CHAPTER 4

# Linear Regression with One Regressor

The superintendent of an elementary school district must decide whether to hire additional teachers, and she wants your advice. Hiring the teachers will reduce the number of students per teacher (the student–teacher ratio) by two but will increase the district's expenses. So she asks you: If she cuts class sizes by two, what will the effect be on student performance, as measured by scores on standardized tests?

Now suppose a father tells you that his family wants to move to a town with a good school system. He is interested in a specific school district: Test scores for this district are not publicly available, but the father knows its class size, based on the district's student–teacher ratio. So he asks you: if he tells you the district's class size, could you predict that district's standardized test scores?

These two questions are clearly related: They both pertain to the relation between class size and test scores. Yet they are different. To answer the superintendent's question, you need an estimate of the causal effect of a change in one variable (the student–teacher ratio, $X$) on another (test scores, $Y$). To answer the father's question, you need to know how $X$ relates to $Y$, on average, across school districts so you can use this relation to predict $Y$ given $X$ in a specific district.

These two questions are examples of two different types of questions that arise in econometrics. The first type of questions pertains to **causal inference**: using data to estimate the effect on an outcome of interest of an intervention that changes the value of another variable. The second type of questions concerns **prediction**: using the observed value of some variable to predict the value of another variable.

This chapter introduces the linear regression model relating one variable, $X$, to another, $Y$. This model postulates a linear relationship between $X$ and $Y$. Just as the mean of $Y$ is an unknown characteristic of the population distribution of $Y$, the intercept and slope of the line relating $X$ and $Y$ are unknown characteristics of the population joint distribution of $X$ and $Y$. The econometric problem is to estimate the intercept and slope using a sample of data on these two variables.

Like the differences in means, linear regression is a statistical procedure that can be used for causal inference and for prediction. The two uses, however, place different requirements on the data. Section 3.5 explained how a difference in mean outcomes between a treatment and a control group estimates the causal effect of the treatment when the treatment is randomly assigned in an experiment. When $X$ is continuous, computing differences-in-means no longer works because there are many values $X$ can take on, not just two. If, however, we make the additional assumption that the relation between $X$ and $Y$ is linear, then if $X$ is randomly assigned, we can use linear regression to estimate the causal effect on $Y$ of an intervention that changes $X$. Even if $X$ is not randomly assigned,

however, linear regression gives us a way to predict the value of $Y$ given $X$ by modeling the conditional mean of $Y$ given $X$ as a linear function of $X$. As long as the observation for which $Y$ is to be predicted is drawn from the same population as the data used to estimate the linear regression, the regression line provides a way to predict $Y$ given $X$.

Sections 4.1–4.3 lay out the linear regression model and the least squares estimators of its slope and intercept. In Section 4.4, we turn to requirements on the data for estimation of a causal effect. In essence, the key requirement is that either $X$ is set at random in an experiment or $X$ is as-if randomly set.

Our focus on causal inference continues through Chapter 13. We return to the prediction problem in Chapter 14.

## 4.1 The Linear Regression Model

Return to the father's question: If he tells you the district's class size, could you predict that district's standardized test scores? In Chapter 2, we used the notation $E(Y|X = x)$ to denote the mean of $Y$ given that $X$ takes on the value $x$—that is, the conditional expectation of $Y$ given $X = x$. The easiest starting point for modeling a function of $X$, when $X$ can take on multiple values, is to suppose that it is linear. In the case of test scores and class size, this linear function can be written

$$E(TestScore|ClassSize) = \beta_0 + \beta_{ClassSize} \times ClassSize, \tag{4.1}$$

where $\beta$ is the Greek letter beta, $\beta_0$ is the intercept, and $\beta_{ClassSize}$ is the slope.

If you were lucky enough to know $\beta_0$ and $\beta_{ClassSize}$, you could use Equation (4.1) to answer the father's question. For example, suppose he was looking at a district with a class size of 20 and that $\beta_0 = 720$ and $\beta_{ClassSize} = -0.6$. Then you could answer his question: Given that the class size is 20, you would predict test scores to be $720 - 0.6 \times 20 = 708$.

Equation (4.1) tells you what the test score will be, on average, for districts with class sizes of that value; it does not tell you what specifically the test score will be in any one district. Districts with the same class sizes can nevertheless differ in many ways and in general will have different values of test scores. As a result, if we use Equation (4.1) to make a prediction for a given district, we know that prediction will not be exactly right: The prediction will have an error. Stated mathematically, for any given district the imperfect relationship between class size and test score can be written

$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize + error. \tag{4.2}$$

Equation (4.2) expresses the test score for the district in terms of one component, $\beta_0 + \beta_{ClassSize} \times ClassSize$, that represents the average relationship between class

size and scores in the population of school districts, and a second component that represents the error made using the prediction in Equation (4.1).

Although this discussion has focused on test scores and class size, the idea expressed in Equation (4.2) 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, and let $X_i$ be the average class size in the $i^{\text{th}}$ district, so that Equation (4.1) becomes $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$. Let $u_i$ denote the error made by predicting $Y_i$ using its conditional mean. Then Equation (4.2) can be written more generally as

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{4.3}$$

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 (4.3) instead of $\beta_{ClassSize}$ because this equation is written in terms of a general variable $X$.

Equation (4.3) 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 (4.3), $\beta_0 + \beta_1 X_i$, 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, given the value of $X$, according to this population regression line you would predict the value of the dependent variable, $Y$, to be its conditional mean given $X$. That conditional mean is given by Equation (4.1) which, in the more general notation of Equation (4.3), is $E(Y|X) = \beta_0 + \beta_1 X$.

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 difference in $Y$ associated with a unit difference in $X$. The intercept is the value of the population regression line when $X = 0$; it is the point at which the population regression line intersects the $Y$ axis. In some econometric applications, the intercept has a meaningful economic interpretation. In other applications, the intercept has no real-world meaning; for example, when $X$ is the class size, strictly speaking the intercept is the expected 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 simply as the coefficient that determines the level of the regression line.

The term $u_i$ in Equation (4.3) is the **error term**. In the context of the prediction problem, $u_i$ is the difference between $Y_i$ and its predicted value using the population regression line.

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

KEY CONCEPT 4.1

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

The linear regression model is

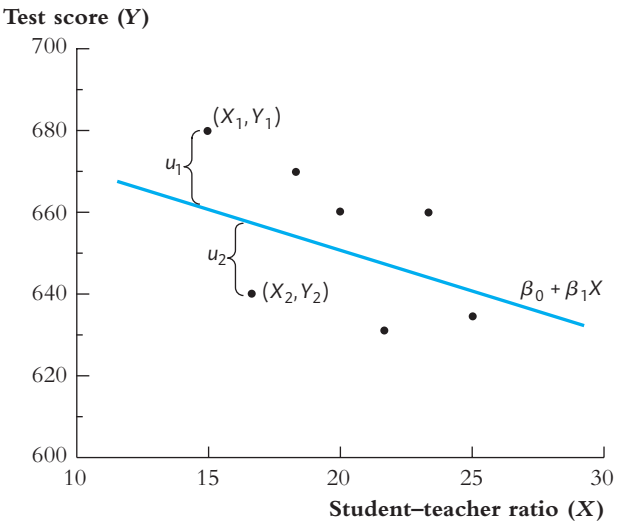
$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

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 the *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*.

**FIGURE 4.1** Scatterplot of Test Score vs. Student–Teacher Ratio (Hypothetical Data)

The scatterplot shows hypothetical observations for seven school districts. The population regression line is $\beta_0 + \beta_1 X$. The vertical distance from the $i^{\text{th}}$ 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 $i^{\text{th}}$ observation.




down ($\beta_1 < 0$), 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.

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 $u_2 < 0$.

## 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 the slope $\beta_1$ of the population regression line are unknown. Therefore, we must use data to estimate these unknown coefficients.

This estimation problem is similar to those faced in Chapter 3. For example, suppose you want to compare the mean earnings of men and women who recently graduated from college. Although the population mean earnings are unknown, we can estimate the population means using a random sample of male and female college graduates. Then the natural estimator of the unknown population mean earnings for women, for example, is the average earnings of the female college graduates in the sample.

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

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

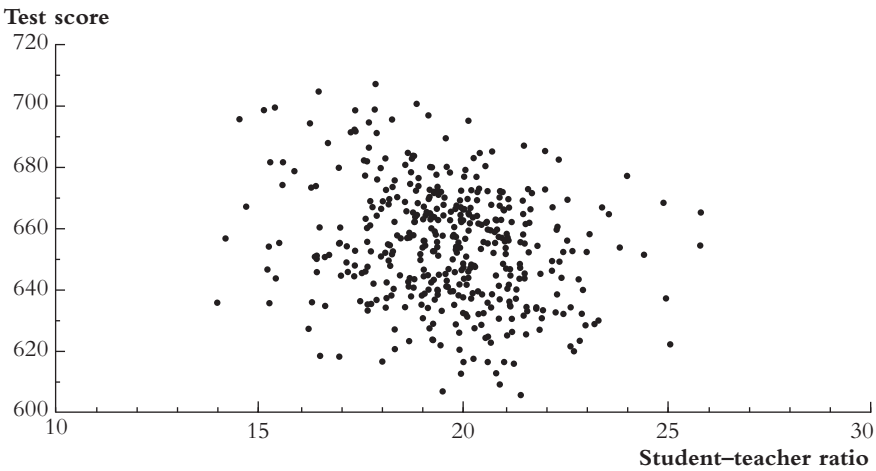
Table 4.1 summarizes the distributions of test scores and class sizes for this sample. The average student–teacher ratio is 19.6 students per teacher, and the standard deviation is 1.9 students per teacher. The 10th percentile of the distribution of

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

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

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

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




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

A scatterplot of these 420 observations on test scores and student–teacher ratios 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 unscientific, and different people would create different estimated lines.

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

## The Ordinary Least Squares Estimator

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

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

The OLS estimator extends this idea to the linear regression model. Let $b_0$ and $b_1$ be some estimators of $\beta_0$ 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

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1X_i)^2. \tag{4.4}$$

The sum of the squared mistakes for the linear regression model in Expression (4.4) 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 (4.4), and the two problems are identical except for the different notation [$m$ in Expression (3.2), $b_0$ in Expression (4.4)]. Just as there is a unique estimator, $\overline{Y}$, that minimizes Expression (3.2), so there is a unique pair of estimators of $\beta_0$ and $\beta_1$ that minimizes Expression (4.4).

The estimators of the intercept and slope that minimize the sum of squared mistakes in Expression (4.4) 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**, also called the **sample regression line** or **sample regression function**, is the straight line constructed using the OLS estimators: $\hat{\beta}_0 + \hat{\beta}_1X$. The **predicted value** of $Y_i$ given $X_i$, based on the OLS regression line, is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1X_i$. The **residual** for the $i^{\text{th}}$ observation is the difference between $Y_i$ and its predicted value: $\hat{u}_i = Y_i - \hat{Y}_i$.

The OLS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$, are sample counterparts of the population coefficients, $\beta_0$ and $\beta_1$. Similarly, the OLS regression line, $\hat{\beta}_0 + \hat{\beta}_1X$, is the sample counterpart of the population regression line, $\beta_0 + \beta_1X$; and the OLS residuals, $\hat{u}_i$, are sample counterparts of the population errors, $u_i$.

You could compute the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ by trying different values of $b_0$ and $b_1$ repeatedly until you find those that minimize the total squared mistakes in Expression (4.4); they are the least squares estimates. This method would be tedious, however. Fortunately, there are formulas, derived by minimizing Expression (4.4) using calculus, that streamline the calculation of the OLS estimators.

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

## 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

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.5}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{4.6}$$

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

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \ldots, n \tag{4.7}$$

$$\hat{u}_i = Y_i - \hat{Y}_i, \quad i = 1, \ldots, n. \tag{4.8}$$

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 estimated intercept is 698.9. Accordingly, the OLS regression line for these 420 observations is

$$\widehat{TestScore} = 698.9 - 2.28 \times STR, \tag{4.9}$$

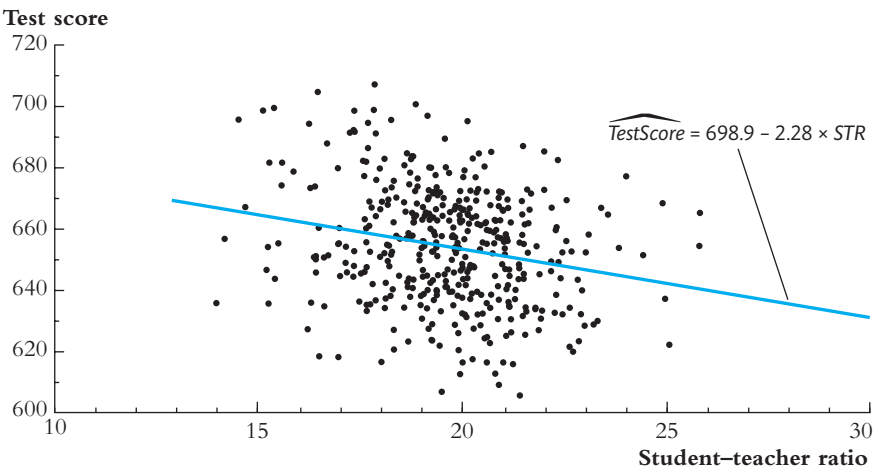
where *TestScore* is the average test score in the district and *STR* is the student–teacher ratio. The "^" over *TestScore* in Equation (4.9) indicates that it is the predicted value based on the OLS regression line. Figure 4.3 plots this OLS regression line superimposed over the scatterplot of the data previously shown in Figure 4.2.

The slope of −2.28 means that when comparing two districts with class sizes that differ by one student per class (that is, *STR* differs by 1), the district with the larger class size has, on average, test scores that are lower by 2.28 points. A difference in the student–teacher ratio of two students per class is, on average, associated with a difference in test scores of 4.56 points $[= -2 \times (-2.28)]$. The negative slope indicates that districts with more students per teacher (larger classes) tend to do worse 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

**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. For two districts with class sizes that differ by one student per class, the district with the larger class has, on average, test scores that are lower by 2.28 points.




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

Is the estimated slope large or small? According to Equation (4.9), for two districts with student-teacher ratios that differ by 2, the predicted value of test scores would differ by 4.56 points. For the California data, this difference of two students per class is large: It is roughly the difference between the median and the $10^{th}$ percentile in Table 4.1. The associated difference in predicted test scores, however, is small compared to the spread of test scores in the data: 4.56 is slightly less than the difference between the median and the $60^{th}$ percentile of test scores. In other words, a difference in class size that is large among these schools is associated with a relatively small difference in predicted test scores.

## Why Use the OLS Estimator?

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

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

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

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

$$R - R_f = \beta(R_m - R_f), \tag{4.10}$$

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 commands a higher expected excess return.

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

The table below gives estimated betas for seven U.S. stocks. Low-risk sellers and producers of consumer staples like Wal-Mart and Coca-Cola have stocks with low betas; riskier stocks have high betas.

| Company | Estimated $\beta$ |
|---|---|
| Wal-Mart (discount retailer) | 0.1 |
| Coca-Cola (soft drinks) | 0.6 |
| Verizon (telecommunications) | 0.7 |
| Google (information technology) | 1.0 |
| General Electric (industrial) | 1.1 |
| Boeing (aircraft) | 1.3 |
| Bank of America (bank) | 1.7 |

*Source:* finance.yahoo.com.

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

The OLS estimators also have desirable theoretical properties. They are analogous to the desirable properties, studied in Section 3.1, of $\overline{Y}$ as an estimator of the population mean. Under the assumptions introduced in Section 4.4, the OLS estimator is unbiased and consistent. The OLS estimator is also efficient among a certain class of unbiased estimators; however, this efficiency result holds under some additional special conditions, and further discussion of this result is deferred until Section 5.5.

# 4.3 Measures of Fit and Prediction Accuracy

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

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

## The $R^2$

The **regression $R^2$** is the fraction of the sample variance of $Y$ explained by (or predicted by) $X$. 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 $\hat{u}_i$:

$$Y_i = \hat{Y}_i + \hat{u}_i. \tag{4.11}$$

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

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

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 \tag{4.12}$$

$$TSS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2. \tag{4.13}$$

Equation (4.12) uses the fact that the sample average OLS predicted value equals $\overline{Y}$ (proven in Appendix 4.3).

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

$$R^2 = \frac{ESS}{TSS}. \tag{4.14}$$

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 (*SSR*)** is the sum of the squared OLS residuals:

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2. \tag{4.15}$$

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

$$R^2 = 1 - \frac{SSR}{TSS}. \tag{4.16}$$

Finally, the $R^2$ of the regression of $Y$ on the single regressor $X$ is the square of the correlation coefficient between $Y$ and $X$ (Exercise 4.12).

The $R^2$ ranges between 0 and 1. If $\hat{\beta}_1 = 0$, then $X_i$ explains none of the variation of $Y_i$, and the predicted value of $Y_i$ is $\hat{Y}_i = \hat{\beta}_0 = \overline{Y}$ [from Equation (4.6)]. In this case, the explained sum of squares is 0 and the sum of squared residuals equals the total sum of squares; thus the $R^2$ is 0. 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 0 (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 value of 0 or 1 but falls somewhere in between. An $R^2$ near 1 indicates that the regressor is good at predicting $Y_i$, while an $R^2$ near 0 indicates that the regressor is not very good at predicting $Y_i$.

## The Standard Error of the Regression

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

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

$$SER = s_{\hat{u}} = \sqrt{s_{\hat{u}}^2}, \text{ where } s_{\hat{u}}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-2}, \tag{4.17}$$

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 0.

The formula for the *SER* in Equation (4.17) is similar to the formula for the sample standard deviation of $Y$ given in Equation (3.7) in Section 3.2, except that $Y_i - \overline{Y}$ in Equation (3.7) is replaced by $\hat{u}_i$ and the divisor in Equation (3.7) is $n - 1$, whereas here it is $n - 2$. The reason for using the divisor $n - 2$ here (instead of $n$) is the same as the reason for using the divisor $n - 1$ in Equation (3.7): It corrects for a slight downward bias introduced because two regression coefficients were estimated. This is called a "degrees of freedom" correction because when 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 5.6.) When $n$ is large, the difference among dividing by $n$, by $n - 1$, or by $n - 2$ is negligible.

## Prediction Using OLS

The predicted value $\hat{Y}_i$ for the $i^{\text{th}}$ observation is the value of $Y$ predicted by the OLS regression line when $X$ takes on its value $X_i$ for that observation. This is called an **in-sample prediction** because the observation for which the prediction is made was also used to estimate the regression coefficients.

In practice, prediction methods are used to predict $Y$ when $X$ is known but $Y$ is not. Such observations are not in the data set used to estimate the coefficients. Prediction for observations *not* in the estimation sample is called **out-of-sample prediction**.

The goal of prediction is to provide accurate out-of-sample predictions. For example, in the father's prediction problem, he was interested in predicting test scores for a district that had not reported them, using that district's student–teacher ratio. In the linear regression model with a single regressor, the predicted value for an out-of-sample observation that takes on the value $X$ is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

Because no prediction is perfect, a prediction should be accompanied by an estimate of its accuracy—that is, by an estimate of how accurate the prediction might reasonably be expected to be. A natural measure of that accuracy is the standard deviation of the out-of-sample prediction error, $Y - \hat{Y}$. Because $Y$ is not known, this out-of-sample standard deviation cannot be estimated directly. If, however, the observation being predicted is drawn from the same population as the data used to estimate the regression coefficients, then the standard deviation of the out-of-sample prediction error can be estimated using the sample standard deviation of the in-sample prediction error, which is the standard error of the regression. A common way to report a prediction and its accuracy is as the prediction $\pm$ the $SER$—that is, $\hat{Y} \pm s_{\hat{u}}$. More refined measures of prediction accuracy are introduced in Chapter 14.

## Application to the Test Score Data

Equation (4.9) reports the regression line, estimated using the California test score data, relating the standardized test score (*TestScore*) to the student–teacher ratio (*STR*). The $R^2$ of this regression is 0.051, or 5.1%, and the *SER* is 18.6.

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

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

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

## 4.4 The Least Squares Assumptions for Causal Inference

In the test score example, the sample regression line, estimated using California district-level data, provides an answer to the father's problem of predicting the test score in a district when he knows its student–teacher ratio but not its test score.

The superintendent, however, is not interested in predicting test scores: She wants to improve them in her district. For that purpose, she needs to know the causal effect on test scores if she were to reduce the student–teacher ratio. Said differently, the superintendent has in mind a very particular definition of $\beta_1$: the causal effect on test scores of an intervention that changes the student–teacher ratio.

When $\beta_1$ is defined to be the causal effect, whether it is well estimated by OLS depends on the nature of the data. As discussed in Section 3.5, the difference in means between the treatment and control groups in an ideal randomized experiment is an unbiased estimator of the causal effect of a binary treatment; that is, if $X$ is randomly assigned, the causal effect of the treatment is $E(Y|X = 1) - E(Y|X = 0)$. The difference in means is a workhorse statistical tool that can be used for many purposes; when $X$ is randomly assigned, it provides an unbiased estimate of the causal effect of a binary treatment. This logic extends to the linear regression model and the least squares estimator.

In this section, we define $\beta_1$ to be the causal effect of a unit change in $X$. Because $X$ can take on multiple values, the causal effect of a given change in $X$, $\Delta x$, is $\beta_1 \Delta x$, where the Greek letter $\Delta$ (delta) stands for "change in." This definition of the coefficient on the variable of interest (for example, $STR$) as its causal effect is maintained through Chapter 13.

This section lays out three mathematical assumptions under which OLS estimates the causal effect. The first assumption translates the idea that $X$ is randomly assigned, or as-if randomly assigned, into the language of linear regression. The other two assumptions are technical ones under which the sampling distributions of the OLS estimators can be approximated by a normal distribution in large samples. These latter two assumptions are extensions of the two assumptions underlying the weak law of large numbers (Key Concept 2.6) and central limit theorem (Key Concept 2.7) for the sample mean $\overline{Y}$: that the data are i.i.d. and that outliers are unlikely.

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

The first least squares assumption translates into the language of regression analysis the requirement that, for estimation of the causal effect, $X$ must be randomly assigned or as-if randomly assigned. To make this translation, we first need to be more specific about what the error term $u_i$ is.

In the test score example, class size is just one of many facets of elementary education. 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, 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 or errors in recording their scores. The error term in the class size regression represents the contribution to test scores made by all these other, omitted factors.

The first **least squares assumption** is that the conditional distribution of $u_i$ given $X_i$ has a mean of 0. 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 0.

***The conditional mean of u in a randomized controlled experiment.*** In a randomized controlled experiment with binary treatment, subjects are randomly assigned to the treatment group $(X = 1)$ or to the control group $(X = 0)$. When random assignment is done using a computer program that uses no information about the subject, $X$ is distributed independently of the subject's personal characteristics, including those that determine $Y$. Because of random assignment, the conditional mean of $u$ given $X$ is 0. Because regression analysis models the conditional mean, $X$ does not need to be distributed independently of all the other factors comprising $u$. However, the mean of $u$ cannot be related to $X$; that is, $E(u_i|X_i) = 0$.

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

***Correlation and conditional mean.*** Recall from Section 2.3 that if the conditional mean of one random variable given another is 0, then the two random variables have 0 covariance and thus are uncorrelated [Equation (2.28)]. 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 (see Figure 3.3). However, if $X_i$ and $u_i$ are correlated, then it must

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

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

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

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

Not all sampling schemes produce i.i.d. observations on $(X_i, Y_i)$. 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 technique (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. Said differently, $X$ is fixed in repeated experiments—that is, repeated draws of the sample. Thus $X_i$ is nonrandom (although the outcome $Y_i$ is random), so the sampling scheme is not i.i.d. The results presented in this chapter developed for i.i.d. regressors are also true if the regressors are nonrandom. The case of a nonrandom regressor is, however, quite special. For example, modern experimental protocols would have the horticulturalist assign the level of $X$ to the different plots using a computerized random number generator, thereby circumventing any possible bias by the horticulturalist (she might use her favorite weeding method for the tomatoes in the sunniest plot). When this modern experimental protocol is used, the level of $X$ is random, and $(X_i, Y_i)$ are i.i.d.

Another example of non-i.i.d. sampling is when observations refer to the same unit of observation over time. For example, we might have data on inventory levels ($Y$) at a firm and the interest rate at which the firm can borrow ($X$), where these data are collected over time from a specific firm; for example, they might be recorded four

times a year (quarterly) for 30 years. This is an example of time series data, and a key feature of time series data is that observations falling close to each other in time are not independent but rather tend to be correlated with each other: If interest rates are low now, they are likely to be low next quarter. This pattern of correlation violates the "independence" part of the i.i.d. assumption. Time series data introduce a set of complications that are best handled after developing the basic tools of regression analysis, so we postpone discussion of time series data until Chapter 15.

## Assumption 3: Large Outliers Are Unlikely

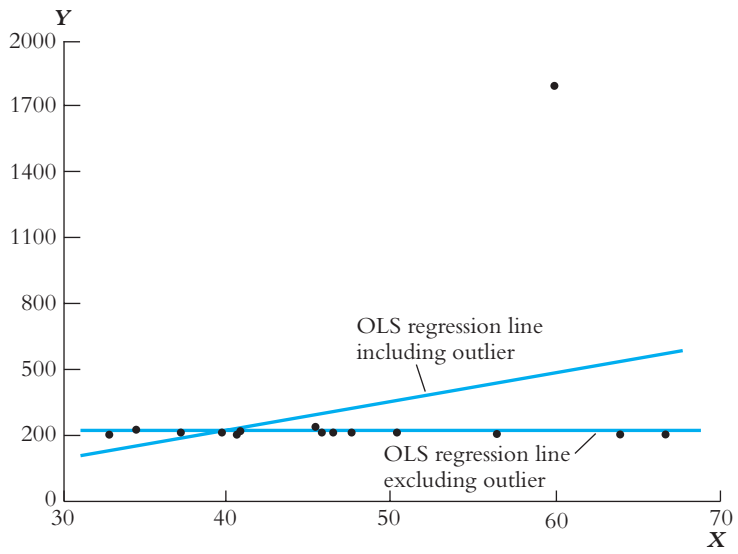
The third least squares assumption is that large outliers—that is, observations with values of $X_i$, $Y_i$, or both that are far outside the usual range of the data—are unlikely. Large outliers can make OLS regression results misleading. This potential sensitivity of OLS to extreme outliers is illustrated in Figure 4.4 using hypothetical data.

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

The assumption of finite kurtosis is used in the mathematics that justify the large-sample approximations to the distributions of the OLS test statistics. For example, we encountered this assumption in Chapter 3 when discussing the consistency of the sample variance. Specifically, Equation (3.9) states that the sample variance is a consistent estimator of the population variance $\sigma_Y^2$ ($s_Y^2 \xrightarrow{p} \sigma_Y^2$). If $Y_1, \ldots, Y_n$ are i.i.d. and the

**FIGURE 4.4 The Sensitivity of OLS to Large Outliers**

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



fourth moment of $Y_i$ is finite, then the law of large numbers in Key Concept 2.6 applies to the average, $\frac{1}{n}\sum_{i=1}^{n} Y_i^2$, a key step in the proof in Appendix 3.3 showing that $s_Y^2$ is consistent.

One source of large outliers is data entry errors, such as a typographical error or incorrectly using different units for different observations. Imagine collecting data on the height of students in meters but inadvertently recording one student's height in centimeters instead. This would create a large outlier in the sample. One way to find outliers is to plot your data. If you decide that an outlier is due to a data entry error, then you can either correct the error or, if that is impossible, drop the observation from your data set.

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

## Use of the Least Squares Assumptions

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

Their first role is mathematical: If these assumptions hold, then, as is shown in the next section, in large samples the OLS estimators are consistent and have sampling distributions that are normal. This large-sample normal distribution underpins methods for testing hypotheses and constructing confidence intervals using the OLS estimators.

KEY CONCEPT 4.3

### The Least Squares Assumptions for Causal Inference

$$Y_i = \beta_0 + \beta_1 X_i + u_i, i = 1, \ldots, n,$$

where $\beta_1$ is the causal effect on $Y$ of $X$, and:

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

Their second role is to organize the circumstances that pose difficulties for OLS estimation of the causal effect $\beta_1$. As we will see, the first least squares assumption is the most important to consider in practice. One reason why the first least squares assumption might not hold in practice is discussed in Chapter 6, and additional reasons are discussed in Section 9.2.

It is also important to consider whether the second assumption holds in an application. Although it plausibly holds in many cross-sectional data sets, the independence assumption is inappropriate for panel and time series data. In those settings, some of the regression methods developed under assumption 2 require modifications. Those modifications are developed in Chapters 10 and 15–17.

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

The assumptions in Key Concept 4.3 apply when the aim is to estimate the causal effect—that is, when $\beta_1$ is the causal effect. Appendix 4.4 lays out a parallel set of least squares assumptions for prediction and discusses their relation to the assumptions in Key Concept 4.3.

## 4.5 The 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. In small samples, these sampling distributions are complicated, but in large samples, they are approximately normal because of the central limit theorem.

***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 $\beta_0$ and $\beta_1$ 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 means 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,

$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1; \tag{4.18}$$

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 4.7.

If the sample is sufficiently large, by the central limit theorem the joint 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 (4.5) 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 a smaller $n$ suffices. This criterion carries over to the more complicated averages appearing in regression analysis. In virtually all modern

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

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)]^2}. \tag{4.19}$$

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

$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{\text{var}(H_i u_i)}{[E(H_i^2)]^2}, \text{ where } H_i = 1 - \left[\frac{\mu_X}{E(X_i^2)}\right]X_i. \tag{4.20}$$

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.

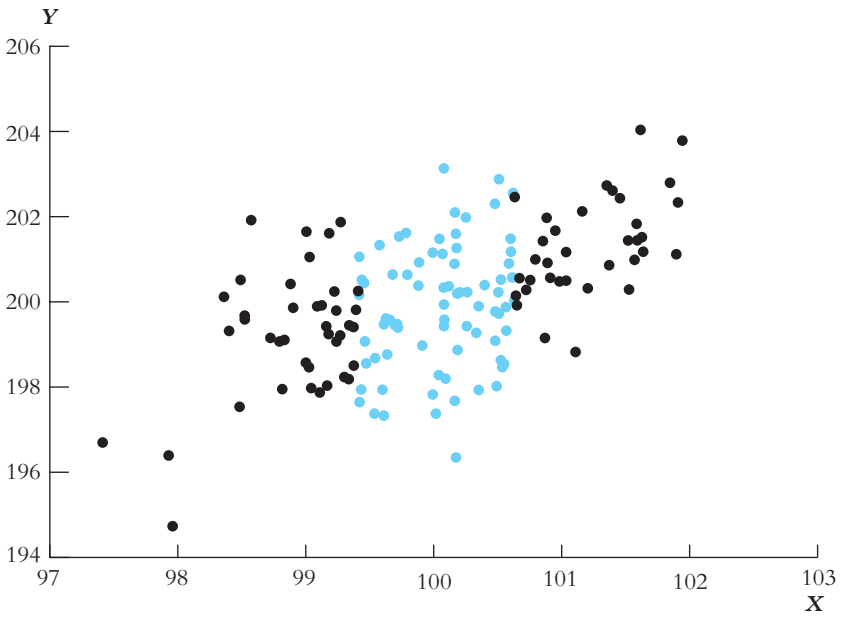
The results in Key Concept 4.4 imply that the OLS estimators are consistent; that is, when the sample size is large and the least squares assumptions hold, $\hat{\beta}_0$ and $\hat{\beta}_1$ will be close to the true population coefficients $\beta_0$ and $\beta_1$ with high probability. This is because the variances $\sigma^2_{\hat{\beta}_0}$ and $\sigma^2_{\hat{\beta}_1}$ of the estimators decrease to 0 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 implication arises because the variance of $\hat{\beta}_1$ in Equation (4.19) is inversely proportional to the square of the variance of $X_i$: the larger is var($X_i$), the larger is the denominator in Equation (4.19) 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 possible through *either* the colored or the black dots—which would you choose? It would be easier to draw a precise line through the black dots, which have a larger variance than the colored dots. Similarly, the larger the variance of $X$, the more precise is $\hat{\beta}_1$.

The distributions in Key Concept 4.4 also imply that the smaller is the variance of the error $u_i$, the smaller is the variance of $\hat{\beta}_1$. This can be seen mathematically in

**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.



Equation (4.19) because $u_i$ enters the numerator, but not denominator, of $\sigma^2_{\hat{\beta}_1}$: If all $u_i$ were smaller by a factor of one-half but the $X$'s did not change, then $\sigma_{\hat{\beta}_1}$ would be smaller by a factor of one-half and $\sigma^2_{\hat{\beta}_1}$ would be smaller by a factor of one-fourth (Exercise 4.13). Stated less mathematically, if the errors are smaller (holding the $X$'s fixed), then the data will have a tighter scatter around the population regression line, so its slope will be estimated more precisely.

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

## 4.6 Conclusion

This chapter has focused on the use of ordinary least squares to estimate the intercept and slope of a population regression line using a sample of $n$ observations on a dependent variable, $Y$, and a single regressor, $X$. The sample regression line, estimated by OLS, can be used to predict $Y$ given a value of $X$. When $\beta_1$ is defined to be the causal effect on $Y$ of a unit change in $X$ and the least squares assumptions for causal inference (Key Concept 4.3) hold, then the OLS estimators of the slope and intercept are unbiased, are consistent, and have a sampling distribution with a variance that is inversely proportional to the sample size $n$. Moreover, if $n$ is large, then the sampling distribution of the OLS estimator is normal.

The first least squares assumption for causal inference is that the error term in the linear regression model has a conditional mean of 0 given the regressor $X$. This assumption holds if $X$ is randomly assigned in an experiment or is as-if randomly assigned in observational data. Under this assumption, the OLS estimator is an unbiased estimator of the causal effect $\beta_1$.

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

The third least squares assumption is that large outliers are unlikely. Stated more formally, $X$ and $Y$ have finite fourth moments (finite kurtosis). This assumption is needed because OLS can be unreliable if there are large outliers. Taken together, the three least squares assumptions imply that the OLS estimator is normally distributed in large samples as described in Key Concept 4.4.

The results in this chapter describe the sampling distribution of the OLS estimator. By themselves, however, these results are not sufficient to test a hypothesis about the value of $\beta_1$ or to construct a confidence interval for $\beta_1$. Doing so requires an estimator of the standard deviation of the sampling distribution—that is, the standard error of the OLS estimator. This step—moving from the sampling distribution of $\hat{\beta}_1$ to its standard error, hypothesis tests, and confidence intervals—is taken in the next chapter.

## Summary

1. The population regression line, $\beta_0 + \beta_1 X$, is the mean of $Y$ as a function of the value of $X$. The slope, $\beta_1$, is the expected difference in $Y$ between two observations with $X$ values that differ by one unit. 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 $\hat{\beta}_0$ and $\hat{\beta}_1$. The predicted value of $Y$ given $X$ is $\hat{\beta}_0 + \hat{\beta}_1 X$.
3. The $R^2$ and standard error of the regression ($SER$) are measures of how close the values of $Y_i$ are to the estimated regression line. The $R^2$ is between 0 and 1, with a larger value indicating that the $Y_i$'s are closer to the line. The standard error of the regression estimates the standard deviation of the regression error.
4. There are three key assumptions for estimating causal effects using the linear regression model: (1) The regression errors, $u_i$, have a mean of 0, conditional on the regressors $X_i$; (2) the sample observations are i.i.d. random draws from the population; and (3) large outliers are unlikely. If these assumptions hold, the OLS estimator $\hat{\beta}_1$ is (1) an unbiased estimator of the causal effect $\beta_1$, (2) consistent, and (3) normally distributed when the sample is large.

## Key Terms

causal inference (143)
prediction (143)
linear regression model with a single regressor (145)
dependent variable (145)
independent variable (145)
regressor (145)
population regression line (145)
population regression function (145)
intercept (145)
slope (145)
coefficients (145)
parameters (145)
error term (145)
ordinary least squares (OLS) estimators (149)
OLS regression line (149)
sample regression line (149)
sample regression function (149)
predicted value (149)
residual (149)
regression $R^2$ (153)
explained sum of squares (ESS) (153)
total sum of squares (TSS) (153)
sum of squared residuals (SSR) (153)
standard error of the regression (SER) (154)
in-sample prediction (155)
out-of-sample prediction (155)
least squares assumptions (157)

## Review the Concepts

**4.1** What is a linear regression model? What is measured by the coefficients of a linear regression model—intercept $\beta_0$ and slope $\beta_1$? What is the ordinary least squares estimator?

**4.2** Explain what is meant by the error term. What assumptions do we make about the error term when estimating an OLS regression?

**4.3** What is meant by the assumption that a paired sample observations of $Y_i$ and $X_i$ are independently and identically distributed? Why is this an important assumption for OLS estimation? When is this assumption likely to be violated?

**4.4** Distinguish between $R^2$ and $SER$. How do each of these measures describe the fit of a regression?

## Exercises

**4.1** Suppose that a researcher, using data on class size ($CS$) and average test scores from 50 third-grade classes, estimates the OLS regression:

$$\widehat{TestScore} = 640.3 - 4.93 \times CS, R^2 = 0.11, SER = 8.7.$$

**a.** A classroom has 25 students. What is the regression's prediction for that classroom's average test score?

**b.** Last year a classroom had 21 students, and this year it has 24 students. What is the regression's prediction for the change in the classroom average test score?

**c.** The sample average class size across the 50 classrooms is 22.8. What is the sample average of the test scores across the 50 classrooms? (*Hint*: Review the formulas for the OLS estimators.)

**d.** What is the sample standard deviation of test scores across the 50 classrooms? (*Hint:* Review the formulas for the $R^2$ and $SER$.)

**4.2** A random sample of 100 20-year-old men is selected from a population and these men's height and weight are recorded. A regression of weight on height yields

$$\widehat{Weight} = -79.24 + 4.16 \times Height, R^2 = 0.72, SER = 12.6,$$

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

**a.** What is the regression's weight prediction for someone who is 64 inches tall? 68 inches tall? 72 inches tall?

**b.** A man has a late growth spurt and grows 2 inches over the course of a year. What is the regression's prediction for the increase in this man's weight?

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

**4.3** A regression of average monthly expenditure (*AME*, measured in dollars) on average monthly income (*AMI*, measured in dollars) using a random sample of college-educated full-time workers earning €100 to €1.5 million yields the following:

$$\widehat{AME} = 710.7 + 8.8 \times AMI, R^2 = 0.030, SER = 540.30$$

**a.** Explain what the coefficient values 710.7 and 8.8 mean.

**b.** The standard error of the regression (*SER*) is 540.30. What are the units of measurement for the *SER*? (Euros? Or is it unit free?)

**c.** The regression $R^2$ is 0.030. What are the units of measurement for the $R^2$? (Euros? Or is $R^2$ unit free?)

**d.** What does the regression predict will be the expenditure of a person with an income of €100? With an income of €200?

**e.** Will the regression give reliable predictions for a person with an income of €2 million? Why or why not?

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

**4.4** Your class is asked to investigate the effect of average temperature on average weekly earnings (*AWE*, measured in dollars) across countries, using the following general regression approach:

$$\widehat{AWE} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{temperature}$$

One of your classmates, Rachel, is an American and decides to analyze the effect of temperature measured in Fahrenheit, while most of the other students analyze the effect of temperature measured in Celsius.

$$X_F = 32 + \frac{9}{5} \times X_C$$

If everything else is the same in Rachel's analysis compared to the other students' analysis, then how will the following quantities differ?

**a.** $\hat{\beta}_0$ (*Hint:* Review Key Concept 2.3)

**b.** $\hat{\beta}_1$

**c.** $R^2$ (*Hint:* $R^2$ is equal to the square of the correlation coefficient, which can be obtained using Equation 2.26)

**4.5** A researcher runs an experiment to measure the impact of a short nap on memory. There are 200 participants and they can take a short nap of either 60 minutes or 75 minutes. After waking up, each participant takes a short test for short-term recall. Each participant is randomly assigned one of the examination times, based on the flip of a coin. Let $Y_i$ denote the number of points scored on the test by the $i^{\text{th}}$ participant ($0 \leq Y_i \leq 100$), let $X_i$ denote the amount of time for which the participant slept prior to taking the test ($X_i = 60$ or 75), and consider the regression model $Y_i = \beta_0 + \beta_i X_i + u_i$.

**a.** Explain what the term $u_i$ represents. Why will different participants have different values of $u_i$?

**b.** What is $E(u_i|X_i)$? Are the estimated coefficients unbiased?

**c.** What concerns might the researcher have about ensuring compliance among participants?

**d.** The estimated regression is $Y_i = 55 + 0.17\,X_i$.

i. Compute the estimated regression's prediction for the average score of participants who slept for 60 minutes before taking the test. Repeat for 75 minutes and 90 minutes.

ii. Compute the estimated gain in score for a participant who is given an additional 5 minutes to nap.

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

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

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

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

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

**4.10** Suppose $Y_i = \beta_0 + \beta_1 X_i + u_i$, where $(X_i, u_i)$ are i.i.d. and $X_i$ is a Bernoulli random variable with $\Pr(X = 1) = 0.30$. When $X = 1, u_i$ is $N(0, 3)$; when $X = 0, u_i$ is $N(0, 2)$.

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

**b.** Derive an expression for large-sample variance of $\hat{\beta}_1$. [*Hint:* Evaluate the terms in Equation (4.19).]

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

**a.** Suppose you know that $\beta_0 = 0$. Derive a formula for the least squares estimator of $\beta_1$.

**b.** Suppose you know that $\beta_0 = 4$. Derive a formula for the least squares estimator of $\beta_1$.

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

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

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

**4.13** Suppose $Y_i = \beta_0 + \beta_1 X_i + \kappa u_i$, where $\kappa$ is a nonzero constant and $(Y_i, X_i)$ satisfy the three least squares assumptions. Show that the large-sample variance of $\hat{\beta}_1$ is given by $\sigma_{\hat{\beta}_1}^2 = \kappa^2 \frac{1}{n} \frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)^2]}$. [*Hint:* This equation is the variance given in Equation (4.19) multiplied by $\kappa^2$.]

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

**4.15** (Requires Appendix 4.4) A sample $(X_i, Y_i), i = 1, \ldots, n$, is collected from a population with $E(Y|X) = \beta_0 + \beta_1 X$ and used to compute the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. You are interested in predicting the value of $Y^{oos}$ from a randomly chosen out-of-sample observation with $X^{oos} = x^{oos}$.

**a.** Suppose the out-of-sample observation is from the same population as the in-sample observations $(X_i, Y_i)$ and is chosen independently of the in-sample observations.

i. Explain why $E(Y^{oos}|X^{oos} = x^{oos}) = \beta_0 + \beta_1 x^{oos}$.

ii. Let $\hat{Y}^{oos} = \hat{\beta}_0 + \hat{\beta}_1 x^{oos}$. Show that $E(\hat{Y}^{oos}|X^{oos} = x^{oos}) = \beta_0 + \beta_1 x^{oos}$.

iii. Let $u^{oos} = Y^{oos} - (\beta_0 + \beta_1 X^{oos})$ and $\hat{u}^{oos} = Y^{oos} - (\hat{\beta}_0 + \hat{\beta}_1 X^{oos})$. Show that $\text{var}(\hat{u}^{oos}) = \text{var}(u^{oos}) + \text{var}(\hat{\beta}_0 + \hat{\beta}_1 X^{oos})$.

**b.** Suppose the out-of-sample observation is drawn from a different population than the in-sample population and that the joint distributions of $X$ and $Y$ differ for the two populations. Continue to let $\beta_0$ and $\beta_1$

be the coefficients of the population regression line for the in-sample population.

i. Does $E(Y^{oos}|X^{oos} = x^{oos}) = \beta_0 + \beta_1 x^{oos}$?

ii. Does $E(\hat{Y}^{oos}|X^{oos} = x^{oos}) = \beta_0 + \beta_1 x^{oos}$?

## Empirical Exercises

**E4.1** On the text website, **http://www.pearsonglobaleditions.com**, you will find the data file **Growth**, which contains data on average growth rates from 1960 through 1995 for 65 countries, along with variables that are potentially related to growth.[1] A detailed description is given in **Growth_Description**, also available on the website. In this exercise, you will investigate the relationship between growth and trade.

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

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

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

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

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

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

**E4.2** On the text website, **http://www.pearsonglobaleditions.com**, you will find the data file **Earnings_and_Height**, which contains data on earnings, height, and other characteristics of a random sample of U.S. workers.[2]

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

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

A detailed description is given in **Earnings_and_Height_Description**, also available on the website. In this exercise, you will investigate the relationship between earnings and height.

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

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

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

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

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

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

i. What is the estimated slope?

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

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

i. What is the estimated slope of the regression?

ii. What is the estimated intercept?

iii. What is the $R^2$?

iv. What is the standard error of the regression?

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

i. What is the estimated slope?

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

**g.** Repeat (f) for male workers.

**h.** Do you think that height is uncorrelated with other factors that cause earning? That is, do you think that the regression error term, $u_i$ has a conditional mean of 0 given *Height* ($X_i$)? (You will investigate this more in the *Earnings* and *Height* exercises in later chapters.)

APPENDIX

## 4.1 The California Test Score Data Set

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

APPENDIX

## 4.2 Derivation of the OLS Estimators

This appendix uses calculus to derive the formulas for the OLS estimators given in Key Concept 4.2. To minimize the sum of squared prediction mistakes $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$ [Equation (4.4)], first take the partial derivatives with respect to $b_0$ and $b_1$:

$$\frac{\partial}{\partial b_0}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i) \text{ and} \tag{4.21}$$

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)X_i. \tag{4.22}$$

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 (4.21) and (4.22) equal 0. Accordingly, setting these derivatives equal to 0, collecting terms, and dividing by $n$ shows that the OLS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy the two equations

$$\overline{Y} - \hat{\beta}_0 - \hat{\beta}_1\overline{X} = 0 \text{ and} \tag{4.23}$$

$$\frac{1}{n}\sum_{i=1}^{n}X_iY_i - \hat{\beta}_0\overline{X} - \hat{\beta}_1\frac{1}{n}\sum_{i=1}^{n}X_i^2 = 0. \tag{4.24}$$

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

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}\,\overline{Y}}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{4.25}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{4.26}$$

Equations (4.25) and (4.26) 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 (4.25) 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 (4.25) is

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) &= \sum_{i=1}^{n}(X_i - \overline{X})[\beta_1(X_i - \overline{X}) + (u_i - \overline{u})] \\ &= \beta_1 \sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{n}(X_i - \overline{X})(u_i - \overline{u}).\end{aligned} \tag{4.27}$$

Now $\sum_{i=1}^{n}(X_i - \overline{X})(u_i - \overline{u}) = \sum_{i=1}^{n}(X_i - \overline{X})u_i - \sum_{i=1}^{n}(X_i - \overline{X})\overline{u} = \sum_{i=1}^{n}(X_i - \overline{X})u_i$, where the final equality follows from the definition of $\overline{X}$, which implies that $\sum_{i=1}^{n}(X_i - \overline{X})\overline{u} = (\sum_{i=1}^{n}X_i - n\overline{X})\overline{u} = 0$. Substituting $\sum_{i=1}^{n}(X_i - \overline{X})(u_i - \overline{u}) = \sum_{i=1}^{n}(X_i - \overline{X})u_i$ into the final expression in Equation (4.27) 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 (4.25) yields

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{4.28}$$

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

The argument that $\hat{\beta}_1$ is unbiased under the first least squares assumption uses the law of iterated expectations [Equation (2.20)]. First, obtain $E(\hat{\beta}_1|X_1, \ldots, X_n)$ by taking the conditional expectation of both sides of Equation (4.28):

$$E(\hat{\beta}_1|X_1, \ldots, X_n) = \beta_1 + E\left[\left.\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}\right| X_1, \ldots, X_n\right]$$
$$= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})E(u_i|X_i, \ldots, X_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}. \quad (4.29)$$

By the second least squares assumption, $u_i$ is distributed independently of $X$ for all observations other than $i$, so $E(u_i|X_1, \ldots, X_n) = E(u_i|X_i)$. By the first least squares assumption, however, $E(u_i|X_i) = 0$. Thus the second term in the final line of Equation (4.29) is 0, from which it follows that $E(\hat{\beta}_1|X_1, \ldots, X_n) = \beta_1$.

Because $\hat{\beta}_1$ is unbiased given $X_1, \ldots, X_n$, it is unbiased after averaging over all samples $X_1, \ldots, X_n$. Thus the unbiasedness of $\hat{\beta}_1$ follows Equation (4.29) and the law of iterated expectations: $E(\hat{\beta}_1) = E[E(\hat{\beta}_1|X_1, \ldots, X_n)] = \beta_1$.

## 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 (4.28).

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 (4.28) 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 0. 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 $\sigma_{\overline{v}}^2 = \sigma_v^2/n$. Therefore 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 (4.28); 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 (4.19).

## Some Additional Algebraic Facts About OLS

The OLS residuals and predicted values satisfy

$$\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0, \tag{4.30}$$

$$\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \overline{Y}, \tag{4.31}$$

$$\sum_{i=1}^{n}\hat{u}_i X_i = 0 \text{ and } s_{\hat{u}X} = 0, \text{ and} \tag{4.32}$$

$$TSS = SSR + ESS. \tag{4.33}$$

Equations (4.30) through (4.33) say that the sample average of the OLS residuals is 0; 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 0; and the total sum of squares is the sum of squared residuals and the explained sum of squares. [The *ESS*, *TSS*, and *SSR* are defined in Equations (4.12), (4.13), and (4.15).]

To verify Equation (4.30), note that the definition of $\hat{\beta}_0$ lets us write the OLS residuals as $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = (Y_i - \overline{Y}) - \hat{\beta}_1(X_i - \overline{X})$; thus

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

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

To verify Equation (4.31), note that $Y_i = \hat{Y}_i + \hat{u}_i$, so $\sum_{i=1}^{n}Y_i = \sum_{i=1}^{n}\hat{Y}_i + \sum_{i=1}^{n}\hat{u}_1 = \sum_{i=1}^{n}\hat{Y}_i$, where the second equality is a consequence of Equation (4.30).

To verify Equation (4.32), note that $\sum_{i=1}^{n}\hat{u}_i = 0$ implies $\sum_{i=1}^{n}\hat{u}_i X_i = \sum_{i=1}^{n}\hat{u}_i(X_i - \overline{X})$, so

$$\begin{aligned}\sum_{i=1}^{n}\hat{u}_i X_i &= \sum_{i=1}^{n}[(Y_i - \overline{Y}) - \hat{\beta}_1(X_i - \overline{X})](X_i - \overline{X}) \\ &= \sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X}) - \hat{\beta}_1\sum_{i=1}^{n}(X_i - \overline{X})^2 = 0,\end{aligned} \tag{4.34}$$

where the final equality in Equation (4.34) is obtained using the formula for $\hat{\beta}_1$ in Equation (4.25). This result, combined with the preceding results, implies that $s_{\hat{u}X} = 0$.

Equation (4.33) follows from the previous results and some algebra:

$$\begin{aligned}TSS &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}) \\ &= SSR + ESS + 2\sum_{i=1}^{n}\hat{u}_i\hat{Y}_i = SSR + ESS,\end{aligned} \tag{4.35}$$

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 The Least Squares Assumptions for Prediction

Section 4.4 provides the least squares assumptions for estimation of a causal effect. There is a parallel set of least squares assumptions for prediction. The difference between the two stems from the difference between the two problems. For estimation of a causal effect, $X$ must be randomly assigned or as-if randomly assigned, which leads to least squares assumption 1 in Key Concept 4.3. In contrast, as discussed in Section 4.3, the goal of prediction is to provide accurate out-of-sample predictions. To do so, the estimated regression line must be relevant to the observation being predicted. This is the case if the data used for estimation and the observation being predicted are drawn from the same population distribution.

For example, return to the superintendent's and father's problems. The superintendent is interested in the causal effect on *TestScore* of a change in *STR*. Ideally, to answer her question we would have data from an experiment in which students were randomly assigned to different size classes. Absent such an experiment, she may or may not be satisfied with the regression of *TestScore* on *STR* using California data—that depends on whether least squares assumption 1 is satisfied where $\beta_1$ is defined to be the causal effect.

In contrast, the father is interested in predicting test scores in a California district that did not report its test scores, so for his purposes he is interested in the population regression line relating *TestScore* and *STR* in California, the slope of which may or may not be the causal effect.

To make this precise, we introduce some additional notation. Let $(X^{oos}, Y^{oos})$ denote the out-of-sample ("oos") observation for which the prediction is to be made, and continue to let $(X_i, Y_i), i = 1, \ldots, n$, be the data used to estimate the regression coefficients. The least squares assumptions for prediction are

$E(Y|X) = \beta_0 + \beta_1 X$ and $u = Y - E(Y|X)$, where

1. $(X^{oos}, Y^{oos})$ are randomly drawn from the same population distribution as $(X_i, Y_i), i = 1, \ldots, n$;
2. $(X_i, Y_i), i = 1, \ldots, n$, are independent and identically distributed (i.i.d.) draws from their joint distribution; and
3. Large outliers are unlikely: $X_i$ and $Y_i$ have nonzero finite fourth moments.

There are two differences between these assumptions and the assumptions in Key Concept 4.3. The first is the definition of $\beta_1$. The best predictor is given by $E(Y|X)$ (where the best predictor is defined in terms of the mean squared prediction error; see Appendix 2.2). With the assumption of linearity, for prediction $\beta_1$ is defined to be the slope of this conditional expectation, which may or may not be the causal effect. Second, because the regression line is estimated using in-sample observations but is used to predict an out-of-sample observation, the first assumption is that these are drawn from the same population.

The second and third assumptions are the same as those for estimation of causal effects in Section 4.4. They ensure that the OLS estimators are consistent for the coefficients of the population prediction line and are normally distributed when $n$ is large.

Under the least squares assumptions for prediction, the OLS predicted value of $Y^{oos}$ is unbiased:

$$\begin{aligned} E(\hat{Y}^{oos}|X^{oos} = x^{oos}) &= E(\hat{\beta}_0 + \hat{\beta}_0 X^{oos}|X^{oos} = x^{oos}) \\ &= E(\hat{\beta}_0) + E(\hat{\beta}_1)x^{oos} \end{aligned} \tag{4.36}$$

where the second equality follows because $(X^{oos},Y^{oos})$ are independent of the observations used to compute the OLS estimators. For the prediction problem, $u$ is defined to be $u = Y - E(Y|X)$, so by definition $E(u_i|X_i) = 0$ and the algebra in Appendix 4.3 applies directly. Thus $E(\hat{\beta}_0) + E(\hat{\beta}_1)x^{oos} = \beta_0 + \beta_1 x^{oos} = E(Y^{oos}|X^{oos} = x^{oos})$. Combining this expression with the first expression in Equation (4.36), we have that $E(Y^{oos} - \hat{Y}^{oos}|X^{oos} = x^{oos}) = 0$; that is, the OLS prediction is unbiased.

The least squares assumptions for prediction also ensure that the regression *SER* estimates the variance of the out-of-sample prediction error, $\hat{u}^{oos} = Y^{oos} - \hat{Y}^{oos}$. To show this, it is useful to write the out-of-sample prediction error as the sum of two terms: the error that would be made were the regression coefficients known and the error made by needing to estimate them. Write $\hat{u}^{oos} = Y^{oos} - (\hat{\beta}_0 + \hat{\beta}_1 X^{oos}) = \beta_0 + \beta_1 X^{oos} + u^{oos} - (\hat{\beta}_0 + \hat{\beta}_1 X^{oos}) = u^{oos} - [(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)X^{oos}]$. Thus $\text{var}(\hat{u}^{oos}) = \text{var}(u^{oos}) + \text{var}(\hat{\beta}_0 + \hat{\beta}_1 X^{oos})$ (Exercise 4.15). The second term in this final expression is the contribution of the estimation error to the out-of-sample prediction error. When the sample size is large, the first term in this final expression is much larger than the second term. Because the in- and out-of-sample observations are from the same population, $\text{var}(u^{oos}) = \text{var}(u_i) = \sigma_u^2$, so the standard deviation of $\hat{u}^{oos}$ is estimated by the *SER*.