CHAPTER 4

# Linear Regression with One Regressor

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

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

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

This chapter describes methods for making statistical inferences about this regression model 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). Moreover, the OLS estimator can be used to test hypotheses about the

population value of the slope—for example, testing the hypothesis that cutting class size has no effect whatsoever on test scores—and to create confidence intervals for the slope.

## 4.1 The Linear Regression Model

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

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

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

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

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

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

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

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

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

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

where $\beta_0$ is the intercept of this straight line, and, as before, $\beta_{ClassSize}$ is the slope. According to Equation (4.3), if you knew $\beta_0$ 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, 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 (quality of their teachers, background of their students, how lucky the students were on test day, etc.). 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 5). 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 $i^{\text{th}}$ district, let $X_i$ be the average class size in the $i^{\text{th}}$ district, and let $u_i$ denote the other factors influencing the test score in the $i^{\text{th}}$ 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 $\beta_0$ is the intercept of this line and $\beta_1$ is the slope. (The general notation "$\beta_1$" 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** $\beta_0$ and the **slope** $\beta_1$ are the **coefficients** of the population regression line, also known as the **parameters** of the population regression line. The slope $\beta_1$ 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, such as the application in Section 4.7, the intercept has a meaningful economic interpretation. In other applications, however, 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 $i^{\text{th}}$ 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.

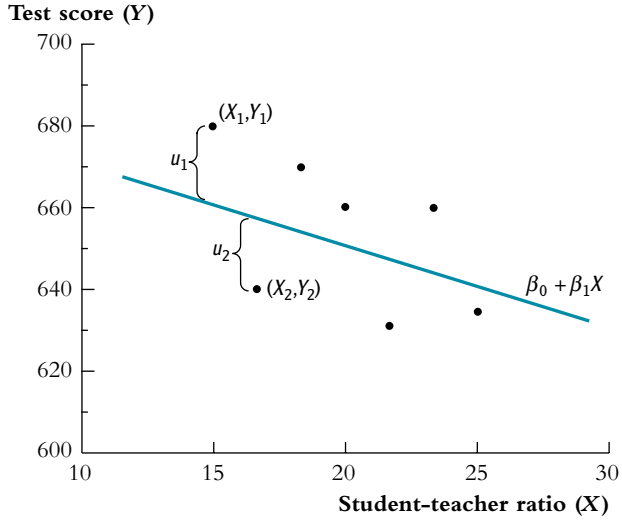
The linear regression model and its terminology are summarized in Key Concept 4.1.

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, that is, $\beta_1 < 0$, which means that districts with lower student-teacher ratios (smaller classes) tend to have higher test scores. The intercept $\beta_0$ has a mathematical meaning as the value of the $Y$ axis intersected by the population regression line, but, as mentioned earlier, it has no real-world meaning in this example.

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

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

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



Key Concept 4.1

## Terminology for the Linear Regression Model with a Single Regressor

The linear regression model is:

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

where:

the subscript $i$ runs over observations, $i = 1, \ldots, n$;
$Y_i$ is the *dependent variable*, the *regressand*, or simply the *left-hand variable*;
$X_i$ is the *independent variable*, the *regressor*, or simply the *right-hand variable*;
$\beta_0 + \beta_1 X$ is the *population regression line* or *population regression function*;
$\beta_0$ is the *intercept* of the population regression line;
$\beta_1$ is the *slope* of the population regression line; and
$u_i$ is the *error term*.

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 $\beta_0$ and slope $\beta_1$ 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 1998 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 $10^{th}$ percentile of the distribution of the student-teacher ratio is 17.3 (that is, only 10% of districts have student-teacher ratios below 17.3), while the district at the $90^{th}$ percentile has a student-teacher ratio of 21.9.

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

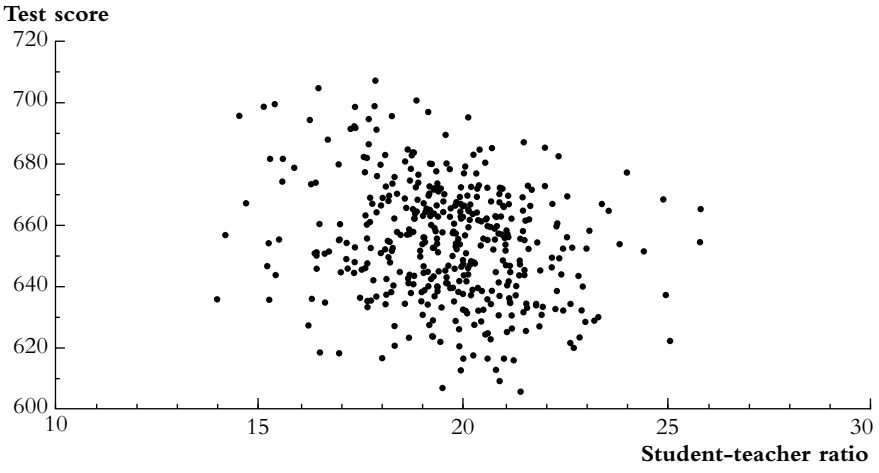
Despite this low correlation, if one could somehow draw a straight line through these data, then the slope of this line would be an estimate of $\beta_{ClassSize}$ based on these data. One way to draw the line would be to take out a pencil and a ruler and to "eyeball" the best line you could. While this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

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

| | | | Percentile | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Average | Standard Deviation | 10% | 25% | 40% | 50% (median) | 60% | 75% | 90% |
| Student-teacher ratio | 19.6 | 1.9 | 17.3 | 18.6 | 19.3 | 19.7 | 20.1 | 20.9 | 21.9 |
| Test score | 654.2 | 19.1 | 630.4 | 640.0 | 649.1 | 654.5 | 659.4 | 666.7 | 679.1 |

**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.




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 $\beta_0$ and $\beta_1$. 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 $i^{\text{th}}$ 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 $\beta_0$ and $\beta_1$ 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 $\beta_0$ and $\beta_1$.

OLS has its own special notation and terminology. The OLS estimator of $\beta_0$ is denoted $\hat{\beta}_0$, and the OLS estimator of $\beta_1$ is denoted $\hat{\beta}_1$. The **OLS regression line** is the straight line constructed using the OLS estimators, that is, $\hat{\beta}_0 + \hat{\beta}_1 X$. The **predicted value** of $Y_i$ given $X_i$, based on the OLS regression line, is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. The **residual** for the $i^{\text{th}}$ observation is the difference between $Y_i$ and its predicted value; that is, the residual is $\hat{u}_i = Y_i - \hat{Y}_i$.

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

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

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

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

$$\widehat{TestScore} = 698.9 - 2.28 \times STR, \tag{4.7}$$

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

## Key Concept 4.2

### The OLS Estimator, Predicted Values, and Residuals

The OLS estimators of the slope $\beta_1$ and the intercept $\beta_0$ are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.8}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{4.9}$$

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

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \; i = 1, \ldots, n \tag{4.10}$$

$$\hat{u}_i = Y_i - \hat{Y}_i, \; i = 1, \ldots, n. \tag{4.11}$$

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 ($\beta_0$), slope ($\beta_1$), and error term ($u_i$).

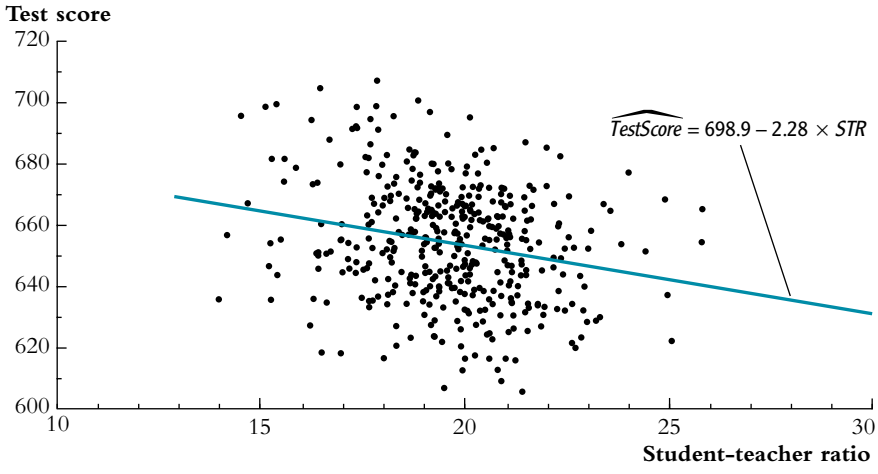
The slope of −2.28 means that an increase in the student-teacher ratio by one student per class is, on average, associated with a decline in districtwide test scores by 2.28 points on the test. A decrease in the student-teacher ratio by 2 students per class is, on average, associated with an increase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

It is now possible to predict the districtwide test score given a value of the student-teacher ratio. For example, for a district with 20 students per teacher, the predicted test score is 698.9 − 2.28 × 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

**FIGURE 4.3 The Estimated Regression Line for the California Data**

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




19.7 and the median test score is 654.5. A reduction of 2 students per class, from 19.7 to 17.7, would move her student-teacher ratio from the 50th percentile to very near the 10th percentile. This is a big change, and she would need to hire many new teachers. How would it affect test scores?

According to Equation (4.7), cutting the student-teacher ratio by 2 is predicted to increase test scores by approximately 4.6 points; if her district's test scores are at the median, 654.5, they are predicted to increase to 659.1. Is this improvement large or small? According to Table 4.1, this improvement would move her district from the median to just short of the 60th percentile. Thus, a decrease in class size that would place her district close to the 10% with the smallest classes would move her test scores from the 50th to the 60th percentile. According to these estimates, at least, cutting the student-teacher ratio by a large amount (2 students per teacher) would help and might be worth doing depending on her budgetary situation, but it would not be a panacea.

What if the superintendent were contemplating a far more radical change, such as reducing the student-teacher ratio from 20 students per teacher to 5? Unfortunately, the estimates in Equation (4.7) 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.

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

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

The table below gives estimated $\beta$'s for six U.S. stocks. Low-risk consumer products firms like Kellogg have stocks with low $\beta$'s; risky high-tech stocks like Microsoft have high $\beta$'s.

| Company | Estimated $\beta$ |
|---|---|
| Kellogg (breakfast cereal) | 0.24 |
| Waste Management (waste disposal) | 0.38 |
| Sprint (long distance telephone) | 0.59 |
| Walmart (discount retailer) | 0.89 |
| Barnes and Noble (book retailer) | 1.03 |
| Best Buy (electronic equipment retailer) | 1.80 |
| Microsoft (software) | 1.83 |

*Source: 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, that paid a \$2.50 dividend during the year and sold on December 31 for \$105, would have a return of $R = [(\$105 - \$100) + \$2.50] / \$100 = 7.5\%$.

## Why Use the OLS Estimator?

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

discussed later in this book) means that you are "speaking the same language" 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. For example, the sample average $\overline{Y}$ is an unbiased estimator of the mean $E(Y)$, that is, $E(\overline{Y}) = \mu_Y$; $\overline{Y}$ is a consistent estimator of $\mu_Y$; and in large samples the sampling distribution of $\overline{Y}$ is approximately normal (Section 3.1). The OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ also have these properties. Under a general set of assumptions (stated in Section 4.3), $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and consistent estimators of $\beta_0$ and $\beta_1$ and their sampling distribution is approximately normal. These results are discussed in Section 4.4.

An additional desirable theoretical property of $\overline{Y}$ is that it is efficient among estimators that are linear functions of $Y_1, \ldots, Y_n$: it has the smallest variance of all estimators that are weighted averages of $Y_1, \ldots, Y_n$ (Section 3.1). A similar result is also true of the OLS estimator, but this result requires an additional assumption beyond those in Section 4.3 so we defer its discussion to Section 4.9.

## 4.3 The Least Squares Assumptions

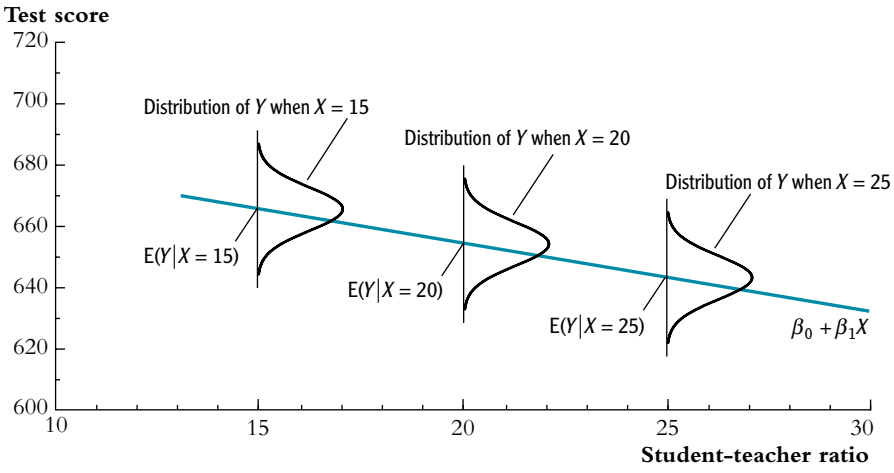
This section presents a set of three assumptions on the linear regression model and the sampling scheme under which OLS provides an appropriate estimator of the unknown regression coefficients, $\beta_0$ and $\beta_1$. Initially these assumptions might appear abstract. They do, however, have natural interpretations, and understanding these assumptions is essential for understanding when OLS will—and will not—give useful estimates of the regression coefficients.

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

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

This is illustrated in Figure 4.4. The population regression is the relationship that holds on average between class size and test scores in the population, and the error term $u_i$ represents the other factors that lead test scores at a given district to differ from the prediction based on the population regression line. As shown in Figure 4.4, at a given value of class size, say 20 students per class, sometimes these

**FIGURE 4.4 The Conditional Probability Distributions and the Population Regression Line**




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

other factors lead to better performance than predicted ($u_i > 0$) and sometimes to worse ($u_i < 0$), 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 around 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, $E(u_i | X_i) = 0$.

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 4.3).

***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 (2.25)). Thus, the conditional mean assumption $E(u_i | X_i) = 0$ implies that $X_i$ and $u_i$ are uncorrelated, or $\text{corr}(X_i, u_i) = 0$. Because correlation is a measure of linear association, this implication does not go the other way; even if $X_i$ and $u_i$ are uncorrelated, the conditional

mean of $u_i$ given $X_i$ might be nonzero. However, if $X_i$ and $u_i$ are correlated, then it must be the case that $E(u_i|X_i)$ is nonzero. It is therefore often convenient to discuss the conditional mean assumption in terms of possible correlation between $X_i$ and $u_i$. If $X_i$ and $u_i$ are correlated, then the conditional mean assumption is violated.

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

The second least squares assumption is that $(X_i, Y_i)$, $i = 1, \ldots, n$ are independently and identically distributed (i.i.d.) across observations. As discussed in Section 2.5 (Key Concept 2.5), this is a statement about how the sample is drawn. If the observations are drawn by simple random sampling from a single large population, then $(X_i, Y_i)$, $i = 1, \ldots, n$ are i.i.d. For example, let $X$ be the age of a worker and $Y$ be his or her earnings, and imagine drawing a person at random from the population of workers. That randomly drawn person will have a certain age and earnings (that is, $X$ and $Y$ will take on some values). If a sample of $n$ workers is drawn from this population, then $(X_i, Y_i)$, $i = 1, \ldots, n$, necessarily have the same distribution, and 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 $i^{\text{th}}$ plot and applies the same technique to the $i^{\text{th}}$ 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 (this is discussed further in Chapter 15). 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 defer further discussion of time series analysis to Part IV.

## Assumption #3: $X_i$ and $u_i$ Have Four Moments

The third least squares assumption is that the fourth moments of $X_i$ and $u_i$ are nonzero and finite ($0 < E(X_i^4) < \infty$ and $0 < E(u_i^4) < \infty$) or, equivalently, that the fourth moments of $X_i$ and $Y_i$ are nonzero and finite. This assumption limits the probability of drawing an observation with extremely large values of $X_i$ or $u_i$. Were we to draw an observation with extremely large $X_i$ or $Y_i$—that is, with $X_i$ or $Y_i$ far outside the normal range of the data—then that observation would be given great importance in an OLS regression and would make the regression results misleading.

The assumption of finite fourth moments is used in the mathematics that justify the large-sample approximations to the distributions of the OLS test statistics. We encountered this assumption in Chapter 3 when discussing the consistency of the sample variance. Specifically, Equation (3.8) states that the sample variance $s_Y^2$ is a consistent estimator of the population variance $\sigma_Y^2$ (that is, that $s_Y^2 \xrightarrow{p} \sigma_Y^2$). If $Y_1, \ldots, Y_n$ are i.i.d. and the fourth moment of $Y_i$ is finite, then the law of large numbers in Key Concept 2.6 applies to the average, $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)^2$, a key step in the proof in Appendix 3.3 showing that $s_Y^2$ is consistent. The role of the fourth moments assumption in the mathematical theory of OLS regression is discussed further in Section 15.3.

One could argue that this assumption is a technical fine point that regularly holds in practice. 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 fourth moments. More generally, commonly used distributions such as the normal have four moments. Still, as a mathematical matter, some distributions have infinite

## Key Concept 4.3

### The Least Squares Assumptions

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

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

fourth moments, and this assumption rules out those distributions. If this assumption holds then it is unlikely that statistical inferences using OLS will be dominated by a few observations.

## Use of the Least Squares Assumptions

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

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 Section 4.10 and Chapter 5, and additional reasons are discussed in Section 7.2.

It is also important to consider whether the second assumption holds in an application. Although it plausibly holds in many cross-sectional data sets, it is inappropriate for time series data. Therefore, the i.i.d. assumption will be replaced by a more appropriate assumption when we discuss regression with time series data in Part IV.

We treat the third assumption as a technical condition that commonly holds in practice so we do not dwell on it further.

## 4.4 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 $Y$, $\mu_Y$. 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 $\mu_Y$, that is, $E(\overline{Y}) = \mu_Y$, so $\overline{Y}$ is an unbiased estimator of $\mu_Y$. 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 $\beta_0$ and slope $\beta_1$ 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 $\beta_0$ and $\beta_1$. In other words, under the least squares assumptions in Key Concept 4.3,

$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1, \tag{4.13}$$

that is, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of $\beta_0$ and $\beta_1$. 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 4.4.

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

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

The results in Key Concept 4.4 imply that the OLS estimators are consistent; that is, when the sample size is large, $\hat{\beta}_0$ and $\hat{\beta}_1$ will be close to the true population coefficients $\beta_0$ and $\beta_1$ 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, $\beta_0$ and $\beta_1$, 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 arises because the variance of $\hat{\beta}_1$ in Equation (4.14) 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 (4.14) so the smaller is $\sigma^2_{\hat{\beta}_1}$. To get a better sense of why this is so, look at Figure 4.5, 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

## Key Concept 4.4

### Large-Sample Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

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

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)]^2}. \tag{4.14}$$

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

$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{\text{var}(H_i u_i)}{[E(H_i^2)]^2}, \quad \text{where } H_i = 1 - \left(\frac{\mu_X}{E(X_i^2)}\right)X_i. \tag{4.15}$$

possible through *either* the colored or the black dots—which would you choose? It would be easier to draw a precise line through the black dots, which have a larger variance than the colored dots. Similarly, the larger the variance of $X$, the more precise is $\hat{\beta}_1$.

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

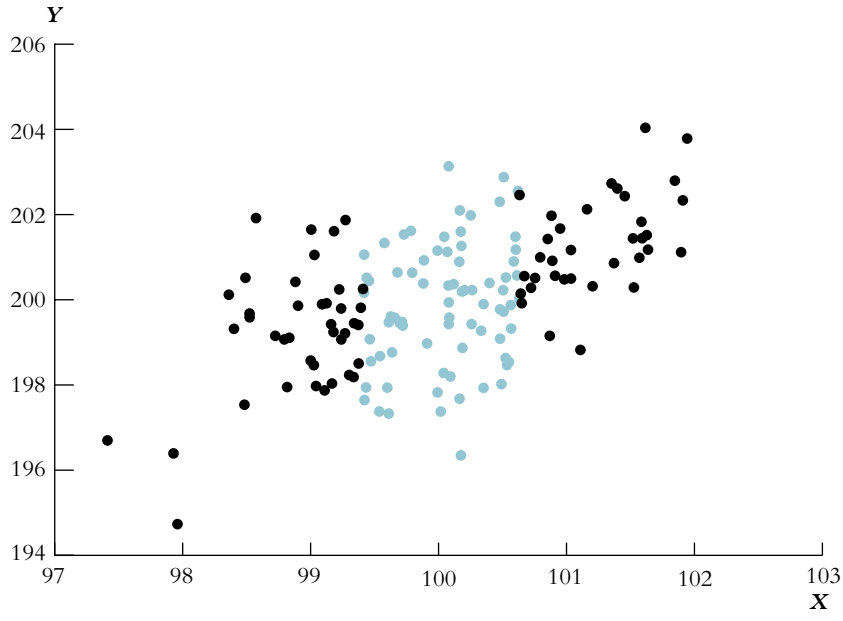
## 4.5 Testing Hypotheses About One of the Regression Coefficients

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 test scores, so that reducing them further 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, that the slope $\beta_{ClassSize}$ of the population regression line is zero. Is there, the superintendent asks, evidence in your sample of 420 observations on California school districts that

**FIGURE 4.5 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.




this slope is nonzero? Can you reject the taxpayer's hypothesis that $\beta_{ClassSize} = 0$, or must you accept it, at least tentatively pending further new evidence?

This section discusses tests of hypotheses about the slope $\beta_1$ or intercept $\beta_0$ of the population regression line. We start by discussing two-sided tests of the slope $\beta_1$ in detail, then turn to one-sided tests and to tests of hypotheses regarding the intercept $\beta_0$.

## Two-Sided Hypotheses Concerning $\beta_1$

The general approach to testing hypotheses about these coefficients 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 $\mu_{Y,0}$ cyan 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 4.5; 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 $\Phi$ 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 than the hypothesized value at the 5% significance level.

***Testing hypotheses about the slope $\beta_1$.*** 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 $\beta_1$ 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 $\beta_1$ takes on some specific value, $\beta_{1,0}$. Under the two-sided alternative, $\beta_1$ does not equal $\beta_{1,0}$. That is, the **null hypothesis** and the **two-sided alternative hypothesis** are

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_1: \beta_1 \neq \beta_{1,0} \text{ (two-sided alternative)}. \tag{4.16}$$

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,

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}, \tag{4.17}$$

## General Form of the *t*-Statistic

**Key Concept 4.5**

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

$$t = \frac{\text{estimator} - \text{hypothesized value}}{\text{standard error of the estimator}}. \tag{4.18}$$

where

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}. \tag{4.19}$$

The estimator of the variance in Equation (4.19) is discussed in Appendix 4.4. 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**,

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}. \tag{4.20}$$

The third step is to compute the ***p*-value**, that is, 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,

$$\begin{aligned} p\text{-value} &= \Pr_{H_0}\left[|\hat{\beta}_1 - \beta_{1,0}| > |\hat{\beta}_1^{act} - \beta_{1,0}|\right] \\ &= \Pr_{H_0}\left[\left|\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right| > \left|\frac{\hat{\beta}_1^{act} - \beta_{1,0}}{SE(\hat{\beta}_1)}\right|\right] = \Pr_{H_0}(|t| > |t^{act}|), \end{aligned} \tag{4.21}$$

where, $\Pr_{H_0}$ denotes the probability computed under the null hypthesis, 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,

$$p\text{-value} = \Pr(|Z| > |t^{act}|) = 2\Phi(-|t^{act}|). \tag{4.22}$$

A small value of the $p$-value, say less than 5%, provides evidence against the null hypothesis in the sense that the chance of obtaining a value of $\hat{\beta}_1$ by pure random variation from one sample to the next is less than 5% if in fact the null hypothesis is correct. 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 value of the $t$-statistic to $\pm 1.96$, 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 4.6.

***Application to test scores.*** The OLS estimator of the slope coefficient, estimated using the 420 observations in Figure 4.2 and reported in Equation (4.7), is $-2.28$. Its standard error is 0.52, that is, $SE(\hat{\beta}_1) = 0.52$. Thus, to test the null hypothesis that $\beta_{ClassSize} = 0$, we construct the $t$-statistic using Equation (4.20); accordingly, $t^{act} = (-2.28 - 0)/0.52 = -4.38$.

This $t$-statistic exceeds the 1% two-sided critical value of 2.58, so the null hypothesis is rejected in favor of the two-sided alternative at the 1% significance level. Alternatively, we can compute the $p$-value associated with $t = -4.38$. This probability is the area in the tails of standard normal distribution, as shown in Figure 4.6. This probability is extremely small, approximately .00001, or .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 .001%. Because this event is so unlikely, it is reasonable to conclude that the null hypothesis is false.

## One-Sided Hypothesis Concerning $\beta_1$

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 $\beta_1$ could be either larger or smaller than $\beta_{1,0}$. Sometimes, however, it 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, $\beta_1$ 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

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_1: \beta_1 < \beta_{1,0}, \text{ (one-sided alternative).} \tag{4.23}$$

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

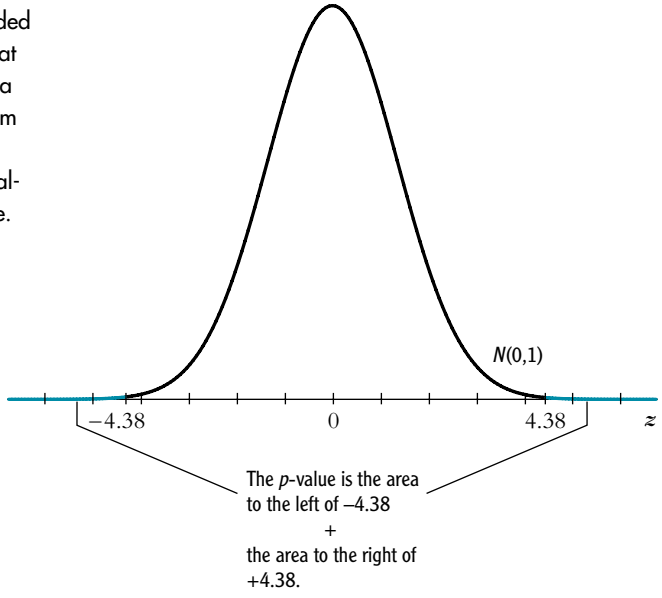
1. Compute the standard error of $\hat{\beta}_1$, $SE(\hat{\beta}_1)$ (Equation (4.17)).
2. Compute the $t$-statistic (Equation (4.20).
3. Compute the $p$-value (Equation (4.22)). Reject the hypothesis at the 5% significance level if the $p$-value is less than .05 or, equivalently, if $|t^{act}| > 1.96$.

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

**Key Concept 4.6**

**FIGURE 4.6 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 $|Z| \geq |t^{act}|$, 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 .00001.




where $\beta_{1,0}$ is the value of $\beta_1$ under the null (0 in the student-teacher ratio example) and the alternative is that $\beta_1$ is less than $\beta_{1,0}$. If the alternative is that $\beta_1$ is greater than $\beta_{1,0}$, the inequality in Equation (4.23) 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 (4.23), 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 $t^{act} < -1.645$.

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

$$p\text{-value} = \Pr(Z < t^{act}) = \Phi(t^{act}) \quad (p\text{-value, one-sided left-tail test}). \quad (4.24)$$

If the alternative hypothesis is that $\beta_1$ is greater than $\beta_{1,0}$, the inequalities in Equations (4.23) and (4.24) 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 when there is a clear reason for $\beta_1$ being on a certain side of the null value $\beta_{1,0}$ under the alternative. This reason could stem 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 (4.23)) is $t^{act} = -4.38$. This 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 .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 $\beta_0$

This discussion has focused on testing hypotheses about the slope, $\beta_1$. Occasionally, however, the hypothesis concerns the intercept, $\beta_0$. The null hypothesis concerning the intercept and the two-sided alternative are

$$H_0: \beta_0 = \beta_{0,0} \text{ vs. } H_1: \beta_0 \neq \beta_{0,0} \quad (\text{two-sided alternative}). \quad (4.25)$$

The general approach to testing this null hypothesis consists of the three steps in Key Concept 4.6, applied to $\beta_0$ (the formula for the standard error of $\hat{\beta}_0$ is given in Appendix 4.4). 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 to 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.

## 4.6 Confidence Intervals for a Regression Coefficient

Because any statistical estimate of the slope $\beta_1$ necessarily has sampling uncertainty, we cannot determine the true value of $\beta_1$ exactly from a sample of data. It is, however, possible to use the OLS estimator and its standard error to construct a confidence interval for the slope $\beta_1$ or for the intercept $\beta_0$.

***Confidence interval for $\beta_1$.*** Recall that a **95% confidence interval** for $\beta_1$ 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 $\beta_1$; that is, in 95% of possible samples that might be drawn, the confidence interval will contain the true value of $\beta_1$. 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 $\beta_1$ in only 5% of all possible samples; that is, in 95% of all possible samples the true value of $\beta_1$ will *not* be rejected. Because the 95% confidence interval (as defined in the first definition) is the set of all values of $\beta_1$ that are *not* rejected at the 5% significance level, it follows that the true value of $\beta_1$ 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 $\beta_1$ (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 $\beta_1$ that are not rejected. But constructing the $t$-statistic for all values of $\beta_1$ 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 $\hat{\beta}_1 \pm 1.96SE(\hat{\beta}_1)$. That is, the 95% confidence interval for $\beta_1$ 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 $\beta_1$ is summarized as Key Concept 4.7.

***Confidence interval for $\beta_0$.*** A 95% confidence interval for $\beta_0$ is constructed as in Key Concept 4.7, 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 (4.7), yielded $\hat{\beta}_0 = 698.7$ 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, we will henceforth include them when reporting OLS regression lines in parentheses below the estimated coefficients:

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR. \tag{4.26}$$

The 95% two-sided confidence interval for $\beta_1$ is $\{-2.28 \pm 1.96 \times 0.52\}$, that is, $-3.30 \leq \beta_1 \leq -1.26$. The value $\beta_1 = 0$ is not contained in this confidence interval, so (as we knew already from Section 4.5) 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 $\beta_1$ 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, $\Delta x$. The predicted change in $Y$ associated with this change in $X$ is $\beta_1 \Delta x$. The population slope $\beta_1$ is unknown, but because we can construct a confidence interval for $\beta_1$, we can construct a confidence interval for the predicted effect $\beta_1 \Delta x$. Because one end of a 95% confidence interval for $\beta_1$ is $\hat{\beta}_1 - 1.96SE(\hat{\beta}_1)$, the predicted effect of the change $\Delta x$ using this estimate of $\beta_1$ is $(\hat{\beta}_1 - 1.96SE(\hat{\beta}_1)) \times \Delta x$. The other end of the confidence interval

### Confidence Intervals for $\beta_1$

Key Concept 4.7

A 95% two-sided confidence interval for $\beta_1$ is an interval that contains the true value of $\beta_1$ with a 95% probability; that is, it contains the true value of $\beta_1$ in 95% of all possible randomly drawn samples. Equivalently, it is also the set of values of $\beta_1$ that cannot be rejected by a 5% two-sided hypothesis test. When the sample size is large, it is constructed as

$$95\% \text{ confidence interval for } \beta_1 = (\hat{\beta}_1 - 1.96SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96SE(\hat{\beta}_1)). \quad (4.27)$$

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 $\Delta x$ can be expressed as

$$\begin{aligned} &95\% \text{ confidence interval for } \beta_1 \Delta x = \\ &(\hat{\beta}_1 \Delta x - 1.96SE(\hat{\beta}_1) \times \Delta x, \hat{\beta}_1 \Delta x + 1.96SE(\hat{\beta}_1) \times \Delta x). \end{aligned} \quad (4.28)$$

For example, our hypothetical superintendent is contemplating reducing the student-teacher ratio by 2. Because the 95% confidence interval for $\beta_1$ 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.

## 4.7 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 $\beta_1$, 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:

$$D_i = \begin{cases} 1 \text{ if the student-teacher ratio in } i^{\text{th}} \text{ district} < 20 \\ 0 \text{ if the student-teacher ratio in } i^{\text{th}} \text{ district} \geq 20. \end{cases} \tag{4.29}$$

The population regression model with $D_i$ as the regressor is

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \; i = 1, \ldots, n. \tag{4.30}$$

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 $\beta_1$ 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 $\beta_1$ as the slope in Equation (4.30); instead we will simply refer to $\beta_1$ as the **coefficient multiplying $D_i$** in this regression or, more compactly, the **coefficient on $D_i$**.

If $\beta_1$ in Equation (4.30) is not a slope, then what is it? The best way to interpret $\beta_0$ and $\beta_1$ 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 (4.30) becomes

$$Y_i = \beta_0 + u_i \quad (D_i = 0). \tag{4.31}$$

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

$$Y_i = \beta_0 + \beta_1 + u_i \quad (D_i = 1). \tag{4.32}$$

Thus, when $D_i = 1$, $E(Y_i | D_i = 1) = \beta_0 + \beta_1$; 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 $\beta_0$ is the population mean of $Y_i$ when $D_i = 0$, the difference $(\beta_0 + \beta_1) - \beta_0 = \beta_1$ is the differ-

ence between these two means. In other words, $\beta_1$ is the difference between the conditional expectation of $Y_i$ when $D_i = 1$ and when $D_i = 0$, or $\beta_1 = E(Y_i | D_i = 1) - E(Y_i | D_i = 0)$. In the test score example, $\beta_1$ 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 $\beta_1$ is the difference in the population means, it makes sense that the OLS estimator $\beta_1$ 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 $\beta_1$ in Equation (4.30) 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 4.5. 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 $\beta_1$, constructed as $\hat{\beta}_1 \pm 1.96 SE(\hat{\beta}_1)$ as described in Section 4.6, 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 (4.29) estimated by OLS using the 420 observations in Figure 4.2, yields

$$\widehat{TestScore} = \underset{(1.3)}{650.0} + \underset{(1.8)}{7.4}D \tag{4.33}$$

where the standard errors of the OLS estimates of the coefficients $\beta_0$ and $\beta_1$ 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. Thus the difference between the sample average test scores for the two groups is 7.4. This is the OLS estimate of $\beta_1$, 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 $\beta_1$: $t = 7.4/1.8 = 4.04$. This 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.

## 4.8 The $R^2$ and the Standard Error of the Regression

The $R^2$ and the standard error of the regression are two measures of 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 variation in $X_i$. 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 $\hat{u}_i$:

$$Y_i = \hat{Y}_i + \hat{u}_i. \tag{4.34}$$

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**, or ***ESS***, is the sum of squared deviations of the predicted values of $Y_i$, $\hat{Y}_i$, from their average, and the **total sum of squares**, or ***TSS***, is the sum of squared deviations of $Y_i$ from its average:

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 \text{ and} \tag{4.35}$$

$$TSS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2, \tag{4.36}$$

where Equation (4.35) uses the fact that $\overline{Y}$ equals the sample average OLS predicted value (proven in Appendix 4.3).

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

$$R^2 = \frac{ESS}{TSS}. \tag{4.37}$$

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:

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2. \tag{4.38}$$

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

$$R^2 = 1 - \frac{SSR}{TSS}. \tag{4.39}$$

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

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

## The Standard Error of the Regression

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

$$SER = s_{\hat{u}}, \text{ where } s_{\hat{u}}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2 = \frac{SSR}{n-2}, \tag{4.40}$$

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 (4.40) is the same as 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 ($\beta_0$ and $\beta_1$), 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 15.4.) When $n$ is large, the difference between dividing by $n$, by $n - 1$, or by $n - 2$ is negligible.

## 4.9 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 4.7 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 4.7 the variance of $u_i$ given $X_i = x$ increases with $x$, so that the errors in Figure 4.7 are heteroskedastic.

The definitions of heteroskedasticity and homoskedasticity are summarized in Key Concept 4.8.

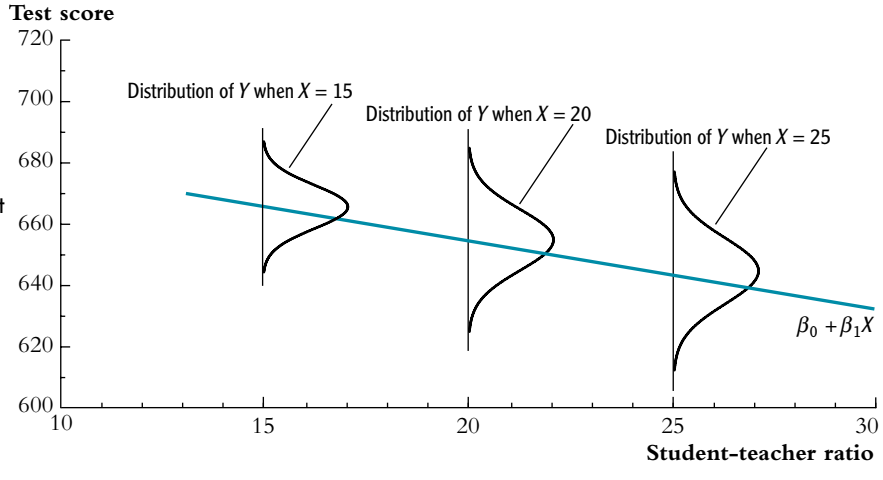
***Example.*** These terms are a mouthful and the definitions might seem abstract. To help clarify them with an example, we digress from the student-teacher ratio/test score problem and instead return to the example of earnings of male versus female college graduates considered in Section 3.5. Let $MALE_i$ be a binary variable that equals 1 for male college graduates and equals 0 for female graduates. The binary variable regression model relating someone's earnings to his or her gender is

$$Earnings_i = \beta_0 + \beta_1 MALE_i + u_i \tag{4.41}$$

for $i = 1, \ldots, n$. Because the regressor is binary, $\beta_1$ 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.

**FIGURE 4.7 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 4.8**

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

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 (4.41) as two separate equations, one for men and one for women:

$$Earnings_i = \beta_0 + u_i \text{ (women) and} \tag{4.42}$$

$$Earnings_i = \beta_0 + \beta_1 + u_i \text{ (men).} \tag{4.43}$$

Thus, for women, $u_i$ is the deviation of the $i^{\text{th}}$ woman's earnings from the population mean earnings for women ($\beta_0$), and for men, $u_i$ is the deviation of the $i^{\text{th}}$ 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.*** If the least squares assumptions in Key Concept 4.3 hold and in addition 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 $X_1, \ldots, X_n$. That is, the OLS estimators have the smallest variance of all unbiased estimators that are weighted averages of $Y_1, \ldots, Y_n$. In other words, if, in addition to the least squares assumptions, the errors are homoskedastic, then the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are the **best linear unbiased estimators**, or **BLUE**. This result was stated for the sample average $\overline{Y}$ in Key Concept 3.3 and it extends to OLS under homoskedasticity. This result, which is known as the Gauss-Markov theorem, is proven in Chapter 15.

If the errors are heteroskedastic, then OLS is no longer BLUE. In theory, if the errors are heteroskedastic then it is possible to construct an estimator that has a smaller variance than the OLS estimator. This method is called **weighted least squares**, in which the observations are weighted 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, applied to this weighted regression, is BLUE. Although theoretically elegant, the problem with weighted least squares in practice is that you must know how the conditional variance of $u_i$ actually depends on $X_i$, which is rarely known in applications. Because weighted least squares is mainly of theoretical interest, we defer further discussion to Chapter 15.

***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 $\hat{\beta}_1$. These formulas are given in Appendix 4.4. 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, discussed in footnote 1 in Section 3.4.

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 $\pm 1.96$ 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 (4.19) and (4.59) 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 (i.e., those based on Equations (4.19) and (4.59)) lead to statistical inferences that are valid whether or not the errors are heteroskedastic, they are called **heteroskedasticity-robust 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. In other words, the variance of the error term in Equation (4.42) for women is plausibly less than the variance of the error term in Equation (4.43) 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 (4.41) 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.

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 use the 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 the empirical examples in this book employ heteroskedasticity-robust standard errors unless explicitly stated otherwise.[2]

## 4.10 Conclusion

Return for a moment to the problem that started this chapter, 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 2 fewer students per teacher have, on average, test scores that are 4.6 points higher. This corresponds to moving a district at the $50^{th}$ percentile of the distribution of test scores to approximately the $60^{th}$ percentile.

[2]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.

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 $\beta_1$ as large as we did) purely by random variation over potential samples is exceedingly small, approximately 0.001%. A 95% confidence interval for $\beta_1$ is $-3.30 \leq \beta_1 \leq -1.26$.

We have made considerable progress towards answering the superintendent's question. Yet, a nagging concern remains. We estimated 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? We have found that 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 small 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 in fact 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 Chapter 5.

## 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, $\beta_1$, is the expected change in $Y$ associated with a 1-unit change in $X$. The intercept, $\beta_0$, determines the level (or height) of the regression

line. Key Concept 4.1 summarizes the terminology of the population linear regression model.

2. The population regression line can be estimated using sample observations $(Y_i, X_i)$, $i = 1, \ldots, n$ by ordinary least squares (OLS). The OLS estimators of the regression intercept and slope are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$.
3. 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) the random variables have four moments. 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.
4. 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 $\pm 1.96$ standard errors.
5. 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.
6. 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.
7. In general the error $u_i$ is heteroskedastic, that is, the variance of $u_i$ at a given value of $X_i$, $\text{var}(u_i | X_i = x)$ depends on $x$. A special case is when the error is homoskedastic, that is, $\text{var}(u_i | X_i = x)$ is constant. Homoskedasticity-only standard errors do not produce valid statistical inferences when the errors are heteroskedastic, but heteroskedasticity-robust standard errors do.

## Key Terms

linear regression model with a single regressor (94)
dependent variable (94)
independent variable (94)
regressor (94)
population regression line (94)
population regression function (94)
population intercept and slope (94)
population coefficients (94)
parameters (94)
error term (94)
ordinary least squares (OLS) estimator (99)
OLS regression line (99)

predicted value (99)
residual (99)
least squares assumptions (103)
standard error of $\hat{\beta}_1$ (112)
*t*-statistic (113)
*p*-value (113)
confidence interval for $\beta_1$ (117)
confidence level (117)
indicator variable (119)
dummy variable (119)
coeffecient multiplying variable $D_1$ (120)
coeffecient on $D_1$ (120)
regression $R^2$ (122)
explained sum of squares (*ESS*) (122)
total sum of squares (*TSS*) (122)
sum of squared residuals (*SSR*) (123)
standard error of the regression (*SER*) (123)
heteroskedasticity and homoskedasticity (124)
best linear unbiased estimator (BLUE) (127)
weighted least squares (127)
homoskedasticity-only standard errors (128)
heteroskedasticity-robust standard error (128)

## Review the Concepts

**4.1** Explain the difference between $\hat{\beta}_1$ and $\beta_1$; 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** 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 $(Y_i, X_i)$, $i = 1, \ldots, n$.

**4.3** Explain how you could use a regression model to estimate the wage gender gap using the data from Section 3.5. What are the dependent and independent variables?

**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

Solutions to exercises denoted by * can be found on the text website at www.aw.com/stock_watson.

***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,

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

**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.** Construct a 95% confidence interval for $\beta_1$, the regression slope coefficient.

**d.** 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?

**e.** 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.)

**f.** 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 researcher, using wage data on 250 randomly selected male workers and 280 female workers, estimates the OLS regression,

$$\widehat{Wage} = 12.68 + 2.79\ Male,\ R^2 = 0.06,\ SER = 3.10$$
$$(0.18)\quad (0.84)$$

where *Wage* is measured in \$/hour and *Male* is a binary variable that is equal to one 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 zero? (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 one if the person is female and zero if the person a male. What are the regression estimates calculated from this regression?

$\widehat{Wage}$ = _____ + _____ *Female*, $R^2$ = _____, *SER* = _____.

*__4.3__ Show that the first least squares assumption, $E(u_i|X_i) = 0$, implies that $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$.

**4.4** Show t hat $\hat{\beta}_0$ is an unbiased estimator of $\beta_0$. (Hint: use the fact that $\hat{\beta}_1$ is unbiased, which is shown in Appendix 4.3).

**4.5** Suppose that a random sample of 200 20-year-old men is selected from a population and their height and weight is recorded. A regression of weight on height yields:

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

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

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

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

**c.** Construct a 99% confidence interval for the weight gain in (b).

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

**4.6** Starting from Equation (4.15), derive the variance of $\hat{\beta}_0$ under homoskedasticity given in Equation (4.61) in Appendix 4.4.

## APPENDIX 4.1 The California Test Score Data Set

The California Standardized Testing and Reporting data set contains data on test performance, school characteristics, and student demographic backgrounds. The data used here are from all 420 K–6 and K–8 districts in California with data available for 1998 and 1999.

Test scores are the average of the reading and math scores on the Stanford 9 Achievement Test, a standardized test administered to fifth-grade students. School characteristics (averaged across the district) include enrollment, number of teachers (measured as "full-time-equivalents"), number of computers per classroom, and expenditures per student. The student-teacher ratio used here is the number of full-time equivalent teachers in the district, divided by the number of students. Demographic variables for the students also are averaged across the district. The demographic variables include the percentage of students in the public assistance program CalWorks (formerly AFDC), the percentage of students that qualify for a reduced price lunch, and the percentage of students that 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 $b_1$:

$$\frac{\partial}{\partial b_0}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i) \text{ and} \tag{4.44}$$

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)X_i. \tag{4.45}$$

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 (4.44) and (4.45) 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,

$$\overline{Y} - \hat{\beta}_0 - \hat{\beta}_1\overline{X} = 0 \text{ and} \tag{4.46}$$

$$\frac{1}{n}\sum_{i=1}^{n}X_iY_i - \hat{\beta}_0\overline{X} - \hat{\beta}_1\frac{1}{n}\sum_{i=1}^{n}X_i^2 = 0. \tag{4.47}$$

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

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}\,\overline{Y}}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{4.48}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{4.49}$$

Equations (4.48) and (4.49) 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 (4.48) 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 (4.48) is

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) &= \sum_{i=1}^{n}(X_i - \overline{X})[\beta_1(X_i - \overline{X}) + (u_i - \overline{u})] \\ &= \beta_1\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{n}(X_i - \overline{X})(u_i - \overline{u}).\end{aligned} \tag{4.50}$$

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 $\sum_{i=1}^{n}(X_i - \overline{X})\overline{u} = \left[\sum_{i=1}^{n} X_i - n\overline{X}\right]\overline{u} = 0$. 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 (4.50)

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 (4.48) yields

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{4.51}$$

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

The expectation of $\hat{\beta}_1$ is obtained by taking the expectation of both sides of Equation (4.51). Thus,

$$\begin{aligned} E(\hat{\beta}_1) &= \beta_1 + E\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}\right] \\ &= \beta_1 + E\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})E(u_i \mid X_1, \ldots, X_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}\right] = \beta_1, \end{aligned} \tag{4.52}$$

where the second equality in Equation (4.52) 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 $E(u_i \mid X_1, \ldots, X_n) = E(u_i \mid X_i)$. By the first least squares assumption, however, $E(u_i \mid X_i) = 0$. Thus, the numerator in the final term in Equation (4.52) is zero, so $E(\hat{\beta}_1) = \beta_1$; that is, the OLS estimator 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 (4.51).

First consider the numerator of this term. Because $\overline{X}$ is consistent, if the sample size is large, $\overline{X}$ is nearly equal to $\mu_X$. Thus, to a close approximation, the term in the numerator of Equation (4.51) 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 (4.51); 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 (4.14).

## Some Additional Algebraic Facts About OLS

The OLS residuals and predicted values satisfy:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i = 0, \tag{4.53}$$

$$\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \overline{Y}, \tag{4.54}$$

$$\sum_{i=1}^{n} \hat{u}_i X_i = 0 \text{ and } s_{\hat{u}X} = 0, \text{ and} \tag{4.55}$$

$$TSS = SSR + ESS. \tag{4.56}$$

Equations (4.53) through (4.56) 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 the sum of squared residuals and the explained sum of squares (the *ESS*, *TSS*, and *SSR* are defined in Equations (4.35), (4.36), and (4.38)).

To verify Equation (4.53), 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 definition 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 (4.54), note that $Y_i = \hat{Y}_i + \hat{u}_i$, so $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i + \sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} \hat{Y}_i$, where the second equality is a consequence of Equation (4.53).

To verify Equation (4.55), 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

$$\begin{aligned}\sum_{i=1}^{n}\hat{u}_i X_i &= \sum_{i=1}^{n}\left[(Y_i - \overline{Y}) - \hat{\beta}_1(X_i - \overline{X})\right](X_i - \overline{X}) \\ &= \sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X}) - \hat{\beta}_1\sum_{i=1}^{n}(X_i - \overline{X})^2 = 0,\end{aligned} \tag{4.57}$$

where the final equality in Equation (4.57) is obtained using the formula for $\hat{\beta}_1$ in Equation (4.48). This result, combined with the preceding results, implies that $s_{\hat{u}X} = 0$.

Equation (4.56) follows from the previous results and some algebra:

$$\begin{aligned}TSS &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}) \\ &= SSR + ESS + 2\sum_{i=1}^{n}\hat{u}_i\hat{Y}_i = SSR + ESS,\end{aligned} \tag{4.58}$$

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.

APPENDIX 4.4

# 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 (4.19) is obtained by replacing the population variances in Equation (4.14) by the corresponding sample variances, with a modification. The variance in the numerator of Equation (4.14) 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 4.8. The variance in the denominator is estimated by $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Replacing $\text{var}[(X_i - \mu_X)u_i]$ and $\text{var}(X_i)$ in Equation (4.14) by these two estimators yields $\hat{\sigma}^2_{\hat{\beta}_1}$ in Equation(4.19). The consistency of heteroskedasticity-robust standard errors is discussed in Section 15.3.

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

$$\hat{\sigma}^2_{\hat{\beta}_0} = \frac{1}{n}\frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{H}_i^2\hat{u}_i^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{H}_i^2\right)^2}, \tag{4.59}$$

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, that is, $\text{var}(u_i|X_i) = \sigma_u^2$. If the errors are homoskedastic, the formulas in Key Concept 4.4 simplify to

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma_u^2}{n\sigma_X^2} \text{ and} \tag{4.60}$$

$$\sigma^2_{\hat{\beta}_0} = \frac{E(X_i^2)}{n\sigma_X^2}\sigma_u^2. \tag{4.61}$$

To derive Equation (4.60), write the numerator in Equation (4.14) as $\text{var}[(X_i - \mu_X)u_i] = E(\{(X_i - \mu_X)u_i - E[(X_i - \mu_X)u_i]\}^2) = E\{[(X_i - \mu_X)u_i]^2\} = E[(X_i - \mu_X)^2u_i^2] = E[(X_i - \mu_X)^2\text{var}(u_i|X_i)]$, 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 $\text{var}(u_i|X_i) = \sigma_u^2$ so $E[(X_i - \mu_X)^2\text{var}(u_i|X_i)] = \sigma_u^2\ E[(X_i - \mu_X)^2] = \sigma_u^2\ \sigma_X^2$. The result in Equation (4.60) follows by substituting this expression into the numerator of Equation (4.14) and simplifying. A similar calculation yields Equation (4.61).

## 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 (4.60) and (4.61), and by estimating the variance of $u_i$ by the square of the *SER*. The homoskedasticity-only estimators of these variances are

$$\tilde{\sigma}^2_{\hat{\beta}_1} = \frac{s^2_{\hat{u}}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \text{ (homoskedasticity-only) and} \tag{4.62}$$

$$\tilde{\sigma}^2_{\hat{\beta}_0} = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)s^2_{\hat{u}}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \text{ (homoskedasticity-only),} \tag{4.63}$$

where $s^2_{\hat{u}}$ is given in Equation (4.40). The homoskedasticity-only standard errors are the square roots of $\tilde{\sigma}^2_{\hat{\beta}_0}$ and $\tilde{\sigma}^2_{\hat{\beta}_1}$.