CHAPTER 2

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 will 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, representing 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; and 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

$$y = \beta_0 + \beta_1 x + u. \qquad (2.1)$$

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 in applications that are purely about prediction and not causality. Our terminology for simple regression is summarized in Table 2.1.

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, sometimes called the *constant term*, also has its uses, although it is rarely central to an analysis.

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

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 ethic, 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?

Section 2.5 will show that 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, we can always make one assumption about u. 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,

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

Assumption (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 should 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 interpreting 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 assumption as

$$\mathrm{E}(u|x) = \mathrm{E}(u). \qquad \textbf{2.6}$$

Equation (2.6) says that the average value of the unobservables is the same across all slices of the population determined by the value of x and that the common average is necessarily equal to the average of u over the entire population. When assumption (2.6) holds, we say that u is **mean independent** of x. (Of course, mean independence is implied by full independence between u and x, an assumption often used in basic probability and statistics.) When we combine mean independence with assumption (2.5), we obtain the **zero conditional mean assumption**, $\mathrm{E}(u|x) = 0$. It is critical to remember that equation (2.6) is the assumption with impact; assumption (2.5) essentially defines the intercept, β_0.

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 $\mathrm{E}(abil|8)$ denotes the average ability for the group of all people with eight years of education, and $\mathrm{E}(abil|16)$ denotes the average ability among people in the population with sixteen 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 relying on simple regression analysis.

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). Then

$$score = \beta_0 + \beta_1 attend + u. \qquad (2.7)$$

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

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.

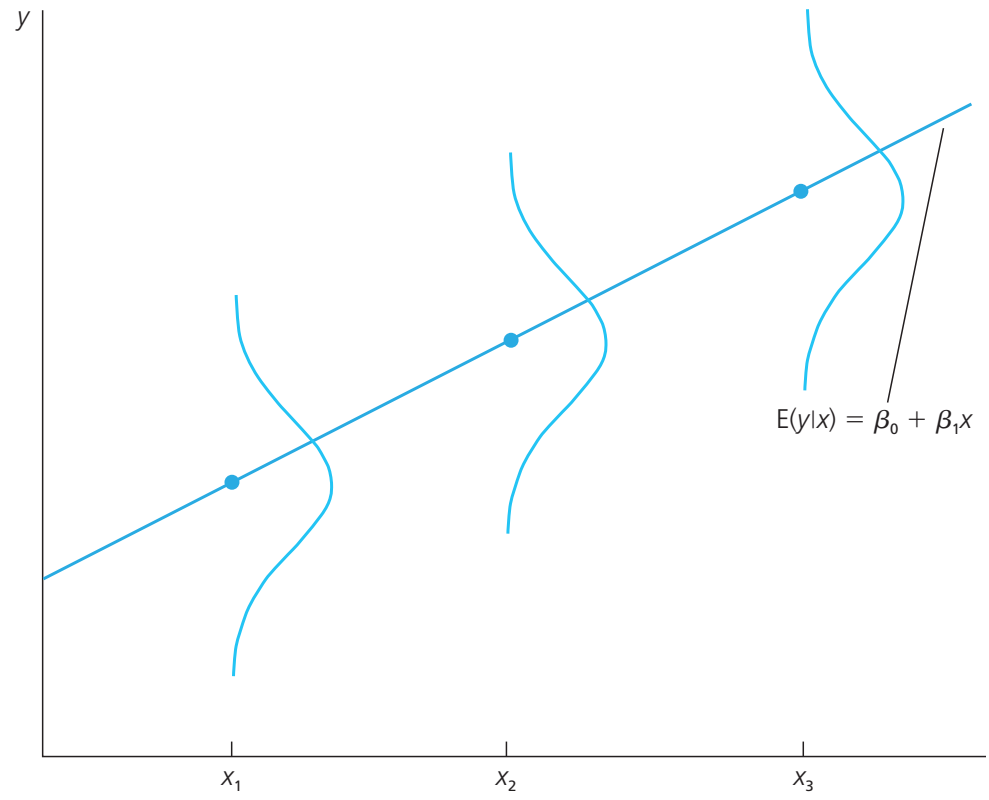
The zero conditional mean assumption gives β_1 another interpretation that is often useful. Taking the expected value of (2.1) conditional on x and using $E(u|x) = 0$ gives

$$E(y|x) = \beta_0 + \beta_1 x. \qquad (2.8)$$

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 *expected 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.

FIGURE 2.1

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



It is important to understand that equation (2.8) tells us how the *average* value of y changes with x; it does not say that y equals $\beta_0 + \beta_1 x$ for all units in the population. For example, suppose that x is the high school grade point average and y is the college GPA, and we happen to know that $E(colGPA|hsGPA) = 1.5 + 0.5\, hsGPA$. [Of course, in practice, we never know the population intercept and slope, but it is useful to pretend momentarily that we do to understand the nature of equation (2.8).] This GPA equation tells us the *average* college GPA among all students who have a given high school GPA. So suppose that $hsGPA = 3.6$. Then the average *colGPA* for all high school graduates who attend college with $hsGPA = 3.6$ is $1.5 + 0.5(3.6) = 3.3$. We are certainly *not* saying that every student with $hsGPA = 3.6$ will have a 3.3 college GPA; this is clearly false. The PRF gives us a relationship between the average level of y at different levels of x. Some students with $hsGPA = 3.6$ will have a college GPA higher than 3.3, and some will have a lower college GPA. Whether the actual *colGPA* is above or below 3.3 depends on the unobservable factors in u, and those differ among students even within the slice of the population with $hsGPA = 3.6$.

Given the zero conditional mean assumption $E(u|x) = 0$, it is useful to view equation (2.1) as breaking y into two components. The piece $\beta_0 + \beta_1 x$, which represents $E(y|x)$, is 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. In Chapter 3, when we introduce more than one explanatory variable, we will discuss how to determine how large the systematic part is relative to the unsystematic part.

In the next section, we will use assumptions (2.5) and (2.6) to motivate estimators of β_0 and β_1 given a random sample of data. The zero conditional mean assumption also plays a crucial role in the statistical analysis in Section 2.6.

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. Because 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 because it contains all factors affecting y_i other than x_i.

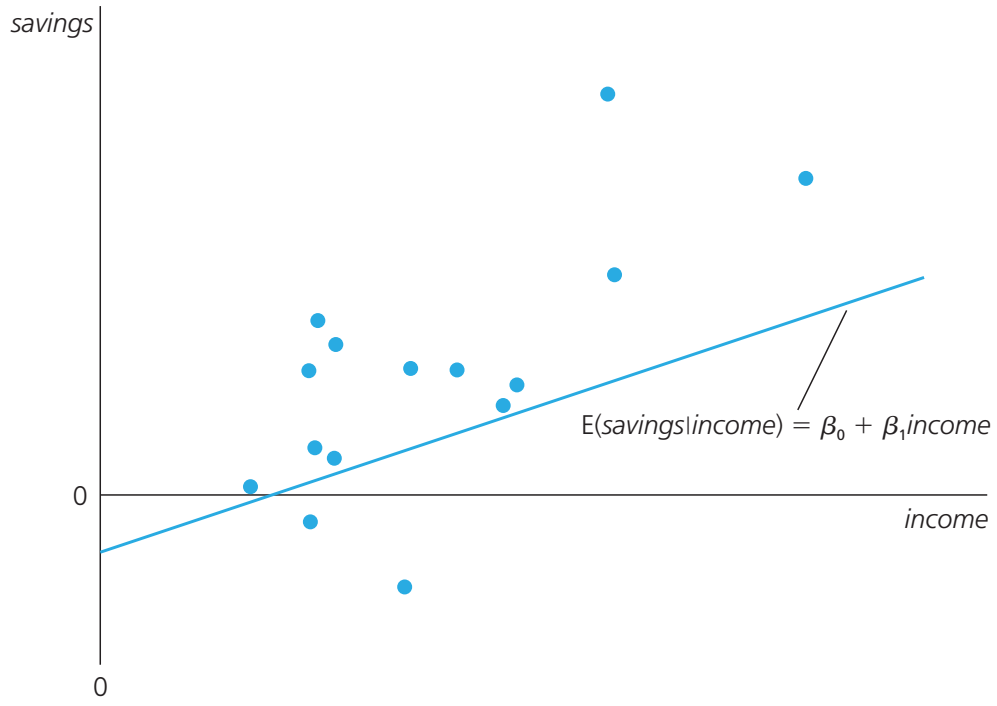
As an example, x_i might be the annual income and y_i the annual savings for family i during a particular year. If we have collected data on fifteen families, then $n = 15$. A scatterplot of such a data set is given in Figure 2.2, along with the (necessarily fictitious) population regression function.

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 is uncorrelated

FIGURE 2.2

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



with x. Therefore, we see that u has zero expected value and that the *covariance* between x and u is zero:

$$E(u) = 0 \tag{2.10}$$

and

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

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

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

and

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

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 estimators 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):

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

and

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

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

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

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 $\bar{x}$:

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

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

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

the estimated slope is

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

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.

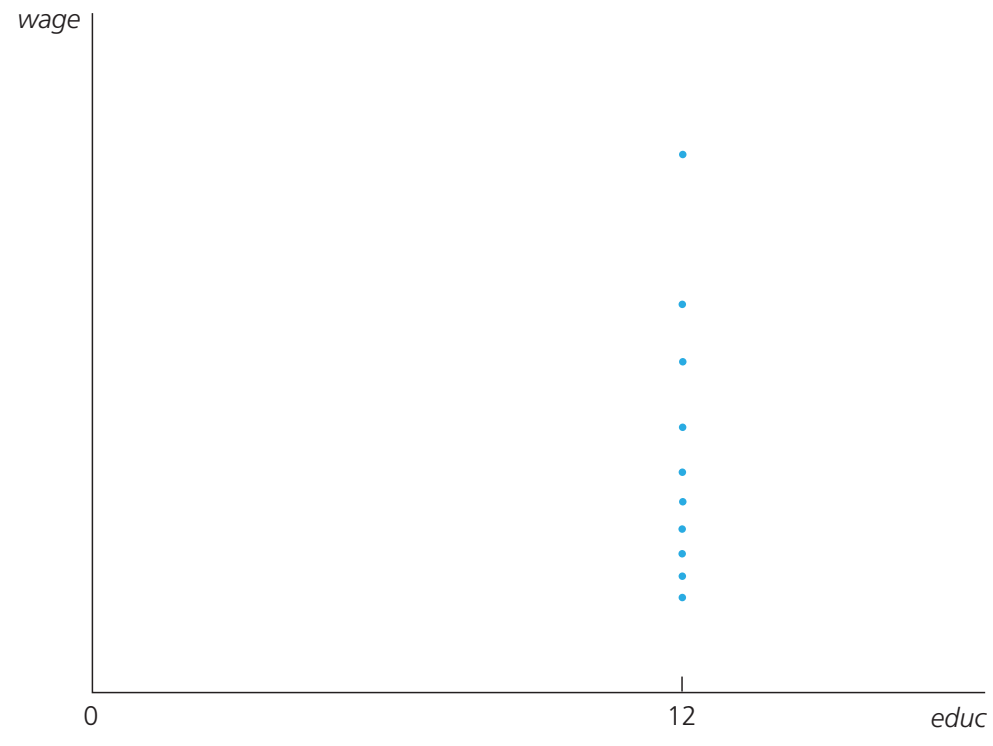
Although the method for obtaining (2.17) and (2.19) is motivated by (2.6), the only assumption needed to compute the estimates for a particular sample is (2.18). This is hardly an assumption at all: (2.18) is true provided the x_i in the sample are not all equal to the same value. If (2.18) fails, then 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 estimates can be computed.

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 $x = x_i$ as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \qquad \textbf{2.20}$$

FIGURE 2.3

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



This is the value we predict for y when $x = x_i$ for the given intercept and slope. 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:

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

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.

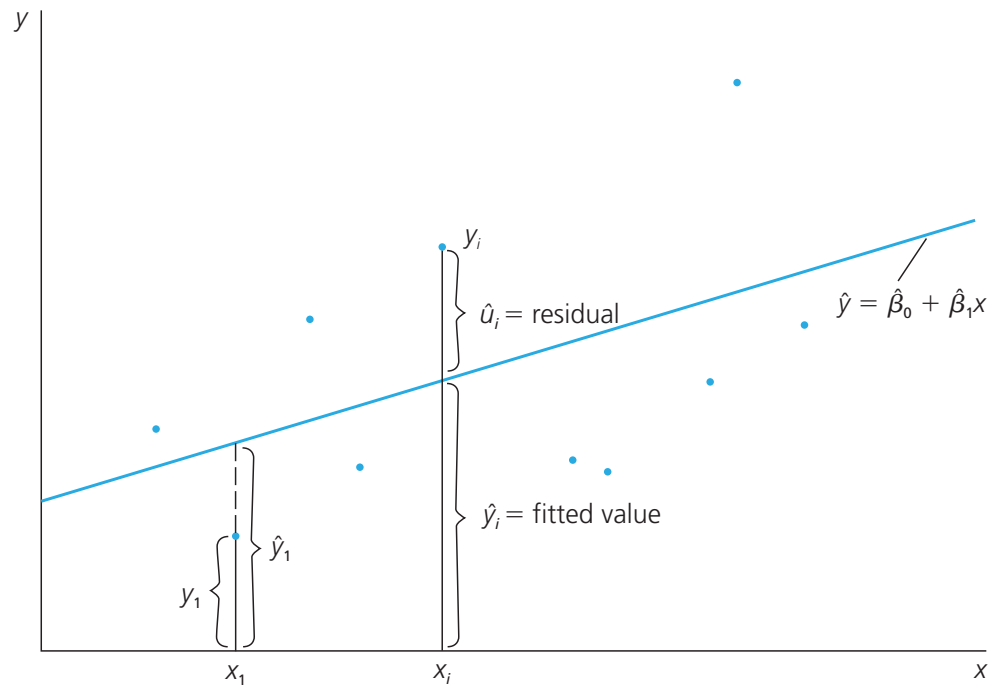
Now, suppose we choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to make the **sum of squared residuals**,

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

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.

FIGURE 2.4

Fitted values and residuals.



When we view ordinary least squares as minimizing the sum of squared residuals, it is natural to ask: Why not minimize some other function of the residuals, such as the absolute values of the residuals? In fact, as we will discuss in the more advanced Section 9.4, minimizing the sum of the absolute values of the residuals is sometimes very useful. But it does have some drawbacks. First, we cannot obtain formulas for the resulting estimators; given a data set, the estimates must be obtained by numerical optimization routines. As a consequence, the statistical theory for estimators that minimize the sum of the absolute residuals is very complicated. Minimizing other functions of the residuals, say, the sum of the residuals each raised to the fourth power, has similar drawbacks. (We would never choose our estimates to minimize, say, the sum of the residuals themselves, as residuals large in magnitude but with opposite signs would tend to cancel out.) With OLS, we will be able to derive unbiasedness, consistency, and other important statistical properties relatively easily. Plus, as the motivation in equations (2.13) and (2.14) suggests, and as we will see in Section 2.5, OLS is suited for estimating the parameters appearing in the conditional mean function (2.8).

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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{2.23}$$

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 $\mathrm{E}(y|x) = \beta_0 + \beta_1 x$. It is important to remember that the PRF is something fixed, but unknown, in the population. Because 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 econometrics 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 $y = 1{,}452.6$ indicates a salary of \$1,452,600. Let x be the average return on 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%.

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%, with the smallest and largest values being 0.5 and 56.3%, respectively.

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

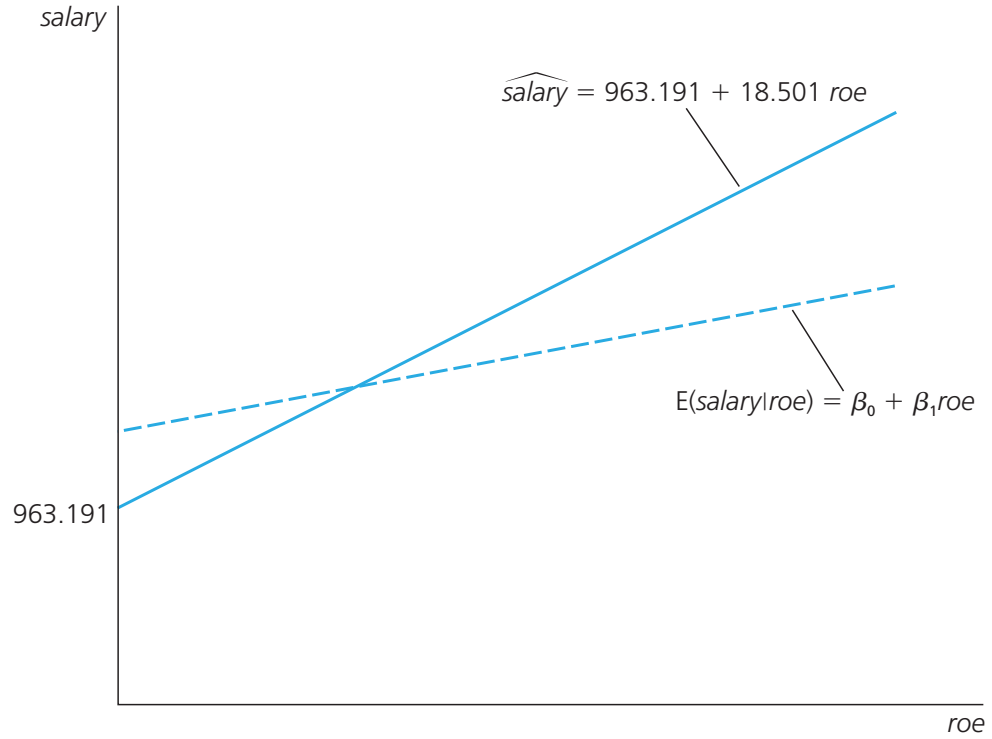
$$\widehat{salary} = 963.191 + 18.501\, roe, \qquad (2.26)$$

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 $\widehat{salary} = 963.191 + 18.501(30) = 1{,}518{,}221$, which is just over \$1.5 million. However, this does *not* mean that a particular CEO whose firm had a $roe = 30$ earns \$1,518,221. Many other factors 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 $\widehat{salary} = 963.191 + 18.501\ roe$ and the (unknown) population regression function.



Example 2.4

[Wage and Education]

For the population of people in the workforce 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 \$19.06 in 2003 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 $-90¢$ an hour. This, of course, is silly. It turns out that only 18 people in the sample of 526 have less than eight years of education. Consequently, it

is not surprising that the regression line does poorly at very low levels of education. 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 2003 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¢ 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 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

$$\widehat{voteA} = 26.81 + 0.464\, shareA. \qquad (2.28)$$

This means that if Candidate A's share of spending increases by one percentage point, Candidate A receives almost one-half a percentage point (0.464) more of the total vote. Whether or not this is a causal effect is unclear, but it is not unbelievable. If *shareA* = 50, *voteA* is predicted to be about 50, or half the vote.

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 Computer Exercise C2.3, 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 Terminology

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*

$$y \text{ on } x, \qquad (2.29)$$

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 Properties of OLS on Any Sample of Data

In the previous section, we went through the algebra of deriving the formulas for the OLS intercept and slope estimates. In this section, we cover some further algebraic properties of the fitted OLS regression line. The best way to think about these properties is to remember that they hold, by construction, for *any* sample of data. The harder task—considering the properties of OLS across all possible random samples of data—is postponed until 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
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) showed 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. These figures do not exactly agree because 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 to little useful effect.

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

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

From property (1), 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

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

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

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

R^2 is the ratio of the explained variation compared to the total variation; 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, because 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:

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

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

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 percent of the variation in salaries for this sample of 209 CEOs. That means that 98.7 percent of the salary variations for these CEOs is left unexplained! This lack of explanatory power may not be too surprising because many other characteristics of both the firm and the individual CEO 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), $R^2 = 0.856$. Thus, the share of campaign expenditures explains over 85% of the variation in the election outcomes for this sample. This is a 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 percentage (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·*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*?

$$\widehat{salary} = 963.191 + 1{,}850.1\, roedec. \qquad \textbf{2.41}$$

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 $\Delta\widehat{salary} = 1{,}850.1(0.01) = 18.501$, 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 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 percentage 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.

Probably a better characterization of how wage changes with education is that each year of education increases wage by a constant *percentage*. For example, an increase in education from 5 years to 6 years increases wage by, say, 8% (ceteris paribus), and an increase in education from 11 to 12 years also increases wage by 8%. A model that gives (approximately) a constant percentage effect is

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

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

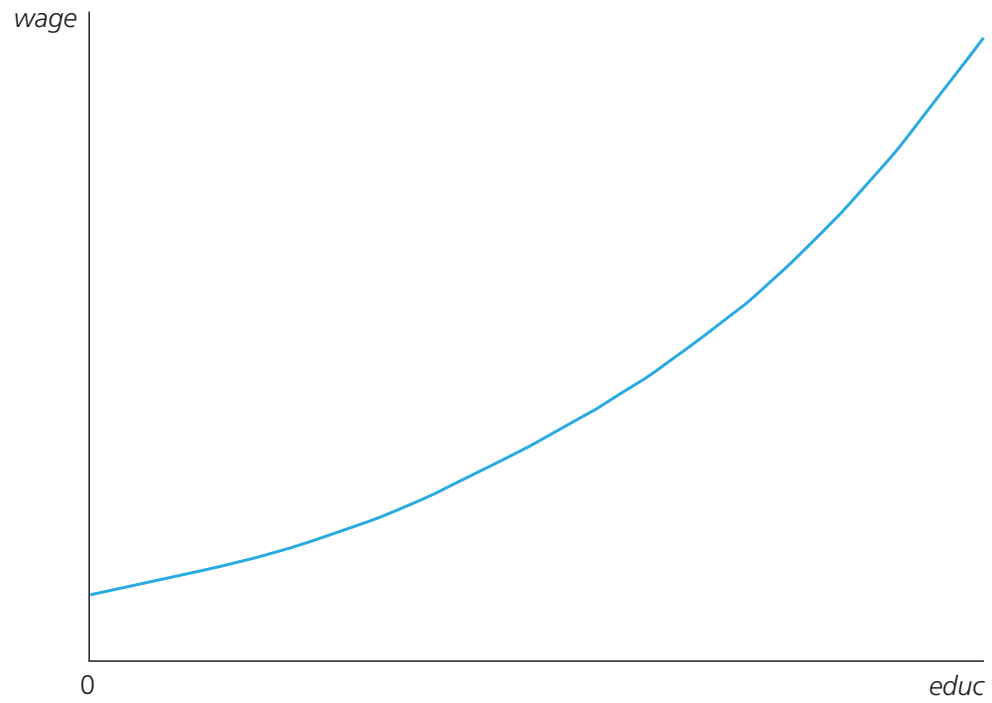
$$\%\Delta wage \approx (100 \cdot \beta_1)\Delta educ. \qquad \textbf{2.43}$$

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 $u = 0$.

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 $educ$.

FIGURE 2.6

***wage* = exp(β_0 + β_1*educ*), with $\beta_1 > 0$.**



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: $\widehat{wage}$ increases by 8.3% 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, because it gives the predicted log(*wage*), when *educ* = 0. The *R*-squared shows that *educ* explains about 18.6% 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

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

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

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

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

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

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

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	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

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 2.10 and 2.11, 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.

Whereas 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.

Plenty of models *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

$$y = \beta_0 + \beta_1 x + u, \qquad (2.47)$$

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 have a random sample of size *n*, $\{(x_i, y_i): i = 1, 2, \ldots, n\}$, following the population model in equation (2.47).

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.

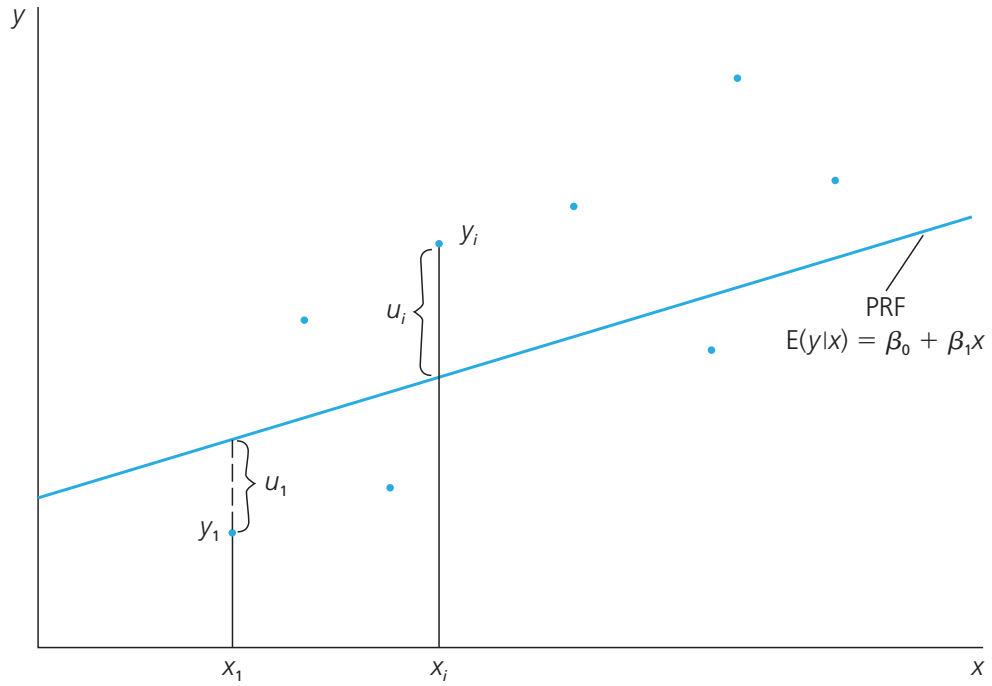
We can write (2.47) in terms of the random sample as

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, 2, \ldots, n, \qquad (2.48)$$

where u_i is the error or disturbance for observation *i* (for example, person *i*, firm *i*, city *i*, and so on). Thus, u_i contains the unobservables for observation *i* that 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

FIGURE 2.7

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



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.

As we already saw in Section 2.2, the OLS slope and intercept estimates are not defined unless we have some sample variation in the explanatory variable. We now add variation in the x_i to our list of assumptions.

Assumption SLR.3 (Sample Variation in the Explanatory Variable)

The sample outcomes on x, namely, $\{x_i, i = 1, \ldots, n\}$, are not all the same value.

This is a very weak assumption—certainly not worth emphasizing, but needed nevertheless. If x varies in the population, random samples on x will typically contain variation, unless the population variation is minimal or the sample size is small. Simple inspection of summary statistics on x_i reveals whether Assumption SLR.3 fails: if the sample standard deviation of x_i is zero, then Assumption SLR.3 fails; otherwise, it holds.

Finally, 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.4 (Zero Conditional Mean)

The error u has an expected value of zero given any value of the explanatory variable. In other words,

$$E(u|x) = 0.$$

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*, which we think of as follows. 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 $x_1, x_2, \ldots, x_n$. Then another sample of y is obtained, again using the same $x_1, x_2, \ldots, x_n$. 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.4.

Now, we are ready to show that the OLS estimators are unbiased. To this end, we use 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) to write the OLS slope estimator in equation (2.19) as

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

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\text{SST}_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\text{SST}_x}, \qquad \textbf{2.50}$$

where we have defined the total variation in x_i as $\text{SST}_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$ 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

$$\sum_{i=1}^{n}(x_i - \bar{x})\beta_0 + \sum_{i=1}^{n}(x_i - \bar{x})\beta_1 x_i + \sum_{i=1}^{n}(x_i - \bar{x})u_i \qquad \textbf{2.51}$$
$$= \beta_0 \sum_{i=1}^{n}(x_i - \bar{x}) + \beta_1 \sum_{i=1}^{n}(x_i - \bar{x})x_i + \sum_{i=1}^{n}(x_i - \bar{x})u_i.$$

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

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\text{SST}_x} = \beta_1 + (1/\text{SST}_x)\sum_{i=1}^{n} d_i u_i, \qquad \textbf{2.52}$$

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,

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

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. Because SST_x and d_i are functions only of the x_i, they are nonrandom in the conditioning. Therefore, from (2.52), and keeping the conditioning on $\{x_1, x_2, \ldots, x_n\}$ implicit, we have

$$\text{E}(\hat{\beta}_1) = \beta_1 + \text{E}\left[(1/\text{SST}_x)\sum_{i=1}^{n} d_i u_i\right] = \beta_1 + (1/\text{SST}_x)\sum_{i=1}^{n}\text{E}(d_i u_i)$$

$$= \beta_1 + (1/\text{SST}_x)\sum_{i=1}^{n} d_i \text{E}(u_i) = \beta_1 + (1/\text{SST}_x)\sum_{i=1}^{n} d_i \cdot 0 = \beta_1,$$

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.4. Since unbiasedness holds for any outcome on $\{x_1, x_2, \ldots, x_n\}$, unbiasedness also holds without conditioning on $\{x_1, x_2, \ldots, x_n\}$.

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.4. 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. 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.

As we have already discussed, Assumption SLR.3 almost always holds in interesting regression applications. Without it, we cannot even obtain the OLS estimates.

The assumption we should concentrate on for now is SLR.4. If SLR.4 holds, the OLS estimators are unbiased. Likewise, if SLR.4 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

$$math10 = \beta_0 + \beta_1 lnchprg + u, \qquad (2.54)$$

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 1992–1993 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. Because 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, 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)

The error u has the same variance given any value of the explanatory variable. In other words,

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

We must emphasize that the homoskedasticity assumption is quite distinct from the zero conditional mean assumption, $\text{E}(u|x) = 0$. Assumption SLR.4 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 $\text{E}(u|x) = \text{E}(u) = 0$ and $\text{Var}(u|x) = \sigma^2$. But independence is sometimes too strong of an assumption.

Because $\text{Var}(u|x) = \text{E}(u^2|x) - [\text{E}(u|x)]^2$ and $\text{E}(u|x) = 0$, $\sigma^2 = \text{E}(u^2|x)$, which means σ^2 is also the *unconditional* expectation of u^2. Therefore, $\sigma^2 = \text{E}(u^2) = \text{Var}(u)$, because $\text{E}(u) = 0$. 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.4 and SLR.5 in terms of the conditional mean and conditional variance of y:

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

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

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 $\beta_1 > 0$.

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

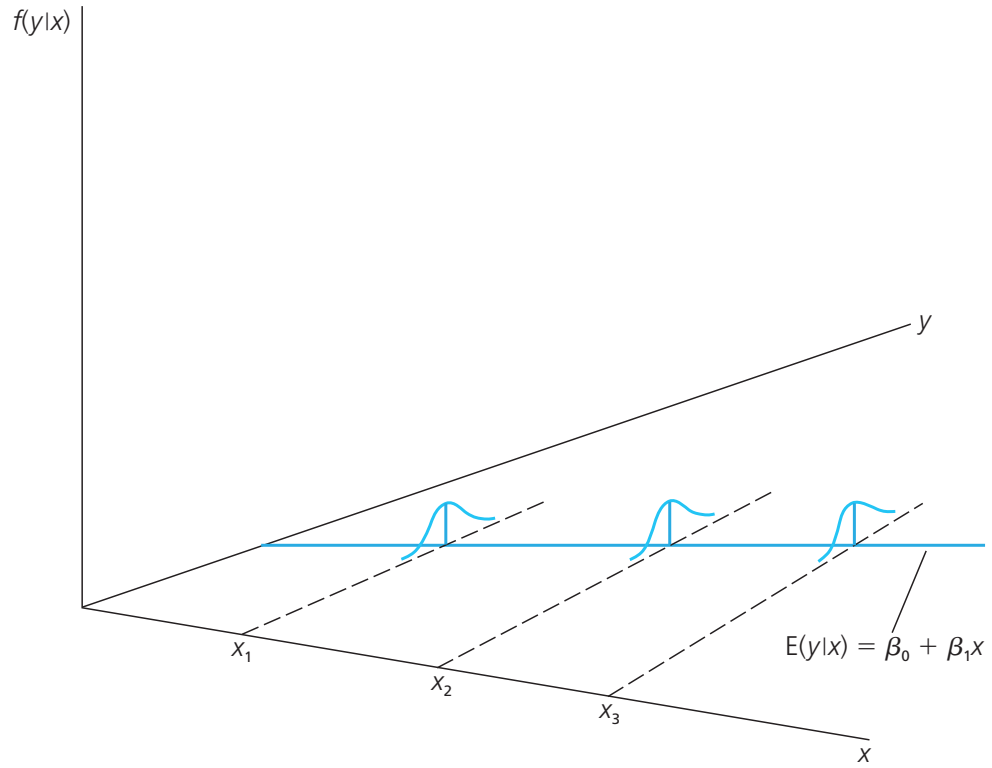
Example 2.13

[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 $\text{E}(u|educ) = 0$, and this implies $\text{E}(wage|educ) = \beta_0 + \beta_1 educ$. 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 estimating—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 fewer 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 2.8

The simple regression model under homoskedasticity.



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,

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2/\text{SST}_x, \tag{2.57}$$

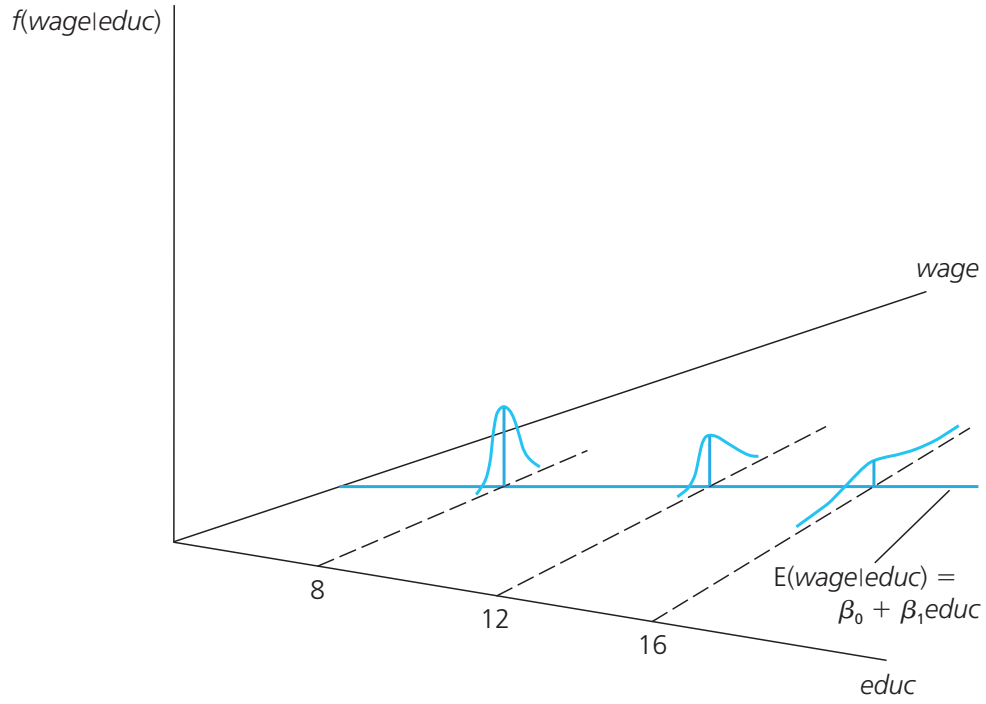
and

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

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

FIGURE 2.9

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



PROOF: We derive the formula for $\text{Var}(\hat{\beta}_1)$, leaving the other derivation as Problem 2.10. The starting point is equation (2.52): $\hat{\beta}_1 = \beta_1 + (1/\text{SST}_x)\sum_{i=1}^{n} d_i u_i$. Because β_1 is just a constant, and we are conditioning on the x_i, SST_x 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

$$\text{Var}(\hat{\beta}_1) = (1/\text{SST}_x)^2\text{Var}\left(\sum_{i=1}^{n} d_i u_i\right) = (1/\text{SST}_x)^2\left(\sum_{i=1}^{n} d_i^2\text{Var}(u_i)\right)$$

$$= (1/\text{SST}_x)^2\left(\sum_{i=1}^{n} d_i^2\sigma^2\right) \qquad [\text{since } \text{Var}(u_i) = \sigma^2 \text{ for all } i]$$

$$= \sigma^2(1/\text{SST}_x)^2\left(\sum_{i=1}^{n} d_i^2\right) = \sigma^2(1/\text{SST}_x)^2\text{SST}_x = \sigma^2/\text{SST}_x,$$

which is what we wanted to show.

Equations (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 $\{x_1, x_2, \ldots, x_n\}$, SST_x. 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 $\text{E}(y|x)$ 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 $\text{E}(y|x)$ 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 β_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, $\text{sd}(\hat{\beta}_1) = \sigma/\sqrt{\text{SST}_x}$, where σ is the square root of σ^2, and $\sqrt{\text{SST}_x}$ is the square root of SST_x.

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 *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

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

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, we 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). Because 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:

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

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 as 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

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \qquad \textbf{2.62}$$

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 because 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 $\text{sd}(\hat{\beta}_1) = \sigma/\sqrt{\text{SST}_x}$, the natural estimator of $\text{sd}(\hat{\beta}_1)$ is

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

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 true 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 settings 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

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

where the tildes over $\tilde{\beta}_1$ and $\tilde{y}$ 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:

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

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

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

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

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

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 and (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.

The Gauss-Markov Assumptions for Simple Regression

For convenience, we summarize the **Gauss-Markov assumptions** that we used in this chapter. It is important to remember that only SLR.1 through SLR.4 are needed to show $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased. We added the homoskedasticity assumption, SLR.5, to obtain the usual OLS variance formulas (2.57) and (2.58).

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

$$y = \beta_0 + \beta_1 x + u,$$

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

Assumption SLR.2 (Random Sampling)

We have a random sample of size n, $\{(x_i,y_i): i = 1, 2, \ldots, n\}$, following the population model in Assumption SLR.1.

Assumption SLR.3 (Sample Variation in the Explanatory Variable)

The sample outcomes on x, namely, $\{x_i, \text{i} = 1, \ldots, n\}$, are not all the same value.

Assumption SLR.4 (Zero Conditional Mean)

The error u has an expected value of zero given any value of the explanatory variable. In other words,

$$\text{E}(u|x) = 0.$$

Assumption SLR.5 (Homoskedasticity)

The error u has the same variance given any value of the explanatory variable. In other words,

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

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