CHAPTER 4

Linear Regression with One Regressor

A state implements tough new penalties on drunk drivers: What is the effect on highway fatalities? A school district cuts the size of its elementary school classes: What is the effect on its students' standardized test scores? You successfully complete one more year of college classes: What is the effect on your future earnings?

All three of these questions are about the unknown effect of changing one variable, X (X being penalties for drunk driving, class size, or years of schooling), on another variable, Y (Y being highway deaths, student test scores, or earnings).

This chapter introduces the linear regression model relating one variable, X, to another, Y. This model postulates a linear relationship between X and Y: the slope of the line relating X and Y is the effect of a one-unit change in X on Y. Just as the mean of Y is an unknown characteristic of the population distribution of Y, the slope of the line relating X and Y is an unknown characteristic of the population joint distribution of X and Y. The econometric problem is to estimate this slope—that is, to estimate the effect on Y of a unit change in X—using a sample of data on these two variables.

This chapter describes methods for estimating this slope using a random sample of data on X and Y. For instance, using data on class sizes and test scores from different school districts, we show how to estimate the expected effect on test scores of reducing class sizes by, say, one student per class. The slope and the intercept of the line relating X and Y can be estimated by a method called ordinary least squares (OLS).

4.1 The Linear Regression Model

The superintendent of an elementary school district must decide whether to hire additional teachers and she wants your advice. If she hires the teachers, she will reduce the number of students per teacher (the student–teacher ratio) by two. She faces a tradeoff. Parents want smaller classes so that their children can receive more individualized attention. But hiring more teachers means spending more money, which is not to the liking of those paying the bill! So she asks you: If she cuts class sizes, what will the effect be on student performance?

In many school districts, student performance is measured by standardized tests, and the job status or pay of some administrators can depend in part on how well their students do on these tests. We therefore sharpen the superintendent's question: If she reduces the average class size by two students, what will the effect be on standardized test scores in her district?

A precise answer to this question requires a quantitative statement about changes. If the superintendent *changes* the class size by a certain amount, what would she expect the *change* in standardized test scores to be? We can write this as a mathematical relationship using the Greek letter beta, $\beta_{ClassSize}$, where the subscript "ClassSize" distinguishes the effect of changing the class size from other effects. Thus,

$$\beta_{ClassSize} = \frac{\text{change in TestScore}}{\text{change in ClassSize}} = \frac{\Delta\text{TestScore}}{\Delta\text{ClassSize}}, \tag{4.1}$$

where the Greek letter Δ (delta) stands for "change in." That is, $\beta_{ClassSize}$ is the change in the test score that results from changing the class size, divided by the change in the class size.

If you were lucky enough to know $\beta_{ClassSize}$, you would be able to tell the superintendent that decreasing class size by one student would change districtwide test scores by $\beta_{ClassSize}$. You could also answer the superintendent's actual question, which concerned changing class size by two students per class. To do so, rearrange Equation (4.1) so that

$$\Delta TestScore = \beta_{ClassSize} \times \Delta ClassSize. \tag{4.2}$$

Suppose that $\beta_{ClassSize} = -0.6$. Then a reduction in class size of two students per class would yield a predicted change in test scores of $(-0.6) \times (-2) = 1.2$; that is, you would predict that test scores would *rise* by 1.2 points as a result of the *reduction* in class sizes by two students per class.

Equation (4.1) is the definition of the slope of a straight line relating test scores and class size. This straight line can be written

$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize. \tag{4.3}$$

where β_0 is the intercept of this straight line, and, as before, $\beta_{ClassSize}$ is the slope. According to Equation (4.3), if you knew β_0 and $\beta_{ClassSize}$, not only would you be able to determine the *change* in test scores at a district associated with a *change* in class size, but you also would be able to predict the average test score itself for a given class size.

When you propose Equation (4.3) to the superintendent, she tells you that something is wrong with this formulation. She points out that class size is just one of many facets of elementary education, and that two districts with the same class sizes will have different test scores for many reasons. One district might have better teachers or it might use better textbooks. Two districts with comparable class sizes, teachers, and textbooks still might have very different student populations; perhaps one district has more immigrants (and thus fewer native English speakers) or wealthier families. Finally, she points out that even if two districts are the same in all these ways, they might have different test scores for essentially random reasons having to do with the performance of the individual students on the day of the test. She is right, of course; for all these reasons, Equation (4.3) will not hold exactly for all districts. Instead, it should be viewed as a statement about a relationship that holds *on average* across the population of districts.

A version of this linear relationship that holds for *each* district must incorporate these other factors influencing test scores, including each district's unique characteristics (for example, quality of their teachers, background of their students, how lucky the students were on test day). One approach would be to list the most important factors and to introduce them explicitly into Equation (4.3) (an idea we return to in Chapter 6). For now, however, we simply lump all these "other factors" together and write the relationship for a given district as

$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize + \text{other factors}. \tag{4.4}$$

Thus, the test score for the district is written in terms of one component, $\beta_0 + \beta_{ClassSize} \times ClassSize$, that represents the average effect of class size on scores in the population of school districts and a second component that represents all other factors.

Although this discussion has focused on test scores and class size, the idea expressed in Equation (4.4) is much more general, so it is useful to introduce more

general notation. Suppose you have a sample of n districts. Let Y_i be the average test score in the i^{th} district, let X_i be the average class size in the i^{th} district, and let u_i denote the other factors influencing the test score in the i^{th} district. Then Equation (4.4) can be written more generally as

$$Y_i = \beta_0 + \beta_1 X_i + u_i. \tag{4.5}$$

for each district, (that is, $i = 1, \ldots, n$), where β_0 is the intercept of this line and β_1 is the slope. [The general notation "β_1" is used for the slope in Equation (4.5) instead of "$\beta_{ClassSize}$" because this equation is written in terms of a general variable X_i.]

Equation (4.5) is the **linear regression model with a single regressor**, in which Y is the **dependent variable** and X is the **independent variable** or the **regressor**.

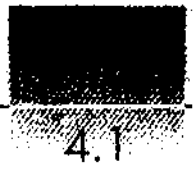
The first part of Equation (4.5), $\beta_0 + \beta_1 X_i$, is the **population regression line** or the **population regression function**. This is the relationship that holds between Y and X on average over the population. Thus, if you knew the value of X, according to this population regression line you would predict that the value of the dependent variable, Y, is $\beta_0 + \beta_1 X$.

The **intercept** β_0 and the **slope** β_1 are the **coefficients** of the population regression line, also known as the **parameters** of the population regression line. The slope β_1 is the change in Y associated with a unit change in X. The intercept is the value of the population regression line when $X = 0$; it is the point at which the population regression line intersects the Y axis. In some econometric applications, the intercept has a meaningful economic interpretation. In other applications, the intercept has no real-world meaning; for example, when X is the class size, strictly speaking the intercept is the predicted value of test scores when there are no students in the class! When the real-world meaning of the intercept is nonsensical it is best to think of it mathematically as the coefficient that determines the level of the regression line.

The term u_i in Equation (4.5) is the **error term**. The error term incorporates all of the factors responsible for the difference between the i^{th} district's average test score and the value predicted by the population regression line. This error term contains all the other factors besides X that determine the value of the dependent variable, Y, for a specific observation, i. In the class size example, these other factors include all the unique features of the i^{th} district that affect the performance of its students on the test, including teacher quality, student economic background, luck, and even any mistakes in grading the test.

The linear regression model and its terminology are summarized in Key Concept 4.1.

Terminology for the Linear Regression Model with a Single Regressor



The linear regression model is

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

where

the subscript i runs over observations, $i = 1, \ldots, n$;

Y_i is the *dependent variable*, the *regressand*, or simply the *left-hand variable*;

X_i is the *independent variable*, the *regressor*, or simply the *right-hand variable*;

$\beta_0 + \beta_1 X$ is the *population regression line* or *population regression function*;

β_0 is the *intercept* of the population regression line;

β_1 is the *slope* of the population regression line; and

u_i is the *error term*.

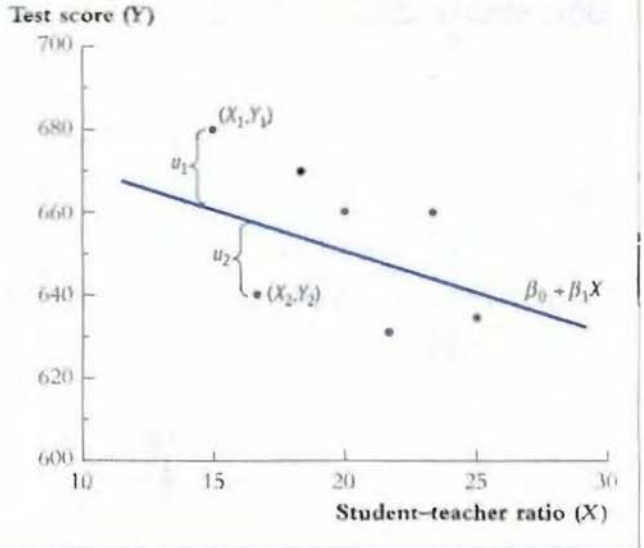
Figure 4.1 summarizes the linear regression model with a single regressor for seven hypothetical observations on test scores (Y) and class size (X). The population regression line is the straight line $\beta_0 + \beta_1 X$. The population regression line slopes down ($\beta_1 < 0$), which means that districts with lower student–teacher ratios (smaller classes) tend to have higher test scores. The intercept β_0 has a mathematical meaning as the value of the Y axis intersected by the population regression line, but, as mentioned earlier, it has no real-world meaning in this example.

Because of the other factors that determine test performance, the hypothetical observations in Figure 4.1 do not fall exactly on the population regression line. For example, the value of Y for district #1, Y_1, is above the population regression line. This means that test scores in district #1 were better than predicted by the population regression line, so the error term for that district, u_1, is positive. In contrast, Y_2 is below the population regression line, so test scores for that district were worse than predicted, and $u_2 < 0$.

Now return to your problem as advisor to the superintendent: What is the expected effect on test scores of reducing the student–teacher ratio by two students per teacher? The answer is easy: The expected change is $(-2) \times \beta_{ClassSize}$. But what is the value of $\beta_{ClassSize}$?

FIGURE 4.1 Scatter Plot of Test Score vs. Student–Teacher Ratio (Hypothetical Data)

The scatterplot shows hypothetical observations for seven school districts. The population regression line is $\beta_0 + \beta_1 X$. The vertical distance from the i^{th} point to the population regression line is $Y_i - (\beta_0 + \beta_1 X_i)$, which is the population error term u_i for the i^{th} observation.



4.2 Estimating the Coefficients of the Linear Regression Model

In a practical situation, such as the application to class size and test scores, the intercept β_0 and slope β_1 of the population regression line are unknown. Therefore, we must use data to estimate the unknown slope and intercept of the population regression line.

This estimation problem is similar to others you have faced in statistics. For example, suppose you want to compare the mean earnings of men and women who recently graduated from college. Although the population mean earnings are unknown, we can estimate the population means using a random sample of male and female college graduates. Then the natural estimator of the unknown population mean earnings for women, for example, is the average earnings of the female college graduates in the sample.

The same idea extends to the linear regression model. We do not know the population value of $\beta_{ClassSize}$, the slope of the unknown population regression line relating X (class size) and Y (test scores). But just as it was possible to learn about the population mean using a sample of data drawn from that

TABLE 4.1 Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1998

			Percentile						
	Average	Standard Deviation	10%	25%	40%	50% (median)	60%	75%	90%
Student–teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	665.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

population, so is it possible to learn about the population slope $\beta_{ClassSize}$ using a sample of data.

The data we analyze here consist of test scores and class sizes in 1999 in 420 California school districts that serve kindergarten through eighth grade. The test score is the districtwide average of reading and math scores for fifth graders. Class size can be measured in various ways. The measure used here is one of the broadest, which is the number of students in the district divided by the number of teachers—that is, the districtwide student–teacher ratio. These data are described in more detail in Appendix 4.1.

Table 4.1 summarizes the distributions of test scores and class sizes for this sample. The average student–teacher ratio is 19.6 students per teacher and the standard deviation is 1.9 students per teacher. The 10th percentile of the distribution of the student–teacher ratio is 17.3 (that is, only 10% of districts have student–teacher ratios below 17.3), while the district at the 90th percentile has a student–teacher ratio of 21.9.

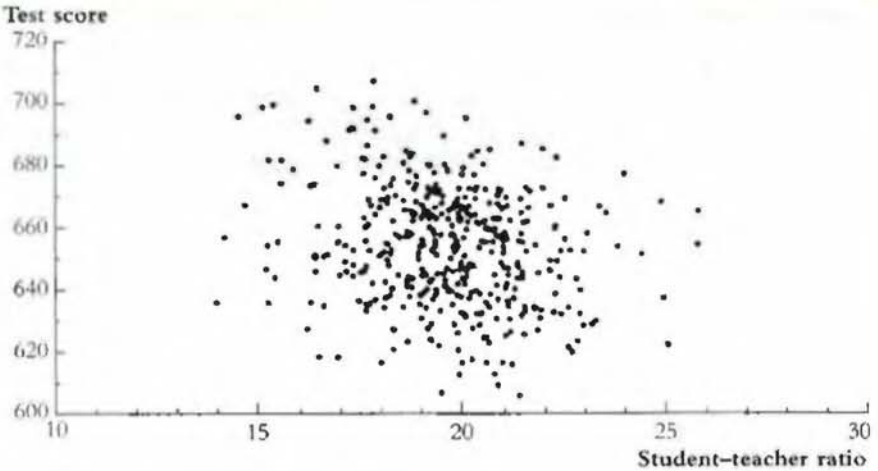
A scatterplot of these 420 observations on test scores and the student–teacher ratio is shown in Figure 4.2. The sample correlation is -0.23, indicating a weak negative relationship between the two variables. Although larger classes in this sample tend to have lower test scores, there are other determinants of test scores that keep the observations from falling perfectly along a straight line.

Despite this low correlation, if one could somehow draw a straight line through these data, then the slope of this line would be an estimate of $\beta_{ClassSize}$ based on these data. One way to draw the line would be to take out a pencil and a ruler and to "eyeball" the best line you could. While this method is easy, it is very unscientific and different people will create different estimated lines.

How, then, should you choose among the many possible lines? By far the most common way is to choose the line that produces the "least squares" fit to these data—that is, to use the ordinary least squares (OLS) estimator.

FIGURE 4.2 Scatterplot of Test Score vs. Student–Teacher Ratio (California School District Data)

Data from 420 California school districts. There is a weak negative relationship between the student–teacher ratio and test scores: The sample correlation is −0.23.



The Ordinary Least Squares Estimator

The OLS estimator chooses the regression coefficients so that the estimated regression line is as close as possible to the observed data, where closeness is measured by the sum of the squared mistakes made in predicting Y given X.

As discussed in Section 3.1, the sample average, $\overline{Y}$, is the least squares estimator of the population mean, $E(Y)$; that is, $\overline{Y}$ minimizes the total squared estimation mistakes $\sum_{i=1}^{n}(Y_i - m)^2$ among all possible estimators m [see expression (3.2)].

The OLS estimator extends this idea to the linear regression model. Let b_0 and b_1 be some estimators of β_0 and β_1. The regression line based on these estimators is $b_0 + b_1X$, so the value of Y_i predicted using this line is $b_0 + b_1X_i$. Thus, the mistake made in predicting the i^{th} observation is $Y_i - (b_0 + b_1X_i) = Y_i - b_0 - b_1X_i$. The sum of these squared prediction mistakes over all n observations is

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2. \tag{4.6}$$

The sum of the squared mistakes for the linear regression model in expression (4.6) is the extension of the sum of the squared mistakes for the problem of estimating the mean in expression (3.2). In fact, if there is no regressor, then b_1 does not enter expression (4.6) and the two problems are identical except for the different notation [m in expression (3.2), b_0 in expression (4.6)]. Just as there is a unique estimator, $\overline{Y}$, that minimizes the expression (3.2), so is there a unique pair of estimators of β_0 and β_1 that minimize expression (4.6).

THE OLS ESTIMATOR, PREDICTED VALUES, AND RESIDUALS

KEY CONCEPT 4.2

The OLS estimators of the slope β_1 and the intercept β_0 are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.7}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}. \tag{4.8}$$

The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,\ i = 1, \ldots, n \tag{4.9}$$

$$\hat{u}_i = Y_i - \hat{Y}_i,\ i = 1, \ldots, n. \tag{4.10}$$

The estimated intercept ($\hat{\beta}_0$), slope ($\hat{\beta}_1$), and residual ($\hat{u}_i$) are computed from a sample of n observations of X_i and Y_i, $i = 1, \ldots, n$. These are estimates of the unknown true population intercept (β_0), slope (β_1), and error term (u_i).

The estimators of the intercept and slope that minimize the sum of squared mistakes in expression (4.6) are called the **ordinary least squares (OLS) estimators** of β_0 and β_1.

OLS has its own special notation and terminology. The OLS estimator of β_0 is denoted $\hat{\beta}_0$, and the OLS estimator of β_1 is denoted $\hat{\beta}_1$. The **OLS regression line** is the straight line constructed using the OLS estimators: $\hat{\beta}_0 + \hat{\beta}_1 X$. The **predicted value** of Y_i given X_i, based on the OLS regression line, is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. The **residual** for the i^{th} observation is the difference between Y_i and its predicted value: $\hat{u}_i = Y_i - \hat{Y}_i$.

You could compute the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ by trying different values of b_0 and b_1 repeatedly until you find those that minimize the total squared mistakes in expression (4.6); they are the least squares estimates. This method would be quite tedious, however. Fortunately there are formulas, derived by minimizing expression (4.6) using calculus, that streamline the calculation of the OLS estimators.

The OLS formulas and terminology are collected in Key Concept 4.2. These formulas are implemented in virtually all statistical and spreadsheet programs. These formulas are derived in Appendix 4.2

OLS Estimates of the Relationship Between Test Scores and the Student–Teacher Ratio

When OLS is used to estimate a line relating the student–teacher ratio to test scores using the 420 observations in Figure 4.2, the estimated slope is −2.28 and the estimated intercept is 698.9. Accordingly, the OLS regression line for these 420 observations is

$$\widehat{TestScore} = 698.9 - 2.28 \times STR, \tag{4.11}$$

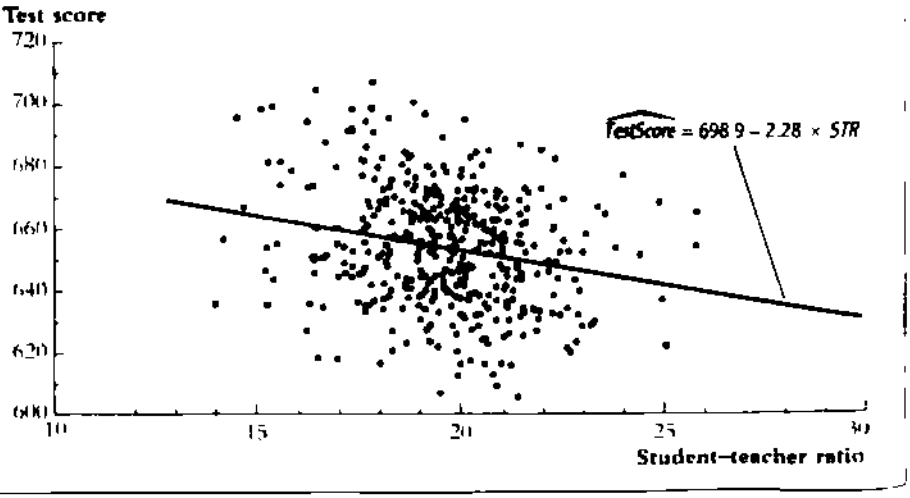
where *TestScore* is the average test score in the district and *STR* is the student–teacher ratio. The symbol "^" over *TestScore* in Equation (4.7) indicates that this is the predicted value based on the OLS regression line. Figure 4.3 plots this OLS regression line superimposed over the scatterplot of the data previously shown in Figure 4.2.

The slope of −2.28 means that an increase in the student–teacher ratio by one student per class is, on average, associated with a decline in districtwide test scores by 2.28 points on the test. A decrease in the student–teacher ratio by 2 students per class is, on average, associated with an increase in test scores of 4.56 points [= −2 × (−2.28)]. The negative slope indicates that more students per teacher (larger classes) is associated with poorer performance on the test.

It is now possible to predict the districtwide test score given a value of the student–teacher ratio. For example, for a district with 20 students per teacher, the

FIGURE 4.3 The Estimated Regression Line for the California Data

The estimated regression line shows a negative relationship between test scores and the student–teacher ratio. If class sizes fall by 1 student, the estimated regression predicts that test scores will increase by 2.28 points.



predicted test score is $698.9 - 2.28 \times 20 = 653.3$. Of course, this prediction will not be exactly right because of the other factors that determine a district's performance. But the regression line does give a prediction (the OLS prediction) of what test scores would be for that district, based on their student–teacher ratio, absent those other factors.

Is this estimate of the slope large or small? To answer this, we return to the superintendent's problem. Recall that she is contemplating hiring enough teachers to reduce the student–teacher ratio by 2. Suppose her district is at the median of the California districts. From Table 4.1, the median student–teacher ratio is 19.7 and the median test score is 654.5. A reduction of 2 students per class, from 19.7 to 17.7, would move her student–teacher ratio from the 50th percentile to very near the 10th percentile. This is a big change, and she would need to hire many new teachers. How would it affect test scores?

According to Equation (4.11), cutting the student–teacher ratio by 2 is predicted to increase test scores by approximately 4.6 points; if her district's test scores are at the median, 654.5, they are predicted to increase to 659.1. Is this improvement large or small? According to Table 4.1, this improvement would move her district from the median to just short of the 60th percentile. Thus, a decrease in class size that would place her district close to the 10% with the smallest classes would move her test scores from the 50th to the 60th percentile. According to these estimates, at least, cutting the student–teacher ratio by a large amount (2 students per teacher) would help and might be worth doing depending on her budgetary situation, but it would not be a panacea.

What if the superintendent were contemplating a far more radical change, such as reducing the student–teacher ratio from 20 students per teacher to 5? Unfortunately, the estimates in Equation (4.11) would not be very useful to her. This regression was estimated using the data in Figure 4.2, and as the figure shows, the smallest student–teacher ratio in these data is 14. These data contain no information on how districts with extremely small classes perform, so these data alone are not a reliable basis for predicting the effect of a radical move to such an extremely low student–teacher ratio.

Why Use the OLS Estimator?

There are both practical and theoretical reasons to use the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. Because OLS is the dominant method used in practice, it has become the common language for regression analysis throughout economics, finance (see the box), and the social sciences more generally. Presenting results using OLS (or its variants discussed later in this book) means that you are "speaking the same language"

The "Beta" of a Stock

A fundamental idea of modern finance is that an investor needs a financial incentive to take a risk. Said differently, the expected return[1] on a risky investment, R, must exceed the return on a safe, or risk-free, investment, R_f. Thus the expected excess return, $R - R_f$, on a risky investment, like owning stock in a company, should be positive.

At first it might seem like the risk of a stock should be measured by its variance. Much of that risk, however, can be reduced by holding other stocks in a "portfolio"—in other words, by diversifying your financial holdings. This means that the right way to measure the risk of a stock is not by its *variance* but rather by its *covariance* with the market.

The capital asset pricing model (CAPM) formalizes this idea. According to the CAPM, the expected excess return on an asset is proportional to the expected excess return on a portfolio of all available assets (the "market portfolio"). That is, the CAPM says that

$$R - R_f = \beta(R_m - R_f). \tag{4.12}$$

where R_m is the expected return on the market portfolio and β is the coefficient in the population regression of $R - R_f$ on $R_m - R_f$. In practice, the risk-free return is often taken to be the rate of interest on short-term U.S. government debt. According to the CAPM, a stock with a $\beta < 1$ has less risk than the market portfolio and therefore has a lower expected excess return than the market portfolio. In contrast, a stock with a $\beta > 1$ is riskier than the market portfolio and thus comands a higher expected excess return.

The "beta" of a stock has become a workhorse of the investment industry, and you can obtain estimated β's for hundreds of stocks on investment firm Web sites. Those β's typically are estimated by OLS regression of the actual excess return on the stock against the actual excess return on a broad market index.

The table below gives estimated β's for six U.S. stocks. Low-risk consumer products firms like Kellogg have stocks with low β's; riskier technology stocks have high β's.

Company	Estimated β
Kellogg (breakfast cereal)	−0.03
Wal-Mart (discount retailer)	0.65
Waste Management (waste disposal)	0.70
Sprint Nextel (telecommunications)	0.78
Barnes and Noble (book retailer)	1.02
Microsoft (software)	1.27
Best Buy (electronic equipment retailer)	2.15
Amazon (online retailer)	2.65

Source: SmartMoney.com

[1] The return on an investment is the change in its price plus any payout (dividend) from the investment as a percentage of its initial price. For example, a stock bought on January 1 for \$100, which then paid a \$2.50 dividend during the year and sold on December 31 for \$105, would have a return of $R = [(\$105 - \$100) + \$2.50]/\$100 = 7.5\%$.

as other economists and statisticians. The OLS formulas are built into virtually all spreadsheet and statistical software packages, making OLS easy to use.

The OLS estimators also have desirable theoretical properties. These are analogous to the desirable properties studied in Section 3.1, of $\bar{Y}$ as an estimator of the population mean. Under the assumptions introduced in Section 4.4, the OLS

estimator is unbiased and consistent. The OLS estimator is also efficient among a certain class of unbiased estimators; however, this efficiency result holds under some additional special conditions, and further discussion of this result is deferred until Section 5.5.

4.3 Measures of Fit

Having estimated a linear regression, you might wonder how well that regression line describes the data. Does the regressor account for much or for little of the variation in the dependent variable? Are the observations tightly clustered around the regression line, or are they spread out?

The R^2 and the standard error of the regression measure how well the OLS regression line fits the data. The R^2 ranges between 0 and 1 and measures the fraction of the variance of Y_i that is explained by X_i. The standard error of the regression measures how far Y_i typically is from its predicted value.

The R^2

The **regression R^2** is the fraction of the sample variance of Y_i explained by (or predicted by) X_i. The definitions of the predicted value and the residual (see Key Concept 4.2) allow us to write the dependent variable Y_i as the sum of the predicted value, $\hat{Y}_i$, plus the residual $\hat{u}_i$:

$$Y_i = \hat{Y}_i + \hat{u}_i. \tag{4.13}$$

In this notation, the R^2 is the ratio of the sample variance of $\hat{Y}_i$ to the sample variance of Y_i.

Mathematically, the R^2 can be written as the ratio of the explained sum of squares to the total sum of squares. The **explained sum of squares (*ESS*)** is the sum of squared deviations of the predicted values of Y_i, $\hat{Y}_i$, from their average, and the **total sum of squares (*TSS*)** is the sum of squared deviations of Y_i from its average:

$$ESS = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \tag{4.14}$$

$$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2. \tag{4.15}$$

Equation (4.14) uses the fact that the sample average OLS predicted value equals $\overline{Y}$ (proven in Appendix 4.3).

The R^2 is the ratio of the explained sum of squares to the total sum of squares

$$R^2 = \frac{ESS}{TSS}. \tag{4.16}$$

Alternatively, the R^2 can be written in terms of the fraction of the variance of Y_i *not* explained by X_i. The **sum of squared residuals**, or ***SSR***, is the sum of the squared OLS residuals:

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2. \tag{4.17}$$

It is shown in Appendix 4.3 that $TSS = ESS + SSR$. Thus the R^2 also can be expressed as 1 minus the ratio of the sum of squared residuals to the total sum of squares:

$$R^2 = 1 - \frac{SSR}{TSS}. \tag{4.18}$$

Finally, the R^2 of the regression of Y on the single regressor X is the square of the correlation coefficient between Y and X.

The R^2 ranges between 0 and 1. If $\hat{\beta}_1 = 0$, then X_i explains none of the variation of Y_i and the predicted value of Y_i based on the regression is just the sample average of Y_i. In this case, the explained sum of squares is zero and the sum of squared residuals equals the total sum of squares; thus the R^2 is zero. In contrast, if X_i explains all of the variation of Y_i, then $Y_i = \hat{Y}_i$ for all i and every residual is zero (that is, $\hat{u}_i = 0$), so that $ESS = TSS$ and $R^2 = 1$. In general, the R^2 does not take on the extreme values of 0 or 1 but falls somewhere in between. An R^2 near 1 indicates that the regressor is good at predicting Y_i, while an R^2 near 0 indicates that the regressor is not very good at predicting Y_i.

The Standard Error of the Regression

The **standard error of the regression** (***SER***) is an estimator of the standard deviation of the regression error u_i. The units of u_i and Y_i are the same, so the SER is a measure of the spread of the observations around the regression line, measured in the units of the dependent variable. For example, if the units of the dependent variable are dollars, then the SER measures the magnitude of a typical deviation from the regression line—that is, the magnitude of a typical regression error—in dollars.

Because the regression errors $u_1, \ldots, u_n$ are unobserved, the *SER* is computed using their sample counterparts, the OLS residuals $\hat{u}_1, \ldots, \hat{u}_n$. The formula for the *SER* is

$$SER = s_{\hat{u}}, \text{ where } s_{\hat{u}}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-2}, \tag{4.19}$$

where the formula for $s_{\hat{u}}^2$ uses the fact (proven in Appendix 4.3) that the sample average of the OLS residuals is zero.

The formula for the *SER* in Equation (4.19) is similar to the formula for the sample standard deviation of Y given in Equation (3.7) in Section 3.2, except that $Y_i - \overline{Y}$ in Equation (3.7) is replaced by $\hat{u}_i$, and the divisor in Equation (3.7) is $n - 1$, whereas here it is $n - 2$. The reason for using the divisor $n - 2$ here (instead of n) is the same as the reason for using the divisor $n - 1$ in Equation (3.7): It corrects for a slight downward bias introduced because two regression coefficients were estimated. This is called a "degrees of freedom" correction: because two coefficients were estimated (β_0 and β_1), two "degrees of freedom" of the data were lost, so the divisor in this factor is $n - 2$. (The mathematics behind this is discussed in Section 5.6.) When n is large, the difference between dividing by n, by $n - 1$, or by n 2 is negligible.

Application to the Test Score Data

Equation (4.11) reports the regression line, estimated using the California test score data, relating the standardized test score (*TestScore*) to the student–teacher ratio (*STR*). The R^2 of this regression is 0.051, or 5.1%, and the *SER* is 18.6.

The R^2 of 0.051 means that the regressor *STR* explains 5.1% of the variance of the dependent variable *TestScore*. Figure 4.3 superimposes this regression line on the scatterplot of the *TestScore* and *STR* data. As the scatterplot shows, the student–teacher ratio explains some of the variation in test scores, but much variation remains unaccounted for.

The *SER* of 18.6 means that standard deviation of the regression residuals is 18.6, where the units are points on the standardized test. Because the standard deviation is a measure of spread, the *SER* of 18.6 means that there is a large spread of the scatterplot in Figure 4.3 around the regression line as measured in points on the test. This large spread means that predictions of test scores made using only the student–teacher ratio for that district will often be wrong by a large amount.

What should we make of this low R^2 and large *SER*? The fact that the R^2 of this regression is low (and the *SER* is large) does not, by itself, imply that this

regression is either "good" or "bad." What the low R^2 *does* tell us is that other important factors influence test scores. These factors could include differences in the student body across districts, differences in school quality unrelated to the student–teacher ratio, or luck on the test. The low R^2 and high SER do not tell us what these factors are, but they do indicate that the student–teacher ratio alone explains only a small part of the variation in test scores in these data.

4.4 The Least Squares Assumptions

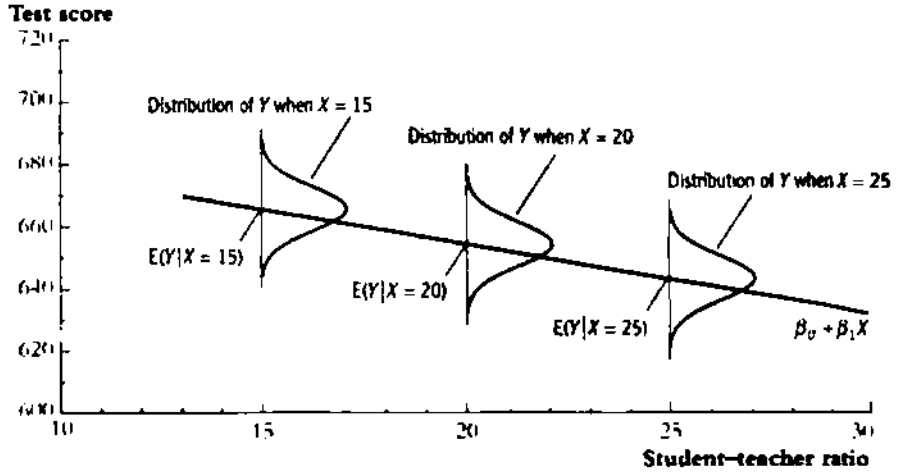
This section presents a set of three assumptions on the linear regression model and the sampling scheme under which OLS provides an appropriate estimator of the unknown regression coefficients, β_0 and β_1. Initially these assumptions might appear abstract. They do, however, have natural interpretations, and understanding these assumptions is essential for understanding when OLS will—and will not—give useful estimates of the regression coefficients.

Assumption #1: The Conditional Distribution of u_i Given X_i Has a Mean of Zero

The first **least squares assumption** is that the conditional distribution of u_i given X_i has a mean of zero. This assumption is a formal mathematical statement about the "other factors" contained in u_i and asserts that these other factors are unrelated to X_i in the sense that, given a value of X_i, the mean of the distribution of these other factors is zero.

This is illustrated in Figure 4.4. The population regression is the relationship that holds on average between class size and test scores in the population, and the error term u_i represents the other factors that lead test scores at a given district to differ from the prediction based on the population regression line. As shown in Figure 4.4, at a given value of class size, say 20 students per class, sometimes these other factors lead to better performance than predicted ($u_i > 0$) and sometimes to worse performance ($u_i < 0$), but on average over the population the prediction is right. In other words, given $X_i = 20$, the mean of the distribution of u_i is zero. In Figure 4.4, this is shown as the distribution of u_i being centered on the population regression line at $X_i = 20$ and, more generally, at other values x of X_i as well. Said differently, the distribution of u_i, conditional on $X_i = x$, has a mean of zero; stated mathematically, $E(u_i|X_i = x) = 0$ or, in somewhat simpler notation, $E(u_i|X_i) = 0$.

FIGURE 4.4 The Conditional Probability Distributions and the Population Regression Line



The figure shows the conditional probability of test scores for districts with class sizes of 15, 20, and 25 students. The mean of the conditional distribution of test scores, given the student–teacher ratio, $E(Y|X)$, is the population regression line $\beta_0 + \beta_1 X$. At a given value of X, Y is distributed around the regression line and the error, $u = Y - (\beta_0 + \beta_1 X)$, has a conditional mean of zero for all values of X.

As shown in Figure 4.4, the assumption that $E(u_i|X_i) = 0$ is equivalent to assuming that the population regression line is the conditional mean of Y_i given X_i (a mathematical proof of this is left as Exercise 4.6).

The conditional mean of u in a randomized controlled experiment. In a randomized controlled experiment, subjects are randomly assigned to the treatment group ($X = 1$) or to the control group ($X = 0$). The random assignment typically is done using a computer program that uses no information about the subject, ensuring that X is distributed independently of all personal characteristics of the subject. Random assignment makes X and u independent, which in turn implies that the conditional mean of u given X is zero.

In observational data, X is not randomly assigned in an experiment. Instead, the best that can be hoped for is that X is *as if* randomly assigned, in the precise sense that $E(u_i|X_i) = 0$. Whether this assumption holds in a given empirical application with observational data requires careful thought and judgment, and we return to this issue repeatedly.

Correlation and conditional mean. Recall from Section 2.3 that if the conditional mean of one random variable given another is zero, then the two random variables have zero covariance and thus are uncorrelated [Equation (2.27)]. Thus, the conditional mean assumption $E(u_i|X_i) = 0$ implies that X_i and u_i are uncorrelated, or $\text{corr}(X_i, u_i) = 0$. Because correlation is a measure of linear association, this implication does not go the other way; even if X_i and u_i are uncorrelated, the conditional mean of u_i given X_i might be nonzero. However, if X_i and u_i are correlated, then it must be the case that $E(u_i|X_i)$ is nonzero. It is therefore often convenient to discuss the conditional mean assumption in terms of possible correlation between X_i and u_i. If X_i and u_i are correlated, then the conditional mean assumption is violated.

Assumption #2: (X_i, Y_i), $i = 1, \ldots, n$ Are Independently and Identically Distributed

The second least squares assumption is that (X_i, Y_i), $i = 1, \ldots, n$ are independently and identically distributed (i.i.d.) across observations. As discussed in Section 2.5 (Key Concept 2.5), this is a statement about how the sample is drawn. If the observations are drawn by simple random sampling from a single large population, then (X_i, Y_i), $i = 1, \ldots, n$ are i.i.d. For example, let X be the age of a worker and Y be his or her earnings, and imagine drawing a person at random from the population of workers. That randomly drawn person will have a certain age and earnings (that is, X and Y will take on some values). If a sample of n workers is drawn from this population, then (X_i, Y_i), $i = 1, \ldots, n$, necessarily have the same distribution. If they are drawn at random they are also distributed independently from one observation to the next; that is, they are i.i.d.

The i.i.d. assumption is a reasonable one for many data collection schemes. For example, survey data from a randomly chosen subset of the population typically can be treated as i.i.d.

Not all sampling schemes produce i.i.d. observations on (X_i, Y_i), however. One example is when the values of X are not drawn from a random sample of the population but rather are set by a researcher as part of an experiment. For example, suppose a horticulturalist wants to study the effects of different organic weeding methods (X) on tomato production (Y) and accordingly grows different plots of tomatoes using different organic weeding techniques. If she picks the techniques (the level of X) to be used on the i^{th} plot and applies the same technique to the i^{th} plot in all repetitions of the experiment, then the value of X_i does not change from one sample to the next. Thus X_i is nonrandom (although the outcome Y_i is random), so the sampling scheme is not i.i.d. The results presented in this chapter

developed for i.i.d. regressors are also true if the regressors are nonrandom. The case of a nonrandom regressor is, however, quite special. For example, modern experimental protocols would have the horticulturalist assign the level of X to the different plots using a computerized random number generator, thereby circumventing any possible bias by the horticulturalist (she might use her favorite weeding method for the tomatoes in the sunniest plot). When this modern experimental protocol is used, the level of X is random and (X_i, Y_i) are i.i.d.

Another example of non-i.i.d. sampling is when observations refer to the same unit of observation over time. For example, we might have data on inventory levels (Y) at a firm and the interest rate at which the firm can borrow (X), where these data are collected over time from a specific firm; for example, they might be recorded four times a year (quarterly) for 30 years. This is an example of time series data, and a key feature of time series data is that observations falling close to each other in time are not independent but rather tend to be correlated with each other; if interest rates are low now, they are likely to be low next quarter. This pattern of correlation violates the "independence" part of the i.i.d. assumption. Time series data introduce a set of complications that are best handled after developing the basic tools of regression analysis.

Assumption #3: Large Outliers Are Unlikely

The third least squares assumption is that large outliers—that is, observations with values of X_i and/or Y_i far outside the usual range of the data—are unlikely. Large outliers can make OLS regression results misleading. This potential sensitivity of OLS to extreme outliers is illustrated in Figure 4.5 using hypothetical data.

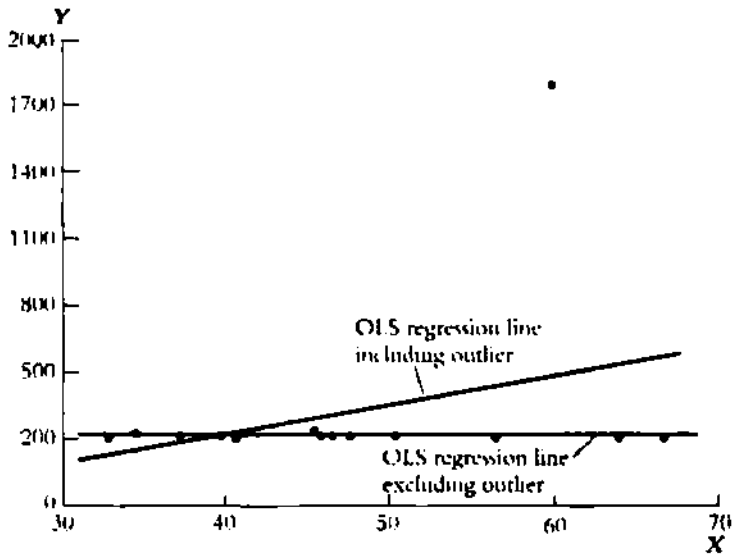
In this book, the assumption that large outliers are unlikely is made mathematically precise by assuming that X and Y have nonzero finite fourth moments: $0 < E(X_i^4) < \infty$ and $0 < E(Y_i^4) < \infty$. Another way to state this assumption is that X and Y have finite kurtosis.

The assumption of finite kurtosis is used in the mathematics that justify the large-sample approximations to the distributions of the OLS test statistics. We encountered this assumption in Chapter 3 when discussing the consistency of the sample variance. Specifically, Equation (3.9) states that the sample variance s_Y^2 is a consistent estimator of the population variance σ_Y^2 ($s_Y^2 \xrightarrow{p} \sigma_Y^2$). If $Y_1, \ldots, Y_n$ are i.i.d. and the fourth moment of Y_i is finite, then the law of large numbers in Key Concept 2.6 applies to the average, $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)^2$, a key step in the proof in Appendix 3.3 showing that s_Y^2 is consistent.

One source of large outliers is data entry errors, such as a typographical error or incorrectly using different units for different observations. Imagine collecting

FIGURE 4.5 The Sensitivity of OLS to Large Outliers

This hypothetical data set has one outlier. The OLS regression line estimated with the outlier shows a strong positive relationship between X and Y, but the OLS regression line estimated without the outlier shows no relationship.



data on the height of students in meters, but inadvertently recording one student's height in centimeters instead. One way to find outliers is to plot your data. If you decide that an outlier is due to a data entry error, then you can either correct the error or, if that is impossible, drop the observation from your data set.

Data entry errors aside, the assumption of finite kurtosis is a plausible one in many applications with economic data. Class size is capped by the physical capacity of a classroom; the best you can do on a standardized test is to get all the questions right and the worst you can do is to get all the questions wrong. Because class size and test scores have a finite range, they necessarily have finite kurtosis. More generally, commonly used distributions such as the normal distribution have four moments. Still, as a mathematical matter, some distributions have infinite fourth moments, and this assumption rules out those distributions. If this assumption holds then it is unlikely that statistical inferences using OLS will be dominated by a few observations.

Use of the Least Squares Assumptions

The three least squares assumptions for the linear regression model are summarized in Key Concept 4.3. The least squares assumptions play twin roles, and we return to them repeatedly throughout this textbook.

THE LEAST SQUARES ASSUMPTIONS

4.3

$$Y_i = \beta_0 + \beta_1 X_i + u_i, i = 1, \ldots, n, \text{ where}$$

1. The error term u_i has conditional mean zero given X_i: $E(u_i|X_i) = 0$;
2. $(X_i, Y_i), i = 1, \ldots, n$ are independent and identically distributed (i.i.d.) draws from their joint distribution; and
3. Large outliers are unlikely: X_i and Y_i have nonzero finite fourth moments.

Their first role is mathematical: If these assumptions hold, then, as is shown in the next section, in large samples the OLS estimators have sampling distributions that are normal. In turn, this large-sample normal distribution lets us develop methods for hypothesis testing and constructing confidence intervals using the OLS estimators.

Their second role is to organize the circumstances that pose difficulties for OLS regression. As we will see, the first least squares assumption is the most important to consider in practice. One reason why the first least squares assumption might not hold in practice is discussed in Chapter 6, and additional reasons are discussed in Section 9.2.

It is also important to consider whether the second assumption holds in an application. Although it plausibly holds in many cross-sectional data sets, the independence assumption is inappropriate for time series data. Therefore, the regression methods developed under assumption 2 require modification for some applications with time series data.

The third assumption serves as a reminder that OLS, just like the sample mean, can be sensitive to large outliers. If your data set contains large outliers, you should examine those outliers carefully to make sure those observations are correctly recorded and belong in the data set.

4.5 Sampling Distribution of the OLS Estimators

Because the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are computed from a randomly drawn sample, the estimators themselves are random variables with a probability distribution—the sampling distribution—that describes the values they could take over

different possible random samples. This section presents these sampling distributions. In small samples, these distributions are complicated, but in large samples, they are approximately normal because of the central limit theorem.

The Sampling Distribution of the OLS Estimators

Review of the sampling distribution of $\overline{Y}$. Recall the discussion in Sections 2.5 and 2.6 about the sampling distribution of the sample average, $\overline{Y}$, an estimator of the unknown population mean of Y, μ_Y. Because $\overline{Y}$ is calculated using a randomly drawn sample, $\overline{Y}$ is a random variable that takes on different values from one sample to the next; the probability of these different values is summarized in its sampling distribution. Although the sampling distribution of $\overline{Y}$ can be complicated when the sample size is small, it is possible to make certain statements about it that hold for all n. In particular, the mean of the sampling distribution is μ_Y, that is, $E(\overline{Y}) = \mu_Y$, so $\overline{Y}$ is an unbiased estimator of μ_Y. If n is large, then more can be said about the sampling distribution. In particular, the central limit theorem (Section 2.6) states that this distribution is approximately normal.

The sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$. These ideas carry over to the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ of the unknown intercept β_0 and slope β_1 of the population regression line. Because the OLS estimators are calculated using a random sample, $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables that take on different values from one sample to the next; the probability of these different values is summarized in their sampling distributions.

Although the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ can be complicated when the sample size is small, it is possible to make certain statements about it that hold for all n. In particular, the mean of the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are β_0 and β_1. In other words, under the least squares assumptions in Key Concept 4.3,

$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1. \tag{4.20}$$

that is, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1. The proof that $\hat{\beta}_1$ is unbiased is given in Appendix 4.3 and the proof that $\hat{\beta}_0$ is unbiased is left as Exercise 4.7.

If the sample is sufficiently large, by the central limit theorem the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ is well approximated by the bivariate normal distribution (Section 2.4.). This implies that the marginal distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are normal in large samples.

LARGE-SAMPLE DISTRIBUTIONS OF $\hat{\beta}_0$ AND $\hat{\beta}_1$

KEY CONCEPT 4.4

If the least squares assumptions in Key Concept 4.3 hold, then in large samples $\hat{\beta}_0$ and $\hat{\beta}_1$ have a jointly normal sampling distribution. The large-sample normal distribution of $\hat{\beta}_1$ is $N(\beta_1, \sigma^2_{\hat{\beta}_1})$, where the variance of this distribution, $\sigma^2_{\hat{\beta}_1}$, is

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)]^2}. \tag{4.21}$$

The large-sample normal distribution of $\hat{\beta}_0$ is $N(\beta_0, \sigma^2_{\hat{\beta}_0})$, where

$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{\text{var}(H_i u_i)}{[E(H_i^2)]^2}, \text{ where } H_i = 1 - \left(\frac{\mu_X}{E(X_i^2)}\right)X_i. \tag{4.22}$$

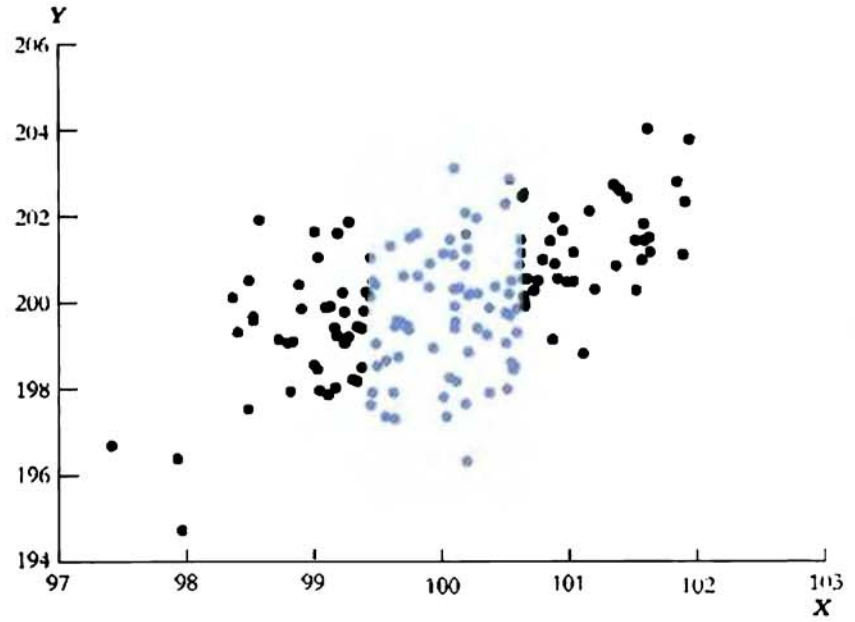
This argument invokes the central limit theorem. Technically, the central limit theorem concerns the distribution of averages (like $\overline{Y}$). If you examine the numerator in Equation (4.7) for $\hat{\beta}_1$, you will see that it, too, is a type of average—not a simple average, like $\overline{Y}$, but an average of the product, $(Y_i - \overline{Y})(X_i - \overline{X})$. As discussed further in Appendix 4.3, the central limit theorem applies to this average so that, like the simpler average $\overline{Y}$, it is normally distributed in large samples.

The normal approximation to the distribution of the OLS estimators in large samples is summarized in Key Concept 4.4. (Appendix 4.3 summarizes the derivation of these formulas.) A relevant question in practice is how large n must be for these approximations to be reliable. In Section 2.6 we suggested that $n = 100$ is sufficiently large for the sampling distribution of $\overline{Y}$ to be well approximated by a normal distribution, and sometimes smaller n suffices. This criterion carries over to the more complicated averages appearing in regression analysis. In virtually all modern econometric applications $n > 100$, so we will treat the normal approximations to the distributions of the OLS estimators as reliable unless there are good reasons to think otherwise.

The results in Key Concept 4.4 imply that the OLS estimators are consistent—that is, when the sample size is large, $\hat{\beta}_0$ and $\hat{\beta}_1$ will be close to the true population coefficients β_0 and β_1 with high probability. This is because the variances $\sigma^2_{\hat{\beta}_0}$ and $\sigma^2_{\hat{\beta}_1}$ of the estimators decrease to zero as n increases (n appears in the denominator of the formulas for the variances), so the distribution of the OLS estimators will be tightly concentrated around their means, β_0 and β_1, when n is large.

FIGURE 4.6 The Variance of $\hat{\beta}_1$ and the Variance of X

The colored dots represent a set of X_i's with a small variance. The black dots represent a set of X_i's with a large variance. The regression line can be estimated more accurately with the black dots than with the colored dots.



Another implication of the distributions in Key Concept 4.4 is that, in general, the larger the variance of X_i, the smaller the variance $\sigma^2_{\hat{\beta}_1}$ of $\hat{\beta}_1$. Mathematically, this arises because the variance of $\hat{\beta}_1$ in Equation (4.21) is inversely proportional to the square of the variance of X_i: the larger is var(X_i), the larger is the denominator in Equation (4.21) so the smaller is $\sigma^2_{\hat{\beta}_1}$. To get a better sense of why this is so, look at Figure 4.6, which presents a scatterplot of 150 artificial data points on X and Y. The data points indicated by the colored dots are the 75 observations closest to $\overline{X}$. Suppose you were asked to draw a line as accurately as possible through *either* the colored or the black dots—which would you choose? It would be easier to draw a precise line through the black dots, which have a larger variance than the colored dots. Similarly, the larger the variance of X, the more precise is $\hat{\beta}_1$.

The normal approximation to the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ is a powerful tool. With this approximation in hand, we are able to develop methods for making inferences about the true population values of the regression coefficients using only a sample of data.

4.6 Conclusion

This chapter has focused on the use of ordinary least squares to estimate the intercept and slope of a population regression line using a sample of n observations on a dependent variable, Y, and a single regressor, X. There are many ways to draw a straight line through a scatterplot, but doing so using OLS has several virtues. If the least squares assumptions hold, then the OLS estimators of the slope and intercept are unbiased, are consistent, and have a sampling distribution with a variance that is inversely proportional to the sample size n. Moreover, if n is large, then the sampling distribution of the OLS estimator is normal.

These important properties of the sampling distribution of the OLS estimator hold under the three least squares assumptions.

The first assumption is that the error term in the linear regression model has a conditional mean of zero, given the regressor X. This assumption implies that the OLS estimator is unbiased.

The second assumption is that (X_i, Y_i) are i.i.d., as is the case if the data are collected by simple random sampling. This assumption yields the formula, presented in Key Concept 4.4, for the variance of the sampling distribution of the OLS estimator.

The third assumption is that large outliers are unlikely. Stated more formally, X and Y have finite fourth moments (finite kurtosis). The reason for this assumption is that OLS can be unreliable if there are large outliers.

The results in this chapter describe the sampling distribution of the OLS estimator. By themselves, however, these results are not sufficient to test a hypothesis about the value of β_j or to construct a confidence interval for β_1. Doing so requires an estimator of the standard deviation of the sampling distribution—that is, the standard error of the OLS estimator. This step—moving from the sampling distribution of $\hat{\beta}_1$ to its standard error, hypothesis tests, and confidence intervals—is taken in the next chapter.

Summary

1. The population regression line, $\beta_0 + \beta_1 X$, is the mean of Y as a function of the value of X. The slope, β_1, is the expected change in Y associated with a 1-unit change in X. The intercept, β_0, determines the level (or height) of the regression line. Key Concept 4.1 summarizes the terminology of the population linear regression model.

2. The population regression line can be estimated using sample observations (Y_i, X_i), $i = 1, \ldots, n$ by ordinary least squares (OLS). The OLS estimators of the regression intercept and slope are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$.
3. The R^2 and standard error of the regression (SER) are measures of how close the values of Y_i are to the estimated regression line. The R^2 is between 0 and 1, with a larger value indicating that the Y_i's are closer to the line. The standard error of the regression is an estimator of the standard deviation of the regression error.
4. There are three key assumptions for the linear regression model: (1) The regression errors, u_i, have a mean of zero conditional on the regressors X_i; (2) the sample observations are i.i.d. random draws from the population; and (3) large outliers are unlikely. If these assumptions hold, the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are (1) unbiased; (2) consistent; and (3) normally distributed when the sample is large.

Key Terms

linear regression model with a single regressor (114)
dependent variable (114)
independent variable (114)
regressor (114)
population regression line (114)
population regression function (114)
population intercept and slope (114)
population coefficients (114)
parameters (114)
error term (114)
ordinary least squares (OLS) estimator (119)
OLS regression line (119)
predicted value (119)
residual (119)
regression R^2 (123)
explained sum of squares (ESS) (123)
total sum of squares (TSS) (123)
sum of squared residuals (SSR) (124)
standard error of the regression (SER) (124)
least squares assumptions (126)

Review the Concepts

4.1 Explain the difference between $\hat{\beta}_1$ and β_1; between the residual $\hat{u}_i$ and the regression error u_i; and between the OLS predicted value $\hat{Y}_i$ and $E(Y_i|X_i)$.

4.2 For each least squares assumption, provide an example in which the assumption is valid, and then provide an example in which the assumption fails.

4.3 Sketch a hypothetical scatterplot of data for an estimated regression with $R^2 = 0.9$. Sketch a hypothetical scatterplot of data for a regression with $R^2 = 0.5$.

Exercises

4.1 Suppose that a researcher, using data on class size (CS) and average test scores from 100 third-grade classes, estimates the OLS regression.

$$\widehat{TestScore} = 520.4 - 5.82 \times CS, R^2 = 0.08, SER = 11.5.$$

a. A classroom has 22 students. What is the regression's prediction for that classroom's average test score?

b. Last year a classroom had 19 students, and this year it has 23 students. What is the regression's prediction for the change in the classroom average test score?

c. The sample average class size across the 100 classrooms is 21.4. What is the sample average of the test scores across the 100 classrooms? (*Hint:* Review the formulas for the OLS estimators.)

d. What is the sample standard deviation of test scores across the 100 classrooms? (*Hint:* Review the formulas for the R^2 and SER.)

4.2 Suppose that a random sample of 200 twenty-year-old men is selected from a population and that these men's height and weight are recorded. A regression of weight on height yields

$$\widehat{Weight} = -99.41 + 3.94 \times Height, R^2 = 0.81, SER = 10.2,$$

where *Weight* is measured in pounds and *Height* is measured in inches.

a. What is the regression's weight prediction for someone who is 70 inches tall? 65 inches tall? 74 inches tall?

b. A man has a late growth spurt and grows 1.5 inches over the course of a year. What is the regression's prediction for the increase in this man's weight?

c. Suppose that instead of measuring weight and height in pounds and inches, these variable are measured in centimeters and kilograms. What are the regression estimates from this new centimeter-kilogram regression? (Give all results, estimated coefficients, R^2, and SER.)

4.3 A regression of average weekly earnings (AWE, measured in dollars) on age (measured in years) using a random sample of college-educated full-time workers aged 25–65 yields the following:

$$\widehat{AWE} = 696.7 + 9.6 \times Age, R^2 = 0.023, SER = 624.1.$$

a. Explain what the coefficient values 696.7 and 9.6 mean.

b. The standard error of the regression (SER) is 624.1. What are the units of measurement for the SER (dollars? years? or is SER unit-free)?

c. The regression R^2 is 0.023. What are the units of measurement for the R^2 (dollars? years? or is R^2 unit-free)?

d. What is the regression's predicted earnings for a 25-year-old worker? A 45-year-old worker?

e. Will the regression give reliable predictions for a 99-year-old worker? Why or why not?

f. Given what you know about the distribution of earnings, do you think it is plausible that the distribution of errors in the regression is normal? (*Hint:* Do you think that the distribution is symmetric or skewed? What is the smallest value of earnings, and is it consistent with a normal distribution?)

g. The average age in this sample is 41.6 years. What is the average value of AWE in the sample? (*Hint:* Review Key Concept 4.2.)

4.4 Read the box "The 'Beta' of a Stock" in Section 4.2.

a. Suppose that the value of β is greater than 1 for a particular stock. Show that the variance of $(R - R_f)$ for this stock is greater than the variance of $(R_m - R_f)$.

b. Suppose that the value of β is less than 1 for a particular stock. Is it possible that variance of $(R - R_f)$ for this stock is greater than the variance of $(R_m - R_f)$? (*Hint:* Don't forget the regression error.)

c. In a given year, the rate of return on 3-month Treasury bills is 3.5% and the rate of return on a large diversified portfolio of stocks (the S&P 500) is 7.3%. For each company listed in the table at the end of the box, use the estimated value of β to estimate the stock's expected rate of return.

4.5 A professor decides to run an experiment to measure the effect of time pressure on final exam scores. He gives each of the 400 students in his course the same final exam, but some students have 90 minutes to complete the exam while others have 120 minutes. Each student is randomly assigned one of the examination times based on the flip of a coin. Let Y_i denote the number of points scored on the exam by the i^{th} student ($0 \leq Y_i \leq 100$), let X_i denote the amount of time that the student has to complete the exam ($X_i = 90$ or 120), and consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$.

a. Explain what the term u_i represents. Why will different students have different values of u_i?

b. Explain why $E(u_i|X_i) = 0$ for this regression model.

c. Are the other assumptions in Key Concept 4.3 satisfied? Explain.

d. The estimated regression is $\hat{Y}_i = 49 + 0.24\,X_i$.

i. Compute the estimated regression's prediction for the average score of students given 90 minutes to complete the exam; 120 minutes; and 150 minutes.

ii. Compute the estimated gain in score for a student who is given an additional 10 minutes on the exam.

4.6 Show that the first least squares assumption, $E(u_i|X_i) = 0$, implies that $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$.

4.7 Show that $\hat{\beta}_0$ is an unbiased estimator of β_0. (*Hint:* Use the fact that $\hat{\beta}_1$ is unbiased, which is shown in Appendix 4.3.)

4.8 Suppose that all of the regression assumptions in Key Concept 4.3 are satisfied except that the first assumption is replaced with $E(u_i|X_i) = 2$. Which parts of Key Concept 4.4 continue to hold? Which change? Why? (Is $\hat{\beta}_1$ normally distributed in large samples with mean and variance given in Key Concept 4.4? What about $\hat{\beta}_0$?)

4.9 a. A linear regression yields $\hat{\beta}_1 = 0$. Show that $R^2 = 0$.

b. A linear regression yields $R^2 = 0$. Does this imply that $\hat{\beta}_1 = 0$?

4.10 Suppose that $Y_i = \beta_0 + \beta_1 X_i + u_i$, where (X_i, u_i) are i.i.d., and X_i is a Bernoulli random variable with $\Pr(X = 1) = 0.20$. When $X = 1$, u_i is $N(0, 4)$; when $X = 0$, u_i is $N(0, 1)$.

a. Show that the regression assumptions in Key Concept 4.3 are satisfied.

b. Derive an expression for the large-sample variance of $\hat{\beta}_1$. [*Hint:* Evaluate the terms in Equation (4.21).]

4.11 Consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$.

a. Suppose you know that $\beta_0 = 0$. Derive a formula for the least squares estimator of β_1.

b. Suppose you know that $\beta_0 = 4$. Derive a formula for the least squares estimator of β_1.

4.12 a. Show that the regression R^2 in the regression of Y on X is the squared value of the sample correlation between X and Y. That is, show that $R^2 = r_{XY}^2$.

b. Show that the R^2 from the regression of Y on X is the same as the R^2 from the regression of X on Y.

Empirical Exercises

E4.1 On the text Web site (**www.aw-bc.com/stock_watson**), you will find a data file **CPS04** that contains an extended version of the data set used in Table 3.1 for 2004. It contains data for full-time, full-year workers, age 25–34, with a high school diploma or B.A./B.S. as their highest degree. A detailed description is given in **CPS04_Description**, also available on the Web site. (These are the same data as in **CPS92_04** but are limited to the year 2004.) In this exercise you will investigate the relationship between a worker's age and earnings. (Generally, older workers have more job experience, leading to higher productivity and earnings.)

a. Run a regression of average hourly earnings (AHE) on age (Age). What is the estimated intercept? What is the estimated slope? Use the estimated regression to answer this question: How much do earnings increase as workers age by one year?

b. Bob is a 26-year-old worker. Predict Bob's earnings using the estimated regression. Alexis is a 30-year-old worker. Predict Alexis's earnings using the estimated regression.

c. Does age account for a large fraction of the variance in earnings across individuals? Explain.

E4.2 On the text Web site (**www.aw-bc.com/stock_watson**), you will find a data file **TeachingRatings** that contains data on course evaluations, course characteristics, and professor characteristics for 463 courses at the University of Texas at Austin.[1] A detailed description is given in **TeachingRatings_Description**, also available on the Web site. One of the characteristics is an index of the professor's "beauty" as rated by a panel of six judges. In this exercise you will investigate how course evaluations are related to the professor's beauty.

[1]These data were provided by Professor Daniel Hamermesh of the University of Texas at Austin and were used in his paper with Amy Parker, "Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity," *Economics of Education Review*, August 2005, 24(4): pp. 369–376.

a. Construct a scatterplot of average course evaluations (*Course_Eval*) on the professor's beauty (*Beauty*). Does there appear to be a relationship between the variables?

b. Run a regression of average course evaluations (*Course_Eval*) on the professor's beauty (*Beauty*). What is the estimated intercept? What is the estimated slope? Explain why the estimated intercept is equal to the sample mean of *Course_Eval*. (*Hint:* What is the sample mean of *Beauty*?)

c. Professor Watson has an average value of *Beauty*, while Professor Stock's value of *Beauty* is one standard deviation above the average. Predict Professor Stock's and Professor Watson's course evaluations.

d. Comment on the size of the regression's slope. Is the estimated effect of *Beauty* on *Course_Eval* large or small? Explain what you mean by "large" and "small."

e. Does *Beauty* explain a large fraction of the variance in evaluations across courses? Explain.

E4.3 On the text Web site (**www.aw-bc.com/stock_watson**), you will find a data file **CollegeDistance** that contains data from a random sample of high school seniors interviewed in 1980 and re-interviewed in 1986. In this exercise you will use these data to investigate the relationship between the number of completed years of education for young adults and the distance from each student's high school to the nearest four-year college. (Proximity to college lowers the cost of education, so that students who live closer to a four-year college should, on average, complete more years of higher education.) A detailed description is given in **CollegeDistance_Description**, also available on the Web site.[2]

a. Run a regression of years of completed education (*ED*) on distance to the nearest college (*Dist*), where *Dist* is measured in tens of miles. (For example, *Dist* = 2 means that the distance is 20 miles.) What is the estimated intercept? What is the estimated slope? Use the estimated regression to answer this question: How does the average value of years of completed schooling change when colleges are built close to where students go to high school?

[2]These data were provided by Professor Cecilia Rouse of Princeton University and were used in her paper "Democratization or Diversion? The Effect of Community Colleges on Educational Attainment," *Journal of Business and Economic Statistics*, April 1995, 12(2): pp 217–224.

b. Bob's high school was 20 miles from the nearest college. Predict Bob's years of completed education using the estimated regression. How would the prediction change if Bob lived 10 miles from the nearest college?

c. Does distance to college explain a large fraction of the variance in educational attainment across individuals? Explain.

d. What is the value of the standard error of the regression? What are the units for the standard error (meters, grams, years, dollars, cents, or something else)?

E4.4 On the text Web site (**www.aw-bc.com/stock_watson**), you will find a data file **Growth** that contains data on average growth rates over 1960–1995 for 65 countries, along with variables that are potentially related to growth. A detailed description is given in **Growth_Description**, also available on the Web site. In this exercise you will investigate the relationship between growth and trade.[3]

a. Construct a scatterplot of average annual growth rate (*Growth*) on the average trade share (*TradeShare*). Does there appear to be a relationship between the variables?

b. One country, Malta, has a trade share much larger than the other countries. Find Malta on the scatterplot. Does Malta look like an outlier?

c. Using all observations, run a regression of *Growth* on *TradeShare*. What is the estimated slope? What is the estimated intercept? Use the regression to predict the growth rate for a country with trade share of 0.5 and with a trade share equal to 1.0.

d. Estimate the same regression excluding the data from Malta. Answer the same questions in (c).

e. Where is Malta? Why is the Malta trade share so large? Should Malta be included or excluded from the analysis?

[3]These data were provided by Professor Ross Levine of Brown University and were used in his paper with Thorsten Beck and Norman Loayza, "Finance and the Sources of Growth," *Journal of Financial Economics*, 2000, 58, 261–300.

APPENDIX 4.1

The California Test Score Data Set

The California Standardized Testing and Reporting data set contains data on test performance, school characteristics, and student demographic backgrounds. The data used here are from all 420 K–6 and K–8 districts in California with data available for 1998 and 1999. Test scores are the average of the reading and math scores on the Stanford 9 Achievement Test, a standardized test administered to fifth-grade students. School characteristics (averaged across the district) include enrollment, number of teachers (measured as "full-time equivalents"), number of computers per classroom, and expenditures per student. The student–teacher ratio used here is the number of students in the district, divided by the number of full-time equivalent teachers. Demographic variables for the students also are averaged across the district. The demographic variables include the percentage of students who are in the public assistance program CalWorks (formerly AFDC), the percentage of students who qualify for a reduced price lunch, and the percentage of students who are English learners (that is, students for whom English is a second language). All of these data were obtained from the California Department of Education (www.cde.ca.gov).

APPENDIX 4.2

Derivation of the OLS Estimators

This appendix uses calculus to derive the formulas for the OLS estimators given in Key Concept 4.2. To minimize the sum of squared prediction mistakes $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$ [Equation (4.6)], first take the partial derivatives with respect to b_0 and b_1:

$$\frac{\partial}{\partial b_0}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i) \text{ and} \tag{4.23}$$

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)X_i. \tag{4.24}$$

The OLS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$, are the values of b_0 and b_1 that minimize $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$ or, equivalently, the values of b_0 and b_1 for which the derivatives in Equations (4.23)

and (4.24) equal zero. Accordingly, setting these derivatives equal to zero, collecting terms, and dividing by n shows that the OLS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy the two equations

$$\overline{Y} - \hat{\beta}_0 - \hat{\beta}_1 \overline{X} = 0 \text{ and} \tag{4.25}$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \overline{X} - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} X_i^2 = 0. \tag{4.26}$$

Solving this pair of equations for $\hat{\beta}_0$ and $\hat{\beta}_1$ yields

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}\,\overline{Y}}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{4.27}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{4.28}$$

Equations (4.27) and (4.28) are the formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ given in Key Concept 4.2; the formula $\hat{\beta}_1 = s_{XY}/s_X^2$ is obtained by dividing the numerator and denominator in Equation (4.27) by $n - 1$.

APPENDIX 4.3 Sampling Distribution of the OLS Estimator

In this appendix, we show that the OLS estimator $\hat{\beta}_1$ is unbiased and, in large samples, has the normal sampling distribution given in Key Concept 4.4.

Representation of $\hat{\beta}_1$ in Terms of the Regressors and Errors

We start by providing an expression for $\hat{\beta}_1$ in terms of the regressors and errors. Because $Y_i = \beta_0 + \beta_1 X_i + u_i$, $Y_i - \overline{Y} = \beta_1(X_i - \overline{X}) + u_i - \overline{u}$, so the numerator of the formula for $\hat{\beta}_1$ in Equation (4.27) is

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) &= \sum_{i=1}^{n}(X_i - \overline{X})[\beta_1(X_i - \overline{X}) + (u_i - \overline{u})] \\ &= \beta_1 \sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{n}(X_i - \overline{X})(u_i - \overline{u}).\end{aligned} \tag{4.29}$$

Now $\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^{n}(X_i - \bar{X})u_i - \sum_{i=1}^{n}(X_i - \bar{X})\bar{u} = \sum_{i=1}^{n}(X_i - \bar{X})u_i$, where the final equality follows from the definition of $\bar{X}$, which implies that $\sum_{i=1}^{n}(X_i - \bar{X})\bar{u} = [\sum_{i=1}^{n}X_i - n\bar{X}]\bar{u} = 0$. Substituting $\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^{n}(X_i - \bar{X})u_i$ into the final expression in Equation (4.29) yields $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \beta_1\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(X_i - \bar{X})u_i$. Substituting this expression in turn into the formula for $\hat{\beta}_1$ in Equation (4.27) yields

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}. \tag{4.30}$$

Proof That $\hat{\beta}_1$ Is Unbiased

The expectation of $\hat{\beta}_1$ is obtained by taking the expectation of both sides of Equation (4.30). Thus,

$$\begin{aligned} E(\hat{\beta}_1) &= \beta_1 + E\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \\ &= \beta_1 + E\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})E(u_i|X_1, \ldots, X_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] = \beta_1, \end{aligned} \tag{4.31}$$

where the second equality in Equation (4.31) follows by using the law of iterated expectations (Section 2.3). By the second least squares assumption, u_i is distributed independently of X for all observations other than i, so $E(u_i|X_1, \ldots, X_n) = E(u_i|X_i)$. By the first least squares assumption, however, $E(u_i|X_i) = 0$. It follows that the conditional expectation in large brackets in the second line of Equation (4.31) is zero, so that $E(\hat{\beta}_1 - \beta_1|X_1, \ldots, X_n) = 0$. Equivalently, $E(\hat{\beta}_1|X_1, \ldots, X_n) = \beta_1$; that is, $\hat{\beta}_1$ is conditionally unbiased, given $X_1, \ldots, X_n$. By the law of iterated expectations $E(\hat{\beta}_1 - \beta_1) = E[E(\hat{\beta}_1 - \beta_1|X_1, \ldots, X_n)] = 0$, so that $E(\hat{\beta}_1) = \beta_1$; that is, $\hat{\beta}_1$ is unbiased.

Large-Sample Normal Distribution of the OLS Estimator

The large-sample normal approximation to the limiting distribution of $\hat{\beta}_1$ (Key Concept 4.4) is obtained by considering the behavior of the final term in Equation (4.30).

First consider the numerator of this term. Because $\bar{X}$ is consistent, if the sample size is large, $\bar{X}$ is nearly equal to μ_X. Thus, to a close approximation, the term in the numerator of Equation (4.30) is the sample average $\bar{v}$, where $v_i = (X_i - \mu_X)u_i$. By the first least squares assumption, v_i has a mean of zero. By the second least squares assumption, v_i is i.i.d. The variance of v_i is $\sigma_v^2 = \text{var}[(X_i - \mu_X)u_i]$ which, by the third least squares assumption, is nonzero and finite. Therefore, $\bar{v}$ satisfies all the requirements of the central limit theorem (Key Concept 2.7). Thus, $\bar{v}/\sigma_{\bar{v}}$ is, in large samples, distributed $N(0, 1)$, where $\sigma_{\bar{v}}^2 = \sigma_v^2/n$. Thus the distribution of $\bar{v}$ is well approximated by the $N(0, \sigma_v^2/n)$ distribution.

Next consider the expression in the denominator in Equation (4.30); this is the sample variance of X (except dividing by n rather than $n - 1$, which is inconsequential if n is large). As discussed in Section 3.2 [Equation (3.8)], the sample variance is a consistent estimator of the population variance, so in large samples it is arbitrarily close to the population variance of X.

Combining these two results, we have that, in large samples, $\hat{\beta}_1 - \beta_1 \cong \bar{v}/\text{var}(X_i)$ so that the sampling distribution of $\hat{\beta}_1$ is, in large samples, $N(\beta_1, \sigma_{\hat{\beta}_1}^2)$, where $\sigma_{\hat{\beta}_1}^2 = \text{var}(v)/[\text{var}(X_i)]^2 = \text{var}[(X_i - \mu_X)u_i]/[n[\text{var}(X_i)]^2]$, which is the expression in Equation (4.21).

Some Additional Algebraic Facts About OLS

The OLS residuals and predicted values satisfy:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0, \tag{4.32}$$

$$\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \bar{Y}. \tag{4.33}$$

$$\sum_{i=1}^{n}\hat{u}_i X_i = 0 \text{ and } s_{\hat{u}X} = 0, \text{ and} \tag{4.34}$$

$$TSS = SSR + ESS. \tag{4.35}$$

Equations (4.32) through (4.35) say that the sample average of the OLS residuals is zero; the sample average of the OLS predicted values equals $\bar{Y}$; the sample covariance $s_{\hat{u}X}$ between the OLS residuals and the regressors is zero; and the total sum of squares is the sum of the sum of squared residuals and the explained sum of squares [the ESS, TSS, and SSR are defined in Equations (4.14), (4.15), and (4.17)].

To verify Equation (4.32), note that the definition of $\hat{\beta}_0$ lets us write the OLS residuals as $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = (Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})$; thus

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} (Y_i - \bar{Y}) - \hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X}).$$

But the definition of $\bar{Y}$ and $\bar{X}$ imply that $\sum_{i=1}^{n}(Y_i - \bar{Y}) = 0$ and $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$, so $\sum_{i=1}^{n}\hat{u}_i = 0$.

To verify Equation (4.33), note that $Y_i = \hat{Y}_i + \hat{u}_i$, so $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n}\hat{Y}_i + \sum_{i=1}^{n}\hat{u}_i = \sum_{i=1}^{n}\hat{Y}_i$, where the second equality is a consequence of Equation (4.32).

To verify Equation (4.34), note that $\sum_{i=1}^{n}\hat{u}_i = 0$ implies $\sum_{i=1}^{n}\hat{u}_i X_i = \sum_{i=1}^{n}\hat{u}_i(X_i - \bar{X})$, so

$$\begin{aligned}\sum_{i=1}^{n}\hat{u}_i X_i &= \sum_{i=1}^{n}[(Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})](X_i - \bar{X}) \\ &= \sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) - \hat{\beta}_1 \sum_{i=1}^{n}(X_i - \bar{X})^2 = 0,\end{aligned} \tag{4.36}$$

where the final equality in Equation (4.36) is obtained using the formula for $\hat{\beta}_1$ in Equation (4.27). This result, combined with the preceding results, implies that $s_{\hat{u}X} = 0$.

Equation (4.35) follows from the previous results and some algebra:

$$\begin{aligned}TSS &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \\ &= SSR + ESS + 2\sum_{i=1}^{n}\hat{u}_i\hat{Y}_i = SSR + ESS.\end{aligned} \tag{4.37}$$

where the final equality follows from $\sum_{i=1}^{n}\hat{u}_i\hat{Y}_i = \sum_{i=1}^{n}\hat{u}_i(\hat{\beta}_0 + \hat{\beta}_1 X_i) = \hat{\beta}_0\sum_{i=1}^{n}\hat{u}_i + \hat{\beta}_1\sum_{i=1}^{n}\hat{u}_i X_i = 0$ by the previous results.