Introduction to Econometrics

THIRD EDITION UPDATE

James H. Stock
Harvard University

Mark W. Watson
Princeton University

PEARSON

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 trade-off. 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,

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

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 district-wide 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.

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

Terminology for the Linear Regression Model with a Single Regressor

KEY CONCEPT 4.1

The linear regression model is

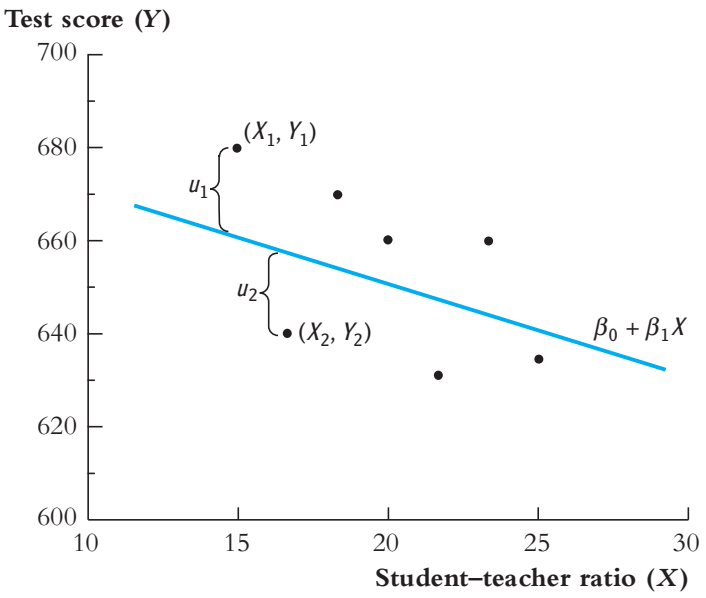
$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

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 the *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 Scatterplot 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.



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 $\beta_{ClassSize}$?

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 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

TABLE 4.1 Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1999

			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	654.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

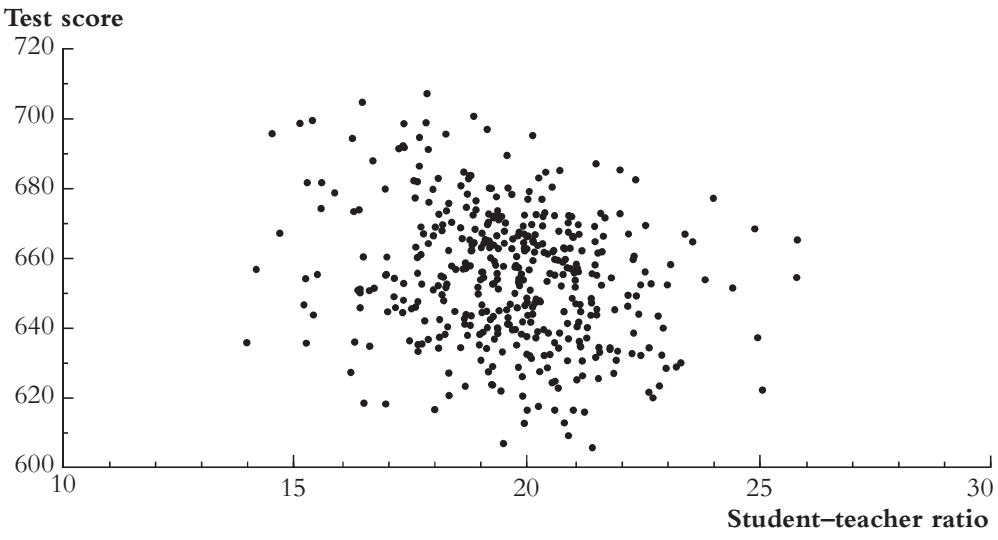
student–teacher ratio is 17.3 (that is, only 10% of districts have student–teacher ratios below 17.3), while the district at the 90th percentile has a student–teacher ratio of 21.9.

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 $\beta_{ClassSize}$

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.



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.

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 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**, also called the **sample regression line** or **sample regression function**, is the straight line constructed using the OLS estimators: $\hat{\beta}_0 + \hat{\beta}_1X$. The **predicted value** of Y_i

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

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

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

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).

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: $\hat{u}_i = Y_i - \hat{Y}_i$.

The OLS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$, are sample counterparts of the population coefficients, β_0 and β_1. Similarly, the OLS regression line $\hat{\beta}_0 + \hat{\beta}_1 X$ is the sample counterpart of the population regression line $\beta_0 + \beta_1 X$, and the OLS residuals $\hat{u}_i$ are sample counterparts of the population errors u_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

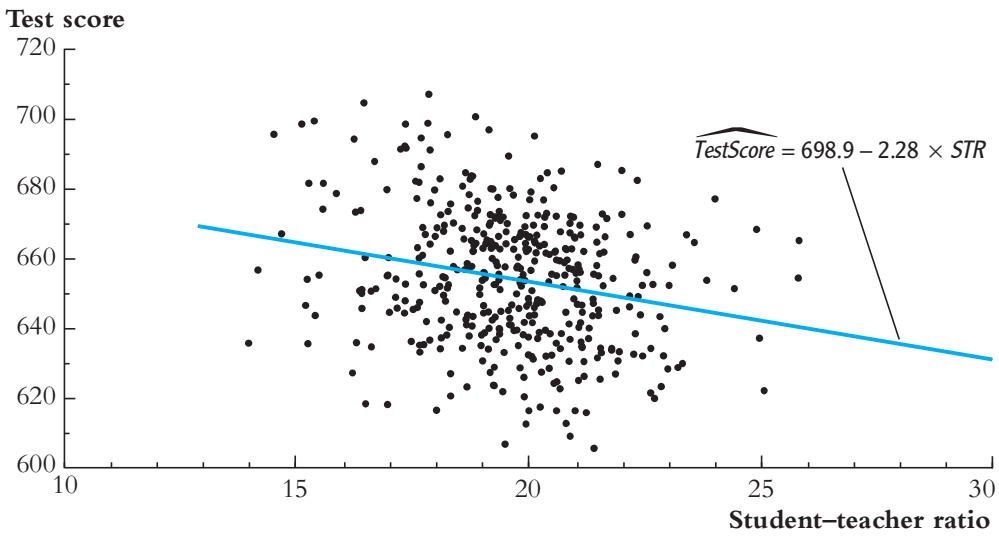
$$\widehat{TestScore} = 698.9 - 2.28 \times STR, \tag{4.11}$$

where *TestScore* is the average test score in the district and *STR* is the student–teacher ratio. The "^" over *TestScore* in Equation (4.11) indicates that it 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 two students per class is, on average, associated with an increase in test scores of 4.56 points $[= -2 \times (-2.28)]$. The negative slope indicates that more students per teacher (larger classes) is associated with poorer performance on the test.

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 one student, the estimated regression predicts that test scores will increase by 2.28 points.



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 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 two 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 (two 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 'Beta' of a Stock" box), and the social sciences more generally. Presenting results

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 commands a higher expected excess return.

The "beta" of a stock has become a workhorse of the investment industry, and you can obtain estimated betas for hundreds of stocks on investment firm websites. Those betas 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 betas for seven U.S. stocks. Low-risk producers of consumer staples like Kellogg have stocks with low betas; riskier stocks have high betas.

Company	Estimated β
Verizon (telecommunications)	0.0
Wal-Mart (discount retailer)	0.3
Kellogg (breakfast cereal)	0.5
Waste Management (waste disposal)	0.6
Google (information technology)	1.0
Ford Motor Company (auto producer)	1.3
Bank of America (bank)	2.2

Source: finance.yahoo.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 $R = [(\$105 - \$100) + \$2.50]/\$100 = 7.5\%$.

using OLS (or its variants discussed later in this book) means that you are "speaking the same language" 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. They are analogous to the desirable properties, studied in Section 3.1, of $\overline{Y}$ 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 value, $\hat{Y}_i$, from its 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 (Exercise 4.12).

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 is $\hat{Y}_i = \hat{\beta}_0 = \overline{Y}$ [from Equation (4.8)]. 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

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

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

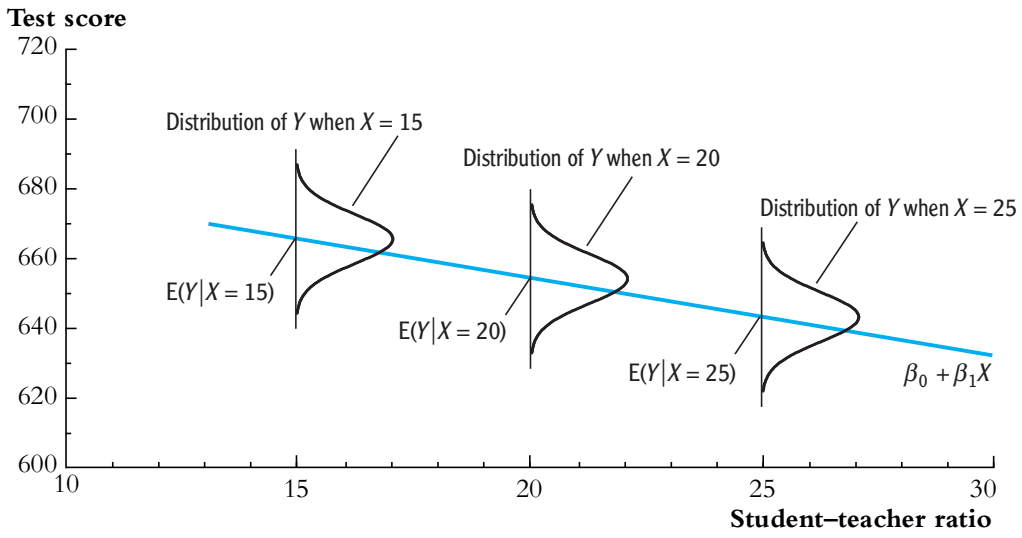
This section presents a set of three assumptions on the linear regression model and the sampling scheme under which OLS provides an appropriate estimator of the unknown regression coefficients, β_0 and β_1. Initially, these assumptions might appear abstract. They do, however, have natural interpretations, 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 of the three **least squares assumptions** 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 assumption 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. 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 assumption 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, so we postpone discussion of time series data until Chapter 14.

Assumption #3: Large Outliers Are Unlikely

The third least squares assumption is that large outliers—that is, observations with values of X_i, Y_i, or both that are 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.

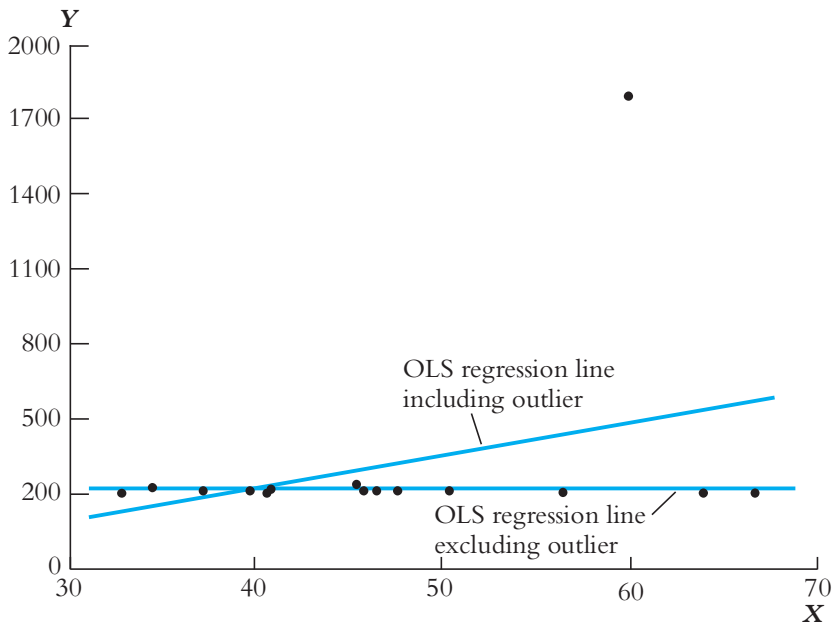
In this book, the assumption that large outliers are unlikely is made mathematically precise by assuming that X and Y have nonzero finite fourth moments: $0 < E(X_i^4) < \infty$ and $0 < E(Y_i^4) < \infty$. Another way to state this assumption is that X and Y have finite kurtosis.

The assumption of finite kurtosis is used in the mathematics that justify the large-sample approximations to the distributions of the OLS test statistics. For example, we encountered this assumption in Chapter 3 when discussing the consistency of the sample variance. Specifically, Equation (3.9) states that the sample variance 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, $\frac{1}{n}\sum_{i=1}^{n} Y_i^2$, 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 data on the height of students in meters, but inadvertently recording one student's

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.



height in centimeters instead. This would create a large outlier in the sample. 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 the assumption of finite fourth moments 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

KEY CONCEPT
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 panel and time series data. Therefore, the regression methods developed under assumption 2 require modification for some applications with time series data. These modifications are developed in Chapters 10 and 14–16.

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.

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

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

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

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.

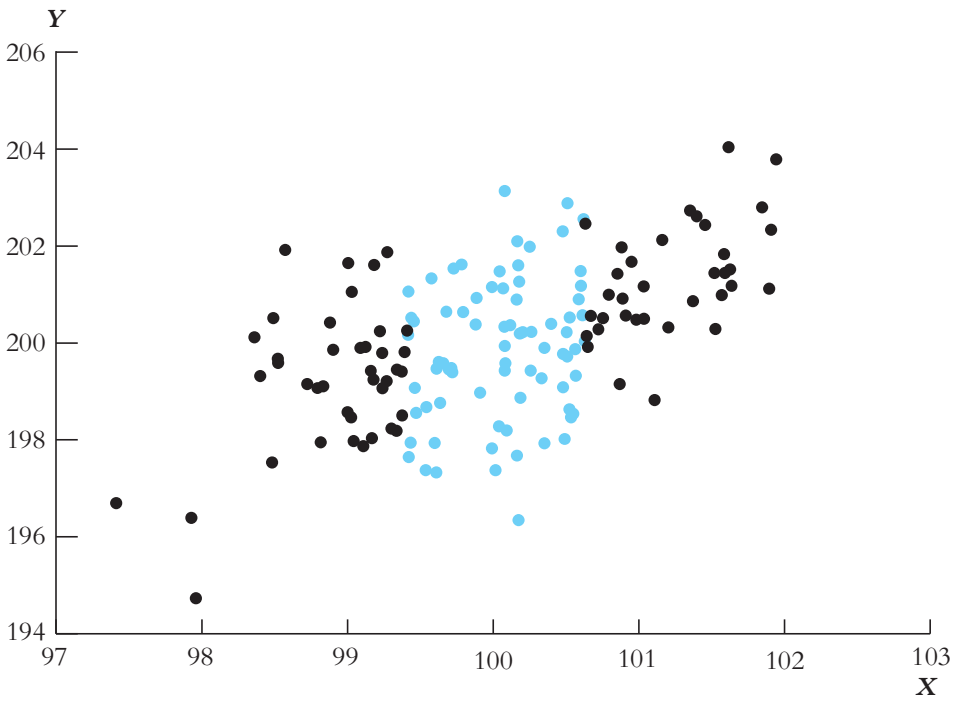
The normal approximation to the distribution of the OLS estimators in large samples is summarized in Key Concept 4.4. (Appendix 4.3 summarizes the derivation of these formulas.) A relevant question in practice is how large n must be for these approximations to be reliable. In Section 2.6, we suggested 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.

Another implication of the distributions in Key Concept 4.4 is that, in general, the larger is the variance of X_i, the smaller is the variance $\sigma^2_{\hat{\beta}_1}$ of $\hat{\beta}_1$. Mathematically, this implication 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 $\text{var}(X_i)$, the larger is the denominator in Equation (4.21) so the smaller is $\sigma^2_{\hat{\beta}_1}$. To get a better sense

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.



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 distributions in Key Concept 4.4 also imply that the smaller is the variance of the error u_i, the smaller is the variance of $\hat{\beta}_1$. This can be seen mathematically in Equation (4.21) because u_i enters the numerator, but not denominator, of $\sigma^2_{\hat{\beta}_1}$: If all u_i were smaller by a factor of one-half but the X's did not change, then $\sigma_{\hat{\beta}_1}$ would be smaller by a factor of one-half and $\sigma^2_{\hat{\beta}_1}$ would be smaller by a factor of one-fourth (Exercise 4.13). Stated less mathematically, if the errors are smaller (holding the X's fixed), then the data will have a tighter scatter around the population regression line so its slope will be estimated more precisely.

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. Taken together, the three least squares assumptions imply that the OLS estimator is normally distributed in large samples as described in Key Concept 4.4.

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_1 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 one-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 $\hat{\beta}_0$ 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 (112)
dependent variable (112)
independent variable (112)
regressor (112)
population regression line (112)
population regression function (112)
population intercept (112)
population slope (112)
population coefficients (112)
parameters (112)
error term (112)
ordinary least squares (OLS) estimators (116)
OLS regression line (116)
sample regression line (116)
sample regression function (116)
predicted value (116)
residual (117)
regression R^2 (121)
explained sum of squares (ESS) (121)
total sum of squares (TSS) (121)
sum of squared residuals (SSR) (122)
standard error of the regression (SER) (122)
least squares assumptions (124)

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 *SER* and R^2 are "measures of fit" for a regression. Explain how *SER* measures the fit of a regression. What are the units of *SER*? Explain how R^2 measures the fit of a regression. What are the units of R^2?

4.4 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 20-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 in. tall? 65 in. tall? 74 in. tall?

b. A man has a late growth spurt and grows 1.5 in. 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 variables 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 does the regression predict will be the earnings for a 25-year-old worker? For 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 $(R_m - R_t)$.

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 $(R_m - R_t)$? (*Hint:* Don't forget the regression error.)

c. In a given year, the rate of return on 3-month Treasury bills is 2.0% and the rate of return on a large diversified portfolio of stocks (the S&P 500) is 5.3%. For each company listed in the table in 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. Repeat for 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.

c. Show that $\hat{\beta}_1 = r_{XY}(s_Y/s_X)$, where r_{XY} is the sample correlation between X and Y, and s_X and s_Y are the sample standard deviations of X and Y.

4.13 Suppose that $Y_i = \beta_0 + \beta_1 X_i + \kappa u_i$, where κ is a nonzero constant and (Y_i, X_i) satisfy the three least squares assumptions. Show that the large sample variance of $\hat{\beta}_1$ is given by $\sigma_{\hat{\beta}_1}^2 = \kappa^2 \frac{1}{n} \frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)^2]}$. [*Hint:* This equation is the variance given in Equation (4.21) multiplied by κ^2.]

4.14 Show that the sample regression line passes through the point $(\overline{X}, \overline{Y})$.

Empirical Exercises

(Only two empirical exercises for this chapter are given in the text, but you can find more on the text website, **http://www.pearsonhighered.com/stock_watson/**.)

E4.1 On the text website, **http://www.pearsonhighered.com/stock_watson/**, you will find the data file **Growth**, which contains data on average growth rates from 1960 through 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 website. In this exercise, you will investigate the relationship between growth and trade.[1]

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 a 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. Plot the estimated regression functions from (c) and (d). Using the scatterplot in (a), explain why the regression function that includes Malta is steeper than the regression function that excludes Malta.

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

E4.2 On the text website, **http://www.pearsonhighered.com/stock_watson/**, you will find the data file **Earnings_and_Height**, which contains data on earnings, height, and other characteristics of a random sample of U.S. workers.[2] A detailed description is given in **Earnings_and_Height_Description**, also available on the website. In this exercise, you will investigate the relationship between earnings and height.

a. What is the median value of height in the sample?

b. i. Estimate average earnings for workers whose height is at most 67 inches.

ii. Estimate average earnings for workers whose height is greater than 67 inches.

[1]These data were provided by Professor Ross Levine of the University of California at Berkeley 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.

[2]These data were provided by Professors Anne Case (Princeton University) and Christina Paxson (Brown University) and were used in their paper "Stature and Status: Height, Ability, and Labor Market Outcomes," *Journal of Political Economy*, 2008, 116(3): 499–532.

iii. On average, do taller workers earn more than shorter workers? How much more? What is a 95% confidence interval for the difference in average earnings?

c. Construct a scatterplot of annual earnings (*Earnings*) on height (*Height*). Notice that the points on the plot fall along horizontal lines. (There are only 23 distinct values of *Earnings*). Why? (*Hint:* Carefully read the detailed data description.)

d. Run a regression of *Earnings* on *Height*.

i. What is the estimated slope?

ii. Use the estimated regression to predict earnings for a worker who is 67 inches tall, for a worker who is 70 inches tall, and for a worker who is 65 inches tall.

e. Suppose height were measured in centimeters instead of inches. Answer the following questions about the *Earnings* on *Height* (in cm) regression.

i. What is the estimated slope of the regression?

ii. What is the estimated intercept?

iii. What is the R^2?

iv. What is the standard error of the regression?

f. Run a regression of *Earnings* on *Height*, using data for female workers only.

i. What is the estimated slope?

ii. A randomly selected woman is 1 inch taller than the average woman in the sample. Would you predict her earnings to be higher or lower than the average earnings for women in the sample? By how much?

g. Repeat (f) for male workers.

h. Do you think that height is uncorrelated with other factors that cause earning? That is, do you think that the regression error term, say u_i, has a conditional mean of zero, given *Height* (X_i)? (You will investigate this more in the *Earnings* and *Height* exercises in later chapters.)

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 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

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

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

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

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

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,

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

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 $\overline{X}$ is consistent, if the sample size is large, $\overline{X}$ is nearly equal to μ_X. Thus, to a close approximation, the term in the numerator of Equation (4.30) is the sample average $\overline{v}$, 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, $\overline{v}$ satisfies all the requirements of the central limit theorem (Key Concept 2.7). Thus $\overline{v}/\sigma_{\overline{v}}$ is, in large samples, distributed $N(0, 1)$, where $\sigma_{\overline{v}}^2 = \sigma_v^2/n$. Thus the distribution of $\overline{v}$ 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, $\hat{\beta}_1 - \beta_1 \cong \overline{v}/\text{var}(X_i)$, so that the sampling distribution of $\hat{\beta}_1$ is, in large samples, $N(\beta_1, \sigma_{\hat{\beta}_1}^2)$, where $\sigma_{\hat{\beta}_1}^2 = \text{var}(\overline{v})/[\text{var}(X_i)]^2 = \text{var}[(X_i - \mu_X)u_i]/\{n[\text{var}(X_i)]^2\}$, which is the expression in Equation (4.21).

Some Additional Algebraic Facts About OLS

The OLS residuals and predicted values satisfy

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

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

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

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

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 $\overline{Y}$; 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 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 $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = (Y_i - \overline{Y}) - \hat{\beta}_1(X_i - \overline{X})$; thus

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

But the definitions of $\overline{Y}$ and $\overline{X}$ imply that $\sum_{i=1}^{n}(Y_i - \overline{Y}) = 0$ and $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$, 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 $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i + \sum_{i=1}^n \hat{u}_1 = \sum_{i=1}^n \hat{Y}_i$, 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 $\sum_{i=1}^n \hat{u}_i X_i = \sum_{i=1}^n \hat{u}_i(X_i - \overline{X})$, so

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

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:

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

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.

5 Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals

This chapter continues the treatment of linear regression with a single regressor. Chapter 4 explained how the OLS estimator $\hat{\beta}_1$ of the slope coefficient β_1 differs from one sample to the next—that is, how $\hat{\beta}_1$ has a sampling distribution. In this chapter, we show how knowledge of this sampling distribution can be used to make statements about β_1 that accurately summarize the sampling uncertainty. The starting point is the standard error of the OLS estimator, which measures the spread of the sampling distribution of $\hat{\beta}_1$. Section 5.1 provides an expression for this standard error (and for the standard error of the OLS estimator of the intercept), then shows how to use $\hat{\beta}_1$ and its standard error to test hypotheses. Section 5.2 explains how to construct confidence intervals for β_1. Section 5.3 takes up the special case of a binary regressor.

Sections 5.1 through 5.3 assume that the three least squares assumptions of Chapter 4 hold. If, in addition, some stronger conditions hold, then some stronger results can be derived regarding the distribution of the OLS estimator. One of these stronger conditions is that the errors are homoskedastic, a concept introduced in Section 5.4. Section 5.5 presents the Gauss–Markov theorem, which states that, under certain conditions, OLS is efficient (has the smallest variance) among a certain class of estimators. Section 5.6 discusses the distribution of the OLS estimator when the population distribution of the regression errors is normal.

5.1 Testing Hypotheses About One of the Regression Coefficients

Your client, the superintendent, calls you with a problem. She has an angry taxpayer in her office who asserts that cutting class size will not help boost test scores, so reducing them is a waste of money. Class size, the taxpayer claims, has no effect on test scores.

The taxpayer's claim can be rephrased in the language of regression analysis. Because the effect on test scores of a unit change in class size is $\beta_{ClassSize}$, the taxpayer is asserting that the population regression line is flat—that is, the slope $\beta_{ClassSize}$ of the population regression line is zero. Is there, the superintendent asks,

General Form of the *t*-Statistic

KEY CONCEPT 5.1

In general, the *t*-statistic has the form

$$t = \frac{\text{estimator} - \text{hypothesized value}}{\text{standard error of the estimator}}. \quad (5.1)$$

evidence in your sample of 420 observations on California school districts that this slope is nonzero? Can you reject the taxpayer's hypothesis that $\beta_{ClassSize} = 0$, or should you accept it, at least tentatively pending further new evidence?

This section discusses tests of hypotheses about the slope β_1 or intercept β_0 of the population regression line. We start by discussing two-sided tests of the slope β_1 in detail, then turn to one-sided tests and to tests of hypotheses regarding the intercept β_0.

Two-Sided Hypotheses Concerning β_1

The general approach to testing hypotheses about the coefficient β_1 is the same as to testing hypotheses about the population mean, so we begin with a brief review.

Testing hypotheses about the population mean. Recall from Section 3.2 that the null hypothesis that the mean of Y is a specific value $\mu_{Y,0}$ can be written as H_0: $E(Y) = \mu_{Y,0}$, and the two-sided alternative is H_1: $E(Y) \neq \mu_{Y,0}$.

The test of the null hypothesis H_0 against the two-sided alternative proceeds as in the three steps summarized in Key Concept 3.6. The first is to compute the standard error of $\overline{Y}$, $SE(\overline{Y})$, which is an estimator of the standard deviation of the sampling distribution of $\overline{Y}$. The second step is to compute the *t*-statistic, which has the general form given in Key Concept 5.1; applied here, the *t*-statistic is $t = (\overline{Y} - \mu_{Y,0})/SE(\overline{Y})$.

The third step is to compute the *p*-value, which is the smallest significance level at which the null hypothesis could be rejected, based on the test statistic actually observed; equivalently, the *p*-value is the probability of obtaining a statistic, by random sampling variation, at least as different from the null hypothesis value as is the statistic actually observed, assuming that the null hypothesis is correct (Key Concept 3.5). Because the *t*-statistic has a standard normal distribution in large samples under the null hypothesis, the *p*-value for a two-sided hypothesis test is $2\Phi(-|t^{act}|)$, where t^{act} is the value of the *t*-statistic actually computed and Φ is the cumulative standard normal distribution tabulated in Appendix Table 1. Alternatively,

the third step can be replaced by simply comparing the t-statistic to the critical value appropriate for the test with the desired significance level. For example, a two-sided test with a 5% significance level would reject the null hypothesis if $|t^{act}| > 1.96$. In this case, the population mean is said to be statistically significantly different from the hypothesized value at the 5% significance level.

Testing hypotheses about the slope β_1. At a theoretical level, the critical feature justifying the foregoing testing procedure for the population mean is that, in large samples, the sampling distribution of $\overline{Y}$ is approximately normal. Because $\hat{\beta}_1$ also has a normal sampling distribution in large samples, hypotheses about the true value of the slope β_1 can be tested using the same general approach.

The null and alternative hypotheses need to be stated precisely before they can be tested. The angry taxpayer's hypothesis is that $\beta_{ClassSize} = 0$. More generally, under the null hypothesis the true population slope β_1 takes on some specific value, $\beta_{1,0}$. Under the two-sided alternative, β_1 does not equal $\beta_{1,0}$. That is, the **null hypothesis** and the **two-sided alternative hypothesis** are

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_1: \beta_1 \neq \beta_{1,0} \quad \text{(two-sided alternative)}. \tag{5.2}$$

To test the null hypothesis H_0, we follow the same three steps as for the population mean.

The first step is to compute the **standard error of $\hat{\beta}_1$,** $SE(\hat{\beta}_1)$. The standard error of $\hat{\beta}_1$ is an estimator of $\sigma_{\hat{\beta}_1}$ the standard deviation of the sampling distribution of $\hat{\beta}_1$. Specifically,

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}, \tag{5.3}$$

where

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}. \tag{5.4}$$

The estimator of the variance in Equation (5.4) is discussed in Appendix (5.1). Although the formula for $\hat{\sigma}^2_{\hat{\beta}_1}$ is complicated, in applications the standard error is computed by regression software so that it is easy to use in practice.

The second step is to compute the ***t*-statistic**,

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}. \tag{5.5}$$

Testing the Hypothesis $\beta_1 = \beta_{1,0}$ Against the Alternative $\beta_1 \neq \beta_{1,0}$

KEY CONCEPT 5.2

1. Compute the standard error of $\hat{\beta}_1$, $SE(\hat{\beta}_1)$ [Equation (5.3)].
2. Compute the t-statistic [Equation (5.5)].
3. Compute the p-value [Equation (5.7)]. Reject the hypothesis at the 5% significance level if the p-value is less than 0.05 or, equivalently, if $|t^{act}| > 1.96$.

The standard error and (typically) the t-statistic and p-value testing $\beta_1 = 0$ are computed automatically by regression software.

The third step is to compute the ***p*-value**, the probability of observing a value of $\hat{\beta}_1$ at least as different from $\beta_{1,0}$ as the estimate actually computed ($\hat{\beta}_1^{act}$), assuming that the null hypothesis is correct. Stated mathematically,

$$\begin{aligned} p\text{-value} &= \Pr_{H_0}\left[|\hat{\beta}_1 - \beta_{1,0}| > |\hat{\beta}_1^{act} - \beta_{1,0}|\right] \\ &= \Pr_{H_0}\left[\left|\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right| > \left|\frac{\hat{\beta}_1^{act} - \beta_{1,0}}{SE(\hat{\beta}_1)}\right|\right] = \Pr_{H_0}(|t| > |t^{act}|), \end{aligned} \quad (5.6)$$

where $\Pr_{H_0}$ denotes the probability computed under the null hypothesis, the second equality follows by dividing by $SE(\hat{\beta}_1)$, and t^{act} is the value of the t-statistic actually computed. Because $\hat{\beta}_1$ is approximately normally distributed in large samples, under the null hypothesis the t-statistic is approximately distributed as a standard normal random variable, so in large samples,

$$p\text{-value} = \Pr(|Z| > |t^{act}|) = 2\Phi(-|t^{act}|). \quad (5.7)$$

A p-value of less than 5% provides evidence against the null hypothesis in the sense that, under the null hypothesis, the probability of obtaining a value of $\hat{\beta}_1$ at least as far from the null as that actually observed is less than 5%. If so, the null hypothesis is rejected at the 5% significance level.

Alternatively, the hypothesis can be tested at the 5% significance level simply by comparing the absolute value of the t-statistic to 1.96, the critical value for a two-sided test, and rejecting the null hypothesis at the 5% level if $|t^{act}| > 1.96$.

These steps are summarized in Key Concept 5.2.

Reporting regression equations and application to test scores. The OLS regression of the test score against the student–teacher ratio, reported in Equation (4.11), yielded $\hat{\beta}_0 = 698.9$ and $\hat{\beta}_1 = -2.28$. The standard errors of these estimates are $SE(\hat{\beta}_0) = 10.4$ and $SE(\hat{\beta}_1) = 0.52$.

Because of the importance of the standard errors, by convention they are included when reporting the estimated OLS coefficients. One compact way to report the standard errors is to place them in parentheses below the respective coefficients of the OLS regression line:

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR, R^2 = 0.051, SER = 18.6. \qquad (5.8)$$

Equation (5.8) also reports the regression R^2 and the standard error of the regression (SER) following the estimated regression line. Thus Equation (5.8) provides the estimated regression line, estimates of the sampling uncertainty of the slope and the intercept (the standard errors), and two measures of the fit of this regression line (the R^2 and the SER). This is a common format for reporting a single regression equation, and it will be used throughout the rest of this book.

Suppose you wish to test the null hypothesis that the slope β_1 is zero in the population counterpart of Equation (5.8) at the 5% significance level. To do so, construct the t-statistic and compare its absolute value to 1.96, the 5% (two-sided) critical value taken from the standard normal distribution. The t-statistic is constructed by substituting the hypothesized value of β_1 under the null hypothesis (zero), the estimated slope, and its standard error from Equation (5.8) into the general formula in Equation (5.5); the result is $t^{act} = (-2.28 - 0)/0.52 = -4.38$. The absolute value of this t-statistic exceeds the 5% two-sided critical value of 1.96, so the null hypothesis is rejected in favor of the two-sided alternative at the 5% significance level.

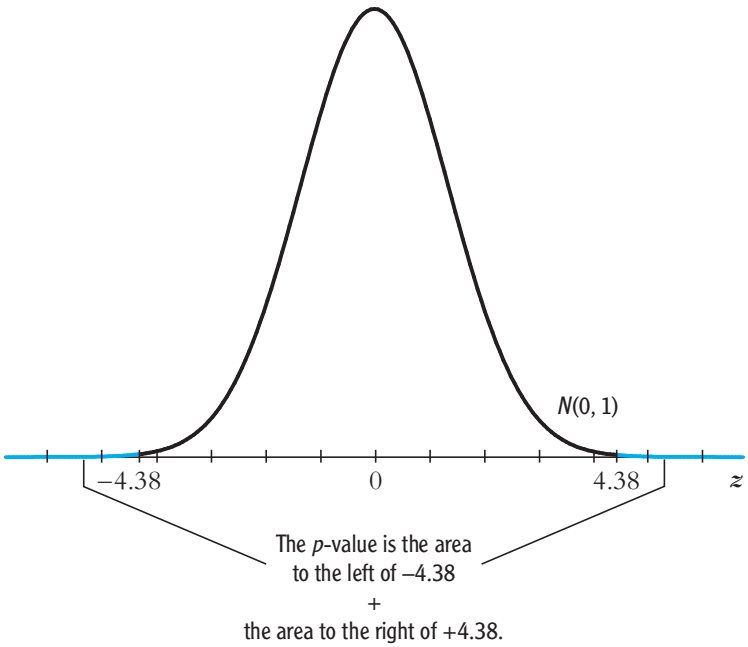
Alternatively, we can compute the p-value associated with $t^{act} = -4.38$. This probability is the area in the tails of standard normal distribution, as shown in Figure 5.1. This probability is extremely small, approximately 0.00001, or 0.001%. That is, if the null hypothesis $\beta_{ClassSize} = 0$ is true, the probability of obtaining a value of $\hat{\beta}_1$ as far from the null as the value we actually obtained is extremely small, less than 0.001%. Because this event is so unlikely, it is reasonable to conclude that the null hypothesis is false.

One-Sided Hypotheses Concerning β_1

The discussion so far has focused on testing the hypothesis that $\beta_1 = \beta_{1,0}$ against the hypothesis that $\beta_1 \neq \beta_{1,0}$. This is a two-sided hypothesis test, because under the alternative β_1 could be either larger or smaller than $\beta_{1,0}$. Sometimes, however, it

FIGURE 5.1 Calculating the *p*-Value of a Two-Sided Test When $t^{act} = -4.38$

The *p*-value of a two-sided test is the probability that $|Z| > |t^{act}|$ where Z is a standard normal random variable and t^{act} is the value of the *t*-statistic calculated from the sample. When $t^{act} = -4.38$, the *p*-value is only 0.00001.



is appropriate to use a one-sided hypothesis test. For example, in the student–teacher ratio/test score problem, many people think that smaller classes provide a better learning environment. Under that hypothesis, β_1 is negative: Smaller classes lead to higher scores. It might make sense, therefore, to test the null hypothesis that $\beta_1 = 0$ (no effect) against the one-sided alternative that $\beta_1 < 0$.

For a one-sided test, the null hypothesis and the one-sided alternative hypothesis are

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_1: \beta_1 < \beta_{1,0} \quad \text{(one-sided alternative)}. \tag{5.9}$$

where $\beta_{1,0}$ is the value of β_1 under the null (0 in the student–teacher ratio example) and the alternative is that β_1 is less than $\beta_{1,0}$. If the alternative is that β_1 is greater than $\beta_{1,0}$, the inequality in Equation (5.9) is reversed.

Because the null hypothesis is the same for a one- and a two-sided hypothesis test, the construction of the *t*-statistic is the same. The only difference between a one- and two-sided hypothesis test is how you interpret the *t*-statistic. For the one-sided alternative in Equation (5.9), the null hypothesis is rejected against the one-sided alternative for large negative, but not large positive, values of the *t*-statistic: Instead of rejecting if $|t^{act}| > 1.96$, the hypothesis is rejected at the 5% significance level if $t^{act} < -1.64$.

The p-value for a one-sided test is obtained from the cumulative standard normal distribution as

$$p\text{-value} = \Pr(Z < t^{act}) = \Phi(t^{act}) \quad (p\text{-value, one-sided left-tail test}). \quad (5.10)$$

If the alternative hypothesis is that β_1 is greater than $\beta_{1,0}$, the inequalities in Equations (5.9) and (5.10) are reversed, so the p-value is the right-tail probability, $\Pr(Z > t^{act})$.

When should a one-sided test be used? In practice, one-sided alternative hypotheses should be used only when there is a clear reason for doing so. This reason could come from economic theory, prior empirical evidence, or both. However, even if it initially seems that the relevant alternative is one-sided, upon reflection this might not necessarily be so. A newly formulated drug undergoing clinical trials actually could prove harmful because of previously unrecognized side effects. In the class size example, we are reminded of the graduation joke that a university's secret of success is to admit talented students and then make sure that the faculty stays out of their way and does as little damage as possible. In practice, such ambiguity often leads econometricians to use two-sided tests.

Application to test scores. The t-statistic testing the hypothesis that there is no effect of class size on test scores [so $\beta_{1,0} = 0$ in Equation (5.9)] is $t^{act} = -4.38$. This value is less than -2.33 (the critical value for a one-sided test with a 1% significance level), so the null hypothesis is rejected against the one-sided alternative at the 1% level. In fact, the p-value is less than 0.0006%. Based on these data, you can reject the angry taxpayer's assertion that the negative estimate of the slope arose purely because of random sampling variation at the 1% significance level.

Testing Hypotheses About the Intercept β_0

This discussion has focused on testing hypotheses about the slope, β_1. Occasionally, however, the hypothesis concerns the intercept β_0. The null hypothesis concerning the intercept and the two-sided alternative are

$$H_0\colon \beta_0 = \beta_{0,0} \text{ vs. } H_1\colon \beta_0 \neq \beta_{0,0} \quad (\text{two-sided alternative}). \quad (5.11)$$

The general approach to testing this null hypothesis consists of the three steps in Key Concept 5.2 applied to β_0 (the formula for the standard error of $\hat{\beta}_0$ is given in Appendix 5.1). If the alternative is one-sided, this approach is modified as was discussed in the previous subsection for hypotheses about the slope.

Hypothesis tests are useful if you have a specific null hypothesis in mind (as did our angry taxpayer). Being able to accept or reject this null hypothesis based on the statistical evidence provides a powerful tool for coping with the uncertainty inherent in using a sample to learn about the population. Yet, there are many times that no single hypothesis about a regression coefficient is dominant, and instead one would like to know a range of values of the coefficient that are consistent with the data. This calls for constructing a confidence interval.

5.2 Confidence Intervals for a Regression Coefficient

Because any statistical estimate of the slope β_1 necessarily has sampling uncertainty, we cannot determine the true value of β_1 exactly from a sample of data. It is possible, however, to use the OLS estimator and its standard error to construct a confidence interval for the slope β_1 or for the intercept β_0.

Confidence interval for β_1. Recall from the discussion of confidence intervals in Section 3.3 that a 95% **confidence interval for β_1** has two equivalent definitions. First, it is the set of values that cannot be rejected using a two-sided hypothesis test with a 5% significance level. Second, it is an interval that has a 95% probability of containing the true value of β_1; that is, in 95% of possible samples that might be drawn, the confidence interval will contain the true value of β_1. Because this interval contains the true value in 95% of all samples, it is said to have a **confidence level** of 95%.

The reason these two definitions are equivalent is as follows. A hypothesis test with a 5% significance level will, by definition, reject the true value of β_1 in only 5% of all possible samples; that is, in 95% of all possible samples, the true value of β_1 will *not* be rejected. Because the 95% confidence interval (as defined in the first definition) is the set of all values of β_1 that are *not* rejected at the 5% significance level, it follows that the true value of β_1 will be contained in the confidence interval in 95% of all possible samples.

As in the case of a confidence interval for the population mean (Section 3.3), in principle a 95% confidence interval can be computed by testing all possible values of β_1 (that is, testing the null hypothesis $\beta_1 = \beta_{1,0}$ for all values of $\beta_{1,0}$) at the 5% significance level using the t-statistic. The 95% confidence interval is then the collection of all the values of β_1 that are not rejected. But constructing the t-statistic for all values of β_1 would take forever.

An easier way to construct the confidence interval is to note that the t-statistic will reject the hypothesized value $\beta_{1,0}$ whenever $\beta_{1,0}$ is outside the range

KEY CONCEPT
5.3

Confidence Interval for β_1

A 95% two-sided confidence interval for β_1 is an interval that contains the true value of β_1 with a 95% probability; that is, it contains the true value of β_1 in 95% of all possible randomly drawn samples. Equivalently, it is the set of values of β_1 that cannot be rejected by a 5% two-sided hypothesis test. When the sample size is large, it is constructed as

$$95\% \text{ confidence interval for } \beta_1 = [\hat{\beta}_1 - 1.96SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96SE(\hat{\beta}_1)]. \quad (5.12)$$

$\hat{\beta}_1 \pm 1.96SE(\hat{\beta}_1)$. This implies that the 95% confidence interval for β_1 is the interval $[\hat{\beta}_1 - 1.96SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96SE(\hat{\beta}_1)]$. This argument parallels the argument used to develop a confidence interval for the population mean.

The construction of a confidence interval for β_1 is summarized as Key Concept 5.3.

Confidence interval for β_0. A 95% confidence interval for β_0 is constructed as in Key Concept 5.3, with $\hat{\beta}_0$ and $SE(\hat{\beta}_0)$ replacing $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$.

Application to test scores. The OLS regression of the test score against the student–teacher ratio, reported in Equation (5.8), yielded $\hat{\beta}_1 = -2.28$ and $SE(\hat{\beta}_1) = 0.52$. The 95% two-sided confidence interval for β_1 is $\{-2.28 \pm 1.96 \times 0.52\}$, or $-3.30 \leq \beta_1 \leq -1.26$. The value $\beta_1 = 0$ is not contained in this confidence interval, so (as we knew already from Section 5.1) the hypothesis $\beta_1 = 0$ can be rejected at the 5% significance level.

Confidence intervals for predicted effects of changing X. The 95% confidence interval for β_1 can be used to construct a 95% confidence interval for the predicted effect of a general change in X.

Consider changing X by a given amount, Δx. The predicted change in Y associated with this change in X is $\beta_1 \Delta x$. The population slope β_1 is unknown, but because we can construct a confidence interval for β_1, we can construct a confidence interval for the predicted effect $\beta_1 \Delta x$. Because one end of a 95% confidence interval for β_1 is $\hat{\beta}_1 - 1.96SE(\hat{\beta}_1)$, the predicted effect of the change Δx using this estimate of β_1 is $[\hat{\beta}_1 - 1.96SE(\hat{\beta}_1)] \times \Delta x$. The other end of the confidence

interval is $\hat{\beta}_1 + 1.96SE(\hat{\beta}_1)$, and the predicted effect of the change using that estimate is $[\hat{\beta}_1 + 1.96SE(\hat{\beta}_1)] \times \Delta x$. Thus a 95% confidence interval for the effect of changing x by the amount Δx can be expressed as

$$\begin{aligned} &95\% \text{ confidence interval for } \beta_1 \Delta x \\ &= [(\hat{\beta}_1 - 1.96SE(\hat{\beta}_1))\Delta x, (\hat{\beta}_1 + 1.96SE(\hat{\beta}_1))\Delta x]. \end{aligned} \tag{5.13}$$

For example, our hypothetical superintendent is contemplating reducing the student–teacher ratio by 2. Because the 95% confidence interval for β_1 is $[-3.30, -1.26]$, the effect of reducing the student–teacher ratio by 2 could be as great as $-3.30 \times (-2) = 6.60$ or as little as $-1.26 \times (-2) = 2.52$. Thus decreasing the student–teacher ratio by 2 is predicted to increase test scores by between 2.52 and 6.60 points, with a 95% confidence level.

5.3 Regression When X Is a Binary Variable

The discussion so far has focused on the case that the regressor is a continuous variable. Regression analysis can also be used when the regressor is binary—that is, when it takes on only two values, 0 or 1. For example, X might be a worker's gender (=1 if female, = 0 if male), whether a school district is urban or rural (= 1 if urban, = 0 if rural), or whether the district's class size is small or large (= 1 if small, = 0 if large). A binary variable is also called an **indicator variable** or sometimes a **dummy variable**.

Interpretation of the Regression Coefficients

The mechanics of regression with a binary regressor are the same as if it is continuous. The interpretation of β_1, however, is different, and it turns out that regression with a binary variable is equivalent to performing a difference of means analysis, as described in Section 3.4.

To see this, suppose you have a variable D_i that equals either 0 or 1, depending on whether the student–teacher ratio is less than 20:

$$D_i = \begin{cases} 1 \text{ if the student–teacher ratio in } i^{\text{th}} \text{ district} < 20 \\ 0 \text{ if the student–teacher ratio in } i^{\text{th}} \text{ district} \geq 20 \end{cases}. \tag{5.14}$$

The population regression model with D_i as the regressor is

$$Y_i = \beta_0 + \beta_1 D_i + u_i, i = 1, \ldots, n. \tag{5.15}$$

This is the same as the regression model with the continuous regressor X_i except that now the regressor is the binary variable D_i. Because D_i is not continuous, it is not useful to think of β_1 as a slope; indeed, because D_i can take on only two values, there is no "line," so it makes no sense to talk about a slope. Thus we will not refer to β_1 as the slope in Equation (5.15); instead we will simply refer to β_1 as the **coefficient multiplying D_i** in this regression or, more compactly, the **coefficient on D_i**.

If β_1 in Equation (5.15) is not a slope, what is it? The best way to interpret β_0 and β_1 in a regression with a binary regressor is to consider, one at a time, the two possible cases, $D_i = 0$ and $D_i = 1$. If the student–teacher ratio is high, then $D_i = 0$ and Equation (5.15) becomes

$$Y_i = \beta_0 + u_i \quad (D_i = 0). \tag{5.16}$$

Because $E(u_i \mid D_i) = 0$, the conditional expectation of Y_i when $D_i = 0$ is $E(Y_i \mid D_i = 0) = \beta_0$; that is, β_0 is the population mean value of test scores when the student–teacher ratio is high. Similarly, when $D_i = 1$,

$$Y_i = \beta_0 + \beta_1 + u_i \quad (D_i = 1). \tag{5.17}$$

Thus, when $D_i = 1$, $E(Y_i \mid D_i = 1) = \beta_0 + \beta_1$; that is, $\beta_0 + \beta_1$ is the population mean value of test scores when the student–teacher ratio is low.

Because $\beta_0 + \beta_1$ is the population mean of Y_i when $D_i = 1$ and β_0 is the population mean of Y_i when $D_i = 0$, the difference $(\beta_0 + \beta_1) - \beta_0 = \beta_1$ is the difference between these two means. In other words, β_1 is the difference between the conditional expectation of Y_i when $D_i = 1$ and when $D_i = 0$, or $\beta_1 = E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0)$. In the test score example, β_1 is the difference between mean test score in districts with low student–teacher ratios and the mean test score in districts with high student–teacher ratios.

Because β_1 is the difference in the population means, it makes sense that the OLS estimator β_1 is the difference between the sample averages of Y_i in the two groups, and, in fact, this is the case.

Hypothesis tests and confidence intervals. If the two population means are the same, then β_1 in Equation (5.15) is zero. Thus the null hypothesis that the two population means are the same can be tested against the alternative hypothesis that they differ by testing the null hypothesis $\beta_1 = 0$ against the alternative $\beta_1 \neq 0$. This hypothesis can be tested using the procedure outlined in Section 5.1. Specifically, the null hypothesis can be rejected at the 5% level against the two-sided

alternative when the OLS t-statistic $t = \hat{\beta}_1 / SE(\hat{\beta}_1)$ exceeds 1.96 in absolute value. Similarly, a 95% confidence interval for β_1, constructed as $\hat{\beta}_1 \pm 1.96SE(\hat{\beta}_1)$. as described in Section 5.2, provides a 95% confidence interval for the difference between the two population means.

Application to test scores. As an example, a regression of the test score against the student–teacher ratio binary variable D defined in Equation (5.14) estimated by OLS using the 420 observations in Figure 4.2 yields

$$\widehat{TestScore} = \underset{(1.3)}{650.0} + \underset{(1.8)}{7.4D}, R^2 = 0.037, SER = 18.7, \tag{5.18}$$

where the standard errors of the OLS estimates of the coefficients β_0 and β_1 are given in parentheses below the OLS estimates. Thus the average test score for the subsample with student–teacher ratios greater than or equal to 20 (that is, for which $D = 0$) is 650.0, and the average test score for the subsample with student–teacher ratios less than 20 (so $D = 1$) is $650.0 + 7.4 = 657.4$. The difference between the sample average test scores for the two groups is 7.4. This is the OLS estimate of β_1, the coefficient on the student–teacher ratio binary variable D.

Is the difference in the population mean test scores in the two groups statistically significantly different from zero at the 5% level? To find out, construct the t-statistic on β_1: $t = 7.4/1.8 = 4.04$. This value exceeds 1.96 in absolute value, so the hypothesis that the population mean test scores in districts with high and low student–teacher ratios is the same can be rejected at the 5% significance level.

The OLS estimator and its standard error can be used to construct a 95% confidence interval for the true difference in means. This is $7.4 \pm 1.96 \times 1.8 = (3.9, 10.9)$. This confidence interval excludes $\beta_1 = 0$, so that (as we know from the previous paragraph) the hypothesis $\beta_1 = 0$ can be rejected at the 5% significance level.

5.4 Heteroskedasticity and Homoskedasticity

Our only assumption about the distribution of u_i conditional on X_i is that it has a mean of zero (the first least squares assumption). If, furthermore, the *variance* of this conditional distribution does not depend on X_i, then the errors are said to be homoskedastic. This section discusses homoskedasticity, its theoretical implications, the simplified formulas for the standard errors of the OLS estimators that arise if the errors are homoskedastic, and the risks you run if you use these simplified formulas in practice.

What Are Heteroskedasticity and Homoskedasticity?

Definitions of heteroskedasticity and homoskedasticity. The error term u_i is **homoskedastic** if the variance of the conditional distribution of u_i given X_i is constant for $i = 1, \ldots, n$ and in particular does not depend on X_i. Otherwise, the error term is **heteroskedastic**.

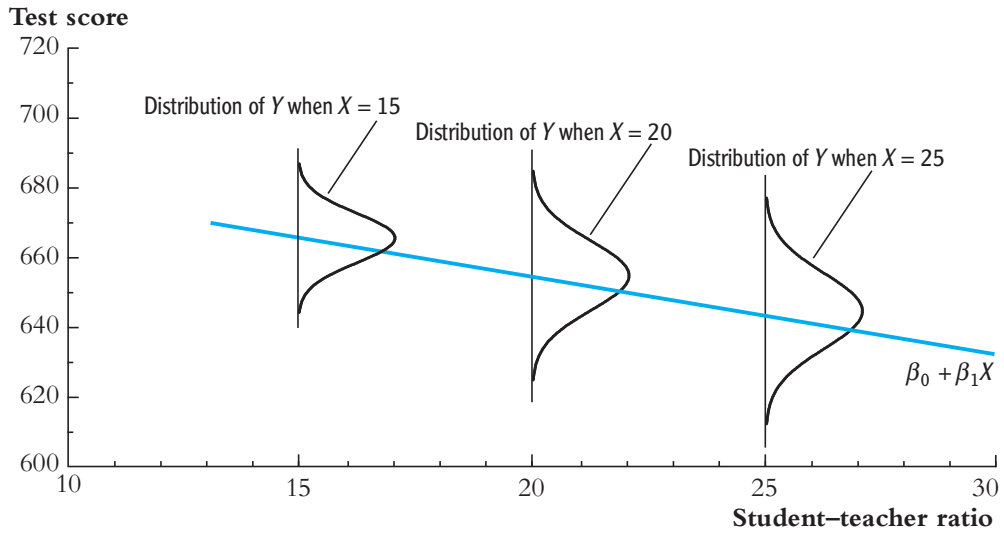
As an illustration, return to Figure 4.4. The distribution of the errors u_i is shown for various values of x. Because this distribution applies specifically for the indicated value of x, this is the conditional distribution of u_i given $X_i = x$. As drawn in that figure, all these conditional distributions have the same spread; more precisely, the variance of these distributions is the same for the various values of x. That is, in Figure 4.4, the conditional variance of u_i given $X_i = x$ does not depend on x, so the errors illustrated in Figure 4.4 are homoskedastic.

In contrast, Figure 5.2 illustrates a case in which the conditional distribution of u_i spreads out as x increases. For small values of x, this distribution is tight, but for larger values of x, it has a greater spread. Thus in Figure 5.2 the variance of u_i given $X_i = x$ increases with x, so that the errors in Figure 5.2 are heteroskedastic.

The definitions of heteroskedasticity and homoskedasticity are summarized in Key Concept 5.4.

FIGURE 5.2 An Example of Heteroskedasticity

Like Figure 4.4, this shows the conditional distribution of test scores for three different class sizes. Unlike Figure 4.4, these distributions become more spread out (have a larger variance) for larger class sizes. Because the variance of the distribution of u given X, var($u|X$), depends on X, u is heteroskedastic.



Heteroskedasticity and Homoskedasticity

KEY CONCEPT
5.4

The error term u_i is homoskedastic if the variance of the conditional distribution of u_i given X_i, $var(u_i|X_i = x)$, is constant for $i = 1, \ldots, n$ and in particular does not depend on x. Otherwise, the error term is heteroskedastic.

Example. These terms are a mouthful, and the definitions might seem abstract. To help clarify them with an example, we digress from the student–teacher ratio/test score problem and instead return to the example of earnings of male versus female college graduates considered in the box in Chapter 3 "The Gender Gap in Earnings of College Graduates in the United States." Let $MALE_i$ be a binary variable that equals 1 for male college graduates and equals 0 for female graduates. The binary variable regression model relating a college graduate's earnings to his or her gender is

$$Earnings_i = \beta_0 + \beta_1 MALE_i + u_i \tag{5.19}$$

for $i = 1, \ldots, n$. Because the regressor is binary, β_1 is the difference in the population means of the two groups—in this case, the difference in mean earnings between men and women who graduated from college.

The definition of homoskedasticity states that the variance of u_i does not depend on the regressor. Here the regressor is $MALE_i$, so at issue is whether the variance of the error term depends on $MALE_i$. In other words, is the variance of the error term the same for men and for women? If so, the error is homoskedastic; if not, it is heteroskedastic.

Deciding whether the variance of u_i depends on $MALE_i$ requires thinking hard about what the error term actually is. In this regard, it is useful to write Equation (5.19) as two separate equations, one for men and one for women:

$$Earnings_i = \beta_0 + u_i \quad \text{(women) and} \tag{5.20}$$

$$Earnings_i = \beta_0 + \beta_1 + u_i \quad \text{(men).} \tag{5.21}$$

Thus, for women, u_i is the deviation of the i^{th} woman's earnings from the population mean earnings for women (β_0), and for men, u_i is the deviation of the i^{th} man's earnings from the population mean earnings for men ($\beta_0 + \beta_1$). It follows that the statement "the variance of u_i does not depend on $MALE$" is equivalent to the

statement "the variance of earnings is the same for men as it is for women." In other words, in this example, the error term is homoskedastic if the variance of the population distribution of earnings is the same for men and women; if these variances differ, the error term is heteroskedastic.

Mathematical Implications of Homoskedasticity

The OLS estimators remain unbiased and asymptotically normal. Because the least squares assumptions in Key Concept 4.3 place no restrictions on the conditional variance, they apply to both the general case of heteroskedasticity and the special case of homoskedasticity. Therefore, the OLS estimators remain unbiased and consistent even if the errors are homoskedastic. In addition, the OLS estimators have sampling distributions that are normal in large samples even if the errors are homoskedastic. Whether the errors are homoskedastic or heteroskedastic, the OLS estimator is unbiased, consistent, and asymptotically normal.

Efficiency of the OLS estimator when the errors are homoskedastic. If the least squares assumptions in Key Concept 4.3 hold and the errors are homoskedastic, then the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are efficient among all estimators that are linear in $Y_1, \ldots, Y_n$ and are unbiased, conditional on $X_1, \ldots, X_n$. This result, which is called the Gauss–Markov theorem, is discussed in Section 5.5.

Homoskedasticity-only variance formula. If the error term is homoskedastic, then the formulas for the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ in Key Concept 4.4 simplify. Consequently, if the errors are homoskedastic, then there is a specialized formula that can be used for the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$. The **homoskedasticity-only standard error** of $\hat{\beta}_1$, derived in Appendix (5.1), is $SE(\hat{\beta}_1) = \sqrt{\tilde{\sigma}^2_{\hat{\beta}_1}}$ where $\tilde{\sigma}^2_{\hat{\beta}_1}$ is the homoskedasticity-only estimator of the variance of $\hat{\beta}_1$:

$$\tilde{\sigma}^2_{\hat{\beta}_1} = \frac{s^2_{\hat{u}}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad \text{(homoskedasticity-only)}, \tag{5.22}$$

where $s^2_{\hat{u}}$ is given in Equation (4.19). The homoskedasticity-only formula for the standard error of $\hat{\beta}_0$ is given in Appendix (5.1). In the special case that X is a binary variable, the estimator of the variance of $\hat{\beta}_1$ under homoskedasticity (that is, the square of the standard error of $\hat{\beta}_1$ under homoskedasticity) is the so-called pooled variance formula for the difference in means, given in Equation (3.23).

Because these alternative formulas are derived for the special case that the errors are homoskedastic and do not apply if the errors are heteroskedastic, they

will be referred to as the "homoskedasticity-only" formulas for the variance and standard error of the OLS estimators. As the name suggests, if the errors are heteroskedastic, then the homoskedasticity-only standard errors are inappropriate. Specifically, if the errors are heteroskedastic, then the *t*-statistic computed using the homoskedasticity-only standard error does not have a standard normal distribution, even in large samples. In fact, the correct critical values to use for this homoskedasticity-only *t*-statistic depend on the precise nature of the heteroskedasticity, so those critical values cannot be tabulated. Similarly, if the errors are heteroskedastic but a confidence interval is constructed as ± 1.96 homoskedasticity-only standard errors, in general the probability that this interval contains the true value of the coefficient is not 95%, even in large samples.

In contrast, because homoskedasticity is a special case of heteroskedasticity, the estimators $\hat{\sigma}^2_{\hat{\beta}_1}$ and $\hat{\sigma}^2_{\hat{\beta}_0}$ of the variances of $\hat{\beta}_1$ and $\hat{\beta}_0$ given in Equations (5.4) and (5.26) produce valid statistical inferences whether the errors are heteroskedastic or homoskedastic. Thus hypothesis tests and confidence intervals based on those standard errors are valid whether or not the errors are heteroskedastic. Because the standard errors we have used so far [that is, those based on Equations (5.4) and (5.26)] lead to statistical inferences that are valid whether or not the errors are heteroskedastic, they are called **heteroskedasticity-robust standard errors**. Because such formulas were proposed by Eicker (1967), Huber (1967), and White (1980), they are also referred to as Eicker–Huber–White standard errors.

What Does This Mean in Practice?

Which is more realistic, heteroskedasticity or homoskedasticity? The answer to this question depends on the application. However, the issues can be clarified by returning to the example of the gender gap in earnings among college graduates. Familiarity with how people are paid in the world around us gives some clues as to which assumption is more sensible. For many years—and, to a lesser extent, today—women were not found in the top-paying jobs: There have always been poorly paid men, but there have rarely been highly paid women. This suggests that the distribution of earnings among women is tighter than among men (see the box in Chapter 3 "The Gender Gap in Earnings of College Graduates in the United States"). In other words, the variance of the error term in Equation (5.20) for women is plausibly less than the variance of the error term in Equation (5.21) for men. Thus the presence of a "glass ceiling" for women's jobs and pay suggests that the error term in the binary variable regression model in Equation (5.19) is heteroskedastic. Unless there are compelling reasons to the contrary—and we can think of none—it makes sense to treat the error term in this example as heteroskedastic.

The Economic Value of a Year of Education: Homoskedasticity or Heteroskedasticity?

On average, workers with more education have higher earnings than workers with less education. But if the best-paying jobs mainly go to the college educated, it might also be that the *spread* of the distribution of earnings is greater for workers with more education. Does the distribution of earnings spread out as education increases?

This is an empirical question, so answering it requires analyzing data. Figure 5.3 is a scatterplot of the hourly earnings and the number of years of education for a sample of 2829 full-time workers in the United States in 2012, ages 29 and 30, with between 6 and 18 years of education. The data come from the March 2013 Current Population Survey, which is described in Appendix 3.1.

Figure 5.3 has two striking features. The first is that the mean of the distribution of earnings increases with the number of years of education. This increase is summarized by the OLS regression line,

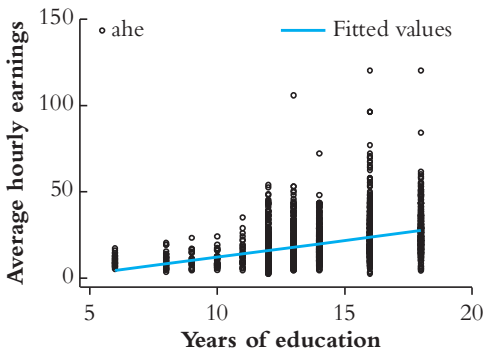
$$\widehat{Earnings} = -7.29 + 1.93 Years\ Education, \quad (1.10)\ (0.08)$$
$$R^2 = 0.162, SER = 10.29. \qquad (5.23)$$

This line is plotted in Figure 5.3. The coefficient of 1.93 in the OLS regression line means that, on average, hourly earnings increase by \$1.93 for each additional year of education. The 95% confidence interval for this coefficient is $1.93 \pm 1.96 \times 0.08$, or 1.77 to 2.09.

The second striking feature of Figure 5.3 is that the spread of the distribution of earnings increases with the years of education. While some workers with many years of education have low-paying jobs, very few workers with low levels of education have high-paying jobs. This can be quantified by looking at the spread of the residuals around the OLS regression line. For workers with ten years of education, the standard deviation of the residuals is \$4.32; for workers with a high school diploma, this standard deviation is \$7.80; and for workers with a college degree, this standard deviation increases to \$12.46. Because these standard deviations differ for different levels of education, the variance of the residuals in the regression of Equation (5.23) depends on the value of the regressor (the years of education); in other words, the regression errors are heteroskedastic. In real-world terms, not all college graduates will be earning \$50 per hour by the time they are 29, but some will, and workers with only ten years of education have no shot at those jobs.

FIGURE 5.3 Scatterplot of Hourly Earnings and Years of Education for 29- to 30-Year-Olds in the United States in 2012

Hourly earnings are plotted against years of education for 2,829 full-time 29- to 30-year-old workers. The spread around the regression line increases with the years of education, indicating that the regression errors are heteroskedastic.



As this example of modeling earnings illustrates, heteroskedasticity arises in many econometric applications. At a general level, economic theory rarely gives any reason to believe that the errors are homoskedastic. It therefore is prudent to assume that the errors might be heteroskedastic unless you have compelling reasons to believe otherwise.

Practical implications. The main issue of practical relevance in this discussion is whether one should use heteroskedasticity-robust or homoskedasticity-only standard errors. In this regard, it is useful to imagine computing both, then choosing between them. If the homoskedasticity-only and heteroskedasticity-robust standard errors are the same, nothing is lost by using the heteroskedasticity-robust standard errors; if they differ, however, then you should use the more reliable ones that allow for heteroskedasticity. The simplest thing, then, is always to use the heteroskedasticity-robust standard errors.

For historical reasons, many software programs report homoskedasticity-only standard errors as their default setting, so it is up to the user to specify the option of heteroskedasticity-robust standard errors. The details of how to implement heteroskedasticity-robust standard errors depend on the software package you use.

All of the empirical examples in this book employ heteroskedasticity-robust standard errors unless explicitly stated otherwise.[1]

*5.5 The Theoretical Foundations of Ordinary Least Squares

As discussed in Section 4.5, the OLS estimator is unbiased, is consistent, has a variance that is inversely proportional to n, and has a normal sampling distribution when the sample size is large. In addition, under certain conditions the OLS estimator is more efficient than some other candidate estimators. Specifically, if the least squares assumptions hold and if the errors are homoskedastic, then the OLS estimator has the smallest variance of all conditionally unbiased estimators that are linear functions of $Y_1, \ldots, Y_n$. This section explains and discusses this result, which is a consequence of the Gauss–Markov theorem. The section concludes

[1]In case this book is used in conjunction with other texts, it might be helpful to note that some textbooks add homoskedasticity to the list of least squares assumptions. As just discussed, however, this additional assumption is not needed for the validity of OLS regression analysis as long as heteroskedasticity-robust standard errors are used.

*This section is optional and is not used in later chapters.

with a discussion of alternative estimators that are more efficient than OLS when the conditions of the Gauss–Markov theorem do not hold.

Linear Conditionally Unbiased Estimators and the Gauss–Markov Theorem

If the three least squares assumptions (Key Concept 4.3) hold and if the error is homoskedastic, then the OLS estimator has the smallest variance, conditional on $X_1, \ldots, X_n$, among all estimators in the class of linear conditionally unbiased estimators. In other words, the OLS estimator is the **B**est **L**inear conditionally **U**nbiased **E**stimator—that is, it is BLUE. This result is an extension of the result, summarized in Key Concept 3.3, that the sample average $\overline{Y}$ is the most efficient estimator of the population mean among the class of all estimators that are unbiased and are linear functions (weighted averages) of $Y_1, \ldots, Y_n$.

Linear conditionally unbiased estimators. The class of linear conditionally unbiased estimators consists of all estimators of β_1 that are linear functions of $Y_1, \ldots, Y_n$ and that are unbiased, conditional on $X_1, \ldots, X_n$. That is, if $\widetilde{\beta}_1$ is a linear estimator, then it can be written as

$$\widetilde{\beta}_1 = \sum_{i=1}^{n} a_i Y_i \quad (\widetilde{\beta}_1 \text{ is linear}), \tag{5.24}$$

where the weights $a_1, \ldots, a_n$ can depend on $X_1, \ldots, X_n$ but *not* on $Y_1, \ldots, Y_n$. The estimator $\widetilde{\beta}_1$ is conditionally unbiased if the mean of its conditional sampling distribution, given $X_1, \ldots, X_n$, is β_1. That is, the estimator $\widetilde{\beta}_1$ is conditionally unbiased if

$$E(\widetilde{\beta}_1 \mid X_1, \ldots, X_n) = \beta_1 \quad (\widetilde{\beta}_1 \text{ is conditionally unbiased}). \tag{5.25}$$

The estimator $\widetilde{\beta}_1$ is a linear conditionally unbiased estimator if it can be written in the form of Equation (5.24) (it is linear) and if Equation (5.25) holds (it is conditionally unbiased). It is shown in Appendix 5.2 that the OLS estimator is linear and conditionally unbiased.

The Gauss–Markov theorem. The **Gauss–Markov theorem** states that, under a set of conditions known as the Gauss–Markov conditions, the OLS estimator $\hat{\beta}_1$ has the smallest conditional variance, given $X_1, \ldots, X_n$, of all linear conditionally unbiased estimators of β_1; that is, the OLS estimator is BLUE. The Gauss–Markov conditions, which are stated in Appendix 5.2, are implied by the three least

The Gauss–Markov Theorem for $\hat{\beta}_1$

KEY CONCEPT
5.5

If the three least squares assumptions in Key Concept 4.3 hold *and* if errors are homoskedastic, then the OLS estimator $\hat{\beta}_1$ is the **B**est (most efficient) **L**inear conditionally **U**nbiased **E**stimator (**BLUE**).

squares assumptions plus the assumption that the errors are homoskedastic. Consequently, if the three least squares assumptions hold and the errors are homoskedastic, then OLS is BLUE. The Gauss–Markov theorem is stated in Key Concept 5.5 and proven in Appendix 5.2.

Limitations of the Gauss–Markov theorem. The Gauss–Markov theorem provides a theoretical justification for using OLS. However, the theorem has two important limitations. First, its conditions might not hold in practice. In particular, if the error term is heteroskedastic—as it often is in economic applications—then the OLS estimator is no longer BLUE. As discussed in Section 5.4, the presence of heteroskedasticity does not pose a threat to inference based on heteroskedasticity-robust standard errors, but it does mean that OLS is no longer the efficient linear conditionally unbiased estimator. An alternative to OLS when there is heteroskedasticity of a known form, called the weighted least squares estimator, is discussed below.

The second limitation of the Gauss–Markov theorem is that even if the conditions of the theorem hold, there are other candidate estimators that are not linear and conditionally unbiased; under some conditions, these other estimators are more efficient than OLS.

Regression Estimators Other Than OLS

Under certain conditions, some regression estimators are more efficient than OLS.

The weighted least squares estimator. If the errors are heteroskedastic, then OLS is no longer BLUE. If the nature of the heteroskedasticity is known—specifically, if the conditional variance of u_i given X_i is known up to a constant factor of proportionality—then it is possible to construct an estimator that has a smaller variance than the OLS estimator. This method, called **weighted least squares (WLS)**, weights the i^{th} observation by the inverse of the square root of the conditional variance of u_i given X_i. Because of this weighting, the errors in this weighted regression are homoskedastic, so OLS, when applied to the weighted data, is BLUE.

Although theoretically elegant, the practical problem with weighted least squares is that you must know how the conditional variance of u_i depends on X_i, something that is rarely known in econometric applications. Weighted least squares is therefore used far less frequently than OLS, and further discussion of WLS is deferred to Chapter 17.

The least absolute deviations estimator. As discussed in Section 4.3, the OLS estimator can be sensitive to outliers. If extreme outliers are not rare, then other estimators can be more efficient than OLS and can produce inferences that are more reliable. One such estimator is the least absolute deviations (LAD) estimator, in which the regression coefficients β_0 and β_1 are obtained by solving a minimization problem like that in Equation (4.6) except that the absolute value of the prediction "mistake" is used instead of its square. That is, the LAD estimators of β_0 and β_1 are the values of b_0 and b_1 that minimize $\sum_{i=1}^{n}|Y_i - b_0 - b_1X_i|$. The LAD estimator is less sensitive to large outliers in u than is OLS.

In many economic data sets, severe outliers in u are rare, so use of the LAD estimator, or other estimators with reduced sensitivity to outliers, is uncommon in applications. Thus the treatment of linear regression throughout the remainder of this text focuses exclusively on least squares methods.

*5.6 Using the *t*-Statistic in Regression When the Sample Size Is Small

When the sample size is small, the exact distribution of the t-statistic is complicated and depends on the unknown population distribution of the data. If, however, the three least squares assumptions hold, the regression errors are homoskedastic, *and* the regression errors are normally distributed, then the OLS estimator is normally distributed and the homoskedasticity-only t-statistic has a Student t distribution. These five assumptions—the three least squares assumptions, that the errors are homoskedastic, and that the errors are normally distributed—are collectively called the **homoskedastic normal regression assumptions**.

The *t*-Statistic and the Student *t* Distribution

Recall from Section 2.4 that the Student t distribution with m degrees of freedom is defined to be the distribution of $Z/\sqrt{W/m}$, where Z is a random variable with a standard normal distribution, W is a random variable with a chi-squared distribution

*This section is optional and is not used in later chapters.

with m degrees of freedom, and Z and W are independent. Under the null hypothesis, the t-statistic computed using the homoskedasticity-only standard error can be written in this form.

The details of the calculation is presented in Sections 17.4 and 18.4, but the main ideas are as follows. The homoskedasticity-only t-statistic testing $\beta_1 = \beta_{1,0}$ is $\tilde{t} = (\hat{\beta}_1 - \beta_{1,0})/\tilde{\sigma}_{\hat{\beta}_1}$, where $\tilde{\sigma}^2_{\hat{\beta}_1}$ is defined in Equation (5.22). Under the homoskedastic normal regression assumptions, Y has a normal distribution, conditional on $X_1, \ldots, X_n$. As discussed in Section 5.5, the OLS estimator is a weighted average of $Y_1, \ldots, Y_n$, where the weights depend on $X_1, \ldots, X_n$ [see Equation (5.32) in Appendix 5.2]. Because a weighted average of independent normal random variables is normally distributed, $\hat{\beta}_1$ has a normal distribution, conditional on $X_1, \ldots, X_n$. Thus $(\hat{\beta}_1 - \beta_{1,0})$ has a normal distribution under the null hypothesis, conditional on $X_1, \ldots, X_n$. In addition, sections 17.4 and 18.4 show that the (normalized) homoskedasticity-only variance estimator has a chi-squared distribution with $n - 2$ degrees of freedom, divided by $n - 2$, and $\tilde{\sigma}^2_{\hat{\beta}_1}$ and $\hat{\beta}_1$ are independently distributed. Consequently, the homoskedasticity-only t-statistic has a Student t distribution with $n - 2$ degrees of freedom.

This result is closely related to a result discussed in Section 3.5 in the context of testing for the equality of the means in two samples. In that problem, if the two population distributions are normal with the same variance and if the t-statistic is constructed using the pooled standard error formula [Equation (3.23)], then the (pooled) t-statistic has a Student t distribution. When X is binary, the homoskedasticity-only standard error for $\hat{\beta}_1$ simplifies to the pooled standard error formula for the difference of means. It follows that the result of Section 3.5 is a special case of the result that if the homoskedastic normal regression assumptions hold, then the homoskedasticity-only regression t-statistic has a Student t distribution (see Exercise 5.10).

Use of the Student *t* Distribution in Practice

If the regression errors are homoskedastic and normally distributed and if the homoskedasticity-only t-statistic is used, then critical values should be taken from the Student t distribution (Appendix Table 2) instead of the standard normal distribution. Because the difference between the Student t distribution and the normal distribution is negligible if n is moderate or large, this distinction is relevant only if the sample size is small.

In econometric applications, there is rarely a reason to believe that the errors are homoskedastic and normally distributed. Because sample sizes typically are large, however, inference can proceed as described in Section 5.1 and 5.2—that is, by first computing heteroskedasticity-robust standard errors and then by using the standard normal distribution to compute p-values, hypothesis tests, and confidence intervals.

5.7 Conclusion

Return for a moment to the problem that started Chapter 4: the superintendent who is considering hiring additional teachers to cut the student–teacher ratio. What have we learned that she might find useful?

Our regression analysis, based on the 420 observations for 1998 in the California test score data set, showed that there was a negative relationship between the student–teacher ratio and test scores: Districts with smaller classes have higher test scores. The coefficient is moderately large, in a practical sense: Districts with two fewer students per teacher have, on average, test scores that are 4.6 points higher. This corresponds to moving a district at the 50th percentile of the distribution of test scores to approximately the 60th percentile.

The coefficient on the student–teacher ratio is statistically significantly different from 0 at the 5% significance level. The population coefficient might be 0, and we might simply have estimated our negative coefficient by random sampling variation. However, the probability of doing so (and of obtaining a t-statistic on β_1 as large as we did) purely by random variation over potential samples is exceedingly small, approximately 0.001%. A 95% confidence interval for β_1 is $-3.30 \leq \beta_1 \leq -1.26$.

This result represents considerable progress toward answering the superintendent's question yet a nagging concern remains. There is a negative relationship between the student–teacher ratio and test scores, but is this relationship necessarily the *causal* one that the superintendent needs to make her decision? Districts with lower student–teacher ratios have, on average, higher test scores. But does this mean that reducing the student–teacher ratio will, in fact, increase scores?

There is, in fact, reason to worry that it might not. Hiring more teachers, after all, costs money, so wealthier school districts can better afford smaller classes. But students at wealthier schools also have other advantages over their poorer neighbors, including better facilities, newer books, and better-paid teachers. Moreover, students at wealthier schools tend themselves to come from more affluent families and thus have other advantages not directly associated with their school. For example, California has a large immigrant community; these immigrants tend to be poorer than the overall population, and in many cases, their children are not native English speakers. It thus might be that our negative estimated relationship between test scores and the student–teacher ratio is a consequence of large classes being found in conjunction with many other factors that are, in fact, the real cause of the lower test scores.

These other factors, or "omitted variables," could mean that the OLS analysis done so far has little value to the superintendent. Indeed, it could be misleading:

Changing the student–teacher ratio alone would not change these other factors that determine a child's performance at school. To address this problem, we need a method that will allow us to isolate the effect on test scores of changing the student–teacher ratio, *holding these other factors constant*. That method is multiple regression analysis, the topic of Chapters 6 and 7.

Summary

1. Hypothesis testing for regression coefficients is analogous to hypothesis testing for the population mean: Use the *t*-statistic to calculate the *p*-values and either accept or reject the null hypothesis. Like a confidence interval for the population mean, a 95% confidence interval for a regression coefficient is computed as the estimator ± 1.96 standard errors.
2. When X is binary, the regression model can be used to estimate and test hypotheses about the difference between the population means of the "$X = 0$" group and the "$X = 1$" group.
3. In general, the error u_i is heteroskedastic—that is, the variance of u_i at a given value of X_i, $\text{var}(u_i \mid X_i = x)$, depends on x. A special case is when the error is homoskedastic—that is, $\text{var}(u_i \mid X_i = x)$ is constant. Homoskedasticity-only standard errors do not produce valid statistical inferences when the errors are heteroskedastic, but heteroskedasticity-robust standard errors do.
4. If the three least squares assumption hold *and* if the regression errors are homoskedastic, then, as a result of the Gauss–Markov theorem, the OLS estimator is BLUE.
5. If the three least squares assumptions hold, if the regression errors are homoskedastic, *and* if the regression errors are normally distributed, then the OLS *t*-statistic computed using homoskedasticity-only standard errors has a Student *t* distribution when the null hypothesis is true. The difference between the Student *t* distribution and the normal distribution is negligible if the sample size is moderate or large.

Key Terms

null hypothesis (148)
two-sided alternative hypothesis (148)
standard error of $\hat{\beta}_1$ (148)
t-statistic (148)
p-value (149)
confidence interval for β_1 (153)
confidence level (153)

indicator variable (155)
dummy variable (155)
coefficient multiplying D_i (156)
coefficient on D_i (156)
heteroskedasticity and homoskedasticity (158)
homoskedasticity-only standard errors (160)
heteroskedasticity-robust standard error (161)
Gauss–Markov theorem (164)
best linear unbiased estimator (BLUE) (165)
weighted least squares (165)
homoskedastic normal regression assumptions (166)
Gauss–Markov conditions (179)

Review the Concepts

5.1 Outline the procedures for computing the p-value of a two-sided test of $H_0: \mu_Y = 0$ using an i.i.d. set of observations $Y_i, i = 1, \ldots, n$. Outline the procedures for computing the p-value of a two-sided test of $H_0: \beta_1 = 0$ in a regression model using an i.i.d. set of observations $(Y_i, X_i), i = 1, \ldots, n$.

5.2 Explain how you could use a regression model to estimate the wage gender gap using the data on earnings of men and women. What are the dependent and independent variables?

5.3 Define *homoskedasticity* and *heteroskedasticity*. Provide a hypothetical empirical example in which you think the errors would be heteroskedastic and explain your reasoning.

5.4 Consider the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$, where Y_i denotes a worker's average hourly earnings (measured in dollars) and X_i is a binary (or indicator) variable that is equal to 1 if the worker has a college degree and is equal to 0 otherwise. Suppose that $\beta_1 = 8.1$. Explain what this value means. Include the units of β_1 in your answer.

Exercises

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

$$\widehat{TestScore} = \underset{(20.4)}{520.4} - \underset{(2.21)}{5.82} \times CS, R^2 = 0.08, SER = 11.5.$$

a. Construct a 95% confidence interval for β_1, the regression slope coefficient.

b. Calculate the p-value for the two-sided test of the null hypothesis H_0: $\beta_1 = 0$. Do you reject the null hypothesis at the 5% level? At the 1% level?

c. Calculate the p-value for the two-sided test of the null hypothesis H_0: $\beta_1 = -5.6$. Without doing any additional calculations, determine whether -5.6 is contained in the 95% confidence interval for β_1.

d. Construct a 99% confidence interval for β_0.

5.2 Suppose that a researcher, using wage data on 250 randomly selected male workers and 280 female workers, estimates the OLS regression

$$\widehat{Wage} = \underset{(0.23)}{12.52} + \underset{(0.36)}{2.12} \times Male, R^2 = 0.06, SER = 4.2,$$

where *Wage* is measured in dollars per hour and *Male* is a binary variable that is equal to 1 if the person is a male and 0 if the person is a female. Define the wage gender gap as the difference in mean earnings between men and women.

a. What is the estimated gender gap?

b. Is the estimated gender gap significantly different from 0? (Compute the p-value for testing the null hypothesis that there is no gender gap.)

c. Construct a 95% confidence interval for the gender gap.

d. In the sample, what is the mean wage of women? Of men?

e. Another researcher uses these same data but regresses *Wages* on *Female*, a variable that is equal to 1 if the person is female and 0 if the person a male. What are the regression estimates calculated from this regression?

$$\widehat{Wage} = \_\_\_ + \_\_\_ \times Female, R^2 = \_\_\_, SER = \_\_\_.$$

5.3 Suppose that a random sample of 200 20-year-old men is selected from a population, and their heights and weights are recorded. A regression of weight on height yields

$$\widehat{Weight} = \underset{(2.15)}{-99.41} + \underset{(0.31)}{3.94} \times Height, R^2 = 0.81, SER = 10.2,$$

where *Weight* is measured in pounds, and *Height* is measured in inches. A man has a late growth spurt and grows 1.5 inches over the course of a year. Construct a 99% confidence interval for the person's weight gain.

5.4 Read the box "The Economic Value of a Year of Education: Homoskedasticity or Heteroskedasticity?" in Section 5.4. Use the regression reported in Equation (5.23) to answer the following.

a. A randomly selected 30-year-old worker reports an education level of 16 years. What is the worker's expected average hourly earnings?

b. A high school graduate (12 years of education) is contemplating going to a community college for a 2-year degree. How much is this worker's average hourly earnings expected to increase?

c. A high school counselor tells a student that, on average, college graduates earn $10 per hour more than high school graduates. Is this statement consistent with the regression evidence? What range of values is consistent with the regression evidence?

5.5 In the 1980s, Tennessee conducted an experiment in which kindergarten students were randomly assigned to "regular" and "small" classes and given standardized tests at the end of the year. (Regular classes contained approximately 24 students, and small classes contained approximately 15 students.) Suppose that, in the population, the standardized tests have a mean score of 925 points and a standard deviation of 75 points. Let *SmallClass* denote a binary variable equal to 1 if the student is assigned to a small class and equal to 0 otherwise. A regression of *TestScore* on *SmallClass* yields

$$\widehat{TestScore} = \underset{(1.6)}{918.0} + \underset{(2.5)}{13.9} \times SmallClass, R^2 = 0.01, SER = 74.6.$$

a. Do small classes improve test scores? By how much? Is the effect large? Explain.

b. Is the estimated effect of class size on test scores statistically significant? Carry out a test at the 5% level.

c. Construct a 99% confidence interval for the effect of *SmallClass* on *Test Score*.

5.6 Refer to the regression described in Exercise 5.5.

a. Do you think that the regression errors are plausibly homoskedastic? Explain.

b. $SE(\hat{\beta}_1)$ was computed using Equation (5.3). Suppose that the regression errors were homoskedastic: Would this affect the validity of the confidence interval constructed in Exercise 5.5(c)? Explain.

5.7 Suppose that (Y_i, X_i) satisfy the least squares assumptions in Key Concept 4.3. A random sample of size $n = 250$ is drawn and yields

$$\underset{(3.1)}{\hat{Y} = 5.4 + 3.2X}, \underset{(1.5)}{R^2 = 0.26}, SER = 6.2.$$

a. Test H_0: $\beta_1 = 0$ vs. H_1: $\beta_1 \neq 0$ at the 5% level.

b. Construct a 95% confidence interval for β_1.

c. Suppose you learned that Y_i and X_i were independent. Would you be surprised? Explain.

d. Suppose that Y_i and X_i are independent and many samples of size $n = 250$ are drawn, regressions estimated, and (a) and (b) answered. In what fraction of the samples would H_0 from (a) be rejected? In what fraction of samples would the value $\beta_1 = 0$ be included in the confidence interval from (b)?

5.8 Suppose that (Y_i, X_i) satisfy the least squares assumptions in Key Concept 4.3 and, in addition, u_i is $N(0, \sigma_u^2)$ and is independent of X_i. A sample of size $n = 30$ yields

$$\underset{(10.2)}{\hat{Y} = 43.2 + 61.5X}, \underset{(7.4)}{R^2 = 0.54}, SER = 1.52,$$

where the numbers in parentheses are the homoskedastic-only standard errors for the regression coefficients.

a. Construct a 95% confidence interval for β_0.

b. Test H_0: $\beta_1 = 55$ vs. H_1: $\beta_1 \neq 55$ at the 5% level.

c. Test H_0: $\beta_1 = 55$ vs. H_1: $\beta_1 > 55$ at the 5% level.

5.9 Consider the regression model

$$Y_i = \beta X_i + u_i,$$

where u_i and X_i satisfy the least squares assumptions in Key Concept 4.3. Let $\overline{\beta}$ denote an estimator of β that is constructed as $\overline{\beta} = \overline{Y}/\overline{X}$, where $\overline{Y}$ and $\overline{X}$ are the sample means of Y_i and X_i, respectively.

a. Show that $\overline{\beta}$ is a linear function of $Y_1, Y_2, \ldots, Y_n$.

b. Show that $\overline{\beta}$ is conditionally unbiased.

5.10 Let X_i denote a binary variable and consider the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$. Let $\overline{Y}_0$ denote the sample mean for observations with $X = 0$ and let $\overline{Y}_1$ denote the sample mean for observations with $X = 1$. Show that $\hat{\beta}_0 = \overline{Y}_0$, $\hat{\beta}_0 + \hat{\beta}_1 = \overline{Y}_1$, and $\hat{\beta}_1 = \overline{Y}_1 - \overline{Y}_0$.

5.11 A random sample of workers contains $n_m = 120$ men and $n_w = 131$ women. The sample average of men's weekly earnings $[\overline{Y}_m = (1/n_m)\sum_{i=1}^{n_m} Y_{m,i}]$ is \$523.10, and the sample standard deviation $[s_m = \sqrt{\frac{1}{n_m - 1}\sum_{i=1}^{n_m}(Y_{m,i} - \overline{Y}_m)^2}]$ is \$68.10. The corresponding values for women are $\overline{Y}_w = \$485.10$ and $s_w = \$51.10$. Let *Women* denote an indicator variable that is equal to 1 for women and 0 for men and suppose that all 251 observations are used in the regression $Y_i = \beta_0 + \beta_1 Women_i + u_i$. Find the OLS estimates of β_0 and β_1 and their corresponding standard errors.

5.12 Starting from Equation (4.22), derive the variance of $\hat{\beta}_0$ under homoskedasticity given in Equation (5.28) in Appendix 5.1.

5.13 Suppose that (Y_i, X_i) satisfy the least squares assumptions in Key Concept 4.3 and, in addition, u_i is $N(0, \sigma_u^2)$ and is independent of X_i.

a. Is $\hat{\beta}_1$ conditionally unbiased?

b. Is $\hat{\beta}_1$ the best linear conditionally unbiased estimator of β_1?

c. How would your answers to (a) and (b) change if you assumed only that (Y_i, X_i) satisfied the least squares assumptions in Key Concept 4.3 and $\text{var}(u_i \mid X_i = x)$ is constant?

d. How would your answers to (a) and (b) change if you assumed only that (Y_i, X_i) satisfied the least squares assumptions in Key Concept 4.3?

5.14 Suppose that $Y_i = \beta X_i + u_i$, where (u_i, X_i) satisfy the Gauss–Markov conditions given in Equation (5.31).

a. Derive the least squares estimator of β and show that it is a linear function of $Y_1, \ldots, Y_n$.

b. Show that the estimator is conditionally unbiased.

c. Derive the conditional variance of the estimator.

d. Prove that the estimator is BLUE.

5.15 A researcher has two independent samples of observations on (Y_i, X_i). To be specific, suppose that Y_i denotes earnings, X_i denotes years of schooling, and the independent samples are for men and women. Write the regression for men as $Y_{m,i} = \beta_{m,0} + \beta_{m,1}X_{m,i} + u_{m,i}$ and the regression for women as $Y_{w,i} = \beta_{w,0} + \beta_{w,1}X_{w,i} + u_{w,i}$. Let $\hat{\beta}_{m,1}$ denote the OLS estimator constructed using the sample of men, $\hat{\beta}_{w,1}$ denote the OLS estimator constructed from the sample of women, and $SE(\hat{\beta}_{m,1})$ and $SE(\hat{\beta}_{w,1})$ denote the corresponding standard errors. Show that the standard error of $\hat{\beta}_{m,1} - \hat{\beta}_{w,1}$ is given by $SE(\hat{\beta}_{m,1} - \hat{\beta}_{w,1}) = \sqrt{[SE(\hat{\beta}_{m,1})]^2 + [SE(\hat{\beta}_{w,1})]^2}$.

Empirical Exercises

(Only three empirical exercises for this chapter are given in the text, but you can find more on the text website, **http://www.pearsonhighered.com/stock_watson/**.)

E5.1 Use the data set **Earnings_and_Height** described in Empirical Exercise 4.2 to carry out the following exercises.

a. Run a regression of *Earnings* on *Height*.

i. Is the estimated slope statistically significant?

ii. Construct a 95% confidence interval for the slope coefficient.

b. Repeat (a) for women.

c. Repeat (a) for men.

d. Test the null hypothesis that the effect of height on earnings is the same for men and women. (*Hint:* See Exercise 5.15.)

e. One explanation for the effect on height on earnings is that some professions require strength, which is correlated with height. Does the effect of height on earnings disappear when the sample is restricted to occupations in which strength is unlikely to be important?

E5.2 Using the data set **Growth** described in Empirical Exercise 4.1, but excluding the data for Malta, run a regression of *Growth* on *TradeShare*.

a. Is the estimated regression slope statistically significant? This is, can you reject the null hypothesis H_0: $\beta_1 = 0$ vs. a two-sided alternative hypothesis at the 10%, 5%, or 1% significance level?

b. What is the p-value associated with the coefficient's t-statistic?

c. Construct a 90% confidence interval for β_1.

E5.3 On the text website, **http://www.pearsonhighered.com/stock_watson/**, you will find the data file **Birthweight_Smoking**, which contains data for a random sample of babies born in Pennsylvania in 1989. The data include the baby's birth weight together with various characteristics of the mother, including whether she smoked during the pregnancy.[2] A detailed description is given in **Birthweight_Smoking_Description**, also available on the website. In this exercise you will investigate the relationship between birth weight and smoking during pregnancy.

a. In the sample:

i. What is the average value of *Birthweight* for all mothers?

ii. For mothers who smoke?

iii. For mothers who do not smoke?

b. i. Use the data in the sample to estimate the difference in average birth weight for smoking and nonsmoking mothers.

ii. What is the standard error for the estimated difference in (i)?

iii. Construct a 95% confidence interval for the difference in the average birth weight for smoking and nonsmoking mothers.

c. Run a regression of *Birthweight* on the binary variable *Smoker*.

i. Explain how the estimated slope and intercept are related to your answers in parts (a) and (b).

ii. Explain how the $SE(\hat{\beta}_1)$ is related to your answer in b(ii).

iii. Construct a 95% confidence interval for the effect of smoking on birth weight.

d. Do you think smoking is uncorrelated with other factors that cause low birth weight? That is, do you think that the regression error term, say u_i, has a conditional mean of zero, given *Smoking* (X_i)? (You will investigate this further in *Birthweight* and *Smoking* exercises in later chapters.)

[2]These data were provided by Professors Douglas Almond (Columbia University), Ken Chay (Brown University), and David Lee (Princeton University) and were used in their paper "The Costs of Low Birth Weight," *Quarterly Journal of Economics*, August 2005, 120(3): 1031–1083.

APPENDIX

5.1 Formulas for OLS Standard Errors

This appendix discusses the formulas for OLS standard errors. These are first presented under the least squares assumptions in Key Concept 4.3, which allow for heteroskedasticity; these are the "heteroskedasticity-robust" standard errors. Formulas for the variance of the OLS estimators and the associated standard errors are then given for the special case of homoskedasticity.

Heteroskedasticity-Robust Standard Errors

The estimator $\hat{\sigma}^2_{\hat{\beta}_1}$ defined in Equation (5.4) is obtained by replacing the population variances in Equation (4.21) by the corresponding sample variances, with a modification. The variance in the numerator of Equation (4.21) is estimated by $\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2$, where the divisor $n - 2$ (instead of n) incorporates a degrees-of-freedom adjustment to correct for downward bias, analogously to the degrees-of-freedom adjustment used in the definition of the SER in Section 4.3. The variance in the denominator is estimated by $(1/n)\sum_{i=1}^{n}(X_i - \overline{X})^2$. Replacing $\text{var}[(X_i - \mu_X)u_i]$ and $\text{var}(X_i)$ in Equation (4.21) by these two estimators yields $\hat{\sigma}^2_{\hat{\beta}_1}$ in Equation (5.4). The consistency of heteroskedasticity-robust standard errors is discussed in Section 17.3.

The estimator of the variance of $\hat{\beta}_0$ is

$$\hat{\sigma}^2_{\hat{\beta}_0} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{H}_i^2\hat{u}_i^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{H}_i^2\right)^2}, \tag{5.26}$$

where $\hat{H}_i = 1 - (\overline{X}/\frac{1}{n}\sum_{i=1}^{n}X_i^2)X_i$. The standard error of $\hat{\beta}_0$ is $SE(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_0}}$. The reasoning behind the estimator $\hat{\sigma}^2_{\hat{\beta}_0}$ is the same as behind $\hat{\sigma}^2_{\hat{\beta}_1}$ and stems from replacing population expectations with sample averages.

Homoskedasticity-Only Variances

Under homoskedasticity, the conditional variance of u_i given X_i is a constant: $\text{var}(u_i \mid X_i) = \sigma_u^2$. If the errors are homoskedastic, the formulas in Key Concept 4.4 simplify to

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma_u^2}{n\sigma_X^2} \text{ and} \tag{5.27}$$

$$\sigma^2_{\hat{\beta}_0} = \frac{E(X_i^2)}{n\sigma_X^2}\sigma_u^2. \tag{5.28}$$

To derive Equation (5.27), write the numerator in Equation (4.21) as $\text{var}[(X_i - \mu_X)u_i] = E(\{(X_i - \mu_X)u_i - E[(X_i - \mu_X)u_i]\}^2) = E\{[(X_i - \mu_X)u_i]^2\} = E[(X_i - \mu_X)^2 u_i^2] = E[(X_i - \mu_X)^2 \text{var}(u_i \mid X_i)]$, where the second equality follows because $E[(X_i - \mu_X)u_i] = 0$ (by the first least squares assumption) and where the final equality follows from the law of iterated expectations (Section 2.3). If u_i is homoskedastic, then $\text{var}(u_i \mid X_i) = \sigma_u^2$, so $E[(X_i - \mu_X)^2 \text{var}(u_i \mid X_i)] = \sigma_u^2 E[(X_i - \mu_X)^2] = \sigma_u^2 \sigma_X^2$. The result in Equation (5.27) follows by substituting this expression into the numerator of Equation (4.21) and simplifying. A similar calculation yields Equation (5.28).

Homoskedasticity-Only Standard Errors

The homoskedasticity-only standard errors are obtained by substituting sample means and variances for the population means and variances in Equations (5.27) and (5.28) and by estimating the variance of u_i by the square of the *SER*. The homoskedasticity-only estimators of these variances are

$$\tilde{\sigma}^2_{\hat{\beta}_1} = \frac{s^2_{\hat{u}}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad \text{(homoskedasticity-only) and} \tag{5.29}$$

$$\tilde{\sigma}^2_{\hat{\beta}_0} = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) s^2_{\hat{u}}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad \text{(homoskedasticity-only),} \tag{5.30}$$

where $s^2_{\hat{u}}$ is given in Equation (4.19). The homoskedasticity-only standard errors are the square roots of $\tilde{\sigma}^2_{\hat{\beta}_0}$ and $\tilde{\sigma}^2_{\hat{\beta}_1}$.

APPENDIX

5.2 The Gauss–Markov Conditions and a Proof of the Gauss–Markov Theorem

As discussed in Section 5.5, the Gauss–Markov theorem states that if the Gauss–Markov conditions hold, then the OLS estimator is the best (most efficient) conditionally linear unbiased estimator (is BLUE). This appendix begins by stating the Gauss–Markov conditions and showing that they are implied by the three least squares condition plus homoskedasticity.

We next show that the OLS estimator is a linear conditionally unbiased estimator. Finally, we turn to the proof of the theorem.

The Gauss–Markov Conditions

The three Gauss–Markov conditions are

$$\begin{aligned}
&\text{(i)} \quad E(u_i|X_1, \ldots, X_n) = 0 \\
&\text{(ii)} \quad \text{var}(u_i|X_1, \ldots, X_n) = \sigma_u^2, \quad 0 < \sigma_u^2 < \infty \\
&\text{(iii)} \quad E(u_i u_j|X_1, \ldots, X_n) = 0, i \neq j,
\end{aligned} \tag{5.31}$$

where the conditions hold for $i, j = 1, \ldots, n$. The three conditions, respectively, state that u_i has mean zero, that u_i has a constant variance, and that the errors are uncorrelated for different observations, where all these statements hold conditionally on all observed X's $(X_1, \ldots, X_n)$.

The **Gauss–Markov conditions** are implied by the three least squares assumptions (Key Concept 4.3), plus the additional assumptions that the errors are homoskedastic. Because the observations are i.i.d. (Assumption 2), $E(u_i|X_1, \ldots, X_n) = E(u_i|X_i)$, and by Assumption 1, $E(u_i|X_i) = 0$; thus condition (i) holds. Similarly, by Assumption 2, $\text{var}(u_i|X_1, \ldots, X_n) = \text{var}(u_i|X_i)$, and because the errors are assumed to be homoskedastic, $\text{var}(u_i|X_i) = \sigma_u^2$, which is constant. Assumption 3 (nonzero finite fourth moments) ensures that $0 < \sigma_u^2 < \infty$, so condition (ii) holds. To show that condition (iii) is implied by the least squares assumptions, note that $E(u_i u_j|X_1, \ldots, X_n) = E(u_i u_j|X_i, X_j)$ because (X_i, Y_i) are i.i.d. by Assumption 2. Assumption 2 also implies that $E(u_i u_j|X_i, X_j) = E(u_i|X_i)\,E(u_j|X_j)$ for $i \neq j$; because $E(u_i|X_i) = 0$ for all i, it follows that $E(u_i u_j|X_1, \ldots, X_n) = 0$ for all $i \neq j$, so condition (iii) holds. Thus the least squares assumptions in Key Concept 4.3, plus homoskedasticity of the errors, imply the Gauss–Markov conditions in Equation (5.31).

The OLS Estimator $\hat{\beta}_1$ Is a Linear Conditionally Unbiased Estimator

To show that $\hat{\beta}_1$ is linear, first note that, because $\sum_{i=1}^n (X_i - \overline{X}) = 0$ (by the definition of $\overline{X}$), $\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^n (X_i - \overline{X})Y_i - \overline{Y}\sum_{i=1}^n (X_i - \overline{X}) = \sum_{i=1}^n (X_i - \overline{X})Y_i$. Substituting this result into the formula for $\hat{\beta}_1$ in Equation (4.7) yields

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{j=1}^n (X_j - \overline{X})^2} = \sum_{i=1}^n \hat{a}_i Y_i, \text{ where } \hat{a}_i = \frac{(X_i - \overline{X})}{\sum_{j=1}^n (X_j - \overline{X})^2} \tag{5.32}$$

Because the weights $\hat{a}_i$, $i = 1, \ldots, n$, in Equation (5.32) depend on $X_1, \ldots, X_n$ but not on $Y_1, \ldots, Y_n$, the OLS estimator $\hat{\beta}_1$ is a linear estimator.

Under the Gauss–Markov conditions, $\hat{\beta}_1$ is conditionally unbiased, and the variance of the conditional distribution of $\hat{\beta}_1$, given $X_1, \ldots, X_n$, is

$$\text{var}(\hat{\beta}_1 \mid X_1, \ldots, X_n) = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{5.33}$$

The result that $\hat{\beta}_1$ is conditionally unbiased was previously shown in Appendix 4.3.

Proof of the Gauss–Markov Theorem

We start by deriving some facts that hold for all linear conditionally unbiased estimators—that is, for all estimators $\tilde{\beta}_1$ satisfying Equations (5.24) and (5.25). Substituting $Y_i = \beta_0 + \beta_1 X_i + u_i$ into $\tilde{\beta}_1 = \sum_{i=1}^{n} a_i Y_i$ and collecting terms, we have that

$$\tilde{\beta}_1 = \beta_0\left(\sum_{i=1}^{n} a_i\right) + \beta_1\left(\sum_{i=1}^{n} a_i X_i\right) + \sum_{i=1}^{n} a_i u_i. \tag{5.34}$$

By the first Gauss–Markov condition, $E(\sum_{i=1}^{n} a_i u_i | X_1, \ldots, X_n) = \sum_{i=1}^{n} a_i E(u_i | X_1, \ldots, X_n) = 0$; thus taking conditional expectations of both sides of Equation (5.34) yields $E(\tilde{\beta}_1 | X_1, \ldots, X_n) = \beta_0(\sum_{i=1}^{n} a_i) + \beta_1(\sum_{i=1}^{n} a_i X_i)$. Because $\tilde{\beta}_1$ is conditionally unbiased by assumption, it must be that $\beta_0(\sum_{i=1}^{n} a_i) + \beta_1(\sum_{i=1}^{n} a_i X_i) = \beta_1$, but for this equality to hold for all values of β_0 and β_1, it must be the case that, for $\tilde{\beta}_1$ to be conditionally unbiased,

$$\sum_{i=1}^{n} a_i = 0 \text{ and } \sum_{i=1}^{n} a_i X_i = 1. \tag{5.35}$$

Under the Gauss–Markov conditions, the variance of $\tilde{\beta}_1$, conditional on $X_1, \ldots, X_n$, has a simple form. Substituting Equation (5.35) into Equation (5.34) yields $\tilde{\beta}_1 - \beta_1 = \sum_{i=1}^{n} a_i u_i$. Thus $\text{var}(\tilde{\beta}_1 | X_1, \ldots, X_n) = \text{var}(\sum_{i=1}^{n} a_i u_i | X_1, \ldots, X_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \text{cov}(u_i, u_j | X_1, \ldots, X_n)$; applying the second and third Gauss–Markov conditions, the cross terms in the double summation vanish, and the expression for the conditional variance simplifies to

$$\text{var}(\tilde{\beta}_1 \mid X_1, \ldots, X_n) = \sigma_u^2 \sum_{i=1}^{n} a_i^2. \tag{5.36}$$

Note that Equations (5.35) and (5.36) apply to $\hat{\beta}_1$ with weights $a_i = \hat{a}_i$, given in Equation (5.32).

We now show that the two restrictions in Equation (5.35) and the expression for the conditional variance in Equation (5.36) imply that the conditional variance of $\tilde{\beta}_1$ exceeds the conditional variance of $\hat{\beta}_1$ unless $\tilde{\beta}_1 = \hat{\beta}_1$. Let $a_i = \hat{a}_i + d_i$, so $\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} (\hat{a}_i + d_i)^2 = \sum_{i=1}^{n} \hat{a}_i^2 + 2\sum_{i=1}^{n} \hat{a}_i d_i + \sum_{i=1}^{n} d_i^2$.

Using the definition of $\hat{a}_i$ in Equation (5.32), we have that

$$\sum_{i=1}^{n}\hat{a}_i d_i = \frac{\sum_{i=1}^{n}(X_i - \overline{X})d_i}{\sum_{j=1}^{n}(X_j - \overline{X})^2} = \frac{\sum_{i=1}^{n} d_i X_i - \overline{X}\sum_{i=1}^{n} d_i}{\sum_{j=1}^{n}(X_j - \overline{X})^2}$$

$$= \frac{\left(\sum_{i=1}^{n} a_i X_i - \sum_{i=1}^{n}\hat{a}_i X_i\right) - \overline{X}\left(\sum_{i=1}^{n} a_i - \sum_{i=1}^{n}\hat{a}_i\right)}{\sum_{j=1}^{n}(X_j - \overline{X})^2} = 0,$$

where the penultimate equality follows from $d_i = a_i - \hat{a}_i$ and the final equality follows from Equation (5.35) (which holds for both a_i and $\hat{a}_i$). Thus $\sigma_u^2\sum_{i=1}^{n}a_i^2 = \sigma_u^2\sum_{i=1}^{n}\hat{a}_i^2 + \sigma_u^2\sum_{i=1}^{n}d_i^2 = \text{var}(\hat{\beta}_1 \,|\, X_1, \ldots, X_n) + \sigma_u^2\sum_{i=1}^{n}d_i^2$; substituting this result into Equation (5.36) yields

$$\text{var}(\widetilde{\beta}_1 \,|\, X_1, \ldots, X_n) - \text{var}(\hat{\beta}_1 \,|\, X_1, \ldots, X_n) = \sigma_u^2\sum_{i=1}^{n} d_i^2. \tag{5.37}$$

Thus $\widetilde{\beta}_1$ has a greater conditional variance than $\hat{\beta}_1$ if d_i is nonzero for any $i = 1, \ldots, n$. But if $d_i = 0$ for all i, then $a_i = \hat{a}_i$ and $\widetilde{\beta}_1 = \hat{\beta}_1$, which proves that OLS is BLUE.

The Gauss–Markov Theorem When *X* Is Nonrandom

With a minor change in interpretation, the Gauss–Markov theorem also applies to nonrandom regressors; that is, it applies to regressors that do not change their values over repeated samples. Specifically, if the second least squares assumption is replaced by the assumption that $X_1, \ldots, X_n$ are nonrandom (fixed over repeated samples) and $u_1, \ldots, u_n$ are i.i.d., then the foregoing statement and proof of the Gauss–Markov theorem apply directly, except that all of the "conditional on $X_1, \ldots, X_n$" statements are unnecessary because $X_1, \ldots, X_n$ take on the same values from one sample to the next.

The Sample Average Is the Efficient Linear Estimator of *E*(*Y*)

An implication of the Gauss–Markov theorem is that the sample average, $\overline{Y}$, is the most efficient linear estimator of $E(Y_i)$ when $Y_i, \ldots, Y_n$ are i.i.d. To see this, consider the case of regression without an "X" so that the only regressor is the constant regressor $X_{0i} = 1$. Then the OLS estimator $\hat{\beta}_0 = \overline{Y}$. It follows that, under the Gauss–Markov assumptions, $\overline{Y}$ is BLUE. Note that the Gauss–Markov requirement that the error be homoskedastic is automatically satisfied in this case because there is no regressor, so it follows that $\overline{Y}$ is BLUE if $Y_1, \ldots, Y_n$ are i.i.d. This result was stated previously in Key Concept 3.3.