen randomly and their wage and education are both recorded, is representative of how most data sets are obtained for empirical analysis in the social sciences. Once we *assume* that $E(u|x) = 0$, and we have random sampling, nothing is lost in derivations by treating the x_i as nonrandom. The danger is that the fixed in repeated samples assumption *always* implies that u_i and x_i are independent. In deciding when

simple regression analysis is going to produce unbiased estimators, it is critical to think in terms of Assumption SLR.3.

Once we have agreed to condition on the x_i, we need one final assumption for unbiasedness.

ASSUMPTION SLR.4 (SAMPLE VARIATION IN THE INDEPENDENT VARIABLE)

In the sample, the independent variables x_i, $i = 1,2,\ldots,n$, are not all equal to the same constant. This requires some variation in x in the population.

We encountered Assumption SLR.4 when we derived the formulas for the OLS estimators; it is equivalent to $\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0$. Of the four assumptions made, this is the least important because it essentially never fails in interesting applications. If Assumption SLR.4 does fail, we cannot compute the OLS estimators, which means statistical analysis is irrelevant.

Using the fact that $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})y_i$ (see Appendix A), we can write the OLS slope estimator in equation (2.19) as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{2.49}$$

Because we are now interested in the behavior of $\hat{\beta}_1$ across all possible samples, $\hat{\beta}_1$ is properly viewed as a random variable.

We can write $\hat{\beta}_1$ in terms of the population coefficients and errors by substituting the right hand side of (2.48) into (2.49). We have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})y_i}{s_x^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{s_x^2}, \tag{2.50}$$

where we have defined the total variation in x_i as $s_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$ in order to simplify the notation. (This is not quite the sample variance of the x_i because we do not divide by $n - 1$.) Using the algebra of the summation operator, write the numerator of $\hat{\beta}_1$ as

$$\begin{aligned} &\sum_{i=1}^{n} (x_i - \bar{x})\beta_0 + \sum_{i=1}^{n} (x_i - \bar{x})\beta_1 x_i + \sum_{i=1}^{n} (x_i - \bar{x})u_i \\ &\qquad = \beta_0 \sum_{i=1}^{n} (x_i - \bar{x}) + \beta_1 \sum_{i=1}^{n} (x_i - \bar{x})x_i + \sum_{i=1}^{n} (x_i - \bar{x})u_i. \end{aligned} \tag{2.51}$$

As shown in Appendix A, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ and $\sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n}(x_i - \bar{x})^2 = s_x^2$. Therefore, we can write the numerator of $\hat{\beta}_1$ as $\beta_1 s_x^2 + \sum_{i=1}^{n}(x_i - \bar{x})u_i$. Writing this over the denominator gives

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{s_x^2} = \beta_1 + (1/s_x^2)\sum_{i=1}^{n} d_i u_i, \tag{2.52}$$

where $d_i = x_i - \bar{x}$. We now see that the estimator $\hat{\beta}_1$ equals the population slope β_1, plus a term that is a linear combination in the errors $\{u_1,u_2,\ldots,u_n\}$. Conditional on the values of x_i, the randomness in $\hat{\beta}_1$ is due entirely to the errors in the sample. The fact that these errors are generally different from zero is what causes $\hat{\beta}_1$ to differ from β_1.

Using the representation in (2.52), we can prove the first important statistical property of OLS.

THEOREM 2.1 (UNBIASEDNESS OF OLS)
Using Assumptions SLR.1 through SLR.4,

$$\mathrm{E}(\hat{\beta}_0) = \beta_0, \text{ and } \mathrm{E}(\hat{\beta}_1) = \beta_1 \tag{2.53}$$

for any values of β_0 and β_1. In other words, $\hat{\beta}_0$ is unbiased for β_0, and $\hat{\beta}_1$ is unbiased for β_1.

PROOF: In this proof, the expected values are conditional on the sample values of the independent variable. Since s_x^2 and d_i are functions only of the x_i, they are nonrandom in the conditioning. Therefore, from (2.53),

$$\mathrm{E}(\hat{\beta}_1) = \beta_1 + \mathrm{E}[(1/s_x^2)\sum_{i=1}^{n} d_i u_i] = \beta_1 + (1/s_x^2)\sum_{i=1}^{n}\mathrm{E}(d_i u_i)$$

$$= \beta_1 + (1/s_x^2)\sum_{i=1}^{n} d_i \mathrm{E}(u_i) = \beta_1 + (1/s_x^2)\sum_{i=1}^{n} d_i \cdot 0 = \beta_1,$$

where we have used the fact that the expected value of each u_i (conditional on $\{x_1,x_2,\ldots,x_n\}$) is zero under Assumptions SLR.2 and SLR.3.

The proof for $\hat{\beta}_0$ is now straightforward. Average (2.48) across i to get $\bar{y} = \beta_0 + \beta_1\bar{x} + \bar{u}$, and plug this into the formula for $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \beta_0 + \beta_1\bar{x} + \bar{u} - \hat{\beta}_1\bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u}.$$

Then, conditional on the values of the x_i,

$$\mathrm{E}(\hat{\beta}_0) = \beta_0 + \mathrm{E}[(\beta_1 - \hat{\beta}_1)\bar{x}] + \mathrm{E}(\bar{u}) = \beta_0 + \mathrm{E}[(\beta_1 - \hat{\beta}_1)]\bar{x},$$

since $\mathrm{E}(\bar{u}) = 0$ by Assumptions SLR.2 and SLR.3. But, we showed that $\mathrm{E}(\hat{\beta}_1) = \beta_1$, which implies that $\mathrm{E}[(\hat{\beta}_1 - \beta_1)] = 0$. Thus, $\mathrm{E}(\hat{\beta}_0) = \beta_0$. Both of these arguments are valid for any values of β_0 and β_1, and so we have established unbiasedness.

Remember that unbiasedness is a feature of the sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$, which says nothing about the estimate that we obtain for a given sample. We hope that, if the sample we obtain is somehow "typical," then our estimate should be "near" the population value. Unfortunately, it is always possible that we could obtain an unlucky sample that would give us a point estimate far from β_1, and we can *never* know for sure whether this is the case. You may want to review the material on unbiased estimators in Appendix C, especially the simulation exercise in Table C.1 that illustrates the concept of unbiasedness.

Unbiasedness generally fails if any of our four assumptions fail. This means that it is important to think about the veracity of each assumption for a particular application. As we have already discussed, if Assumption SLR.4 fails, then we will not be able to obtain the OLS estimates. Assumption SLR.1 requires that y and x be linearly related, with an additive disturbance. This can certainly fail. But we also know that y and x can be chosen to yield interesting nonlinear relationships. Dealing with the failure of (2.47) requires more advanced methods that are beyond the scope of this text.

Later, we will have to relax Assumption SLR.2, the random sampling assumption, for time series analysis. But what about using it for cross-sectional analysis? Random sampling can fail in a cross section when samples are not representative of the underlying population; in fact, some data sets are constructed by intentionally oversampling different parts of the population. We will discuss problems of nonrandom sampling in Chapters 9 and 17.

The assumption we should concentrate on for now is SLR.3. If SLR.3 holds, the OLS estimators are unbiased. Likewise, if SLR.3 fails, the OLS estimators generally will be *biased*. There are ways to determine the likely direction and size of the bias, which we will study in Chapter 3.

The possibility that x is correlated with u is almost always a concern in simple regression analysis with nonexperimental data, as we indicated with several examples in Section 2.1. Using simple regression when u contains factors affecting y that are also correlated with x can result in *spurious correlation*: that is, we find a relationship between y and x that is really due to other unobserved factors that affect y and also happen to be correlated with x.

EXAMPLE 2.12

(Student Math Performance and the School Lunch Program)

Let *math10* denote the percentage of tenth graders at a high school receiving a passing score on a standardized mathematics exam. Suppose we wish to estimate the effect of the federally funded school lunch program on student performance. If anything, we expect the lunch program to have a positive ceteris paribus effect on performance: all other factors being equal, if a student who is too poor to eat regular meals becomes eligible for the school lunch program, his or her performance should improve. Let *lnchprg* denote the percentage of students who are eligible for the lunch program. Then a simple regression model is

$$math10 = \beta_0 + \beta_1 lnchprg + u, \tag{2.54}$$

where *u* contains school and student characteristics that affect overall school performance. Using the data in MEAP93.RAW on 408 Michigan high schools for the 1992–93 school year, we obtain

$$\widehat{math10} = 32.14 - 0.319\ lnchprg$$

$$n = 408,\ R^2 = 0.171$$

This equation predicts that if student eligibility in the lunch program increases by 10 percentage points, the percentage of students passing the math exam *falls* by about 3.2 percentage points. Do we really believe that higher participation in the lunch program actually *causes* worse performance? Almost certainly not. A better explanation is that the error term *u* in equation (2.54) is correlated with *lnchprg*. In fact, *u* contains factors such as the poverty rate of children attending school, which affects student performance and is highly correlated with eligibility in the lunch program. Variables such as school quality and resources are also contained in *u*, and these are likely correlated with *lnchprg*. It is important to remember that the estimate −0.319 is only for this particular sample, but its sign and magnitude make us suspect that *u* and *x* are correlated, so that simple regression is biased.

In addition to omitted variables, there are other reasons for *x* to be correlated with *u* in the simple regression model. Since the same issues arise in multiple regression analysis, we will postpone a systematic treatment of the problem until then.

Variances of the OLS Estimators

In addition to knowing that the sampling distribution of $\hat{\beta}_1$ is centered about β_1 ($\hat{\beta}_1$ is unbiased), it is important to know how far we can expect $\hat{\beta}_1$ to be away from β_1 on average. Among other things, this allows us to choose the best estimator among all, or at least a broad class of, the unbiased estimators. The measure of spread in the distribution of $\hat{\beta}_1$ (and $\hat{\beta}_0$) that is easiest to work with is the variance or its square root, the standard deviation. (See Appendix C for a more detailed discussion.)

It turns out that the variance of the OLS estimators can be computed under Assumptions SLR.1 through SLR.4. However, these expressions would be somewhat complicated. Instead, we add an assumption that is traditional for cross-sectional analysis. This assumption states that the variance of the unobservable, *u*, conditional on *x*, is constant. This is known as the **homoskedasticity** or "constant variance" assumption.

ASSUMPTION SLR.5 (HOMOSKEDASTICITY)
$\text{Var}(u|x) = \sigma^2$.

We must emphasize that the homoskedasticity assumption is quite distinct from the zero conditional mean assumption, $E(u|x) = 0$. Assumption SLR.3 involves the *expected value* of *u*, while Assumption SLR.5 concerns the *variance* of *u* (both conditional on *x*). Recall that we established the unbiasedness of OLS without Assumption SLR.5: the homoskedasticity assumption plays *no* role in showing that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased. We add Assumption SLR.5 because it simplifies the variance calculations for

$\hat{\beta}_0$ and $\hat{\beta}_1$ and because it implies that ordinary least squares has certain efficiency properties, which we will see in Chapter 3. If we were to assume that u and x are *independent*, then the distribution of u given x does not depend on x, and so $E(u|x) = E(u) = 0$ and $Var(u|x) = \sigma^2$. But independence is sometimes too strong of an assumption.

Because $Var(u|x) = E(u^2|x) - [E(u|x)]^2$ and $E(u|x) = 0$, $\sigma^2 = E(u^2|x)$, which means σ^2 is also the *unconditional* expectation of u^2. Therefore, $\sigma^2 = E(u^2) = Var(u)$, because $E(u) = 0$. In other words, σ^2 is the *unconditional* variance of u, and so σ^2 is often called the **error variance** or disturbance variance. The square root of σ^2, σ, is the standard deviation of the error. A larger σ means that the distribution of the unobservables affecting y is more spread out.

It is often useful to write Assumptions SLR.3 and SLR.5 in terms of the conditional mean and conditional variance of y:

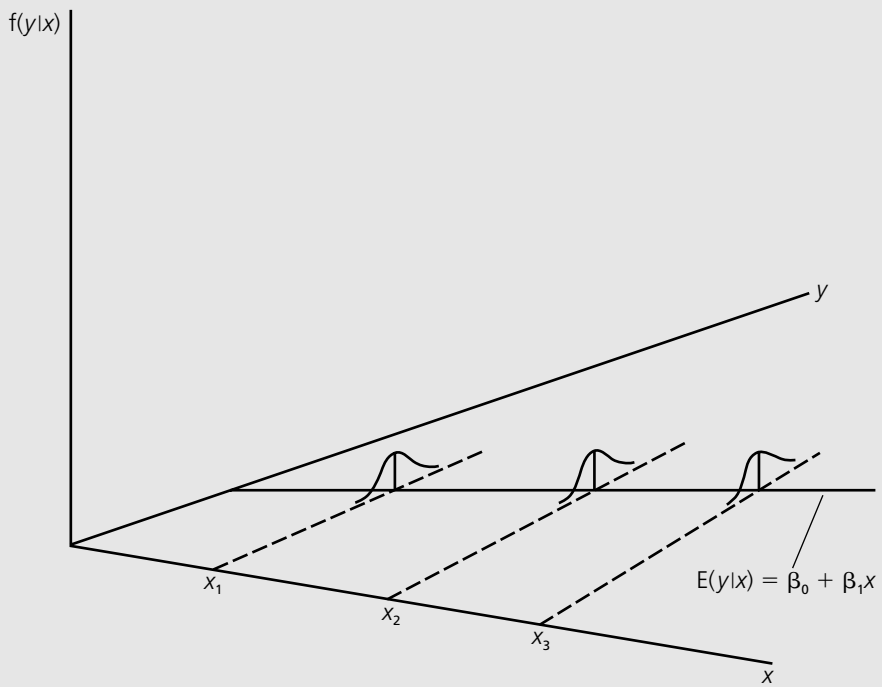
$$E(y|x) = \beta_0 + \beta_1 x. \tag{2.55}$$

$$Var(y|x) = \sigma^2. \tag{2.56}$$

In other words, the conditional expectation of y given x is linear in x, but the variance of y given x is constant. This situation is graphed in Figure 2.8 where $\beta_0 > 0$ and $\beta_1 > 0$.

Figure 2.8

The simple regression model under homoskedasticity.

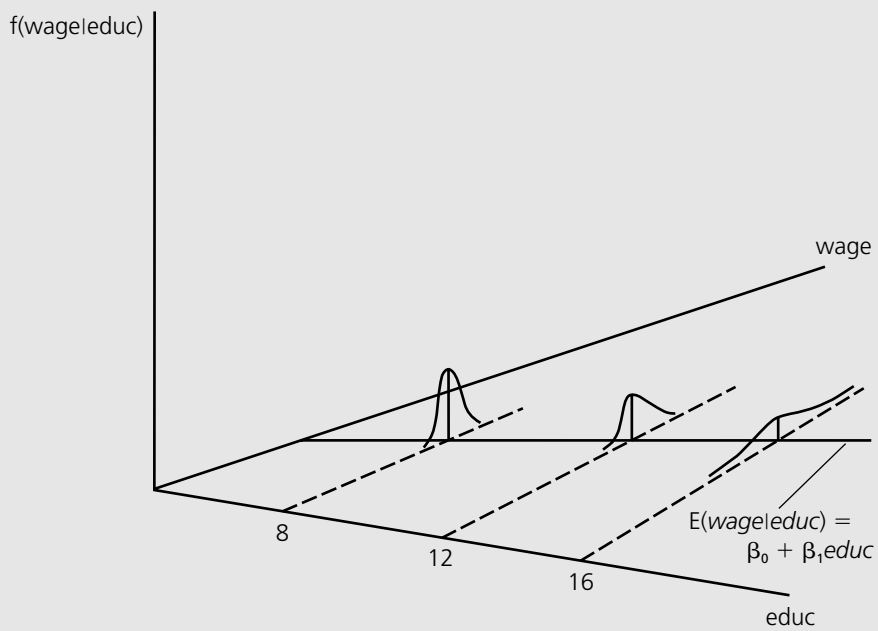


When $\text{Var}(u|x)$ depends on x, the error term is said to exhibit **heteroskedasticity** (or nonconstant variance). Since $\text{Var}(u|x) = \text{Var}(y|x)$, heteroskedasticity is present whenever $\text{Var}(y|x)$ is a function of x.

EXAMPLE 2.13
(Heteroskedasticity in a Wage Equation)

In order to get an unbiased estimator of the ceteris paribus effect of *educ* on *wage*, we must assume that $E(u|educ) = 0$, and this implies $E(wage|educ) = \beta_0 + \beta_1 educ$. If we also make the homoskedasticity assumption, then $\text{Var}(u|educ) = \sigma^2$ does not depend on the level of education, which is the same as assuming $\text{Var}(wage|educ) = \sigma^2$. Thus, while average wage is allowed to increase with education level—it is this rate of increase that we are interested in describing—the *variability* in wage about its mean is assumed to be constant across all education levels. This may not be realistic. It is likely that people with more education have a wider variety of interests and job opportunities, which could lead to more wage variability at higher levels of education. People with very low levels of education have very few opportunities and often must work at the minimum wage; this serves to reduce wage variability at low education levels. This situation is shown in Figure 2.9. Ultimately, whether Assumption SLR.5 holds is an empirical issue, and in Chapter 8 we will show how to test Assumption SLR.5.

Figure 2.9
Var (*wage*|*educ*) increasing with *educ*.



With the homoskedasticity assumption in place, we are ready to prove the following:

THEOREM 2.2 (SAMPLING VARIANCES OF THE OLS ESTIMATORS)
Under Assumptions SLR.1 through SLR.5,

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2/s_x^2 \qquad \textbf{(2.57)}$$

$$\text{Var}(\hat{\beta}_0) = \frac{\sigma^2 n^{-1}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \textbf{(2.58)}$$

where these are conditional on the sample values $\{x_1,\ldots,x_n\}$.

PROOF: We derive the formula for $\text{Var}(\hat{\beta}_1)$, leaving the other derivation as an exercise. The starting point is equation (2.52): $\hat{\beta}_1 = \beta_1 + (1/s_x^2)\sum_{i=1}^{n} d_i u_i$. Since β_1 is just a constant, and we are conditioning on the x_i, s_x^2 and $d_i = x_i - \bar{x}$ are also nonrandom. Furthermore, because the u_i are independent random variables across i (by random sampling), the variance of the sum is the sum of the variances. Using these facts, we have

$$\begin{aligned}
\text{Var}(\hat{\beta}_1) &= (1/s_x^2)^2\text{Var}\left(\sum_{i=1}^{n} d_i u_i\right) = (1/s_x^2)^2\left(\sum_{i=1}^{n} d_i^2\text{Var}(u_i)\right) \\
&= (1/s_x^2)^2\left(\sum_{i=1}^{n} d_i^2\sigma^2\right) \qquad [\text{since } \text{Var}(u_i) = \sigma^2 \text{ for all } i] \\
&= \sigma^2(1/s_x^2)^2\left(\sum_{i=1}^{n} d_i^2\right) = \sigma^2(1/s_x^2)^2 s_x^2 = \sigma^2/s_x^2,
\end{aligned}$$

which is what we wanted to show.

The formulas (2.57) and (2.58) are the "standard" formulas for simple regression analysis, which are invalid in the presence of heteroskedasticity. This will be important when we turn to confidence intervals and hypothesis testing in multiple regression analysis.

For most purposes, we are interested in $\text{Var}(\hat{\beta}_1)$. It is easy to summarize how this variance depends on the error variance, σ^2, and the total variation in $\{x_1,x_2,\ldots,x_n\}$, s_x^2. First, the larger the error variance, the larger is $\text{Var}(\hat{\beta}_1)$. This makes sense since more variation in the unobservables affecting y makes it more difficult to precisely estimate β_1. On the other hand, more variability in the independent variable is preferred: as the variability in the x_i increases, the variance of $\hat{\beta}_1$ decreases. This also makes intuitive

sense since the more spread out is the sample of independent variables, the easier it is to trace out the relationship between $E(y|x)$ and x. That is, the easier it is to estimate β_1. If there is little variation in the x_i, then it can be hard to pinpoint how $E(y|x)$ varies with x. As the sample size increases, so does the total variation in the x_i. Therefore, a larger sample size results in a smaller variance for $\hat{\beta}_1$.

This analysis shows that, if we are interested in $\hat{\beta}_1$, and we have a choice, then we should choose the x_i to be as spread out as possible. This is sometimes possible with experimental data, but rarely do we have this luxury in the social sciences: usually we must take the x_i that we obtain via random sampling. Sometimes we have an opportunity to obtain larger sample sizes, although this can be costly.

QUESTION 2.5

Show that, when estimating β_0, it is best to have $\bar{x} = 0$. What is $\text{Var}(\hat{\beta}_0)$ in this case? (*Hint:* For any sample of numbers, $\sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2$, with equality only if $\bar{x} = 0$.)

For the purposes of constructing confidence intervals and deriving test statistics, we will need to work with the standard deviations of $\hat{\beta}_1$ and $\hat{\beta}_0$, $\text{sd}(\hat{\beta}_1)$ and $\text{sd}(\hat{\beta}_0)$. Recall that these are obtained by taking the square roots of the variances in (2.57) and (2.58). In particular, $\text{sd}(\hat{\beta}_1) = \sigma/s_x$, where σ is the square root of σ^2, and s_x is the square root of s_x^2.

Estimating the Error Variance

The formulas in (2.57) and (2.58) allow us to isolate the factors that contribute to $\text{Var}(\hat{\beta}_1)$ and $\text{Var}(\hat{\beta}_0)$. But these formulas are unknown, except in the extremely rare case that σ^2 is known. Nevertheless, we can use the data to estimate σ^2, which then allows us to estimate $\text{Var}(\hat{\beta}_1)$ and $\text{Var}(\hat{\beta}_0)$.

This is a good place to emphasize the difference between the the *errors* (or disturbances) and the *residuals*, since this distinction is crucial for constructing an estimator of σ^2. Equation (2.48) shows how to write the population model in terms of a randomly sampled observation as $y_i = \beta_0 + \beta_1 x_i + u_i$, where u_i is the error for observation i. We can also express y_i in terms of its fitted value and residual as in equation (2.32): $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$. Comparing these two equations, we see that the error shows up in the equation containing the *population* parameters, β_0 and β_1. On the other hand, the residuals show up in the *estimated* equation with $\hat{\beta}_0$ and $\hat{\beta}_1$. The errors are never observable, while the residuals are computed from the data.

We can use equations (2.32) and (2.48) to write the residuals as a function of the errors:

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = (\beta_0 + \beta_1 x_i + u_i) - \hat{\beta}_0 - \hat{\beta}_1 x_i,$$

or

$$\hat{u}_i = u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1) x_i. \tag{2.59}$$

Although the expected value of $\hat{\beta}_0$ equals β_0, and similarly for $\hat{\beta}_1$, $\hat{u}_i$ is not the same as u_i. The difference between them does have an *expected value* of zero.

Now that we understand the difference between the errors and the residuals, we can

return to estimating σ^2. First, $\sigma^2 = E(u^2)$, so an unbiased "estimator" of σ^2 is $n^{-1}\sum_{i=1}^{n} u_i^2$. Unfortunately, this is not a true estimator, because we do not observe the errors u_i. But, we do have estimates of the u_i, namely the OLS residuals $\hat{u}_i$. If we replace the errors with the OLS residuals, have $n^{-1}\sum_{i=1}^{n} \hat{u}_i^2 = \text{SSR}/n$. This *is* a true estimator, because it gives a computable rule for any sample of data on x and y. One slight drawback to this estimator is that it turns out to be biased (although for large n the bias is small). Since it is easy to compute an unbiased estimator, we use that instead.

The estimator SSR/n is biased essentially because it does not account for two restrictions that must be satisfied by the OLS residuals. These restrictions are given by the two OLS first order conditions:

$$\sum_{i=1}^{n} \hat{u}_i = 0, \sum_{i=1}^{n} x_i\hat{u}_i = 0. \tag{2.60}$$

One way to view these restrictions is this: if we know $n - 2$ of the residuals, we can always get the other two residuals by using the restrictions implied by the first order conditions in (2.60). Thus, there are only $n - 2$ **degrees of freedom** in the OLS residuals [as opposed to n degrees of freedom in the errors. If we replace $\hat{u}_i$ with u_i in (2.60), the restrictions would no longer hold.] The unbiased estimator of σ^2 that we will use makes a degrees-of-freedom adjustment:

$$\hat{\sigma}^2 = \frac{1}{(n-2)}\sum_{i=1}^{n} \hat{u}_i^2 = \text{SSR}/(n-2). \tag{2.61}$$

(This estimator is sometimes denoted s^2, but we continue to use the convention of putting "hats" over estimators.)

THEOREM 2.3 (UNBIASED ESTIMATION OF σ^2)
Under Assumptions SLR.1 through SLR.5,

$$E(\hat{\sigma}^2) = \sigma^2.$$

PROOF: If we average equation (2.59) across all i and use the fact that the OLS residuals average out to zero, we have $0 = \bar{u} - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)\bar{x}$; subtracting this from (2.59) gives $\hat{u}_i = (u_i - \bar{u}) - (\hat{\beta}_1 - \beta_1)(x_i - \bar{x})$. Therefore, $\hat{u}_i^2 = (u_i - \bar{u})^2 + (\hat{\beta}_1 - \beta_1)^2(x_i - \bar{x})^2 - 2(u_i - \bar{u})(\hat{\beta}_1 - \beta_1)(x_i - \bar{x})$. Summing across all i gives $\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (u_i - \bar{u})^2 + (\hat{\beta}_1 - \beta_1)^2\sum_{i=1}^{n} (x_i - \bar{x})^2 - 2(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n} u_i(x_i - \bar{x})$. Now, the expected value of the first term is $(n - 1)\sigma^2$, something that is shown in Appendix C. The expected value of the second term is simply σ^2 because $E[(\hat{\beta}_1 - \beta_1)^2] = \text{Var}(\hat{\beta}_1) = \sigma^2/s_x^2$. Finally, the third term can be written as $2(\hat{\beta}_1 - \beta_1)^2 s_x^2$; taking expectations gives $2\sigma^2$. Putting these three terms together gives $E\left(\sum_{i=1}^{n} \hat{u}_i^2\right) = (n - 1)\sigma^2 + \sigma^2 - 2\sigma^2 = (n - 2)\sigma^2$, so that $E[\text{SSR}/(n - 2)] = \sigma^2$.

If $\hat{\sigma}^2$ is plugged into the variance formulas (2.57) and (2.58), then we have unbiased estimators of $\text{Var}(\hat{\beta}_1)$ and $\text{Var}(\hat{\beta}_0)$. Later on, we will need estimators of the standard deviations of $\hat{\beta}_1$ and $\hat{\beta}_0$, and this requires estimating σ. The natural estimator of σ is

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}, \tag{2.62}$$

and is called the **standard error of the regression** (**SER**). (Other names for $\hat{\sigma}$ are the *standard error of the estimate* and the *root mean squared error*, but we will not use these.) Although $\hat{\sigma}$ is not an unbiased estimator of σ, we can show that it is a *consistent* estimator of σ (see Appendix C), and it will serve our purposes well.

The estimate $\hat{\sigma}$ is interesting since it is an estimate of the standard deviation in the unobservables affecting y; equivalently, it estimates the standard deviation in y after the effect of x has been taken out. Most regression packages report the value of $\hat{\sigma}$ along with the R-squared, intercept, slope, and other OLS statistics (under one of the several names listed above). For now, our primary interest is in using $\hat{\sigma}$ to estimate the standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$. Since $\text{sd}(\hat{\beta}_1) = \sigma/\text{s}_x$, the natural estimator of $\text{sd}(\hat{\beta}_1)$ is

$$\text{se}(\hat{\beta}_1) = \hat{\sigma}/s_x = \hat{\sigma}/\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{1/2};$$

this is called the **standard error of $\hat{\beta}_1$**. Note that $\text{se}(\hat{\beta}_1)$ is viewed as a random variable when we think of running OLS over different samples of y; this is because $\hat{\sigma}$ varies with different samples. For a given sample, $\text{se}(\hat{\beta}_1)$ is a number, just as $\hat{\beta}_1$ is simply a number when we compute it from the given data.

Similarly, $\text{se}(\hat{\beta}_0)$ is obtained from $\text{sd}(\hat{\beta}_0)$ by replacing σ with $\hat{\sigma}$. The standard error of any estimate gives us an idea of how precise the estimator is. Standard errors play a central role throughout this text; we will use them to construct test statistics and confidence intervals for every econometric procedure we cover, starting in Chapter 4.

2.6 REGRESSION THROUGH THE ORIGIN

In rare cases, we wish to impose the restriction that, when $x = 0$, the expected value of y is zero. There are certain relationships for which this is reasonable. For example, if income (x) is zero, then income tax revenues (y) must also be zero. In addition, there are problems where a model that originally has a nonzero intercept is transformed into a model without an intercept.

Formally, we now choose a slope estimator, which we call $\tilde{\beta}_1$, and a line of the form

$$\tilde{y} = \tilde{\beta}_1 x, \tag{2.63}$$

where the tildas over $\tilde{\beta}_1$ and y are used to distinguish this problem from the much more common problem of estimating an intercept along with a slope. Obtaining (2.63) is called **regression through the origin** because the line (2.63) passes through the point $x = 0, \tilde{y} = 0$. To obtain the slope estimate in (2.63), we still rely on the method of ordinary least squares, which in this case minimizes the sum of squared residuals

$$\sum_{i=1}^{n} (y_i - \tilde{\beta}_1 x_i)^2. \tag{2.64}$$

Using calculus, it can be shown that $\tilde{\beta}_1$ must solve the first order condition

$$\sum_{i=1}^{n} x_i(y_i - \tilde{\beta}_1 x_i) = 0. \tag{2.65}$$

From this we can solve for $\tilde{\beta}_1$:

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}, \tag{2.66}$$

provided that not all the x_i are zero, a case we rule out.

Note how $\tilde{\beta}_1$ compares with the slope estimate when we also estimate the intercept (rather than set it equal to zero). These two estimates are the same if, and only if, $\bar{x} = 0$. (See equation (2.49) for $\hat{\beta}_1$.) Obtaining an estimate of β_1 using regression through the origin is not done very often in applied work, and for good reason: if the intercept $\beta_0 \neq 0$ then $\tilde{\beta}_1$ is a biased estimator of β_1. You will be asked to prove this in Problem 2.8.

SUMMARY

We have introduced the simple linear regression model in this chapter, and we have covered its basic properties. Given a random sample, the method of ordinary least squares is used to estimate the slope and intercept parameters in the population model. We have demonstrated the algebra of the OLS regression line, including computation of fitted values and residuals, and the obtaining of predicted changes in the dependent variable for a given change in the independent variable. In Section 2.4, we discussed two issues of practical importance: (1) the behavior of the OLS estimates when we change the units of measurement of the dependent variable or the independent variable; (2) the use of the natural log to allow for constant elasticity and constant semi-elasticity models.

In Section 2.5, we showed that, under the four Assumptions SLR.1 through SLR.4, the OLS estimators are unbiased. The key assumption is that the error term u has zero mean given any value of the independent variable x. Unfortunately, there are reasons to think this is false in many social science applications of simple regression, where the omitted factors in u are often correlated with x. When we add the assumption that the variance of the error given x is constant, we get simple formulas for the sampling variances of the OLS estimators. As we saw, the variance of the slope estimator $\hat{\beta}_1$ increases as the error variance increases, and it decreases when there is more sample variation in the independent variable. We also derived an unbiased estimator for $\sigma^2 = \text{Var}(u)$.

In Section 2.6, we briefly discussed regression through the origin, where the slope estimator is obtained under the assumption that the intercept is zero. Sometimes this is useful, but it appears infrequently in applied work.

Much work is left to be done. For example, we still do not know how to test hypotheses about the population parameters, β_0 and β_1. Thus, although we know that OLS is unbiased for the population parameters under Assumptions SLR.1 through SLR.4, we have no way of drawing inference about the population. Other topics, such as the efficiency of OLS relative to other possible procedures, have also been omitted.

The issues of confidence intervals, hypothesis testing, and efficiency are central to multiple regression analysis as well. Since the way we construct confidence intervals and test statistics is very similar for multiple regression—and because simple regression is a special case of multiple regression—our time is better spent moving on to multiple regression, which is much more widely applicable than simple regression. Our purpose in Chapter 2 was to get you thinking about the issues that arise in econometric analysis in a fairly simple setting.

KEY TERMS

Coefficient of Determination
Constant Elasticity Model
Control Variable
Covariate
Degrees of Freedom
Dependent Variable
Elasticity
Error Term (Disturbance)
Error Variance
Explained Sum of Squares (SSE)
Explained Variable
Explanatory Variable
First Order Conditions
Fitted Value
Heteroskedasticity
Homoskedasticity
Independent Variable
Intercept Parameter
Ordinary Least Squares (OLS)
OLS Regression Line
Population Regression Function (PRF)
Predicted Variable
Predictor Variable
Regressand
Regression Through the Origin
Regressor
Residual
Residual Sum of Squares (SSR)
Response Variable
R-squared
Sample Regression Function (SRF)
Semi-elasticity
Simple Linear Regression Model
Slope Parameter
Standard Error of $\hat{\beta}_1$
Standard Error of the Regression (SER)
Sum of Squared Residuals
Total Sum of Squares (SST)
Zero Conditional Mean Assumption

PROBLEMS

2.1 Let *kids* denote the number of children ever born to a woman, and let *educ* denote years of education for the woman. A simple model relating fertility to years of education is

$$kids = \beta_0 + \beta_1 educ + u,$$

where u is the unobserved error.

(i) What kinds of factors are contained in u? Are these likely to be correlated with level of education?

(ii) Will a simple regression analysis uncover the ceteris paribus effect of education on fertility? Explain.

2.2 In the simple linear regression model $y = \beta_0 + \beta_1 x + u$, suppose that $E(u) \neq 0$. Letting $\alpha_0 = E(u)$, show that the model can always be rewritten with the same slope, but a new intercept and error, where the new error has a zero expected value.

2.3 The following table contains the *ACT* scores and the *GPA* (grade point average) for 8 college students. Grade point average is based on a four-point scale and has been rounded to one digit after the decimal.

Student	*GPA*	*ACT*
1	2.8	21
2	3.4	24
3	3.0	26
4	3.5	27
5	3.6	29
6	3.0	25
7	2.7	25
8	3.7	30

(i) Estimate the relationship between *GPA* and *ACT* using OLS; that is, obtain the intercept and slope estimates in the equation

$$\widehat{GPA} = \hat{\beta}_0 + \hat{\beta}_1 ACT.$$

Comment on the direction of the relationship. Does the intercept have a useful interpretation here? Explain. How much higher is the *GPA* predicted to be, if the *ACT* score is increased by 5 points?

(ii) Compute the fitted values and residuals for each observation and verify that the residuals (approximately) sum to zero.

(iii) What is the predicted value of *GPA* when $ACT = 20$?

(iv) How much of the variation in *GPA* for these 8 students is explained by *ACT*? Explain.

2.4 The data set BWGHT.RAW contains data on births to women in the United States. Two variables of interest are the dependent variable, infant birth weight in ounces (*bwght*), and an explanatory variable, average number of cigarettes the mother smoked

per day during pregnancy (*cigs*). The following simple regression was estimated using data on $n = 1388$ births:

$$\widehat{bwght} = 119.77 - 0.514\, cigs$$

(i) What is the predicted birth weight when $cigs = 0$? What about when $cigs = 20$ (one pack per day)? Comment on the difference.

(ii) Does this simple regression necessarily capture a causal relationship between the child's birth weight and the mother's smoking habits? Explain.

2.5 In the linear consumption function

$$\widehat{cons} = \hat{\beta}_0 + \hat{\beta}_1 inc,$$

the (estimated) *marginal propensity to consume* (MPC) out of income is simply the slope, $\hat{\beta}_1$, while the *average propensity to consume* (APC) is $\widehat{cons}/inc = \hat{\beta}_0/inc + \hat{\beta}_1$. Using observations for 100 families on annual income and consumption (both measured in dollars), the following equation is obtained:

$$\widehat{cons} = -124.84 + 0.853\, inc$$

$$n = 100,\ R^2 = 0.692$$

(i) Interpret the intercept in this equation and comment on its sign and magnitude.

(ii) What is predicted consumption when family income is $30,000?

(iii) With *inc* on the *x*-axis, draw a graph of the estimated MPC and APC.

2.6 Using data from 1988 for houses sold in Andover, MA, from Kiel and McClain (1995), the following equation relates housing price (*price*) to the distance from a recently built garbage incinerator (*dist*):

$$\widehat{\log(price)} = 9.40 + 0.312 \log(dist)$$

$$n = 135,\ R^2 = 0.162$$

(i) Interpret the coefficient on log(*dist*). Is the sign of this estimate what you expect it to be?

(ii) Do you think simple regression provides an unbiased estimator of the ceteris paribus elasticity of *price* with respect to *dist*? (Think about the city's decision on where to put the incinerator.)

(iii) What other factors about a house affect its price? Might these be correlated with distance from the incinerator?

2.7 Consider the savings function

$$sav = \beta_0 + \beta_1 inc + u,\ u = \sqrt{inc} \cdot e,$$

where e is a random variable with $E(e) = 0$ and $Var(e) = \sigma_e^2$. Assume that e is independent of *inc*.

(i) Show that $E(u|inc) = 0$, so that the key zero conditional mean assumption (Assumption SLR.3) is satisfied. [*Hint:* If e is independent of *inc*, then $E(e|inc) = E(e)$.]

(ii) Show that $\text{Var}(u|inc) = \sigma_e^2 inc$, so that the homoskedasticity Assumption SLR.5 is violated. In particular, the variance of *sav* increases with *inc*. [*Hint:* $\text{Var}(e|inc) = \text{Var}(e)$, if e and *inc* are independent.]

(iii) Provide a discussion that supports the assumption that the variance of savings increases with family income.

2.8 Consider the standard simple regression model $y = \beta_0 + \beta_1 x + u$ under Assumptions SLR.1 through SLR.4. Thus, the usual OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased for their respective population parameters. Let $\tilde{\beta}_1$ be the estimator of β_1 obtained by assuming the intercept is zero (see Section 2.6).

(i) Find $\text{E}(\tilde{\beta}_1)$ in terms of the x_i, β_0, and β_1. Verify that $\tilde{\beta}_1$ is unbiased for β_1 when the population intercept (β_0) is zero. Are there other cases where $\tilde{\beta}_1$ is unbiased?

(ii) Find the variance of $\tilde{\beta}_1$. (*Hint:* The variance does not depend on β_0.)

(iii) Show that $\text{Var}(\tilde{\beta}_1) \leq \text{Var}(\hat{\beta}_1)$. [Hint: For any sample of data, $\sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2$, with strict inequality unless $\bar{x} = 0$.]

(iv) Comment on the tradeoff between bias and variance when choosing between $\hat{\beta}_1$ and $\tilde{\beta}_1$.

2.9 (i) Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the intercept and slope from the regression of y_i on x_i, using n observations. Let c_1 and c_2, with $c_2 \neq 0$, be constants. Let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be the intercept and slope from the regression $c_1 y_i$ on $c_2 x_i$. Show that $\tilde{\beta}_1 = (c_1/c_2)\hat{\beta}_1$ and $\tilde{\beta}_0 = c_1\hat{\beta}_0$, thereby verifying the claims on units of measurement in Section 2.4. [*Hint:* To obtain $\tilde{\beta}_1$, plug the scaled versions of x and y into (2.19). Then, use (2.17) for $\tilde{\beta}_0$, being sure to plug in the scaled x and y *and* the correct slope.]

(ii) Now let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be from the regression $(c_1 + y_i)$ on $(c_2 + x_i)$ (with no restriction on c_1 or c_2). Show that $\tilde{\beta}_1 = \hat{\beta}_1$ and $\tilde{\beta}_0 = \hat{\beta}_0 + c_1 - c_2\hat{\beta}_1$.

COMPUTER EXERCISES

2.10 The data in 401K.RAW are a subset of data analyzed by Papke (1995) to study the relationship between participation in a 401(k) pension plan and the generosity of the plan. The variable *prate* is the percentage of eligible workers with an active account; this is the variable we would like to explain. The measure of generosity is the plan match rate, *mrate*. This variable gives the average amount the firm contributes to each worker's plan for each \$1 contribution by the worker. For example, if *mrate* = 0.50, then a \$1 contribution by the worker is matched by a 50¢ contribution by the firm.

(i) Find the average participation rate and the average match rate in the sample of plans.

(ii) Now estimate the simple regression equation

$$\widehat{prate} = \hat{\beta}_0 + \hat{\beta}_1 mrate,$$

and report the results along with the sample size and *R*-squared.

(iii) Interpret the intercept in your equation. Interpret the coefficient on *mrate*.

(iv) Find the predicted *prate* when *mrate* = 3.5. Is this a reasonable prediction? Explain what is happening here.

(v) How much of the variation in *prate* is explained by *mrate*? Is this a lot in your opinion?

2.11 The data set in CEOSAL2.RAW contains information on chief executive officers for U.S. corporations. The variable *salary* is annual compensation, in thousands of dollars, and *ceoten* is prior number of years as company CEO.

(i) Find the average salary and the average tenure in the sample.

(ii) How many CEOs are in their first year as CEO (that is, $ceoten = 0$)? What is the longest tenure as a CEO?

(iii) Estimate the simple regression model

$$\log(salary) = \beta_0 + \beta_1 ceoten + u,$$

and report your results in the usual form. What is the (approximate) predicted percentage increase in salary given one more year as a CEO?

2.12 Use the data in SLEEP75.RAW from Biddle and Hamermesh (1990) to study whether there is a tradeoff between the time spent sleeping per week and the time spent in paid work. We could use either variable as the dependent variable. For concreteness, estimate the model

$$sleep = \beta_0 + \beta_1 totwrk + u,$$

where *sleep* is minutes spent sleeping at night per week and *totwrk* is total minutes worked during the week.

(i) Report your results in equation form along with the number of observations and R^2. What does the intercept in this equation mean?

(ii) If *totwrk* increases by 2 hours, by how much is *sleep* estimated to fall? Do you find this to be a large effect?

2.13 Use the data in WAGE2.RAW to estimate a simple regression explaining monthly salary (*wage*) in terms of IQ score (*IQ*).

(i) Find the average salary and average IQ in the sample. What is the standard deviation of IQ? (IQ scores are standardized so that the average in the population is 100 with a standard deviation equal to 15.)

(ii) Estimate a simple regression model where a one-point increase in *IQ* changes *wage* by a constant dollar amount. Use this model to find the predicted increase in wage for an increase in *IQ* of 15 points. Does *IQ* explain most of the variation in *wage*?

(iii) Now estimate a model where each one-point increase in *IQ* has the same percentage effect on *wage*. If *IQ* increases by 15 points, what is the approximate percentage increase in predicted *wage*?

2.14 For the population of firms in the chemical industry, let *rd* denote annual expenditures on research and development, and let *sales* denote annual sales (both are in millions of dollars).

(i) Write down a model (not an estimated equation) that implies a constant elasticity between *rd* and *sales*. Which parameter is the elasticity?

(ii) Now estimate the model using the data in RDCHEM.RAW. Write out the estimated equation in the usual form. What is the estimated elasticity of *rd* with respect to *sales*? Explain in words what this elasticity means.

APPENDIX 2A

Minimizing the Sum of Squared Residuals

We show that the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ do minimize the sum of squared residuals, as asserted in Section 2.2. Formally, the problem is to characterize the solutions $\hat{\beta}_0$ and $\hat{\beta}_1$ to the minimization problem

$$\min_{b_0,b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2,$$

where b_0 and b_1 are the dummy arguments for the optimization problem; for simplicity, call this function $Q(b_0,b_1)$. By a fundamental result from multivariable calculus (see Appendix A), a necessary condition for $\hat{\beta}_0$ and $\hat{\beta}_1$ to solve the minimization problem is that the partial derivatives of $Q(b_0,b_1)$ with respect to b_0 and b_1 must be zero when evaluated at $\hat{\beta}_0$, $\hat{\beta}_1$: $\partial Q(\hat{\beta}_0,\hat{\beta}_1)/\partial b_0 = 0$ and $\partial Q(\hat{\beta}_0,\hat{\beta}_1)/\partial b_1 = 0$. Using the chain rule from calculus, these two equations become

$$-2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

$$-2 \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

These two equations are just (2.14) and (2.15) multiplied by $-2n$ and, therefore, are solved by the same $\hat{\beta}_0$ and $\hat{\beta}_1$.

How do we know that we have actually minimized the sum of squared residuals? The first order conditions are necessary but not sufficient conditions. One way to verify that we have minimized the sum of squared residuals is to write, for any b_0 and b_1,

$$\begin{aligned} Q(b_0,b_1) &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i + (\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1)x_i)^2 \\ &= \sum_{i=1}^{n} (\hat{u}_i + (\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1)x_i)^2 \\ &= \sum_{i=1}^{n} \hat{u}_i^2 + n(\hat{\beta}_0 - b_0)^2 + (\hat{\beta}_1 - b_1)^2 \sum_{i=1}^{n} x_i^2 + 2(\hat{\beta}_0 - b_0)(\hat{\beta}_1 - b_1) \sum_{i=1}^{n} x_i, \end{aligned}$$

where we have used equations (2.30) and (2.31). The sum of squared residuals does not depend on b_0 or b_1, while the sum of the last three terms can be written as

$$\sum_{i=1}^{n} [(\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1)x_i]^2,$$

as can be verified by straightforward algebra. Because this is a sum of squared terms, it can be at most zero. Therefore, it is smallest when $b_0 = \hat{\beta}_0$ and $b_1 = \hat{\beta}_1$.