Introduction to Econometrics

THIRD EDITION UPDATE

James H. Stock
Harvard University

Mark W. Watson
Princeton University

PEARSON

6 Linear Regression with Multiple Regressors

Chapter 5 ended on a worried note. Although school districts with lower student–teacher ratios tend to have higher test scores in the California data set, perhaps students from districts with small classes have other advantages that help them perform well on standardized tests. Could this have produced misleading results, and, if so, what can be done?

Omitted factors, such as student characteristics, can, in fact, make the ordinary least squares (OLS) estimator of the effect of class size on test scores misleading or, more precisely, biased. This chapter explains this "omitted variable bias" and introduces multiple regression, a method that can eliminate omitted variable bias. The key idea of multiple regression is that if we have data on these omitted variables, then we can include them as additional regressors and thereby estimate the effect of one regressor (the student–teacher ratio) while holding constant the other variables (such as student characteristics).

This chapter explains how to estimate the coefficients of the multiple linear regression model. Many aspects of multiple regression parallel those of regression with a single regressor, studied in Chapters 4 and 5. The coefficients of the multiple regression model can be estimated from data using OLS; the OLS estimators in multiple regression are random variables because they depend on data from a random sample; and in large samples the sampling distributions of the OLS estimators are approximately normal.

6.1 Omitted Variable Bias

By focusing only on the student–teacher ratio, the empirical analysis in Chapters 4 and 5 ignored some potentially important determinants of test scores by collecting their influences in the regression error term. These omitted factors include school characteristics, such as teacher quality and computer usage, and student characteristics, such as family background. We begin by considering an omitted student characteristic that is particularly relevant in California because of its large immigrant population: the prevalence in the school district of students who are still learning English.

By ignoring the percentage of English learners in the district, the OLS estimator of the slope in the regression of test scores on the student–teacher ratio could be biased; that is, the mean of the sampling distribution of the OLS estimator might not equal the true effect on test scores of a unit change in the student–teacher ratio. Here is the reasoning. Students who are still learning English might perform worse on standardized tests than native English speakers. If districts with large classes also have many students still learning English, then the OLS regression of test scores on the student–teacher ratio could erroneously find a correlation and produce a large estimated coefficient, when in fact the true causal effect of cutting class sizes on test scores is small, even zero. Accordingly, based on the analysis of Chapters 4 and 5, the superintendent might hire enough new teachers to reduce the student–teacher ratio by 2, but her hoped-for improvement in test scores will fail to materialize if the true coefficient is small or zero.

A look at the California data lends credence to this concern. The correlation between the student–teacher ratio and the percentage of English learners (students who are not native English speakers and who have not yet mastered English) in the district is 0.19. This small but positive correlation suggests that districts with more English learners tend to have a higher student–teacher ratio (larger classes). If the student–teacher ratio were unrelated to the percentage of English learners, then it would be safe to ignore English proficiency in the regression of test scores against the student–teacher ratio. But because the student–teacher ratio and the percentage of English learners are correlated, it is possible that the OLS coefficient in the regression of test scores on the student–teacher ratio reflects that influence.

Definition of Omitted Variable Bias

If the regressor (the student–teacher ratio) is correlated with a variable that has been omitted from the analysis (the percentage of English learners) and that determines, in part, the dependent variable (test scores), then the OLS estimator will have **omitted variable bias**.

Omitted variable bias occurs when two conditions are true: (1) when the omitted variable is correlated with the included regressor and (2) when the omitted variable is a determinant of the dependent variable. To illustrate these conditions, consider three examples of variables that are omitted from the regression of test scores on the student–teacher ratio.

Example #1: Percentage of English learners. Because the percentage of English learners is correlated with the student–teacher ratio, the first condition for omitted variable bias holds. It is plausible that students who are still learning English will do worse on standardized tests than native English speakers, in which case the percentage of English learners is a determinant of test scores and the second condition for omitted variable bias holds. Thus the OLS estimator in the regression of test scores on the student–teacher ratio could incorrectly reflect the influence of the omitted variable, the percentage of English learners. That is, omitting the percentage of English learners may introduce omitted variable bias.

Example #2: Time of day of the test. Another variable omitted from the analysis is the time of day that the test was administered. For this omitted variable, it is plausible that the first condition for omitted variable bias does not hold but that the second condition does. For example, if the time of day of the test varies from one district to the next in a way that is unrelated to class size, then the time of day and class size would be uncorrelated so the first condition does not hold. Conversely, the time of day of the test could affect scores (alertness varies through the school day), so the second condition holds. However, because in this example the time of day the test is administered is uncorrelated with the student–teacher ratio, the student–teacher ratio could not be incorrectly picking up the "time of day" effect. Thus omitting the time of day of the test does not result in omitted variable bias.

Example #3: Parking lot space per pupil. Another omitted variable is parking lot space per pupil (the area of the teacher parking lot divided by the number of students). This variable satisfies the first but not the second condition for omitted variable bias. Specifically, schools with more teachers per pupil probably have more teacher parking space, so the first condition would be satisfied. However, under the assumption that learning takes place in the classroom, not the parking lot, parking lot space has no direct effect on learning; thus the second condition does not hold. Because parking lot space per pupil is not a determinant of test scores, omitting it from the analysis does not lead to omitted variable bias.

Omitted variable bias is summarized in Key Concept 6.1.

Omitted variable bias and the first least squares assumption. Omitted variable bias means that the first least squares assumption—that $E(u_i | X_i) = 0$, as listed in Key Concept 4.3—is incorrect. To see why, recall that the error term u_i in the linear regression model with a single regressor represents all factors, other than X_i, that are determinants of Y_i. If one of these other factors is correlated with X_i, this means that the error term (which contains this factor) is correlated with X_i. In other words, if an omitted variable is a determinant of Y_i, then it is in the error term, and if it is correlated with X_i, then the error term is correlated with X_i.

Omitted Variable Bias in Regression with a Single Regressor

KEY CONCEPT 6.1

Omitted variable bias is the bias in the OLS estimator that arises when the regressor, X, is correlated with an omitted variable. For omitted variable bias to occur, two conditions must be true:

1. X is correlated with the omitted variable.
2. The omitted variable is a determinant of the dependent variable, Y.

Because u_i and X_i are correlated, the conditional mean of u_i given X_i is nonzero. This correlation therefore violates the first least squares assumption, and the consequence is serious: The OLS estimator is biased. This bias does not vanish even in very large samples, and the OLS estimator is inconsistent.

A Formula for Omitted Variable Bias

The discussion of the previous section about omitted variable bias can be summarized mathematically by a formula for this bias. Let the correlation between X_i and u_i be $\text{corr}(X_i, u_i) = \rho_{Xu}$. Suppose that the second and third least squares assumptions hold, but the first does not because ρ_{Xu} is nonzero. Then the OLS estimator has the limit (derived in Appendix 6.1)

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu}\frac{\sigma_u}{\sigma_X}. \tag{6.1}$$

That is, as the sample size increases, $\hat{\beta}_1$ is close to $\beta_1 + \rho_{Xu}(\sigma_u / \sigma_X)$ with increasingly high probability.

The formula in Equation (6.1) summarizes several of the ideas discussed above about omitted variable bias:

1. Omitted variable bias is a problem whether the sample size is large or small. Because $\hat{\beta}_1$ does not converge in probability to the true value β_1, $\hat{\beta}_1$ is biased and inconsistent; that is, $\hat{\beta}_1$ is not a consistent estimator of β_1 when there is omitted variable bias. The term $\rho_{Xu}(\sigma_u / \sigma_X)$ in Equation (6.1) is the bias in $\hat{\beta}_1$ that persists even in large samples.

The Mozart Effect: Omitted Variable Bias?

A study published in *Nature* in 1993 (Rauscher, Shaw, and Ky, 1993) suggested that listening to Mozart for 10 to 15 minutes could temporarily raise your IQ by 8 or 9 points. That study made big news—and politicians and parents saw an easy way to make their children smarter. For a while, the state of Georgia even distributed classical music CDs to all infants in the state.

What is the evidence for the "Mozart effect"? A review of dozens of studies found that students who take optional music or arts courses in high school do, in fact, have higher English and math test scores than those who don't.[1] A closer look at these studies, however, suggests that the real reason for the better test performance has little to do with those courses. Instead, the authors of the review suggested that the correlation between testing well and taking art or music could arise from any number of things. For example, the academically better students might have more time to take optional music courses or more interest in doing so, or those schools with a deeper music curriculum might just be better schools across the board.

In the terminology of regression, the estimated relationship between test scores and taking optional music courses appears to have omitted variable bias. By omitting factors such as the student's innate ability or the overall quality of the school, studying music appears to have an effect on test scores when in fact it has none.

So is there a Mozart effect? One way to find out is to do a randomized controlled experiment. (As discussed in Chapter 4, randomized controlled experiments eliminate omitted variable bias by randomly assigning participants to "treatment" and "control" groups.) Taken together, the many controlled experiments on the Mozart effect fail to show that listening to Mozart improves IQ or general test performance. For reasons not fully understood, however, it seems that listening to classical music *does* help temporarily in one narrow area: folding paper and visualizing shapes. So the next time you cram for an origami exam, try to fit in a little Mozart, too.

[1]See the fall/winter 2000 issue of *Journal of Aesthetic Education* 34, especially the article by Ellen Winner and Monica Cooper (pp. 11–76) and the one by Lois Hetland (pp. 105–148).

2. Whether this bias is large or small in practice depends on the correlation ρ_{Xu} between the regressor and the error term. The larger $|\rho_{Xu}|$ is, the larger the bias.
3. The direction of the bias in $\hat{\beta}_1$ depends on whether X and u are positively or negatively correlated. For example, we speculated that the percentage of students learning English has a *negative* effect on district test scores (students still learning English have lower scores), so that the percentage of English learners enters the error term with a negative sign. In our data, the fraction of English learners is *positively* correlated with the student–teacher

ratio (districts with more English learners have larger classes). Thus the student–teacher ratio (X) would be *negatively* correlated with the error term (u), so $\rho_{Xu} < 0$ and the coefficient on the student–teacher ratio $\hat{\beta}_1$ would be biased toward a negative number. In other words, having a small percentage of English learners is associated both with *high* test scores and *low* student–teacher ratios, so one reason that the OLS estimator suggests that small classes improve test scores may be that the districts with small classes have fewer English learners.

Addressing Omitted Variable Bias by Dividing the Data into Groups

What can you do about omitted variable bias? Our superintendent is considering increasing the number of teachers in her district, but she has no control over the fraction of immigrants in her community. As a result, she is interested in the effect of the student–teacher ratio on test scores, *holding constant* other factors, including the percentage of English learners. This new way of posing her question suggests that, instead of using data for all districts, perhaps we should focus on districts with percentages of English learners comparable to hers. Among this subset of districts, do those with smaller classes do better on standardized tests?

Table 6.1 reports evidence on the relationship between class size and test scores within districts with comparable percentages of English learners. Districts are divided into eight groups. First, the districts are broken into four categories

TABLE 6.1 Differences in Test Scores for California School Districts with Low and High Student–Teacher Ratios, by the Percentage of English Learners in the District

	Student–Teacher Ratio < 20		Student–Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High STR	
	Average Test Score	*n*	Average Test Score	*n*	Difference	*t*-statistic
All districts	657.4	238	650.0	182	7.4	4.04
Percentage of English learners						
< 1.9%	664.5	76	665.4	27	−0.9	−0.30
1.9–8.8%	665.2	64	661.8	44	3.3	1.13
8.8–23.0%	654.9	54	649.7	50	5.2	1.72
> 23.0%	636.7	44	634.8	61	1.9	0.68

that correspond to the quartiles of the distribution of the percentage of English learners across districts. Second, within each of these four categories, districts are further broken down into two groups, depending on whether the student–teacher ratio is small ($STR < 20$) or large ($STR \geq 20$).

The first row in Table 6.1 reports the overall difference in average test scores between districts with low and high student–teacher ratios, that is, the difference in test scores between these two groups without breaking them down further into the quartiles of English learners. (Recall that this difference was previously reported in regression form in Equation (5.18) as the OLS estimate of the coefficient on D_i in the regression of *TestScore* on D_i, where D_i is a binary regressor that equals 1 if $STR_i < 20$ and equals 0 otherwise.) Over the full sample of 420 districts, the average test score is 7.4 points higher in districts with a low student–teacher ratio than a high one; the *t*-statistic is 4.04, so the null hypothesis that the mean test score is the same in the two groups is rejected at the 1% significance level.

The final four rows in Table 6.1 report the difference in test scores between districts with low and high student–teacher ratios, broken down by the quartile of the percentage of English learners. This evidence presents a different picture. Of the districts with the fewest English learners ($< 1.9\%$), the average test score for those 76 with low student–teacher ratios is 664.5 and the average for the 27 with high student–teacher ratios is 665.4. Thus, for the districts with the fewest English learners, test scores were on average 0.9 points *lower* in the districts with low student–teacher ratios! In the second quartile, districts with low student–teacher ratios had test scores that averaged 3.3 points higher than those with high student–teacher ratios; this gap was 5.2 points for the third quartile and only 1.9 points for the quartile of districts with the most English learners. Once we hold the percentage of English learners constant, the difference in performance between districts with high and low student–teacher ratios is perhaps half (or less) of the overall estimate of 7.4 points.

At first this finding might seem puzzling. How can the overall effect of test scores be twice the effect of test scores within any quartile? The answer is that the districts with the most English learners tend to have *both* the highest student–teacher ratios *and* the lowest test scores. The difference in the average test score between districts in the lowest and highest quartile of the percentage of English learners is large, approximately 30 points. The districts with few English learners tend to have lower student–teacher ratios: 74% (76 of 103) of the districts in the first quartile of English learners have small classes ($STR < 20$), while only 42% (44 of 105) of the districts in the quartile with the most English learners have small classes. So, the districts with the most English learners have both lower test scores and higher student–teacher ratios than the other districts.

This analysis reinforces the superintendent's worry that omitted variable bias is present in the regression of test scores against the student–teacher ratio. By looking within quartiles of the percentage of English learners, the test score differences in the second part of Table 6.1 improve on the simple difference-of-means analysis in the first line of Table 6.1. Still, this analysis does not yet provide the superintendent with a useful estimate of the effect on test scores of changing class size, holding constant the fraction of English learners. Such an estimate can be provided, however, using the method of multiple regression.

6.2 The Multiple Regression Model

The **multiple regression model** extends the single variable regression model of Chapters 4 and 5 to include additional variables as regressors. This model permits estimating the effect on Y_i of changing one variable (X_{1i}) while holding the other regressors (X_{2i}, X_{3i}, and so forth) constant. In the class size problem, the multiple regression model provides a way to isolate the effect on test scores (Y_i) of the student–teacher ratio (X_{1i}) while holding constant the percentage of students in the district who are English learners (X_{2i}).

The Population Regression Line

Suppose for the moment that there are only two independent variables, X_{1i} and X_{2i}. In the linear multiple regression model, the average relationship between these two independent variables and the dependent variable, Y, is given by the linear function

$$E(Y_i | X_{1i} = x_1, X_{2i} = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \tag{6.2}$$

where $E(Y_i | X_{1i} = x_1, X_{2i} = x_2)$ is the conditional expectation of Y_i given that $X_{1i} = x_1$ and $X_{2i} = x_2$. That is, if the student–teacher ratio in the i^{th} district (X_{1i}) equals some value x_1 and the percentage of English learners in the i^{th} district (X_{2i}) equals x_2, then the expected value of Y_i given the student–teacher ratio and the percentage of English learners is given by Equation (6.2).

Equation (6.2) is the **population regression line** or **population regression function** in the multiple regression model. The coefficient β_0 is the **intercept**; the coefficient β_1 is the **slope coefficient of X_{1i}** or, more simply, the **coefficient on X_{1i}**; and the coefficient β_2 is the **slope coefficient of X_{2i}** or, more simply, the **coefficient on X_{2i}**. One or more of the independent variables in the multiple regression model are sometimes referred to as control variables.

The interpretation of the coefficient β_1 in Equation (6.2) is different than it was when X_{1i} was the only regressor: In Equation (6.2), β_1 is the effect on Y of a unit change in X_1, **holding X_2 constant** or **controlling for X_2**.

This interpretation of β_1 follows from the definition that the expected effect on Y of a change in X_1, ΔX_1, holding X_2 constant, is the difference between the expected value of Y when the independent variables take on the values $X_1 + \Delta X_1$ and X_2 and the expected value of Y when the independent variables take on the values X_1 and X_2. Accordingly, write the population regression function in Equation (6.2) as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ and imagine changing X_1 by the amount ΔX_1 while not changing X_2, that is, while holding X_2 constant. Because X_1 has changed, Y will change by some amount, say ΔY. After this change, the new value of Y, $Y + \Delta Y$, is

$$Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2. \tag{6.3}$$

An equation for ΔY in terms of ΔX_1 is obtained by subtracting the equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ from Equation (6.3), yielding $\Delta Y = \beta_1 \Delta X_1$. Rearranging this equation shows that

$$\beta_1 = \frac{\Delta Y}{\Delta X_1} \text{ holding } X_2 \text{ constant.} \tag{6.4}$$

The coefficient β_1 is the effect on Y (the expected change in Y) of a unit change in X_1, holding X_2 fixed. Another phrase used to describe β_1 is the **partial effect** on Y of X_1, holding X_2 fixed.

The interpretation of the intercept in the multiple regression model, β_0, is similar to the interpretation of the intercept in the single-regressor model: It is the expected value of Y_i when X_{1i} and X_{2i} are zero. Simply put, the intercept β_0 determines how far up the Y axis the population regression line starts.

The Population Multiple Regression Model

The population regression line in Equation (6.2) is the relationship between Y and X_1 and X_2 that holds on average in the population. Just as in the case of regression with a single regressor, however, this relationship does not hold exactly because many other factors influence the dependent variable. In addition to the student–teacher ratio and the fraction of students still learning English, for example, test scores are influenced by school characteristics, other student characteristics, and luck. Thus the population regression function in Equation (6.2) needs to be augmented to incorporate these additional factors.

Just as in the case of regression with a single regressor, the factors that determine Y_i in addition to X_{1i} and X_{2i} are incorporated into Equation (6.2) as an "error" term u_i. This error term is the deviation of a particular observation (test scores in the i^{th} district in our example) from the average population relationship. Accordingly, we have

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, i = 1, \ldots, n, \tag{6.5}$$

where the subscript i indicates the i^{th} of the n observations (districts) in the sample.

Equation (6.5) is the **population multiple regression model** when there are two regressors, X_{1i} and X_{2i}.

In regression with binary regressors, it can be useful to treat β_0 as the coefficient on a regressor that always equals 1; think of β_0 as the coefficient on X_{0i}, where $X_{0i} = 1$ for $i = 1, \ldots, n$. Accordingly, the population multiple regression model in Equation (6.5) can alternatively be written as

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \text{ where } X_{0i} = 1, i = 1, \ldots, n. \tag{6.6}$$

The variable X_{0i} is sometimes called the **constant regressor** because it takes on the same value—the value 1—for all observations. Similarly, the intercept, β_0, is sometimes called the **constant term** in the regression.

The two ways of writing the population regression model, Equations (6.5) and (6.6), are equivalent.

The discussion so far has focused on the case of a single additional variable, X_2. In practice, however, there might be multiple factors omitted from the single-regressor model. For example, ignoring the students' economic background might result in omitted variable bias, just as ignoring the fraction of English learners did. This reasoning leads us to consider a model with three regressors or, more generally, a model that includes k regressors. The multiple regression model with k regressors, $X_{1i}, X_{2i}, \ldots, X_{ki}$, is summarized as Key Concept 6.2.

The definitions of homoskedasticity and heteroskedasticity in the multiple regression model are extensions of their definitions in the single-regressor model. The error term u_i in the multiple regression model is **homoskedastic** if the variance of the conditional distribution of u_i given $X_{1i}, \ldots, X_{ki}$, $\text{var}(u_i | X_{1i}, \ldots, X_{ki})$, is constant for $i = 1, \ldots, n$ and thus does not depend on the values of $X_{1i}, \ldots, X_{ki}$. Otherwise, the error term is **heteroskedastic**.

The multiple regression model holds out the promise of providing just what the superintendent wants to know: the effect of changing the student–teacher ratio, holding constant other factors that are beyond her control.

KEY CONCEPT 6.2

The Multiple Regression Model

The multiple regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, i = 1, \ldots, n, \tag{6.7}$$

where

- Y_i is i^{th} observation on the dependent variable; $X_{1i}, X_{2i}, \ldots, X_{ki}$ are the i^{th} observations on each of the k regressors; and u_i is the error term.
- The population regression line is the relationship that holds between Y and the X's on average in the population:

$$E(Y|X_{1i} = x_1, X_{2i} = x_2, \ldots, X_{ki} = x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

- β_1 is the slope coefficient on X_1, β_2 is the coefficient on X_2, and so on. The coefficient β_1 is the expected change in Y_i resulting from changing X_{1i} by one unit, holding constant $X_{2i}, \ldots, X_{ki}$. The coefficients on the other X's are interpreted similarly.
- The intercept β_0 is the expected value of Y when all the X's equal 0. The intercept can be thought of as the coefficient on a regressor, X_{0i}, that equals 1 for all i.

These factors include not just the percentage of English learners, but other measurable factors that might affect test performance, including the economic background of the students. To be of practical help to the superintendent, however, we need to provide her with estimates of the unknown population coefficients $\beta_0, \ldots, \beta_k$ of the population regression model calculated using a sample of data. Fortunately, these coefficients can be estimated using ordinary least squares.

6.3 The OLS Estimator in Multiple Regression

This section describes how the coefficients of the multiple regression model can be estimated using OLS.

The OLS Estimator

Section 4.2 shows how to estimate the intercept and slope coefficients in the single-regressor model by applying OLS to a sample of observations of Y and X. The key idea is that these coefficients can be estimated by minimizing the sum of squared prediction mistakes, that is, by choosing the estimators b_0 and b_1 so as to minimize $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$. The estimators that do so are the OLS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$.

The method of OLS also can be used to estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_k$ in the multiple regression model. Let $b_0, b_1, \ldots, b_k$ be estimates of $\beta_0, \beta_1, \ldots, \beta_k$. The predicted value of Y_i, calculated using these estimates, is $b_0 + b_1X_{1i} + \cdots + b_kX_{ki}$, and the mistake in predicting Y_i is $Y_i - (b_0 + b_1X_{1i} + \cdots + b_kX_{ki}) = Y_i - b_0 - b_1X_{1i} - \cdots - b_kX_{ki}$. The sum of these squared prediction mistakes over all n observations is thus

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1X_{1i} - \cdots - b_kX_{ki})^2. \tag{6.8}$$

The sum of the squared mistakes for the linear regression model in Expression (6.8) is the extension of the sum of the squared mistakes given in Equation (4.6) for the linear regression model with a single regressor.

The estimators of the coefficients $\beta_0, \beta_1, \ldots, \beta_k$ that minimize the sum of squared mistakes in Expression (6.8) are called the **ordinary least squares (OLS) estimators of $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_k}$**. The OLS estimators are denoted $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

The terminology of OLS in the linear multiple regression model is the same as in the linear regression model with a single regressor. The **OLS regression line** is the straight line constructed using the OLS estimators: $\hat{\beta}_0 + \hat{\beta}_1X_1 + \cdots + \hat{\beta}_kX_k$. The **predicted value** of Y_i given $X_{1i}, \ldots, X_{ki}$, based on the OLS regression line, is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1X_{1i} + \cdots + \hat{\beta}_kX_{ki}$. The **OLS residual** for the i^{th} observation is the difference between Y_i and its OLS predicted value; that is, the OLS residual is $\hat{u}_i = Y_i - \hat{Y}_i$.

The OLS estimators could be computed by trial and error, repeatedly trying different values of $b_0, \ldots, b_k$ until you are satisfied that you have minimized the total sum of squares in Expression (6.8). It is far easier, however, to use explicit formulas for the OLS estimators that are derived using calculus. The formulas for the OLS estimators in the multiple regression model are similar to those in Key Concept 4.2 for the single-regressor model. These formulas are incorporated into modern statistical software. In the multiple regression model, the formulas are best expressed and discussed using matrix notation, so their presentation is deferred to Section 18.1.

KEY CONCEPT 6.3

The OLS Estimators, Predicted Values, and Residuals in the Multiple Regression Model

The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the values of $b_0, b_1, \ldots, b_k$ that minimize the sum of squared prediction mistakes $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_{1i} - \cdots - b_kX_{ki})^2$. The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1X_{1i} + \cdots + \hat{\beta}_kX_{ki}, i = 1, \ldots, n, \text{ and} \quad (6.9)$$

$$\hat{u}_i = Y_i - \hat{Y}_i, i = 1, \ldots, n. \quad (6.10)$$

The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ and residual $\hat{u}_i$ are computed from a sample of n observations of $(X_{1i}, \ldots, X_{ki}, Y_i), i = 1, \ldots, n$. These are estimators of the unknown true population coefficients $\beta_0, \beta_1, \ldots, \beta_k$ and error term, u_i.

The definitions and terminology of OLS in multiple regression are summarized in Key Concept 6.3.

Application to Test Scores and the Student–Teacher Ratio

In Section 4.2, we used OLS to estimate the intercept and slope coefficient of the regression relating test scores (*TestScore*) to the student–teacher ratio (*STR*), using our 420 observations for California school districts; the estimated OLS regression line, reported in Equation (4.11), is

$$\widehat{TestScore} = 698.9 - 2.28 \times STR. \quad (6.11)$$

Our concern has been that this relationship is misleading because the student–teacher ratio might be picking up the effect of having many English learners in districts with large classes. That is, it is possible that the OLS estimator is subject to omitted variable bias.

We are now in a position to address this concern by using OLS to estimate a multiple regression in which the dependent variable is the test score (Y_i) and there are two regressors: the student–teacher ratio (X_{1i}) and the percentage of

English learners in the school district (X_{2i}) for our 420 districts ($i = 1, \ldots, 420$). The estimated OLS regression line for this multiple regression is

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 \times PctEL, \tag{6.12}$$

where *PctEL* is the percentage of students in the district who are English learners. The OLS estimate of the intercept ($\hat{\beta}_0$) is 686.0, the OLS estimate of the coefficient on the student–teacher ratio ($\hat{\beta}_1$) is -1.10, and the OLS estimate of the coefficient on the percentage English learners ($\hat{\beta}_2$) is -0.65.

The estimated effect on test scores of a change in the student–teacher ratio in the multiple regression is approximately half as large as when the student–teacher ratio is the only regressor: In the single-regressor equation [Equation (6.11)], a unit decrease in the *STR* is estimated to increase test scores by 2.28 points, but in the multiple regression equation [Equation (6.12)], it is estimated to increase test scores by only 1.10 points. This difference occurs because the coefficient on *STR* in the multiple regression is the effect of a change in *STR*, holding constant (or controlling for) *PctEL*, whereas in the single-regressor regression, *PctEL* is not held constant.

These two estimates can be reconciled by concluding that there is omitted variable bias in the estimate in the single-regressor model in Equation (6.11). In Section 6.1, we saw that districts with a high percentage of English learners tend to have not only low test scores but also a high student–teacher ratio. If the fraction of English learners is omitted from the regression, reducing the student–teacher ratio is estimated to have a larger effect on test scores, but this estimate reflects *both* the effect of a change in the student–teacher ratio *and* the omitted effect of having fewer English learners in the district.

We have reached the same conclusion that there is omitted variable bias in the relationship between test scores and the student–teacher ratio by two different paths: the tabular approach of dividing the data into groups (Section 6.1) and the multiple regression approach [Equation (6.12)]. Of these two methods, multiple regression has two important advantages. First, it provides a quantitative estimate of the effect of a unit decrease in the student–teacher ratio, which is what the superintendent needs to make her decision. Second, it readily extends to more than two regressors so that multiple regression can be used to control for measurable factors other than just the percentage of English learners.

The rest of this chapter is devoted to understanding and using OLS in the multiple regression model. Much of what you learned about the OLS estimator with a single regressor carries over to multiple regression with few or no modifications, so we will focus on that which is new with multiple regression. We begin by discussing measures of fit for the multiple regression model.

6.4 Measures of Fit in Multiple Regression

Three commonly used summary statistics in multiple regression are the standard error of the regression, the regression R^2, and the adjusted R^2 (also known as $\overline{R}^2$). All three statistics measure how well the OLS estimate of the multiple regression line describes, or "fits," the data.

The Standard Error of the Regression (*SER*)

The standard error of the regression (SER) estimates the standard deviation of the error term u_i. Thus the SER is a measure of the spread of the distribution of Y around the regression line. In multiple regression, the SER is

$$SER = s_{\hat{u}} = \sqrt{s_{\hat{u}}^2} \quad \text{where } s_{\hat{u}}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-k-1} \tag{6.13}$$

and where SSR is the sum of squared residuals, $SSR = \sum_{i=1}^{n}\hat{u}_i^2$.

The only difference between the definition in Equation (6.13) and the definition of the SER in Section 4.3 for the single-regressor model is that here the divisor is $n - k - 1$ rather than $n - 2$. In Section 4.3, the divisor $n - 2$ (rather than n) adjusts for the downward bias introduced by estimating two coefficients (the slope and intercept of the regression line). Here, the divisor $n - k - 1$ adjusts for the downward bias introduced by estimating $k + 1$ coefficients (the k slope coefficients plus the intercept). As in Section 4.3, using $n - k - 1$ rather than n is called a degrees-of-freedom adjustment. If there is a single regressor, then $k = 1$, so the formula in Section 4.3 is the same as in Equation (6.13). When n is large, the effect of the degrees-of-freedom adjustment is negligible.

The R^2

The regression $\boldsymbol{R^2}$ is the fraction of the sample variance of Y_i explained by (or predicted by) the regressors. Equivalently, the R^2 is 1 minus the fraction of the variance of Y_i *not* explained by the regressors.

The mathematical definition of the R^2 is the same as for regression with a single regressor:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}, \tag{6.14}$$

where the explained sum of squares is $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ and the total sum of squares is $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$.

In multiple regression, the R^2 increases whenever a regressor is added, unless the estimated coefficient on the added regressor is exactly zero. To see this, think about starting with one regressor and then adding a second. When you use OLS to estimate the model with both regressors, OLS finds the values of the coefficients that minimize the sum of squared residuals. If OLS happens to choose the coefficient on the new regressor to be exactly zero, then the SSR will be the same whether or not the second variable is included in the regression. But if OLS chooses any value other than zero, then it must be that this value reduced the SSR relative to the regression that excludes this regressor. In practice, it is extremely unusual for an estimated coefficient to be exactly zero, so in general the SSR will decrease when a new regressor is added. But this means that the R^2 generally increases (and never decreases) when a new regressor is added.

The "Adjusted R^2"

Because the R^2 increases when a new variable is added, an increase in the R^2 does not mean that adding a variable actually improves the fit of the model. In this sense, the R^2 gives an inflated estimate of how well the regression fits the data. One way to correct for this is to deflate or reduce the R^2 by some factor, and this is what the adjusted R^2, or $\overline{R}^2$, does.

The **adjusted $\boldsymbol{R^2}$**, or $\boldsymbol{\overline{R}^2}$, is a modified version of the R^2 that does not necessarily increase when a new regressor is added. The $\boldsymbol{\overline{R}^2}$ is

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s_{\hat{u}}^2}{s_Y^2}. \tag{6.15}$$

The difference between this formula and the second definition of the R^2 in Equation (6.14) is that the ratio of the sum of squared residuals to the total sum of squares is multiplied by the factor $(n-1)/(n-k-1)$. As the second expression in Equation (6.15) shows, this means that the adjusted R^2 is 1 minus the ratio of the sample variance of the OLS residuals [with the degrees-of-freedom correction in Equation (6.13)] to the sample variance of Y.

There are three useful things to know about the $\overline{R}^2$. First, $(n-1)/(n-k-1)$ is always greater than 1, so $\overline{R}^2$ is always less than R^2.

Second, adding a regressor has two opposite effects on the $\overline{R}^2$. On the one hand, the SSR falls, which increases the $\overline{R}^2$. On the other hand, the factor $(n-1)/(n-k-1)$ increases. Whether the $\overline{R}^2$ increases or decreases depends on which of these two effects is stronger.

Third, the $\overline{R}^2$ can be negative. This happens when the regressors, taken together, reduce the sum of squared residuals by such a small amount that this reduction fails to offset the factor $(n - 1)/(n - k - 1)$.

Application to Test Scores

Equation (6.12) reports the estimated regression line for the multiple regression relating test scores ($TestScore$) to the student–teacher ratio (STR) and the percentage of English learners ($PctEL$). The R^2 for this regression line is $R^2 = 0.426$, the adjusted R^2 is $\overline{R}^2 = 0.424$, and the standard error of the regression is $SER = 14.5$.

Comparing these measures of fit with those for the regression in which $PctEL$ is excluded [Equation (5.8)] shows that including $PctEL$ in the regression increased the R^2 from 0.051 to 0.426. When the only regressor is STR, only a small fraction of the variation in $TestScore$ is explained; however, when $PctEL$ is added to the regression, more than two-fifths (42.6%) of the variation in test scores is explained. In this sense, including the percentage of English learners substantially improves the fit of the regression. Because n is large and only two regressors appear in Equation (6.12), the difference between R^2 and adjusted R^2 is very small ($R^2 = 0.426$ versus $\overline{R}^2 = 0.424$).

The SER for the regression excluding $PctEL$ is 18.6; this value falls to 14.5 when $PctEL$ is included as a second regressor. The units of the SER are points on the standardized test. The reduction in the SER tells us that predictions about standardized test scores are substantially more precise if they are made using the regression with both STR and $PctEL$ than if they are made using the regression with only STR as a regressor.

Using the R^2 and adjusted R^2. The $\overline{R}^2$ is useful because it quantifies the extent to which the regressors account for, or explain, the variation in the dependent variable. Nevertheless, heavy reliance on the $\overline{R}^2$ (or R^2) can be a trap. In applications, "maximize the $\overline{R}^2$" is rarely the answer to any economically or statistically meaningful question. Instead, the decision about whether to include a variable in a multiple regression should be based on whether including that variable allows you better to estimate the causal effect of interest. We return to the issue of how to decide which variables to include—and which to exclude—in Chapter 7. First, however, we need to develop methods for quantifying the sampling uncertainty of the OLS estimator. The starting point for doing so is extending the least squares assumptions of Chapter 4 to the case of multiple regressors.

6.5 The Least Squares Assumptions in Multiple Regression

There are four least squares assumptions in the multiple regression model. The first three are those of Section 4.3 for the single regressor model (Key Concept 4.3), extended to allow for multiple regressors, and these are discussed only briefly. The fourth assumption is new and is discussed in more detail.

Assumption #1: The Conditional Distribution of u_i Given $X_{1i}, X_{2i}, \ldots, X_{ki}$ Has a Mean of Zero

The first assumption is that the conditional distribution of u_i given $X_{1i}, \ldots, X_{ki}$ has a mean of zero. This assumption extends the first least squares assumption with a single regressor to multiple regressors. This assumption means that sometimes Y_i is above the population regression line and sometimes Y_i is below the population regression line, but on average over the population Y_i falls on the population regression line. Therefore, for any value of the regressors, the expected value of u_i is zero. As is the case for regression with a single regressor, this is the key assumption that makes the OLS estimators unbiased. We return to omitted variable bias in multiple regression in Section 7.5.

Assumption #2: $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$, Are i.i.d.

The second assumption is that $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$, are independently and identically distributed (i.i.d.) random variables. This assumption holds automatically if the data are collected by simple random sampling. The comments on this assumption appearing in Section 4.3 for a single regressor also apply to multiple regressors.

Assumption #3: Large Outliers Are Unlikely

The third least squares assumption is that large outliers—that is, observations with values far outside the usual range of the data—are unlikely. This assumption serves as a reminder that, as in single-regressor case, the OLS estimator of the coefficients in the multiple regression model can be sensitive to large outliers.

The assumption that large outliers are unlikely is made mathematically precise by assuming that $X_{1i}, \ldots, X_{ki}$, and Y_i have nonzero finite fourth moments: $0 < E(X_{1i}^4) < \infty, \ldots, 0 < E(X_{ki}^4) < \infty$ and $0 < E(Y_i^4) < \infty$. Another way to state this assumption is that the dependent variable and regressors have finite kurtosis. This assumption is used to derive the properties of OLS regression statistics in large samples.

Assumption #4: No Perfect Multicollinearity

The fourth assumption is new to the multiple regression model. It rules out an inconvenient situation, called perfect multicollinearity, in which it is impossible to compute the OLS estimator. The regressors are said to exhibit **perfect multicollinearity**, (or to be perfectly multicollinear) if one of the regressors is a perfect linear function of the other regressors. The fourth least squares assumption is that the regressors are not perfectly multicollinear.

Why does perfect multicollinearity make it impossible to compute the OLS estimator? Suppose you want to estimate the coefficient on STR in a regression of $TestScore_i$ on STR_i and $PctEL_i$, except that you make a typographical error and accidentally type in STR_i a second time instead of $PctEL_i$; that is, you regress $TestScore_i$ on STR_i and STR_i. This is a case of perfect multicollinearity because one of the regressors (the first occurrence of STR) is a perfect linear function of another regressor (the second occurrence of STR). Depending on how your software package handles perfect multicollinearity, if you try to estimate this regression the software will do one of two things: Either it will drop one of the occurrences of STR or it will refuse to calculate the OLS estimates and give an error message. The mathematical reason for this failure is that perfect multicollinearity produces division by zero in the OLS formulas.

At an intuitive level, perfect multicollinearity is a problem because you are asking the regression to answer an illogical question. In multiple regression, the coefficient on one of the regressors is the effect of a change in that regressor, holding the other regressors constant. In the hypothetical regression of $TestScore$ on STR and STR, the coefficient on the first occurrence of STR is the effect on test scores of a change in STR, holding constant STR. This makes no sense, and OLS cannot estimate this nonsensical partial effect.

The solution to perfect multicollinearity in this hypothetical regression is simply to correct the typo and to replace one of the occurrences of STR with the variable you originally wanted to include. This example is typical: When perfect multicollinearity occurs, it often reflects a logical mistake in choosing the regressors or some previously unrecognized feature of the data set. In general, the solution to perfect multicollinearity is to modify the regressors to eliminate the problem.

Additional examples of perfect multicollinearity are given in Section 6.7, which also defines and discusses imperfect multicollinearity.

The Least Squares Assumptions in the Multiple Regression Model

KEY CONCEPT 6.4

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, i = 1, \ldots, n,$$

where

1. u_i has conditional mean zero given $X_{1i}, X_{2i}, \ldots, X_{ki}$; that is,

$$E(u_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = 0$$

2. $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i), i = 1, \ldots, n$, are independently and identically distributed (i.i.d.) draws from their joint distribution.
3. Large outliers are unlikely: $X_{1i}, \ldots, X_{ki}$ and Y_i have nonzero finite fourth moments.
4. There is no perfect multicollinearity.

The least squares assumptions for the multiple regression model are summarized in Key Concept 6.4.

6.6 The Distribution of the OLS Estimators in Multiple Regression

Because the data differ from one sample to the next, different samples produce different values of the OLS estimators. This variation across possible samples gives rise to the uncertainty associated with the OLS estimators of the population regression coefficients, $\beta_0, \beta_1, \ldots, \beta_k$. Just as in the case of regression with a single regressor, this variation is summarized in the sampling distribution of the OLS estimators.

Recall from Section 4.4 that, under the least squares assumptions, the OLS estimators ($\hat{\beta}_0$ and $\hat{\beta}_1$) are unbiased and consistent estimators of the unknown coefficients (β_0 and β_1) in the linear regression model with a single regressor. In addition, in large samples, the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ is well approximated by a bivariate normal distribution.

These results carry over to multiple regression analysis. That is, under the least squares assumptions of Key Concept 6.4, the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are unbiased and consistent estimators of $\beta_0, \beta_1, \ldots, \beta_k$ in the linear multiple

KEY CONCEPT

6.5

Large-Sample Distribution of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$

If the least squares assumptions (Key Concept 6.4) hold, then in large samples the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are jointly normally distributed and each $\hat{\beta}_j$ is distributed $N(\beta_j, \sigma^2_{\hat{\beta}_j})$, $j = 0, \ldots, k$.

regression model. In large samples, the joint sampling distribution of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ is well approximated by a multivariate normal distribution, which is the extension of the bivariate normal distribution to the general case of two or more jointly normal random variables (Section 2.4).

Although the algebra is more complicated when there are multiple regressors, the central limit theorem applies to the OLS estimators in the multiple regression model for the same reason that it applies to $\overline{Y}$ and to the OLS estimators when there is a single regressor: The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are averages of the randomly sampled data, and if the sample size is sufficiently large, the sampling distribution of those averages becomes normal. Because the multivariate normal distribution is best handled mathematically using matrix algebra, the expressions for the joint distribution of the OLS estimators are deferred to Chapter 18.

Key Concept 6.5 summarizes the result that, in large samples, the distribution of the OLS estimators in multiple regression is approximately jointly normal. In general, the OLS estimators are correlated; this correlation arises from the correlation between the regressors. The joint sampling distribution of the OLS estimators is discussed in more detail for the case that there are two regressors and homoskedastic errors in Appendix (6.2), and the general case is discussed in Section 18.2.

6.7 Multicollinearity

As discussed in Section 6.5, perfect multicollinearity arises when one of the regressors is a perfect linear combination of the other regressors. This section provides some examples of perfect multicollinearity and discusses how perfect multicollinearity can arise, and can be avoided, in regressions with multiple binary regressors. Imperfect multicollinearity arises when one of the regressors is very highly correlated—but not perfectly correlated—with the other regressors. Unlike perfect multicollinearity, imperfect multicollinearity does not prevent estimation of the regression, nor does it imply a logical problem with the choice of regressors. However, it does mean that one or more regression coefficients could be estimated imprecisely.

Examples of Perfect Multicollinearity

We continue the discussion of perfect multicollinearity from Section 6.5 by examining three additional hypothetical regressions. In each, a third regressor is added to the regression of $TestScore_i$ on STR_i and $PctEL_i$ in Equation (6.12).

Example #1: Fraction of English learners. Let $FracEL_i$ be the fraction of English learners in the i^{th} district, which varies between 0 and 1. If the variable $FracEL_i$ were included as a third regressor in addition to STR_i and $PctEL_i$, the regressors would be perfectly multicollinear. The reason is that $PctEL$ is the *percentage* of English learners, so that $PctEL_i = 100 \times FracEL_i$ for every district. Thus one of the regressors ($PctEL_i$) can be written as a perfect linear function of another regressor ($FracEL_i$).

Because of this perfect multicollinearity, it is impossible to compute the OLS estimates of the regression of $TestScore_i$ on STR_i, $PctEL_i$, and $FracEL_i$. At an intuitive level, OLS fails because you are asking, What is the effect of a unit change in the *percentage* of English learners, holding constant the *fraction* of English learners? Because the percentage of English learners and the fraction of English learners move together in a perfect linear relationship, this question makes no sense and OLS cannot answer it.

Example #2: "Not very small" classes. Let NVS_i be a binary variable that equals 1 if the student–teacher ratio in the i^{th} district is "not very small," specifically, NVS_i equals 1 if $STR_i \geq 12$ and equals 0 otherwise. This regression also exhibits perfect multicollinearity, but for a more subtle reason than the regression in the previous example. There are in fact no districts in our data set with $STR_i < 12$; as you can see in the scatterplot in Figure 4.2, the smallest value of STR is 14. Thus $NVS_i = 1$ for all observations. Now recall that the linear regression model with an intercept can equivalently be thought of as including a regressor, X_{0i}, that equals 1 for all i, as shown in Equation (6.6). Thus we can write $NVS_i = 1 \times X_{0i}$ for all the observations in our data set; that is, NVS_i can be written as a perfect linear combination of the regressors; specifically, it equals X_{0i}.

This illustrates two important points about perfect multicollinearity. First, when the regression includes an intercept, then one of the regressors that can be implicated in perfect multicollinearity is the constant regressor X_{0i}. Second, perfect multicollinearity is a statement about the data set you have on hand. While it is possible to imagine a school district with fewer than 12 students per teacher, there are no such districts in our data set so we cannot analyze them in our regression.

Example #3: Percentage of English speakers. Let $PctES_i$ be the percentage of "English speakers" in the i^{th} district, defined to be the percentage of students who are not English learners. Again the regressors will be perfectly multicollinear.

Like the previous example, the perfect linear relationship among the regressors involves the constant regressor X_{0i}: For every district, $PctES_i = 100 - PctEL_i = 100 \times X_{0i} - PctEL_i$, because $X_{0i} = 1$ for all i.

This example illustrates another point: Perfect multicollinearity is a feature of the entire set of regressors. If either the intercept (that is, the regressor X_{0i}) or $PctEL_i$ were excluded from this regression, the regressors would not be perfectly multicollinear.

The dummy variable trap. Another possible source of perfect multicollinearity arises when multiple binary, or dummy, variables are used as regressors. For example, suppose you have partitioned the school districts into three categories: rural, suburban, and urban. Each district falls into one (and only one) category. Let these binary variables be $Rural_i$, which equals 1 for a rural district and equals 0 otherwise; $Suburban_i$; and $Urban_i$. If you include all three binary variables in the regression along with a constant, the regressors will be perfect multicollinearity: Because each district belongs to one and only one category, $Rural_i + Suburban_i + Urban_i = 1 = X_{0i}$, where X_{0i} denotes the constant regressor introduced in Equation (6.6). Thus, to estimate the regression, you must exclude one of these four variables, either one of the binary indicators or the constant term. By convention, the constant term is retained, in which case one of the binary indicators is excluded. For example, if $Rural_i$ were excluded, then the coefficient on $Suburban_i$ would be the average difference between test scores in suburban and rural districts, holding constant the other variables in the regression.

In general, if there are G binary variables, if each observation falls into one and only one category, if there is an intercept in the regression, and if all G binary variables are included as regressors, then the regression will fail because of perfect multicollinearity. This situation is called the **dummy variable trap**. The usual way to avoid the dummy variable trap is to exclude one of the binary variables from the multiple regression, so only $G - 1$ of the G binary variables are included as regressors. In this case, the coefficients on the included binary variables represent the incremental effect of being in that category, relative to the base case of the omitted category, holding constant the other regressors. Alternatively, all G binary regressors can be included if the intercept is omitted from the regression.

Solutions to perfect multicollinearity. Perfect multicollinearity typically arises when a mistake has been made in specifying the regression. Sometimes the mistake is easy to spot (as in the first example) but sometimes it is not (as in the second example). In one way or another, your software will let you know if you make such a mistake because it cannot compute the OLS estimator if you have.

When your software lets you know that you have perfect multicollinearity, it is important that you modify your regression to eliminate it. Some software is unreliable when there is perfect multicollinearity, and at a minimum you will be ceding control over your choice of regressors to your computer if your regressors are perfectly multicollinear.

Imperfect Multicollinearity

Despite its similar name, imperfect multicollinearity is conceptually quite different from perfect multicollinearity. **Imperfect multicollinearity** means that two or more of the regressors are highly correlated in the sense that there is a linear function of the regressors that is highly correlated with another regressor. Imperfect multicollinearity does not pose any problems for the theory of the OLS estimators; indeed, a purpose of OLS is to sort out the independent influences of the various regressors when these regressors are potentially correlated.

If the regressors are imperfectly multicollinear, then the coefficients on at least one individual regressor will be imprecisely estimated. For example, consider the regression of *TestScore* on *STR* and *PctEL*. Suppose we were to add a third regressor, the percentage of the district's residents who are first-generation immigrants. First-generation immigrants often speak English as a second language, so the variables *PctEL* and percentage immigrants will be highly correlated: Districts with many recent immigrants will tend to have many students who are still learning English. Because these two variables are highly correlated, it would be difficult to use these data to estimate the partial effect on test scores of an increase in *PctEL*, holding constant the percentage immigrants. In other words, the data set provides little information about what happens to test scores when the percentage of English learners is low but the fraction of immigrants is high, or vice versa. If the least squares assumptions hold, then the OLS estimator of the coefficient on *PctEL* in this regression will be unbiased; however, it will have a larger variance than if the regressors *PctEL* and percentage immigrants were uncorrelated.

The effect of imperfect multicollinearity on the variance of the OLS estimators can be seen mathematically by inspecting Equation (6.17) in Appendix (6.2), which is the variance of $\hat{\beta}_1$ in a multiple regression with two regressors (X_1 and X_2) for the special case of a homoskedastic error. In this case, the variance of $\hat{\beta}_1$ is inversely proportional to $1 - \rho^2_{X_1,X_2}$, where ρ_{X_1,X_2} is the correlation between X_1 and X_2. The larger the correlation between the two regressors, the closer this term is to zero and the larger is the variance of $\hat{\beta}_1$. More generally, when multiple regressors are imperfectly multicollinear, the coefficients on one or more of these regressors will be imprecisely estimated—that is, they will have a large sampling variance.

Perfect multicollinearity is a problem that often signals the presence of a logical error. In contrast, imperfect multicollinearity is not necessarily an error, but rather just a feature of OLS, your data, and the question you are trying to answer. If the variables in your regression are the ones you meant to include—the ones you chose to address the potential for omitted variable bias—then imperfect multicollinearity implies that it will be difficult to estimate precisely one or more of the partial effects using the data at hand.

6.8 Conclusion

Regression with a single regressor is vulnerable to omitted variable bias: If an omitted variable is a determinant of the dependent variable and is correlated with the regressor, then the OLS estimator of the slope coefficient will be biased and will reflect both the effect of the regressor and the effect of the omitted variable. Multiple regression makes it possible to mitigate omitted variable bias by including the omitted variable in the regression. The coefficient on a regressor, X_1, in multiple regression is the partial effect of a change in X_1, holding constant the other included regressors. In the test score example, including the percentage of English learners as a regressor made it possible to estimate the effect on test scores of a change in the student–teacher ratio, holding constant the percentage of English learners. Doing so reduced by half the estimated effect on test scores of a change in the student–teacher ratio.

The statistical theory of multiple regression builds on the statistical theory of regression with a single regressor. The least squares assumptions for multiple regression are extensions of the three least squares assumptions for regression with a single regressor, plus a fourth assumption ruling out perfect multicollinearity. Because the regression coefficients are estimated using a single sample, the OLS estimators have a joint sampling distribution and therefore have sampling uncertainty. This sampling uncertainty must be quantified as part of an empirical study, and the ways to do so in the multiple regression model are the topic of the next chapter.

Summary

1. Omitted variable bias occurs when an omitted variable (1) is correlated with an included regressor and (2) is a determinant of Y.
2. The multiple regression model is a linear regression model that includes multiple regressors, $X_1, X_2, \ldots, X_k$. Associated with each regressor is a regression coefficient, $\beta_1, \beta_2, \ldots, \beta_k$. The coefficient β_1 is the expected change in Y associated with a one-unit change in X_1, holding the other regressors constant. The other regression coefficients have an analogous interpretation.

3. The coefficients in multiple regression can be estimated by OLS. When the four least squares assumptions in Key Concept 6.4 are satisfied, the OLS estimators are unbiased, consistent, and normally distributed in large samples.
4. Perfect multicollinearity, which occurs when one regressor is an exact linear function of the other regressors, usually arises from a mistake in choosing which regressors to include in a multiple regression. Solving perfect multicollinearity requires changing the set of regressors.
5. The standard error of the regression, the R^2, and the $\overline{R}^2$ are measures of fit for the multiple regression model.

Key Terms

omitted variable bias (183)
multiple regression model (189)
population regression line (189)
population regression function (189)
intercept (189)
slope coefficient of X_{1i} (189)
coefficient on X_{1i} (189)
slope coefficient of X_{2i} (189)
coefficient on X_{2i} (189)
holding X_2 constant (190)
controlling for X_2 (190)
partial effect (190)
population multiple regression model (191)
constant regressor (191)
constant term (191)
homoskedastic (191)
heteroskedastic (191)
ordinary least squares (OLS) estimators of $\beta_0, \beta_1, \ldots, \beta_k$ (193)
OLS regression line (193)
predicted value (193)
OLS residual (193)
R^2 (196)
adjusted R^2 ($\overline{R}^2$) (197)
perfect multicollinearity (200)
dummy variable trap (204)
imperfect multicollinearity (205)

Review the Concepts

6.1 A researcher is interested in the effect on test scores of computer usage. Using school district data like that used in this chapter, she regresses district average test scores on the number of computers per student. Will $\hat{\beta}_1$ be an unbiased estimator of the effect on test scores of increasing the number of computers per student? Why or why not? If you think $\hat{\beta}_1$ is biased, is it biased up or down? Why?

6.2 A multiple regression includes two regressors: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. What is the expected change in Y if X_1 increases by 3 units and X_2 is unchanged? What is the expected change in Y if X_2 decreases by 5 units and X_1 is unchanged? What is the expected change in Y if X_1 increases by 3 units and X_2 decreases by 5 units?

6.3 How does $\overline{R}^2$ differ from R^2? Why is $\overline{R}^2$ useful in a regression model with multiple regressors?

6.4 Explain why two perfectly multicollinear regressors cannot be included in a linear multiple regression. Give two examples of a pair of perfectly multicollinear regressors.

6.5 Explain why it is difficult to estimate precisely the partial effect of X_1, holding X_2 constant, if X_1 and X_2 are highly correlated.

Exercises

The first four exercises refer to the table of estimated regressions on page 209, computed using data for 2012 from the CPS. The data set consists of information on 7440 full-time, full-year workers. The highest educational achievement for each worker was either a high school diploma or a bachelor's degree. The workers' ages ranged from 25 to 34 years. The data set also contains information on the region of the country where the person lived, marital status, and number of children. For the purposes of these exercises, let

AHE = average hourly earnings (in 2012 dollars)
College = binary variable (1 if college, 0 if high school)
Female = binary variable (1 if female, 0 if male)
Age = age (in years)
Ntheast = binary variable (1 if Region = Northeast, 0 otherwise)
Midwest = binary variable (1 if Region = Midwest, 0 otherwise)
South = binary variable (1 if Region = South, 0 otherwise)
West = binary variable (1 if Region = West, 0 otherwise)

6.1 Compute $\overline{R}^2$ for each of the regressions.

6.2 Using the regression results in column (1):

a. Do workers with college degrees earn more, on average, than workers with only high school degrees? How much more?

b. Do men earn more than women, on average? How much more?

6.3 Using the regression results in column (2):

a. Is age an important determinant of earnings? Explain.

b. Sally is a 29-year-old female college graduate. Betsy is a 34-year-old female college graduate. Predict Sally's and Betsy's earnings.

6.4 Using the regression results in column (3):

a. Do there appear to be important regional differences?

b. Why is the regressor *West* omitted from the regression? What would happen if it were included?

Results of Regressions of Average Hourly Earnings on Gender and Education Binary Variables and Other Characteristics, Using 2012 Data from the Current Population Survey

Dependent variable: average hourly earnings (AHE).

Regressor	(1)	(2)	(3)
College (X_1)	8.31	8.32	8.34
Female (X_2)	−3.85	−3.81	−3.80
Age (X_3)		0.51	0.52
Northeast (X_4)			0.18
Midwest (X_5)			−1.23
South (X_6)			−0.43
Intercept	17.02	1.87	2.05
Summary Statistics			
SER	9.79	9.68	9.67
R^2	0.162	0.180	0.182
$\overline{R}^2$			
n	7440	7440	7440

c. Juanita is a 28-year-old female college graduate from the South. Jennifer is a 28-year-old female college graduate from the Midwest. Calculate the expected difference in earnings between Juanita and Jennifer.

6.5 Data were collected from a random sample of 220 home sales from a community in 2013. Let *Price* denote the selling price (in \$1000), *BDR* denote the number of bedrooms, *Bath* denote the number of bathrooms, *Hsize* denote the size of the house (in square feet), *Lsize* denote the lot size (in square feet), *Age* denote the age of the house (in years), and *Poor* denote a binary variable that is equal to 1 if the condition of the house is reported as "poor." An estimated regression yields

$$\widehat{Price} = 119.2 + 0.485BDR + 23.4Bath + 0.156Hsize + 0.002Lsize \\ + 0.090Age - 48.8Poor, \overline{R}^2 = 0.72, SER = 41.5.$$

a. Suppose that a homeowner converts part of an existing family room in her house into a new bathroom. What is the expected increase in the value of the house?

b. Suppose that a homeowner adds a new bathroom to her house, which increases the size of the house by 100 square feet. What is the expected increase in the value of the house?

c. What is the loss in value if a homeowner lets his house run down so that its condition becomes "poor"?

d. Compute the R^2 for the regression.

6.6 A researcher plans to study the causal effect of police on crime, using data from a random sample of U.S. counties. He plans to regress the county's crime rate on the (per capita) size of the county's police force.

a. Explain why this regression is likely to suffer from omitted variable bias. Which variables would you add to the regression to control for important omitted variables?

b. Use your answer to (a) and the expression for omitted variable bias given in Equation (6.1) to determine whether the regression will likely over- or underestimate the effect of police on the crime rate. (That is, do you think that $\hat{\beta}_1 > \beta_1$ or $\hat{\beta}_1 < \beta_1$?)

6.7 Critique each of the following proposed research plans. Your critique should explain any problems with the proposed research and describe how the research plan might be improved. Include a discussion of any additional data that need to be collected and the appropriate statistical techniques for analyzing those data.

a. A researcher is interested in determining whether a large aerospace firm is guilty of gender bias in setting wages. To determine potential bias, the researcher collects salary and gender information for all of the firm's engineers. The researcher then plans to conduct a "difference in means" test to determine whether the average salary for women is significantly less than the average salary for men.

b. A researcher is interested in determining whether time spent in prison has a permanent effect on a person's wage rate. He collects data on a random sample of people who have been out of prison for at least 15 years. He collects similar data on a random sample of people who have never served time in prison. The data set includes information on each person's current wage, education, age, ethnicity, gender, tenure (time in current job), occupation, and union status, as well as whether the person has ever been incarcerated. The researcher plans to estimate the effect of incarceration on wages by regressing wages on an indicator variable for incarceration, including in the regression the other potential determinants of wages (education, tenure, union status, and so on).

6.8 A recent study found that the death rate for people who sleep 6 to 7 hours per night is lower than the death rate for people who sleep 8 or more hours. The 1.1 million observations used for this study came from a random survey of Americans aged 30 to 102. Each survey respondent was tracked for 4 years. The death rate for people sleeping 7 hours was calculated as the ratio of the number of deaths over the span of the study among people sleeping 7 hours to the total number of survey respondents who slept 7 hours. This calculation was then repeated for people sleeping 6 hours and so on. Based on this summary, would you recommend that Americans who sleep 9 hours per night consider reducing their sleep to 6 or 7 hours if they want to prolong their lives? Why or why not? Explain.

6.9 (Y_i, X_{1i}, X_{2i}) satisfy the assumptions in Key Concept 6.4. You are interested in β_1, the causal effect of X_1 on Y. Suppose that X_1 and X_2 are uncorrelated. You estimate β_1 by regressing Y onto X_1 (so that X_2 is not included in the regression). Does this estimator suffer from omitted variable bias? Explain.

6.10 (Y_i, X_{1i}, X_{2i}) satisfy the assumptions in Key Concept 6.4; in addition, $\text{var}(u_i | X_{1i}, X_{2i}) = 4$ and $\text{var}(X_{1i}) = 6$. A random sample of size $n = 400$ is drawn from the population.

a. Assume that X_1 and X_2 are uncorrelated. Compute the variance of $\hat{\beta}_1$. [*Hint:* Look at Equation (6.17) in Appendix 6.2.]

b. Assume that $\text{corr}(X_1, X_2) = 0.5$. Compute the variance of $\hat{\beta}_1$.

c. Comment on the following statements: "When X_1 and X_2 are correlated, the variance of $\hat{\beta}_1$ is larger than it would be if X_1 and X_2 were uncorrelated. Thus, if you are interested in β_1, it is best to leave X_2 out of the regression if it is correlated with X_1."

6.11 (Requires calculus) Consider the regression model

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

for $i = 1, \ldots, n$. (Notice that there is no constant term in the regression.) Following analysis like that used in Appendix (4.2):

a. Specify the least squares function that is minimized by OLS.

b. Compute the partial derivatives of the objective function with respect to b_1 and b_2.

c. Suppose that $\sum_{i=1}^{n} X_{1i}X_{2i} = 0$. Show that $\hat{\beta}_1 = \sum_{i=1}^{n} X_{1i}Y_i / \sum_{i=1}^{n} X_{1i}^2$.

d. Suppose that $\sum_{i=1}^{n} X_{1i}X_{2i} \neq 0$. Derive an expression for $\hat{\beta}_1$ as a function of the data (Y_i, X_{1i}, X_{2i}), $i = 1, \ldots, n$.

e. Suppose that the model includes an intercept: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Show that the least squares estimators satisfy $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2$.

f. As in (e), suppose that the model contains an intercept. Also suppose that $\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2) = 0$. Show that $\hat{\beta}_1 = \sum_{i=1}^{n}(X_{1i} - \overline{X}_1)(Y_i - \overline{Y}) / \sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2$. How does this compare to the OLS estimator of β_1 from the regression that omits X_2?

Empirical Exercises

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

E6.1 Use the **Birthweight_Smoking** data set introduced in Empirical Exercise E5.3 to answer the following questions.

a. Regress *Birthweight* on *Smoker*. What is the estimated effect of smoking on birth weight?

b. Regress *Birthweight* on *Smoker*, *Alcohol*, and *Nprevist*.

i. Using the two conditions in Key Concept 6.1, explain why the exclusion of *Alcohol* and *Nprevist* could lead to omitted variable bias in the regression estimated in (a).

ii. Is the estimated effect of smoking on birth weight substantially different from the regression that excludes *Alcohol* and *Nprevist*? Does the regression in (a) seem to suffer from omitted variable bias?

iii. Jane smoked during her pregnancy, did not drink alcohol, and had 8 prenatal care visits. Use the regression to predict the birth weight of Jane's child.

iv. Compute R^2 and $\overline{R}^2$. Why are they so similar?

c. Estimate the coefficient on *Smoking* for the multiple regression model in (b), using the three-step process in Appendix (6.3) (the Frisch-Waugh theorem). Verify that the three-step process yields the same estimated coefficient for *Smoking* as that obtained in (b).

d. An alternative way to control for prenatal visits is to use the binary variables *Tripre0* through *Tripre3*. Regress *Birthweight* on *Smoker*, *Alcohol*, *Tripre0*, *Tripre2*, and *Tripre3*.

i. Why is *Tripre1* excluded from the regression? What would happen if you included it in the regression?

ii. The estimated coefficient on *Tripre0* is large and negative. What does this coefficient measure? Interpret its value.

iii. Interpret the value of the estimated coefficients on *Tripre2* and *Tripre3*.

iv. Does the regression in (d) explain a larger fraction of the variance in birth weight than the regression in (b)?

E6.2 Using the data set **Growth** described in Empirical Exercise E4.1, but excluding the data for Malta, carry out the following exercises.

a. Construct a table that shows the sample mean, standard deviation, and minimum and maximum values for the series *Growth*, *TradeShare*, *YearsSchool*, *Oil*, *Rev_Coups*, *Assassinations*, and *RGDP60*. Include the appropriate units for all entries.

b. Run a regression of *Growth* on *TradeShare*, *YearsSchool*, *Rev_Coups*, *Assassinations*, and *RGDP60*. What is the value of the coefficient on

Rev_Coups? Interpret the value of this coefficient. Is it large or small in a real-world sense?

c. Use the regression to predict the average annual growth rate for a country that has average values for all regressors.

d. Repeat (c) but now assume that the country's value for *TradeShare* is one standard deviation above the mean.

e. Why is *Oil* omitted from the regression? What would happen if it were included?

APPENDIX

6.1 Derivation of Equation (6.1)

This appendix presents a derivation of the formula for omitted variable bias in Equation (6.1). Equation (4.30) in Appendix (4.3) states

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{6.16}$$

Under the last two assumptions in Key Concept 4.3, $(1/n)\sum_{i=1}^{n}(X_i - \overline{X})^2 \xrightarrow{p} \sigma_X^2$ and $(1/n)\sum_{i=1}^{n}(X_i - \overline{X})u_i \xrightarrow{p} \text{cov}(u_i, X_i) = \rho_{Xu}\sigma_u\sigma_X$. Substitution of these limits into Equation (6.16) yields Equation (6.1).

APPENDIX

6.2 Distribution of the OLS Estimators When There Are Two Regressors and Homoskedastic Errors

Although the general formula for the variance of the OLS estimators in multiple regression is complicated, if there are two regressors ($k = 2$) and the errors are homoskedastic, then the formula simplifies enough to provide some insights into the distribution of the OLS estimators.

Because the errors are homoskedastic, the conditional variance of u_i can be written as $\text{var}(u_i|X_{1i}, X_{2i}) = \sigma_u^2$. When there are two regressors, X_{1i} and X_{2i}, and the error term is homoskedastic, in large samples the sampling distribution of $\hat{\beta}_1$ is $N(\beta_1, \sigma_{\hat{\beta}_1}^2)$ where the variance of this distribution, $\sigma_{\hat{\beta}_1}^2$, is

$$\sigma_{\hat{\beta}_1}^2 = \frac{1}{n}\left(\frac{1}{1 - \rho_{X_1,X_2}^2}\right)\frac{\sigma_u^2}{\sigma_{X_1}^2}, \tag{6.17}$$

where ρ_{X_1, X_2} is the population correlation between the two regressors X_1 and X_2 and $\sigma_{X_1}^2$ is the population variance of X_1.

The variance $\sigma_{\hat{\beta}_1}^2$ of the sampling distribution of $\hat{\beta}_1$ depends on the squared correlation between the regressors. If X_1 and X_2 are highly correlated, either positively or negatively, then ρ_{X_1,X_2}^2 is close to 1, and thus the term $1 - \rho_{X_1,X_2}^2$ in the denominator of Equation (6.17) is small and the variance of $\hat{\beta}_1$ is larger than it would be if ρ_{X_1, X_2} were close to 0.

Another feature of the joint normal large-sample distribution of the OLS estimators is that $\hat{\beta}_1$ and $\hat{\beta}_2$ are in general correlated. When the errors are homoskedastic, the correlation between the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ is the negative of the correlation between the two regressors:

$$\text{corr}(\hat{\beta}_1, \hat{\beta}_2) = -\rho_{X_1, X_2}. \tag{6.18}$$

APPENDIX

6.3 The Frisch–Waugh Theorem

The OLS estimator in multiple regression can be computed by a sequence of shorter regressions. Consider the multiple regression model in Equation (6.7). The OLS estimator of β_1 can be computed in three steps:

1. Regress X_1 on $X_2, X_3, \ldots, X_k$, and let $\widetilde{X}_1$ denote the residuals from this regression;
2. Regress Y on $X_2, X_3, \ldots, X_k$, and let $\widetilde{Y}$ denote the residuals from this regression; and
3. Regress $\widetilde{Y}$ on $\widetilde{X}_1$,

where the regressions include a constant term (intercept). The Frisch-Waugh theorem states that the OLS coefficient in step 3 equals the OLS coefficient on X_1 in the multiple regression model [Equation (6.7)].

This result provides a mathematical statement of how the multiple regression coefficient $\hat{\beta}_1$ estimates the effect on Y of X_1, controlling for the other X's: Because the first two regressions (steps 1 and 2) remove from Y and X_1 their variation associated with the other X's, the third regression estimates the effect on Y of X_1 using what is left over after removing (controlling for) the effect of the other X's. The Frisch-Waugh theorem is proven in Exercise 18.17.

This theorem suggests how Equation (6.17) can be derived from Equation (5.27). Because $\hat{\beta}_1$ is the OLS regression coefficient from the regression of $\widetilde{Y}$ onto $\widetilde{X}_1$, Equation (5.27) suggests that the homoskedasticity-only variance of $\hat{\beta}_1$ is $\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_u}{n\sigma^2_{\widetilde{X}_1}}$, where $\sigma^2_{\widetilde{X}_1}$ is the variance of $\widetilde{X}_1$. Because $\widetilde{X}_1$ is the residual from the regression of X_1 onto X_2 (recall that Equation (6.17) pertains to the model with $k = 2$ regressors), Equation (6.15) implies that $s^2_{\widetilde{X}_1} = (1 - \overline{R}^2_{X_1,X_2})s^2_{X_1}$, where $\overline{R}^2_{X_1,X_2}$ is the adjusted R^2 from the regression of X_1 onto X_2. Equation (6.17) follows from $s^2_{\widetilde{X}_1} \xrightarrow{p} \sigma^2_{\widetilde{X}_1}$, $\overline{R}^2_{X_1,X_2} \xrightarrow{p} \rho^2_{X_1,X_2}$ and $s^2_{X_1} \xrightarrow{p} \sigma^2_{X_1}$.

7 Hypothesis Tests and Confidence Intervals in Multiple Regression

As discussed in Chapter 6, multiple regression analysis provides a way to mitigate the problem of omitted variable bias by including additional regressors, thereby controlling for the effects of those additional regressors. The coefficients of the multiple regression model can be estimated by OLS. Like all estimators, the OLS estimator has sampling uncertainty because its value differs from one sample to the next.

This chapter presents methods for quantifying the sampling uncertainty of the OLS estimator through the use of standard errors, statistical hypothesis tests, and confidence intervals. One new possibility that arises in multiple regression is a hypothesis that simultaneously involves two or more regression coefficients. The general approach to testing such "joint" hypotheses involves a new test statistic, the *F*-statistic.

Section 7.1 extends the methods for statistical inference in regression with a single regressor to multiple regression. Sections 7.2 and 7.3 show how to test hypotheses that involve two or more regression coefficients. Section 7.4 extends the notion of confidence intervals for a single coefficient to confidence sets for multiple coefficients. Deciding which variables to include in a regression is an important practical issue, so Section 7.5 discusses ways to approach this problem. In Section 7.6, we apply multiple regression analysis to obtain improved estimates of the effect on test scores of a reduction in the student–teacher ratio using the California test score data set.

7.1 Hypothesis Tests and Confidence Intervals for a Single Coefficient

This section describes how to compute the standard error, how to test hypotheses, and how to construct confidence intervals for a single coefficient in a multiple regression equation.

Standard Errors for the OLS Estimators

Recall that, in the case of a single regressor, it was possible to estimate the variance of the OLS estimator by substituting sample averages for expectations, which led to the estimator $\hat{\sigma}^2_{\hat{\beta}_1}$ given in Equation (5.4). Under the least squares assumptions, the law of large numbers implies that these sample averages converge to their population counterparts, so, for example, $\hat{\sigma}^2_{\hat{\beta}_1} / \sigma^2_{\hat{\beta}_1} \xrightarrow{p} 1$. The square root of $\hat{\sigma}^2_{\hat{\beta}_1}$ is the standard error of $\hat{\beta}_1$, $SE(\hat{\beta}_1)$, an estimator of the standard deviation of the sampling distribution of $\hat{\beta}_1$.

All this extends directly to multiple regression. The OLS estimator $\hat{\beta}_j$ of the j^{th} regression coefficient has a standard deviation, and this standard deviation is estimated by its standard error, $SE(\hat{\beta}_j)$. The formula for the standard error is most easily stated using matrices (see Section 18.2). The important point is that, as far as standard errors are concerned, there is nothing conceptually different between the single- or multiple-regressor cases. The key ideas—the large-sample normality of the estimators and the ability to estimate consistently the standard deviation of their sampling distribution—are the same whether one has one, two, or 12 regressors.

Hypothesis Tests for a Single Coefficient

Suppose that you want to test the hypothesis that a change in the student–teacher ratio has no effect on test scores, holding constant the percentage of English learners in the district. This corresponds to hypothesizing that the true coefficient β_1 on the student–teacher ratio is zero in the population regression of test scores on *STR* and *PctEL*. More generally, we might want to test the hypothesis that the true coefficient β_j on the j^{th} regressor takes on some specific value, $\beta_{j,0}$. The null value $\beta_{j,0}$ comes either from economic theory or, as in the student–teacher ratio example, from the decision-making context of the application. If the alternative hypothesis is two-sided, then the two hypotheses can be written mathematically as

$$H_0: \beta_j = \beta_{j,0} \text{ vs. } H_1: \beta_j \neq \beta_{j,0} \quad \text{(two-sided alternative).} \tag{7.1}$$

For example, if the first regressor is *STR*, then the null hypothesis that changing the student–teacher ratio has no effect on class size corresponds to the null hypothesis that $\beta_1 = 0$ (so $\beta_{1,0} = 0$). Our task is to test the null hypothesis H_0 against the alternative H_1 using a sample of data.

Key Concept 5.2 gives a procedure for testing this null hypothesis when there is a single regressor. The first step in this procedure is to calculate the standard error of the coefficient. The second step is to calculate the t-statistic using the general formula in Key Concept 5.1. The third step is to compute the p-value of the test using the cumulative normal distribution in Appendix Table 1 or, alternatively, to compare the t-statistic to the critical value corresponding to the

Testing the Hypothesis $\beta_j = \beta_{j,0}$ Against the Alternative $\beta_j \neq \beta_{j,0}$

KEY CONCEPT 7.1

1. Compute the standard error of $\hat{\beta}_j$, $SE(\hat{\beta}_j)$.
2. Compute the t-statistic,

$$t = \frac{\hat{\beta}_j - \beta_{j,0}}{SE(\hat{\beta}_j)}. \tag{7.2}$$

3. Compute the p-value,

$$p\text{-value} = 2\Phi(-|t^{act}|), \tag{7.3}$$

where t^{act} is the value of the t-statistic actually computed. Reject the hypothesis at the 5% significance level if the p-value is less than 0.05 or, equivalently, if $|t^{act}| > 1.96$.

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

desired significance level of the test. The theoretical underpinnings of this procedure are that the OLS estimator has a large-sample normal distribution that, under the null hypothesis, has as its mean the hypothesized true value and that the variance of this distribution can be estimated consistently.

This underpinning is present in multiple regression as well. As stated in Key Concept 6.5, the sampling distribution of $\hat{\beta}_j$ is approximately normal. Under the null hypothesis the mean of this distribution is $\beta_{j,0}$. The variance of this distribution can be estimated consistently. Therefore we can simply follow the same procedure as in the single-regressor case to test the null hypothesis in Equation (7.1).

The procedure for testing a hypothesis on a single coefficient in multiple regression is summarized as Key Concept 7.1. The t-statistic actually computed is denoted t^{act} in this box. However, it is customary to denote this simply as t, and we adopt this simplified notation for the rest of the book.

Confidence Intervals for a Single Coefficient

The method for constructing a confidence interval in the multiple regression model is also the same as in the single-regressor model. This method is summarized as Key Concept 7.2.

KEY CONCEPT 7.2

Confidence Intervals for a Single Coefficient in Multiple Regression

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

$$95\% \text{ confidence interval for } \beta_j = [\hat{\beta}_j - 1.96\,SE(\hat{\beta}_j), \hat{\beta}_j + 1.96\,SE(\hat{\beta}_j)]. \quad (7.4)$$

A 90% confidence interval is obtained by replacing 1.96 in Equation (7.4) with 1.64.

The method for conducting a hypothesis test in Key Concept 7.1 and the method for constructing a confidence interval in Key Concept 7.2 rely on the large-sample normal approximation to the distribution of the OLS estimator $\hat{\beta}_j$. Accordingly, it should be kept in mind that these methods for quantifying the sampling uncertainty are only guaranteed to work in large samples.

Application to Test Scores and the Student–Teacher Ratio

Can we reject the null hypothesis that a change in the student–teacher ratio has no effect on test scores, once we control for the percentage of English learners in the district? What is a 95% confidence interval for the effect on test scores of a change in the student–teacher ratio, controlling for the percentage of English learners? We are now able to find out. The regression of test scores against *STR* and *PctEL*, estimated by OLS, was given in Equation (6.12) and is restated here with standard errors in parentheses below the coefficients:

$$\widehat{TestScore} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10} \times STR - \underset{(0.031)}{0.650} \times PctEL. \quad (7.5)$$

To test the hypothesis that the true coefficient on *STR* is 0, we first need to compute the t-statistic in Equation (7.2). Because the null hypothesis says that the true value of this coefficient is zero, the t-statistic is $t = (-1.10 - 0)/0.43 = -2.54$.

The associated p-value is $2\Phi(-2.54) = 1.1\%$; that is, the smallest significance level at which we can reject the null hypothesis is 1.1%. Because the p-value is less than 5%, the null hypothesis can be rejected at the 5% significance level (but not quite at the 1% significance level).

A 95% confidence interval for the population coefficient on STR is $-1.10 \pm 1.96 \times 0.43 = (-1.95, -0.26)$; that is, we can be 95% confident that the true value of the coefficient is between -1.95 and -0.26. Interpreted in the context of the superintendent's interest in decreasing the student–teacher ratio by 2, the 95% confidence interval for the effect on test scores of this reduction is $(-1.95 \times 2, -0.26 \times 2) = (-3.90, -0.52)$.

Adding expenditures per pupil to the equation. Your analysis of the multiple regression in Equation (7.5) has persuaded the superintendent that, based on the evidence so far, reducing class size will improve test scores in her district. Now, however, she moves on to a more nuanced question. If she is to hire more teachers, she can pay for those teachers either through cuts elsewhere in the budget (no new computers, reduced maintenance, and so on) or by asking for an increase in her budget, which taxpayers do not favor. What, she asks, is the effect on test scores of reducing the student–teacher ratio, holding expenditures per pupil (and the percentage of English learners) constant?

This question can be addressed by estimating a regression of test scores on the student–teacher ratio, total spending per pupil, and the percentage of English learners. The OLS regression line is

$$\widehat{TestScore} = \underset{(15.5)}{649.6} - \underset{(0.48)}{0.29} \times STR + \underset{(1.59)}{3.87} \times Expn - \underset{(0.032)}{0.656} \times PctEL, \quad (7.6)$$

where $Expn$ is total annual expenditures per pupil in the district in thousands of dollars.

The result is striking. Holding expenditures per pupil and the percentage of English learners constant, changing the student–teacher ratio is estimated to have a very small effect on test scores: The estimated coefficient on STR is -1.10 in Equation (7.5) but, after adding $Expn$ as a regressor in Equation (7.6), it is only -0.29. Moreover, the t-statistic for testing that the true value of the coefficient is zero is now $t = (-0.29 - 0)/0.48 = -0.60$, so the hypothesis that the population value of this coefficient is indeed zero cannot be rejected even at the 10% significance level ($|-0.60| < 1.64$). Thus Equation (7.6) provides no evidence that hiring more teachers improves test scores if overall expenditures per pupil are held constant.

One interpretation of the regression in Equation (7.6) is that, in these California data, school administrators allocate their budgets efficiently. Suppose, counterfactually, that the coefficient on *STR* in Equation (7.6) were negative and large. If so, school districts could raise their test scores simply by decreasing funding for other purposes (textbooks, technology, sports, and so on) and transferring those funds to hire more teachers, thereby reducing class sizes while holding expenditures constant. However, the small and statistically insignificant coefficient on *STR* in Equation (7.6) indicates that this transfer would have little effect on test scores. Put differently, districts are already allocating their funds efficiently.

Note that the standard error on *STR* increased when *Expn* was added, from 0.43 in Equation (7.5) to 0.48 in Equation (7.6). This illustrates the general point, introduced in Section 6.7 in the context of imperfect multicollinearity, that correlation between regressors (the correlation between *STR* and *Expn* is −0.62) can make the OLS estimators less precise.

What about our angry taxpayer? He asserts that the population values of *both* the coefficient on the student–teacher ratio (β_1) *and* the coefficient on spending per pupil (β_2) are zero; that is, he hypothesizes that both $\beta_1 = 0$ and $\beta_2 = 0$. Although it might seem that we can reject this hypothesis because the t-statistic testing $\beta_2 = 0$ in Equation (7.6) is $t = 3.87/1.59 = 2.43$, this reasoning is flawed. The taxpayer's hypothesis is a joint hypothesis, and to test it we need a new tool, the F-statistic.

7.2 Tests of Joint Hypotheses

This section describes how to formulate joint hypotheses on multiple regression coefficients and how to test them using an F-statistic.

Testing Hypotheses on Two or More Coefficients

Joint null hypotheses. Consider the regression in Equation (7.6) of the test score against the student–teacher ratio, expenditures per pupil, and the percentage of English learners. Our angry taxpayer hypothesizes that neither the student–teacher ratio nor expenditures per pupil have an effect on test scores, once we control for the percentage of English learners. Because *STR* is the first regressor in Equation (7.6) and *Expn* is the second, we can write this hypothesis mathematically as

$$H_0: \beta_1 = 0 \text{ and } \beta_2 = 0 \text{ vs. } H_1: \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0. \tag{7.7}$$

The hypothesis that *both* the coefficient on the student–teacher ratio (β_1) *and* the coefficient on expenditures per pupil (β_2) are zero is an example of a joint hypothesis on the coefficients in the multiple regression model. In this case, the null hypothesis restricts the value of two of the coefficients, so as a matter of terminology we can say that the null hypothesis in Equation (7.7) imposes two **restrictions** on the multiple regression model: $\beta_1 = 0$ *and* $\beta_2 = 0$.

In general, a **joint hypothesis** is a hypothesis that imposes two or more restrictions on the regression coefficients. We consider joint null and alternative hypotheses of the form

$$H_0\colon \beta_j = \beta_{j,0}, \beta_m = \beta_{m,0}, \ldots, \text{ for a total of } q \text{ restrictions, vs.}$$
$$H_1\colon \text{one or more of the } q \text{ restrictions under } H_0 \text{ does not hold,} \tag{7.8}$$

where $\beta_j, \beta_m, \ldots,$ refer to different regression coefficients and $\beta_{j,0}, \beta_{m,0}, \ldots,$ refer to the values of these coefficients under the null hypothesis. The null hypothesis in Equation (7.7) is an example of Equation (7.8). Another example is that, in a regression with $k = 6$ regressors, the null hypothesis is that the coefficients on the 2^{nd}, 4^{th}, and 5^{th} regressors are zero; that is, $\beta_2 = 0, \beta_4 = 0$, and $\beta_5 = 0$ so that there are $q = 3$ restrictions. In general, under the null hypothesis H_0 there are q such restrictions.

If any one (or more than one) of the equalities under the null hypothesis H_0 in Equation (7.8) is false, then the joint null hypothesis itself is false. Thus the alternative hypothesis is that at least one of the equalities in the null hypothesis H_0 does not hold.

Why can't I just test the individual coefficients one at a time? Although it seems it should be possible to test a joint hypothesis by using the usual t-statistics to test the restrictions one at a time, the following calculation shows that this approach is unreliable. Specifically, suppose that you are interested in testing the joint null hypothesis in Equation (7.6) that $\beta_1 = 0$ and $\beta_2 = 0$. Let t_1 be the t-statistic for testing the null hypothesis that $\beta_1 = 0$ and let t_2 be the t-statistic for testing the null hypothesis that $\beta_2 = 0$. What happens when you use the "one-at-a-time" testing procedure: Reject the joint null hypothesis if either t_1 or t_2 exceeds 1.96 in absolute value?

Because this question involves the two random variables t_1 and t_2, answering it requires characterizing the joint sampling distribution of t_1 and t_2. As mentioned in Section 6.6, in large samples $\hat{\beta}_1$ and $\hat{\beta}_2$ have a joint normal distribution, so under the joint null hypothesis the t-statistics t_1 and t_2 have a bivariate normal distribution, where each t-statistic has mean equal to 0 and variance equal to 1.

First consider the special case in which the t-statistics are uncorrelated and thus are independent. What is the size of the one-at-a-time testing procedure; that is, what is the probability that you will reject the null hypothesis when it is true? More than 5%! In this special case we can calculate the rejection probability of this method exactly. The null is *not* rejected only if both $|t_1| \leq 1.96$ and $|t_2| \leq 1.96$. Because the t-statistics are independent, $Pr(|t_1| \leq 1.96 \text{ and } |t_2| \leq 1.96) = Pr(|t_1| \leq 1.96) \times Pr(|t_2| \leq 1.96) = 0.95^2 = 0.9025 = 90.25\%$. So the probability of rejecting the null hypothesis when it is true is $1 - 0.95^2 = 9.75\%$. This "one at a time" method rejects the null too often because it gives you too many chances: If you fail to reject using the first t-statistic, you get to try again using the second.

If the regressors are correlated, the situation is even more complicated. The size of the "one at a time" procedure depends on the value of the correlation between the regressors. Because the "one at a time" testing approach has the wrong size—that is, its rejection rate under the null hypothesis does not equal the desired significance level—a new approach is needed.

One approach is to modify the "one at a time" method so that it uses different critical values that ensure that its size equals its significance level. This method, called the Bonferroni method, is described in Appendix (7.1). The advantage of the Bonferroni method is that it applies very generally. Its disadvantage is that it can have low power: It frequently fails to reject the null hypothesis when in fact the alternative hypothesis is true.

Fortunately, there is another approach to testing joint hypotheses that is more powerful, especially when the regressors are highly correlated. That approach is based on the F-statistic.

The *F*-Statistic

The ***F*-statistic** is used to test joint hypothesis about regression coefficients. The formulas for the F-statistic are integrated into modern regression software. We first discuss the case of two restrictions, then turn to the general case of q restrictions.

The F-statistic with q = 2 restrictions. When the joint null hypothesis has the two restrictions that $\beta_1 = 0$ and $\beta_2 = 0$, the F-statistic combines the two t-statistics t_1 and t_2 using the formula

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2}t_1t_2}{1 - \hat{\rho}^2_{t_1,t_2}}\right), \tag{7.9}$$

where $\hat{\rho}_{t_1,t_2}$ is an estimator of the correlation between the two t-statistics.

To understand the F-statistic in Equation (7.9), first suppose that we know that the t-statistics are uncorrelated so we can drop the terms involving $\hat{\rho}_{t_1,t_2}$. If so, Equation (7.9) simplifies and $F = \frac{1}{2}(t_1^2 + t_2^2)$; that is, the F-statistic is the average of the squared t-statistics. Under the null hypothesis, t_1 and t_2 are independent standard normal random variables (because the t-statistics are uncorrelated by assumption), so under the null hypothesis F has an $F_{2,\infty}$ distribution (Section 2.4). Under the alternative hypothesis that either β_1 is nonzero or β_2 is nonzero (or both), then either t_1^2 or t_2^2 (or both) will be large, leading the test to reject the null hypothesis.

In general the t-statistics are correlated, and the formula for the F-statistic in Equation (7.9) adjusts for this correlation. This adjustment is made so that, under the null hypothesis, the F-statistic has an $F_{2,\infty}$ distribution in large samples whether or not the t-statistics are correlated.

The F-statistic with q restrictions. The formula for the heteroskedasticity-robust F-statistic testing the q restrictions of the joint null hypothesis in Equation (7.8) is given in Section 18.3. This formula is incorporated into regression software, making the F-statistic easy to compute in practice.

Under the null hypothesis, the F-statistic has a sampling distribution that, in large samples, is given by the $F_{q,\infty}$ distribution. That is, in large samples, under the null hypothesis

$$\text{the } F\text{-statistic is distributed } F_{q,\infty}. \tag{7.10}$$

Thus the critical values for the F-statistic can be obtained from the tables of the $F_{q,\infty}$ distribution in Appendix Table 4 for the appropriate value of q and the desired significance level.

Computing the heteroskedasticity-robust F-statistic in statistical software. If the F-statistic is computed using the general heteroskedasticity-robust formula, its large-n distribution under the null hypothesis is $F_{q,\infty}$ regardless of whether the errors are homoskedastic or heteroskedastic. As discussed in Section 5.4, for historical reasons most statistical software computes homoskedasticity-only standard errors by default. Consequently, in some software packages you must select a "robust" option so that the F-statistic is computed using heteroskedasticity-robust standard errors (and, more generally, a heteroskedasticity-robust estimate of the "covariance matrix"). The homoskedasticity-only version of the F-statistic is discussed at the end of this section.

Computing the p-value using the F-statistic. The p-value of the F-statistic can be computed using the large-sample $F_{q,\infty}$ approximation to its distribution. Let

F^{act} denote the value of the F-statistic actually computed. Because the F-statistic has a large-sample $F_{q,\infty}$ distribution under the null hypothesis, the p-value is

$$p\text{-value} = \Pr[F_{q,\infty} > F^{act}]. \tag{7.11}$$

The p-value in Equation (7.11) can be evaluated using a table of the $F_{q,\infty}$ distribution (or, alternatively, a table of the χ^2_q distribution, because a χ^2_q-distributed random variable is q times an $F_{q,\infty}$-distributed random variable). Alternatively, the p-value can be evaluated using a computer, because formulas for the cumulative chi-squared and F distributions have been incorporated into most modern statistical software.

The "overall" regression F-statistic. The "overall" regression F-statistic tests the joint hypothesis that *all* the slope coefficients are zero. That is, the null and alternative hypotheses are

$$H_0: \beta_1 = 0, \beta_2 = 0, \ldots, \beta_k = 0 \text{ vs. } H_1: \beta_j \neq 0, \text{ at least one } j, j = 1, \ldots, k. \tag{7.12}$$

Under this null hypothesis, none of the regressors explains any of the variation in Y_i, although the intercept (which under the null hypothesis is the mean of Y_i) can be nonzero. The null hypothesis in Equation (7.12) is a special case of the general null hypothesis in Equation (7.8), and the overall regression F-statistic is the F-statistic computed for the null hypothesis in Equation (7.12). In large samples, the overall regression F-statistic has an $F_{k,\infty}$ distribution when the null hypothesis is true.

The F-statistic when q = 1. When $q = 1$, the F-statistic tests a single restriction. Then the joint null hypothesis reduces to the null hypothesis on a single regression coefficient, and the F-statistic is the square of the t-statistic.

Application to Test Scores and the Student–Teacher Ratio

We are now able to test the null hypothesis that the coefficients on *both* the student–teacher ratio *and* expenditures per pupil are zero, against the alternative that at least one coefficient is nonzero, controlling for the percentage of English learners in the district.

To test this hypothesis, we need to compute the heteroskedasticity-robust F-statistic of the test that $\beta_1 = 0$ and $\beta_2 = 0$ using the regression of *TestScore* on *STR*, *Expn*, and *PctEL* reported in Equation (7.6). This F-statistic is 5.43. Under the null hypothesis, in large samples this statistic has an $F_{2,\infty}$ distribution. The 5% critical value of the $F_{2,\infty}$ distribution is 3.00 (Appendix Table 4), and the 1% critical value is 4.61. The value of the F-statistic computed from the data, 5.43, exceeds 4.61, so the null hypothesis is rejected at the 1% level. It is very unlikely that we would have drawn a sample that produced an F-statistic as large as 5.43 if the null hypothesis really were true (the p-value is 0.005). Based on the evidence in Equation (7.6) as summarized in this F-statistic, we can reject the taxpayer's hypothesis that *neither* the student–teacher ratio *nor* expenditures per pupil have an effect on test scores (holding constant the percentage of English learners).

The Homoskedasticity-Only *F*-Statistic

One way to restate the question addressed by the F-statistic is to ask whether relaxing the q restrictions that constitute the null hypothesis improves the fit of the regression by enough that this improvement is unlikely to be the result merely of random sampling variation if the null hypothesis is true. This restatement suggests that there is a link between the F-statistic and the regression R^2: A large F-statistic should, it seems, be associated with a substantial increase in the R^2. In fact, if the error u_i is homoskedastic, this intuition has an exact mathematical expression. Specifically, if the error term is homoskedastic, the F-statistic can be written in terms of the improvement in the fit of the regression as measured either by the decrease in the sum of squared residuals or by the increase in the regression R^2. The resulting F-statistic is referred to as the homoskedasticity-only F-statistic, because it is valid only if the error term is homoskedastic. In contrast, the heteroskedasticity-robust F-statistic computed using the formula in Section 18.3 is valid whether the error term is homoskedastic or heteroskedastic. Despite this significant limitation of the homoskedasticity-only F-statistic, its simple formula sheds light on what the F-statistic is doing. In addition, the simple formula can be computed using standard regression output, such as might be reported in a table that includes regression R^2's but not F-statistics.

The homoskedasticity-only F-statistic is computed using a simple formula based on the sum of squared residuals from two regressions. In the first regression, called the **restricted regression**, the null hypothesis is forced to be true. When the null hypothesis is of the type in Equation (7.8), where all the hypothesized values are zero, the restricted regression is the regression in which those coefficients are set to zero; that is, the relevant regressors are excluded from the regression. In the second regression, called the **unrestricted regression**, the alternative hypothesis is allowed to be true. If the sum of squared residuals is sufficiently smaller in the unrestricted than the restricted regression, then the test rejects the null hypothesis.

The **homoskedasticity-only *F*-statistic** is given by the formula

$$F = \frac{(SSR_{restricted} - SSR_{unrestricted})/q}{SSR_{unrestricted}/(n - k_{unrestricted} - 1)}, \tag{7.13}$$

where $SSR_{restricted}$ is the sum of squared residuals from the restricted regression, $SSR_{unrestricted}$ is the sum of squared residuals from the unrestricted regression, q is the number of restrictions under the null hypothesis, and $k_{unrestricted}$ is the number of regressors in the unrestricted regression. An alternative equivalent formula for the homoskedasticity-only F-statistic is based on the R^2 of the two regressions:

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})(n - k_{unrestricted} - 1)}. \tag{7.14}$$

If the errors are homoskedastic, then the difference between the homoskedasticity-only F-statistic computed using Equation (7.13) or (7.14) and the heteroskedasticity-robust F-statistic vanishes as the sample size n increases. Thus, if the errors are homoskedastic, the sampling distribution of the homoskedasticity-only F-statistic under the null hypothesis is, in large samples, $F_{q,\infty}$.

These formulas are easy to compute and have an intuitive interpretation in terms of how well the unrestricted and restricted regressions fit the data. Unfortunately, the formulas apply only if the errors are homoskedastic. Because homoskedasticity is a special case that cannot be counted on in applications with economic data, or more generally with data sets typically found in the social sciences, in practice the homoskedasticity-only F-statistic is not a satisfactory substitute for the heteroskedasticity-robust F-statistic.

Using the homoskedasticity-only F-statistic when n is small. If the errors are homoskedastic and are i.i.d. normally distributed, then the homoskedasticity-only F-statistic defined in Equations (7.13) and (7.14) has an $F_{q,n-k_{unrestricted}-1}$ distribution under the null hypothesis. Critical values for this distribution, which depend on both q and $n - k_{unrestricted} - 1$, are given in Appendix Table 5. As discussed in Section 2.4, the $F_{q,n-k_{unrestricted}-1}$ distribution converges to the $F_{q,\infty}$ distribution as n increases; for large sample sizes, the differences between the two distributions are negligible. For small samples, however, the two sets of critical values differ.

Application to test scores and the student–teacher ratio. To test the null hypothesis that the population coefficients on STR and $Expn$ are 0, controlling for $PctEL$, we need to compute the R^2 (or SSR) for the restricted and unrestricted regression. The unrestricted regression has the regressors *STR*, *Expn*, and *PctEL*, and is given in Equation (7.6); its R^2 is 0.4366; that is, $R^2_{unrestricted} = 0.4366$. The restricted regression imposes the joint null hypothesis that the true coefficients on *STR* and *Expn* are zero; that is, under the null hypothesis *STR* and *Expn* do not enter the population regression, although *PctEL* does (the null hypothesis does not restrict the coefficient on *PctEL*). The restricted regression, estimated by OLS, is

$$\widehat{TestScore} = \underset{(1.0)}{664.7} - \underset{(0.032)}{0.671} \times PctEL, R^2 = 0.4149, \tag{7.15}$$

so $R^2_{restricted} = 0.4149$. The number of restrictions is $q = 2$, the number of observations is $n = 420$, and the number of regressors in the unrestricted regression is $k = 3$. The homoskedasticity-only F-statistic, computed using Equation (7.14), is

$$F = \frac{(0.4366 - 0.4149)/2}{(1 - 0.4366)(420 - 3 - 1)} = 8.01.$$

Because 8.01 exceeds the 1% critical value of 4.61, the hypothesis is rejected at the 1% level using the homoskedasticity-only test.

This example illustrates the advantages and disadvantages of the homoskedasticity-only F-statistic. Its advantage is that it can be computed using a calculator. Its disadvantage is that the values of the homoskedasticity-only and heteroskedasticity-robust F-statistics can be very different: The heteroskedasticity-robust F-statistic testing this joint hypothesis is 5.43, quite different from the less reliable homoskedasticity-only value of 8.01.

7.3 Testing Single Restrictions Involving Multiple Coefficients

Sometimes economic theory suggests a single restriction that involves two or more regression coefficients. For example, theory might suggest a null hypothesis of the form $\beta_1 = \beta_2$; that is, the effects of the first and second regressor are the same. In this case, the task is to test this null hypothesis against the alternative that the two coefficients differ:

$$H_0: \beta_1 = \beta_2 \text{ vs. } H_1: \beta_1 \neq \beta_2. \tag{7.16}$$

This null hypothesis has a single restriction, so $q = 1$, but that restriction involves multiple coefficients (β_1 and β_2). We need to modify the methods presented so far to test this hypothesis. There are two approaches; which is easier depends on your software.

Approach #1: Test the restriction directly. Some statistical packages have a specialized command designed to test restrictions like Equation (7.16) and the result is an F-statistic that, because $q = 1$, has an $F_{1,\infty}$ distribution under the null hypothesis. (Recall from Section 2.4 that the square of a standard normal random variable has an $F_{1,\infty}$ distribution, so the 95% percentile of the $F_{1,\infty}$ distribution is $1.96^2 = 3.84$.)

Approach #2: Transform the regression. If your statistical package cannot test the restriction directly, the hypothesis in Equation (7.16) can be tested using a trick in which the original regression equation is rewritten to turn the restriction in Equation (7.16) into a restriction on a single regression coefficient. To be concrete, suppose there are only two regressors, X_{1i} and X_{2i}, in the regression, so the population regression has the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i. \tag{7.17}$$

Here is the trick: By subtracting and adding $\beta_2 X_{1i}$, we have that $\beta_1 X_{1i} + \beta_2 X_{2i} = \beta_1 X_{1i} - \beta_2 X_{1i} + \beta_2 X_{1i} + \beta_2 X_{2i} = (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) = \gamma_1 X_{1i} + \beta_2 W_i$, where $\gamma_1 = \beta_1 - \beta_2$ and $W_i = X_{1i} + X_{2i}$. Thus the population regression in Equation (7.17) can be rewritten as

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i. \tag{7.18}$$

Because the coefficient γ_1 in this equation is $\gamma_1 = \beta_1 - \beta_2$, under the null hypothesis in Equation (7.16), $\gamma_1 = 0$, while under the alternative, $\gamma_1 \neq 0$. Thus, by turning Equation (7.17) into Equation (7.18), we have turned a restriction on two regression coefficients into a restriction on a single regression coefficient.

Because the restriction now involves the single coefficient γ_1, the null hypothesis in Equation (7.16) can be tested using the t-statistic method of Section 7.1. In practice, this is done by first constructing the new regressor W_i as the sum of the two original regressors, then estimating the regression of Y_i on X_{1i} and W_i. A 95% confidence interval for the difference in the coefficients $\beta_1 - \beta_2$ can be calculated as $\hat{\gamma}_1 \pm 1.96\, SE(\hat{\gamma}_1)$.

This method can be extended to other restrictions on regression equations using the same trick (see Exercise 7.9).

The two methods (Approaches #1 and #2) are equivalent, in the sense that the F-statistic from the first method equals the square of the t-statistic from the second method.

Extension to $q > 1$. In general, it is possible to have q restrictions under the null hypothesis in which some or all of these restrictions involve multiple coefficients. The F-statistic of Section 7.2 extends to this type of joint hypothesis. The F-statistic can be computed by either of the two methods just discussed for $q = 1$. Precisely how best to do this in practice depends on the specific regression software being used.

7.4 Confidence Sets for Multiple Coefficients

This section explains how to construct a confidence set for two or more regression coefficients. The method is conceptually similar to the method in Section 7.1 for constructing a confidence set for a single coefficient using the t-statistic, except that the confidence set for multiple coefficients is based on the F-statistic.

A **95% confidence set** for two or more coefficients is a set that contains the true population values of these coefficients in 95% of randomly drawn samples. Thus a confidence set is the generalization to two or more coefficients of a confidence interval for a single coefficient.

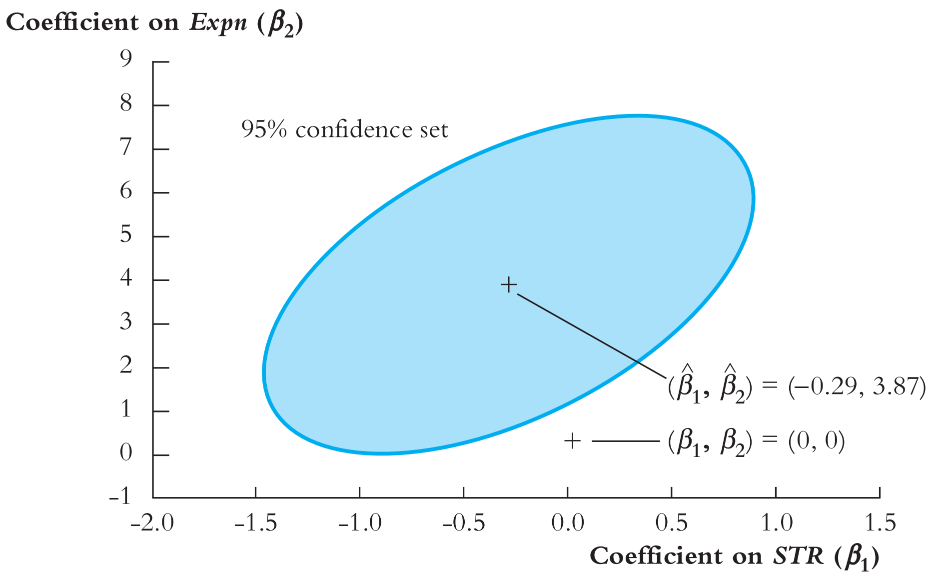
Recall that a 95% confidence interval is computed by finding the set of values of the coefficients that are not rejected using a t-statistic at the 5% significance level. This approach can be extended to the case of multiple coefficients. To make this concrete, suppose you are interested in constructing a confidence set for two coefficients, β_1 and β_2. Section 7.2 showed how to use the F-statistic to test a joint null hypothesis that $\beta_1 = \beta_{1,0}$ and $\beta_2 = \beta_{2,0}$. Suppose you were to test every possible value of $\beta_{1,0}$ and $\beta_{2,0}$ at the 5% level. For each pair of candidates $(\beta_{1,0}, \beta_{2,0})$, you compute the F-statistic and reject it if it exceeds the 5% critical value of 3.00. Because the test has a 5% significance level, the true population values of β_1 and β_2 will not be rejected in 95% of all samples. Thus the set of values not rejected at the 5% level by this F-statistic constitutes a 95% confidence set for β_1 and β_2.

Although this method of trying all possible values of $\beta_{1,0}$ and $\beta_{2,0}$ works in theory, in practice it is much simpler to use an explicit formula for the confidence set. This formula for the confidence set for an arbitrary number of coefficients is based on the formula for the F-statistic. When there are two coefficients, the resulting confidence sets are ellipses.

As an illustration, Figure 7.1 shows a 95% confidence set (confidence ellipse) for the coefficients on the student–teacher ratio and expenditure per pupil, holding constant the percentage of English learners, based on the estimated regression in Equation (7.6). This ellipse does not include the point (0,0). This means that the

FIGURE 7.1 95% Confidence Set for Coefficients on *STR* and *Expn* from Equation (7.6)

The 95% confidence set for the coefficients on *STR* (β_1) and *Expn* (β_2) is an ellipse. The ellipse contains the pairs of values of β_1 and β_2 that cannot be rejected using the *F*-statistic at the 5% significance level. The point $(\beta_1, \beta_2) = (0, 0)$ is not contained in the confidence set, so the null hypothesis H_0: $\beta_1 = 0$ and $\beta_2 = 0$ is rejected at the 5% significance level.



null hypothesis that these two coefficients are both zero is rejected using the *F*-statistic at the 5% significance level, which we already knew from Section 7.2. The confidence ellipse is a fat sausage with the long part of the sausage oriented in the lower-left/upper-right direction. The reason for this orientation is that the estimated correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$ is positive, which in turn arises because the correlation between the regressors *STR* and *Expn* is negative (schools that spend more per pupil tend to have fewer students per teacher).

7.5 Model Specification for Multiple Regression

The job of determining which variables to include in multiple regression—that is, the problem of choosing a regression specification—can be quite challenging, and no single rule applies in all situations. But do not despair, because some useful guidelines are available. The starting point for choosing a regression specification is thinking through the possible sources of omitted variable bias. It is important to rely on your expert knowledge of the empirical problem and to focus on obtaining an unbiased estimate of the causal effect of interest; do not rely solely on purely statistical measures of fit such as the R^2 or $\bar{R}^2$.

Omitted Variable Bias in Multiple Regression

KEY CONCEPT 7.3

Omitted variable bias is the bias in the OLS estimator that arises when one or more included regressors are correlated with an omitted variable. For omitted variable bias to arise, two things must be true:

1. At least one of the included regressors must be correlated with the omitted variable.
2. The omitted variable must be a determinant of the dependent variable, Y.

Omitted Variable Bias in Multiple Regression

The OLS estimators of the coefficients in multiple regression will have omitted variable bias if an omitted determinant of Y_i is correlated with at least one of the regressors. For example, students from affluent families often have more learning opportunities outside the classroom (reading material at home, travel, museum visits, etc.) than do their less affluent peers, which could lead to better test scores. Moreover, if the district is a wealthy one, then the schools will tend to have larger budgets and lower student–teacher ratios. If so, the availability of outside learning opportunities and the student–teacher ratio would be negatively correlated, and the OLS estimate of the coefficient on the student–teacher ratio would pick up the effect of outside learning opportunities, even after controlling for the percentage of English learners. In short, omitting outside learning opportunities (and other variables related to the students' economic background) could lead to omitted variable bias in the regression of test scores on the student–teacher ratio and the percentage of English learners.

The general conditions for omitted variable bias in multiple regression are similar to those for a single regressor: If an omitted variable is a determinant of Y_i and if it is correlated with at least one of the regressors, then the OLS estimator of at least one of the coefficients will have omitted variable bias. The two conditions for omitted variable bias in multiple regression are summarized in Key Concept 7.3.

At a mathematical level, if the two conditions for omitted variable bias are satisfied, then at least one of the regressors is correlated with the error term. This means that the conditional expectation of u_i given $X_{1i}, \ldots, X_{ki}$ is nonzero, so the first least squares assumption is violated. As a result, the omitted variable bias persists even if the sample size is large; that is, omitted variable bias implies that the OLS estimators are inconsistent.

The Role of Control Variables in Multiple Regression

So far, we have implicitly distinguished between a regressor for which we wish to estimate a causal effect—that is, a variable of interest—and control variables. We now discuss this distinction in more detail.

A **control variable** is not the object of interest in the study; rather it is a regressor included to hold constant factors that, if neglected, could lead the estimated causal effect of interest to suffer from omitted variable bias. The least squares assumptions for multiple regression (Section 6.5) treat the regressors symmetrically. In this subsection, we introduce an alternative to the first least squares assumption in which the distinction between a variable of interest and a control variable is explicit. If this alternative assumption holds, the OLS estimator of the effect of interest is unbiased, but the OLS coefficients on control variables are in general biased and do not have a causal interpretation.

For example, consider the potential omitted variable bias arising from omitting outside learning opportunities from a test score regression. Although "outside learning opportunities" is a broad concept that is difficult to measure, those opportunities are correlated with the students' economic background, which can be measured. Thus a measure of economic background can be included in a test score regression to control for omitted income-related determinants of test scores, like outside learning opportunities. To this end, we augment the regression of test scores on *STR* and *PctEL* with the percentage of students receiving a free or subsidized school lunch (*LchPct*). Because students are eligible for this program if their family income is less than a certain threshold (approximately 150% of the poverty line), *LchPct* measures the fraction of economically disadvantaged children in the district. The estimated regression is

$$\widehat{TestScore} = \underset{(5.6)}{700.2} - \underset{(0.27)}{1.00} \times STR - \underset{(0.033)}{0.122} \times PctEL - \underset{(0.024)}{0.547} \times LchPct. \tag{7.19}$$

Including the control variable *LchPct* does not substantially change any conclusions about the class size effect: The coefficient on *STR* changes only slightly from its value of -1.10 in Equation (7.5) to -1.00 in Equation (7.19), and it remains statistically significant at the 1% level.

What does one make of the coefficient on *LchPct* in Equation (7.19)? That coefficient is very large: The difference in test scores between a district with $LchPct = 0\%$

and one with $LchPct = 50\%$ is estimated to be 27.4 points $[= 0.547 \times (50 - 0)]$, approximately the difference between the 75th and 25th percentiles of test scores in Table 4.1. Does this coefficient have a causal interpretation? Suppose that upon seeing Equation (7.19) the superintendent proposed eliminating the reduced-price lunch program so that, for her district, *LchPct* would immediately drop to zero. Would eliminating the lunch program boost her district's test scores? Common sense suggests that the answer is no; in fact, by leaving some students hungry, eliminating the reduced-price lunch program could have the opposite effect. But does it make sense to treat the coefficient on the variable of interest *STR* as causal, but not the coefficient on the control variable *LchPct*?

The distinction between variables of interest and control variables can be made mathematically precise by replacing the first least squares assumption of Key Concept 6.4—that is, the conditional mean-zero assumption—with an assumption called conditional mean independence. Consider a regression with two variables, in which X_{1i} is the variable of interest and X_{2i} is the control variable. **Conditional mean independence** requires that the conditional expectation of u_i given X_{1i} and X_{2i} does not depend on (is independent of) X_{1i}, although it can depend on X_{2i}. That is

$$E(u_i | X_{1i}, X_{2i}) = E(u_i | X_{2i}) \quad \text{(conditional mean independence)}. \tag{7.20}$$

As is shown in Appendix (7.2), under the conditional mean independence assumption in Equation (7.20), the coefficient on X_{1i} has a causal interpretation but the coefficient on X_{2i} does not.

The idea of conditional mean independence is that once you control for X_{2i}, X_{1i} can be treated as if it were randomly assigned, in the sense that the conditional mean of the error term no longer depends on X_{1i}. Including X_{2i} as a control variable makes X_{1i} uncorrelated with the error term so that OLS can estimate the causal effect on Y_{1i} of a change in X_{1i}. The control variable, however, remains correlated with the error term, so the coefficient on the control variable is subject to omitted variable bias and does not have a causal interpretation.

The terminology of control variables can be confusing. The control variable X_{2i} is included because it controls for omitted factors that affect Y_i and are correlated with X_{1i} and because it might (but need not) have a causal effect itself. Thus the coefficient on X_{1i} is the effect on Y_i of X_{1i}, using the control variable X_{2i} both to hold constant the direct effect of X_{2i} and to control for factors correlated with X_{2i}. Because this terminology is awkward, it is conventional simply to say that the coefficient on X_{1i} is the effect on Y_i, controlling for X_{2i}. When a control variable is used, it is controlling both for its own direct causal effect (if any) and for the effect of correlated omitted factors, with the aim of ensuring that conditional mean independence holds.

In the class size example, *LchPct* can be correlated with factors, such as learning opportunities outside school, that enter the error term; indeed, it is *because* of this correlation that *LchPct* is a useful control variable. This correlation between *LchPct* and the error term means that the estimated coefficient on *LchPct* does not have a causal interpretation. What the conditional mean independence assumption requires is that, given the control variables in the regression (*PctEL* and *LchPct*), the mean of the error term does not depend on the student–teacher ratio. Said differently, conditional mean independence says that among schools with the same values of *PctEL* and *LchPct*, class size is "as if" randomly assigned: including *PctEL* and *LchPct* in the regression controls for omitted factors so that *STR* is uncorrelated with the error term. If so, the coefficient on the student–teacher ratio has a causal interpretation even though the coefficient on *LchPct* does not: For the superintendent struggling to increase test scores, there is no free lunch.

Model Specification in Theory and in Practice

In theory, when data are available on the omitted variable, the solution to omitted variable bias is to include the omitted variable in the regression. In practice, however, deciding whether to include a particular variable can be difficult and requires judgment.

Our approach to the challenge of potential omitted variable bias is twofold. First, a core or base set of regressors should be chosen using a combination of expert judgment, economic theory, and knowledge of how the data were collected; the regression using this base set of regressors is sometimes referred to as a **base specification**. This base specification should contain the variables of primary interest and the control variables suggested by expert judgment and economic theory. Expert judgment and economic theory are rarely decisive, however, and often the variables suggested by economic theory are not the ones on which you have data. Therefore the next step is to develop a list of candidate **alternative specifications**, that is, alternative sets of regressors. If the estimates of the coefficients of interest are numerically similar across the alternative specifications, then this provides evidence that the estimates from your base specification are reliable. If, on the other hand, the estimates of the coefficients of interest change substantially across specifications, this often provides evidence that the original specification had omitted variable bias. We elaborate on this approach to model specification in Section 9.2 after studying some tools for specifying regressions.

Interpreting the R^2 and the Adjusted R^2 in Practice

An R^2 or an $\bar{R}^2$ near 1 means that the regressors are good at predicting the values of the dependent variable in the sample, and an R^2 or an $\bar{R}^2$ near 0 means that they are not. This makes these statistics useful summaries of the predictive ability of the regression. However, it is easy to read more into them than they deserve.

There are four potential pitfalls to guard against when using the R^2 or $\bar{R}^2$:

1. ***An increase in the R^2 or $\bar{R}^2$ does not necessarily mean that an added variable is statistically significant.*** The R^2 increases whenever you add a regressor, whether or not it is statistically significant. The $\bar{R}^2$ does not always increase, but if it does, this does not necessarily mean that the coefficient on that added regressor is statistically significant. To ascertain whether an added variable is statistically significant, you need to perform a hypothesis test using the t-statistic.

2. ***A high R^2 or $\bar{R}^2$ does not mean that the regressors are a true cause of the dependent variable.*** Imagine regressing test scores against parking lot area per pupil. Parking lot area is correlated with the student–teacher ratio, with whether the school is in a suburb or a city, and possibly with district income—all things that are correlated with test scores. Thus the regression of test scores on parking lot area per pupil could have a high R^2 and $\bar{R}^2$, but the relationship is not causal (try telling the superintendent that the way to increase test scores is to increase parking space!).

3. ***A high R^2 or $\bar{R}^2$ does not mean that there is no omitted variable bias.*** Recall the discussion of Section 6.1, which concerned omitted variable bias in the regression of test scores on the student–teacher ratio. The R^2 of the regression never came up because it played no logical role in this discussion. Omitted variable bias can occur in regressions with a low R^2, a moderate R^2, or a high R^2. Conversely, a low R^2 does not imply that there necessarily is omitted variable bias.

4. ***A high R^2 or $\bar{R}^2$ does not necessarily mean that you have the most appropriate set of regressors, nor does a low R^2 or $\bar{R}^2$ necessarily mean that you have an inappropriate set of regressors.*** The question of what constitutes the right set of regressors in multiple regression is difficult, and we return to it throughout this textbook. Decisions about the regressors must weigh issues of omitted variable bias, data availability, data quality, and, most importantly, economic theory and the nature of the substantive questions being addressed. None of these questions can be answered simply by having a high (or low) regression R^2 or $\bar{R}^2$.

These points are summarized in Key Concept 7.4.

KEY CONCEPT 7.4

R^2 and $\bar{R}^2$: What They Tell You—and What They Don't

The R^2 and $\bar{R}^2$ tell you whether the regressors are good at predicting, or "explaining," the values of the dependent variable in the sample of data on hand. If the R^2 (or $\bar{R}^2$) is nearly 1, then the regressors produce good predictions of the dependent variable in that sample, in the sense that the variance of the OLS residual is small compared to the variance of the dependent variable. If the R^2 (or $\bar{R}^2$) is nearly 0, the opposite is true.

The R^2 and $\bar{R}^2$ do NOT tell you whether:

1. An included variable is statistically significant,
2. The regressors are a true cause of the movements in the dependent variable,
3. There is omitted variable bias, or
4. You have chosen the most appropriate set of regressors.

7.6 Analysis of the Test Score Data Set

This section presents an analysis of the effect on test scores of the student–teacher ratio using the California data set. Our primary purpose is to provide an example in which multiple regression analysis is used to mitigate omitted variable bias. Our secondary purpose is to demonstrate how to use a table to summarize regression results.

Discussion of the base and alternative specifications. This analysis focuses on estimating the effect on test scores of a change in the student–teacher ratio, holding constant student characteristics that the superintendent cannot control. Many factors potentially affect the average test score in a district. Some of these factors are correlated with the student–teacher ratio, so omitting them from the regression results in omitted variable bias. Because these factors, such as outside learning opportunities, are not directly measured, we include control variables that are correlated with these omitted factors. If the control variables are adequate in the sense that the conditional mean independence assumption holds, then the coefficient on the student–teacher ratio is the effect of a change in the student–teacher ratio, holding constant these other factors.

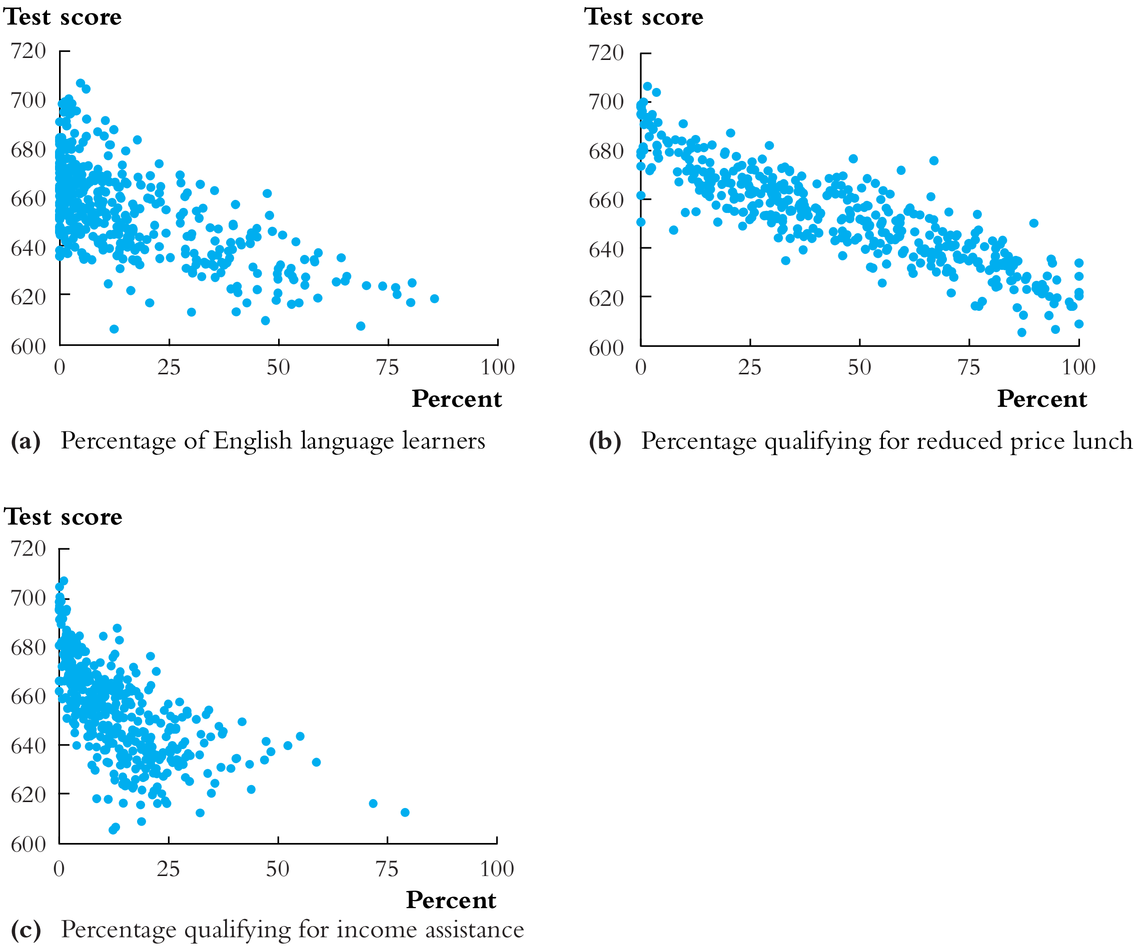
Here we consider three variables that control for background characteristics of the students that could affect test scores: the fraction of students who are still learning English, the percentage of students who are eligible for receiving a subsidized or free lunch at school, and a new variable, the percentage of students in the district whose families qualify for a California income assistance program. Eligibility for this income assistance program depends in part on family income, with a lower (stricter) threshold than the subsidized lunch program. The final two variables thus are different measures of the fraction of economically disadvantaged children in the district (their correlation coefficient is 0.74). Theory and expert judgment do not tell us which of these two variables to use to control for determinants of test scores related to economic background. For our base specification, we use the percentage eligible for a subsidized lunch, but we also consider an alternative specification that uses the fraction eligible for the income assistance program.

Scatterplots of tests scores and these variables are presented in Figure 7.2. Each of these variables exhibits a negative correlation with test scores. The correlation between test scores and the percentage of English learners is -0.64; between test scores and the percentage eligible for a subsidized lunch is -0.87; and between test scores and the percentage qualifying for income assistance is -0.63.

What scale should we use for the regressors? A practical question that arises in regression analysis is what scale you should use for the regressors. In Figure 7.2, the units of the variables are percent, so the maximum possible range of the data is 0 to 100. Alternatively, we could have defined these variables to be a *decimal fraction* rather than a percent; for example, $PctEL$ could be replaced by the *fraction* of English learners, $FracEL (= PctEL/100)$, which would range between 0 and 1 instead of between 0 and 100. More generally, in regression analysis some decision usually needs to be made about the scale of both the dependent and independent variables. How, then, should you choose the scale, or units, of the variables?

The general answer to the question of choosing the scale of the variables is to make the regression results easy to read and to interpret. In the test score application, the natural unit for the dependent variable is the score of the test itself. In the regression of $TestScore$ on STR and $PctEL$ reported in Equation (7.5), the coefficient on $PctEL$ is -0.650. If instead the regressor had been $FracEL$, the regression would have had an identical R^2 and SER; however, the coefficient on $FracEL$ would have been -65.0. In the specification with $PctEL$, the coefficient is the predicted change in test scores for a 1-percentage-point increase in English learners, holding STR constant; in the specification with $FracEL$, the coefficient is the predicted change in test scores for an increase by 1 in the fraction of English learners—that is, for a 100-percentage-point-increase—holding STR constant. Although these two specifications are mathematically equivalent, for the purposes of interpretation the one with $PctEL$ seems, to us, more natural.

FIGURE 7.2 Scatterplots of Test Scores vs. Three Student Characteristics



(a) Percentage of English language learners

(b) Percentage qualifying for reduced price lunch

(c) Percentage qualifying for income assistance

The scatterplots show a negative relationship between test scores and (a) the percentage of English learners (correlation $= -0.64$), (b) the percentage of students qualifying for a reduced price lunch (correlation $= -0.87$); and (c) the percentage qualifying for income assistance (correlation $= -0.63$).

Another consideration when deciding on a scale is to choose the units of the regressors so that the resulting regression coefficients are easy to read. For example, if a regressor is measured in dollars and has a coefficient of 0.00000356, it is easier to read if the regressor is converted to millions of dollars and the coefficient 3.56 is reported.

Tabular presentation of result. We are now faced with a communication problem. What is the best way to show the results from several multiple regressions that contain different subsets of the possible regressors? So far, we have presented

TABLE 7.1 Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28** (0.52)	−1.10* (0.43)	−1.00** (0.27)	−1.31* (0.34)	−1.01* (0.27)
Percent English learners (X_2)		−0.650** (0.031)	−0.122** (0.033)	−0.488** (0.030)	−0.130** (0.036)
Percent eligible for subsidized lunch (X_3)			−0.547* (0.024)		−0.529* (0.038)
Percent on public income assistance (X_4)				−0.790** (0.068)	0.048 (0.059)
Intercept	698.9** (10.4)	686.0** (8.7)	700.2** (5.6)	698.0** (6.9)	700.4** (5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
$\overline{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

These regressions were estimated using the data on K–8 school districts in California, described in Appendix (4.1). Heteroskedasticity-robust standard errors are given in parentheses under coefficients. The individual coefficient is statistically significant at the *5% level or **1% significance level using a two-sided test.

regression results by writing out the estimated regression equations, as in Equations (7.6) and (7.19). This works well when there are only a few regressors and only a few equations, but with more regressors and equations this method of presentation can be confusing. A better way to communicate the results of several regressions is in a table.

Table 7.1 summarizes the results of regressions of the test score on various sets of regressors. Each column summarizes a separate regression. Each regression has the same dependent variable, test score. The entries in the first five rows are the estimated regression coefficients, with their standard errors below them in parentheses. The asterisks indicate whether the t-statistics, testing the hypothesis that the relevant coefficient is zero, is significant at the 5% level (one asterisk) or the 1% level (two asterisks). The final three rows contain summary statistics for the regression (the standard error of the regression, SER, and the adjusted R^2, $\overline{R}^2$) and the sample size (which is the same for all of the regressions, 420 observations).

All the information that we have presented so far in equation format appears as a column of this table. For example, consider the regression of the test score against the student–teacher ratio, with no control variables. In equation form, this regression is

$$\underset{}{\widehat{TestScore}} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR, \ \overline{R}^2 = 0.049, SER = 18.58, n = 420. \tag{7.21}$$

All this information appears in column (1) of Table 7.1. The estimated coefficient on the student–teacher ratio (−2.28) appears in the first row of numerical entries, and its standard error (0.52) appears in parentheses just below the estimated coefficient. The intercept (698.9) and its standard error (10.4) are given in the row labeled "Intercept." (Sometimes you will see this row labeled "constant" because, as discussed in Section 6.2, the intercept can be viewed as the coefficient on a regressor that is always equal to 1.) Similarly, the $\overline{R}^2$ (0.049), the SER (18.58), and the sample size n (420) appear in the final rows. The blank entries in the rows of the other regressors indicate that those regressors are not included in this regression.

Although the table does not report t-statistics, they can be computed from the information provided; for example, the t-statistic testing the hypothesis that the coefficient on the student–teacher ratio in column (1) is zero is $-2.28/0.52 = -4.38$. This hypothesis is rejected at the 1% level, which is indicated by the double asterisk next to the estimated coefficient in the table.

Regressions that include the control variables measuring student characteristics are reported in columns (2) through (5). Column (2), which reports the regression of test scores on the student–teacher ratio and on the percentage of English learners, was previously stated as Equation (7.5).

Column (3) presents the base specification, in which the regressors are the student–teacher ratio and two control variables, the percentage of English learners and the percentage of students eligible for a free lunch.

Columns (4) and (5) present alternative specifications that examine the effect of changes in the way the economic background of the students is measured. In column (4) the percentage of students on income assistance is included as a regressor, and in column (5) both of the economic background variables are included.

Discussion of empirical results. These results suggest three conclusions:

1. Controlling for these student characteristics cuts the effect of the student–teacher ratio on test scores approximately in half. This estimated effect is not very sensitive to which specific control variables are included in the regression. In all cases the coefficient on the student–teacher ratio remains statistically significant at the 5% level. In the four specifications with control variables, regressions (2) through (5), reducing the student–teacher ratio

by one student per teacher is estimated to increase average test scores by approximately 1 point, holding constant student characteristics.

2. The student characteristic variables are potent predictors of test scores. The student–teacher ratio alone explains only a small fraction of the variation in test scores: The $\overline{R}^2$ in column (1) is 0.049. The $\overline{R}^2$ jumps, however, when the student characteristic variables are added. For example, the $\overline{R}^2$ in the base specification, regression (3), is 0.773. The signs of the coefficients on the student demographic variables are consistent with the patterns seen in Figure 7.2: Districts with many English learners and districts with many poor children have lower test scores.

3. The control variables are not always individually statistically significant: In specification (5), the hypothesis that the coefficient on the percentage qualifying for income assistance is zero is not rejected at the 5% level (the t-statistic is -0.82). Because adding this control variable to the base specification (3) has a negligible effect on the estimated coefficient for the student–teacher ratio and its standard error, and because the coefficient on this control variable is not significant in specification (5), this additional control variable is redundant, at least for the purposes of this analysis.

7.7 Conclusion

Chapter 6 began with a concern: In the regression of test scores against the student–teacher ratio, omitted student characteristics that influence test scores might be correlated with the student–teacher ratio in the district, and, if so, the student–teacher ratio in the district would pick up the effect on test scores of these omitted student characteristics. Thus the OLS estimator would have omitted variable bias. To mitigate this potential omitted variable bias, we augmented the regression by including variables that control for various student characteristics (the percentage of English learners and two measures of student economic background). Doing so cuts the estimated effect of a unit change in the student–teacher ratio in half, although it remains possible to reject the null hypothesis that the population effect on test scores, holding these control variables constant, is zero at the 5% significance level. Because they eliminate omitted variable bias arising from these student characteristics, these multiple regression estimates, hypothesis tests, and confidence intervals are much more useful for advising the superintendent than the single-regressor estimates of Chapters 4 and 5.

The analysis in this and the preceding chapter has presumed that the population regression function is linear in the regressors—that is, that the conditional expectation of Y_i given the regressors is a straight line. There is, however, no particular reason to think this is so. In fact, the effect of reducing the student–teacher ratio might be quite different in districts with large classes than in districts that already have small classes. If so, the population regression line is not linear in the X's but rather is a nonlinear function of the X's. To extend our analysis to regression functions that are nonlinear in the X's, however, we need the tools developed in the next chapter.

Summary

1. Hypothesis tests and confidence intervals for a single regression coefficient are carried out using essentially the same procedures used in the one-variable linear regression model of Chapter 5. For example, a 95% confidence interval for β_1 is given by $\hat{\beta}_1 \pm 1.96\, SE(\hat{\beta}_1)$.
2. Hypotheses involving more than one restriction on the coefficients are called joint hypotheses. Joint hypotheses can be tested using an F-statistic.
3. Regression specification proceeds by first determining a base specification chosen to address concern about omitted variable bias. The base specification can be modified by including additional regressors that address other potential sources of omitted variable bias. Simply choosing the specification with the highest R^2 can lead to regression models that do not estimate the causal effect of interest.

Key Terms

restrictions (223)
joint hypothesis (223)
F-statistic (224)
restricted regression (227)
unrestricted regression (227)
homoskedasticity-only F-statistic (228)
95% confidence set (231)
control variable (234)
conditional mean independence (235)
base specification (236)
alternative specifications (236)
Bonferroni test (251)

Review the Concepts

7.1 Explain how you would test the null hypothesis that $\beta_1 = 0$ in the multiple regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Explain how you would test the null hypothesis that $\beta_2 = 0$. Explain how you would test the joint hypothesis that $\beta_1 = 0$ and $\beta_2 = 0$. Why isn't the result of the joint test implied by the results of the first two tests?

7.2 Provide an example of a regression that arguably would have a high value of R^2 but would produce biased and inconsistent estimators of the regression coefficient(s). Explain why the R^2 is likely to be high. Explain why the OLS estimators would be biased and inconsistent.

7.3 What is a control variable, and how does it differ from a variable of interest? Looking at Table 7.1, which variables are control variables? What is the variable of interest? Do coefficients on control variables measure causal effects? Explain.

Exercises

The first six exercises refer to the table of estimated regressions on page 246, computed using data for 2012 from the CPS. The data set consists of information on 7440 full-time, full-year workers. The highest educational achievement for each worker was either a high school diploma or a bachelor's degree. The workers' ages ranged from 25 to 34 years. The data set also contains information on the region of the country where the person lived, marital status, and number of children. For the purposes of these exercises, let

AHE = average hourly earnings (in 2012 dollars)
$College$ = binary variable (1 if college, 0 if high school)
$Female$ = binary variable (1 if female, 0 if male)
Age = age (in years)
$Ntheast$ = binary variable (1 if Region = Northeast, 0 otherwise)
$Midwest$ = binary variable (1 if Region = Midwest, 0 otherwise)
$South$ = binary variable (1 if Region = South, 0 otherwise)
$West$ = binary variable (1 if Region = West, 0 otherwise)

7.1 Add * (5%) and ** (1%) to the table to indicate the statistical significance of the coefficients.

7.2 Using the regression results in column (1):

a. Is the college–high school earnings difference estimated from this regression statistically significant at the 5% level? Construct a 95% confidence interval of the difference.

b. Is the male–female earnings difference estimated from this regression statistically significant at the 5% level? Construct a 95% confidence interval for the difference.

7.3 Using the regression results in column (2):

a. Is age an important determinant of earnings? Use an appropriate statistical test and/or confidence interval to explain your answer.

b. Sally is a 29-year-old female college graduate. Betsy is a 34-year-old female college graduate. Construct a 95% confidence interval for the expected difference between their earnings.

7.4 Using the regression results in column (3):

a. Do there appear to be important regional differences? Use an appropriate hypothesis test to explain your answer.

Results of Regressions of Average Hourly Earnings on Gender and Education Binary Variables and Other Characteristics Using 2012 Data from the Current Population Survey

Dependent variable: average hourly earnings (*AHE*).

Regressor	(1)	(2)	(3)
College (X_1)	8.31 (0.23)	8.32 (0.22)	8.34 (0.22)
Female (X_2)	−3.85 (0.23)	−3.81 (0.22)	−3.80 (0.22)
Age (X_3)		0.51 (0.04)	0.52 (0.04)
Northeast (X_4)			0.18 (0.36)
Midwest (X_5)			−1.23 (0.31)
South (X_6)			−0.43 (0.30)
Intercept	17.02 (0.17)	1.87 (1.18)	2.05 (1.18)
Summary Statistics and Joint Tests			
F-statistic for regional effects $= 0$			7.38
SER	9.79	9.68	9.67
R^2	0.162	0.180	0.182
n	7440	7440	7440

b. Juanita is a 28-year-old female college graduate from the South. Molly is a 28-year-old female college graduate from the West. Jennifer is a 28-year-old female college graduate from the Midwest.

i. Construct a 95% confidence interval for the difference in expected earnings between Juanita and Molly.

ii. Explain how you would construct a 95% confidence interval for the difference in expected earnings between Juanita and Jennifer. (*Hint:* What would happen if you included *West* and excluded *Midwest* from the regression?)

7.5 The regression shown in column (2) was estimated again, this time using data from 1992 (4000 observations selected at random from the March 1993 CPS, converted into 2012 dollars using the consumer price index). The results are

$$\widehat{AHE} = \underset{(1.60)}{1.26} + \underset{(0.33)}{8.66}College - \underset{(0.29)}{4.24}Female + \underset{(0.05)}{0.65}Age,\ SER = 9.57,\ \overline{R}^2 = 0.21.$$

Comparing this regression to the regression for 2012 shown in column (2), was there a statistically significant change in the coefficient on *College*?

7.6 Evaluate the following statement: "In all of the regressions, the coefficient on *Female* is negative, large, and statistically significant. This provides strong statistical evidence of gender discrimination in the U.S. labor market."

7.7 Question 6.5 reported the following regression (where standard errors have been added):

$$\widehat{Price} = \underset{(23.9)}{119.2} + \underset{(2.61)}{0.485}BDR + \underset{(8.94)}{23.4}Bath + \underset{(0.011)}{0.156}Hsize + \underset{(0.00048)}{0.002}Lsize$$
$$+ \underset{(0.311)}{0.090}Age - \underset{(10.5)}{48.8}Poor,\ \overline{R}^2 = 0.72,\ SER = 41.5$$

a. Is the coefficient on *BDR* statistically significantly different from zero?

b. Typically five-bedroom houses sell for much more than two-bedroom houses. Is this consistent with your answer to (a) and with the regression more generally?

c. A homeowner purchases 2000 square feet from an adjacent lot. Construct a 99% confident interval for the change in the value of her house.

d. Lot size is measured in square feet. Do you think that another scale might be more appropriate? Why or why not?

e. The F-statistic for omitting BDR and Age from the regression is $F = 0.08$. Are the coefficients on BDR and Age statistically different from zero at the 10% level?

7.8 Referring to Table 7.1 in the text:

a. Construct the $\bar{R}^2$ for each of the regressions.

b. Construct the homoskedasticity-only F-statistic for testing $\beta_3 = \beta_4 = 0$ in the regression shown in column (5). Is the statistic significant at the 5% level?

c. Test $\beta_3 = \beta_4 = 0$ in the regression shown in column (5) using the Bonferroni test discussed in Appendix 7.1.

d. Construct a 99% confidence interval for β_1 for the regression in column (5).

7.9 Consider the regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Use Approach #2 from Section 7.3 to transform the regression so that you can use a t-statistic to test

a. $\beta_1 = \beta_2$.

b. $\beta_1 + 2\beta_2 = 0$.

c. $\beta_1 + \beta_2 = 1$. (*Hint:* You must redefine the dependent variable in the regression.)

7.10 Equations (7.13) and (7.14) show two formulas for the homoskedasticity-only F-statistic. Show that the two formulas are equivalent.

7.11 A school district undertakes an experiment to estimate the effect of class size on test scores in second-grade classes. The district assigns 50% of its previous year's first graders to small second-grade classes (18 students per classroom) and 50% to regular-size classes (21 students per classroom). Students new to the district are handled differently: 20% are randomly assigned to small classes and 80% to regular-size classes. At the end of the second-grade school year, each student is given a standardized exam. Let Y_i denote the exam score for the i^{th} student, X_{1i} denote a binary variable that equals 1 if the student is assigned to a small class, and X_{2i} denote a binary variable that equals 1 if the student is newly enrolled. Let β_1 denote the causal effect on test scores of reducing class size from regular to small.

a. Consider the regression $Y_i = \beta_0 + \beta_1 X_{1i} + u_i$. Do you think that $E(u_i | X_{1i}) = 0$? Is the OLS estimator of β_1 unbiased and consistent? Explain.

b. Consider the regression $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Do you think that $E(u_i | X_{1i}, X_{2i})$ depends on X_1? Is the OLS estimator of β_1 unbiased and consistent? Explain. Do you think that $E(u_i | X_{1i}, X_{2i})$ depends on X_2? Will the OLS estimator of β_2 provide an unbiased and consistent estimate of the causal effect of transferring to a new school (that is, being a newly enrolled student)? Explain.

Empirical Exercises

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

E7.1 Use the **Birthweight_Smoking** data set introduced in Empirical Exercise E5.3 to answer the following questions. To begin, run three regressions:

(1) *Birthweight* on *Smoker*

(2) *Birthweight* on *Smoker*, *Alcohol*, and *Nprevist*

(3) *Birthweight* on *Smoker*, *Alcohol*, *Nprevist*, and *Unmarried*

a. What is the value of the estimated effect of smoking on birth weight in each of the regressions?

b. Construct a 95% confidence interval for the effect of smoking on birth weight, using each of the regressions.

c. Does the coefficient on *Smoker* in regression (1) suffer from omitted variable bias? Explain.

d. Does the coefficient on *Smoker* in regression (2) suffer from omitted variable bias? Explain.

e. Consider the coefficient on *Unmarried* in regression (3).

i. Construct a 95% confidence interval for the coefficient.

ii. Is the coefficient statistically significant? Explain.

iii. Is the magnitude of the coefficient large? Explain.

iv. A family advocacy group notes that the large coefficient suggests that public policies that encourage marriage will lead, on average, to healthier babies. Do you agree? (*Hint:* Review the discussion of control variables in Section 7.5. Discuss some of the various factors that *Unmarried* may be controlling for and how this affects the interpretation of its coefficient.)

f. Consider the various other control variables in the data set. Which do you think should be included in the regression? Using a table like Table 7.1, examine the robustness of the confidence interval you constructed in (b). What is a reasonable 95% confidence interval for the effect of smoking on birth weight?

E7.2 In the empirical exercises on earning and height in Chapters 4 and 5, you estimated a relatively large and statistically significant effect of a worker's height on his or her earnings. One explanation for this result is omitted variable bias: Height is correlated with an omitted factor that affects earnings. For example, Case and Paxson (2008) suggest that cognitive ability (or intelligence) is the omitted factor. The mechanism they describe is straightforward: Poor nutrition and other harmful environmental factors in utero and in early childhood have, on average, deleterious effects on both cognitive and physical development. Cognitive ability affects earnings later in life and thus is an omitted variable in the regression.

a. Suppose that the mechanism described above is correct. Explain how this leads to omitted variable bias in the OLS regression of *Earnings* on *Height*. Does the bias lead the estimated slope to be too large or too small? [*Hint:* Review Equation (6.1).]

If the mechanism described above is correct, the estimated effect of height on earnings should disappear if a variable measuring cognitive ability is included in the regression. Unfortunately, there isn't a direct measure of cognitive ability in the data set, but the data set does include "years of education" for each individual. Because students with higher cognitive ability are more likely to attend school longer, years of education might serve as a control variable for cognitive ability; in this case, including education in the regression will eliminate, or at least attenuate, the omitted variable bias problem.

Use the years of education variable (*educ*) to construct four indicator variables for whether a worker has less than a high school diploma ($LT\_HS = 1$ if $educ < 12$, 0 otherwise), a high school diploma ($HS = 1$ if $educ = 12$, 0 otherwise), some college ($Some\_Col = 1$ if $12 < educ < 16$, 0 otherwise), or a bachelor's degree or higher ($College = 1$ if $educ \geq 16$, 0 otherwise).

b. Focusing first on women only, run a regression of (1) *Earnings* on *Height* and (2) *Earnings* on *Height*, including *LT_HS*, *HS*, and *Some_Col* as control variables.

i. Compare the estimated coefficient on *Height* in regressions (1) and (2). Is there a large change in the coefficient? Has it changed in a way consistent with the cognitive ability explanation? Explain.

ii. The regression omits the control variable *College*. Why?

iii. Test the joint null hypothesis that the coefficients on the education variables are equal to zero.

iv. Discuss the values of the estimated coefficients on *LT_HS*, *HS*, and *Some_Col*. (Each of the estimated coefficients is negative, and the coefficient on *LT_HS* is more negative than the coefficient on *HS*, which in turn is more negative than the coefficient on *Some_Col*. Why? What do the coefficients measure?)

c. Repeat (b), using data for men.

APPENDIX

7.1 The Bonferroni Test of a Joint Hypothesis

The method of Section 7.2 is the preferred way to test joint hypotheses in multiple regression. However, if the author of a study presents regression results but did not test a joint restriction in which you are interested and if you do not have the original data, then you will not be able to compute the *F*-statistic as in Section 7.2. This appendix describes a way to test joint hypotheses that can be used when you only have a table of regression results. This method is an application of a very general testing approach based on Bonferroni's inequality.

The Bonferroni test is a test of a joint hypothesis based on the *t*-statistics for the individual hypotheses; that is, the Bonferroni test is the one-at-a-time *t*-statistic test of Section 7.2 done properly. The **Bonferroni test** of the joint null hypothesis $\beta_1 = \beta_{1,0}$ and $\beta_2 = \beta_{2,0}$ based on the critical value $c > 0$, uses the following rule:

$$\text{Accept if } |t_1| \leq c \text{ and if } |t_2| \leq c\text{; otherwise, reject} \quad \text{(Bonferroni one-at-a-time } t\text{-statistic test)} \tag{7.22}$$

where t_1 and t_2 are the *t*-statistics that test the restrictions on β_1 and β_2, respectfully.

The trick is to choose the critical value c in such a way that the probability that the one-at-a-time test rejects when the null hypothesis is true is no more than the desired significance level, say 5%. This is done by using Bonferroni's inequality to choose the critical value c to allow both for the fact that two restrictions are being tested and for any possible correlation between t_1 and t_2.

Bonferroni's Inequality

Bonferroni's inequality is a basic result of probability theory. Let A and B be events. Let $A \cap B$ be the event "both A and B" (the intersection of A and B), and let $A \cup B$ be the event "A or B or both" (the union of A and B). Then $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$. Because $\Pr(A \cap B) \geq 0$, it follows that $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$.[1] Now let A be the event that $|t_1| > c$ and B be the event that $|t_2| > c$. Then the inequality $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$ yields

$$\Pr(|t_1| > c \text{ or } |t_2| > c \text{ or both}) \leq \Pr(|t_1| > c) + \Pr(|t_2| > c). \tag{7.23}$$

Bonferroni Tests

Because the event "$|t_1| > c$ or $|t_2| > c$ or both" is the rejection region of the one-at-a-time test, Equation (7.23) leads to a valid critical value for the one-at-a-time test. Under the null hypothesis in large samples, $\Pr(|t_1| > c) = \Pr(|t_2| > c) = \Pr(|Z| > c)$. Thus Equation (7.23) implies that, in large samples, the probability that the one-at-a-time test rejects under the null is

$$\Pr_{H_0}(\text{one-at-a-time test rejects}) \leq 2\Pr(|Z| > c). \tag{7.24}$$

The inequality in Equation (7.24) provides a way to choose a critical value c so that the probability of the rejection under the null hypothesis equals the desired significance level. The Bonferroni approach can be extended to more than two coefficients; if there are q restrictions under the null, the factor of 2 on the right-hand side in Equation (7.24) is replaced by q.

Table 7.2 presents critical values c for the one-at-a-time Bonferroni test for various significance levels and $q = 2, 3$, and 4. For example, suppose the desired significance level is 5% and $q = 2$. According to Table 7.2, the critical value c is 2.241. This critical value is

TABLE 7.2 Bonferroni Critical Values c for the One-at-a-Time t-Statistic Test of a Joint Hypothesis

	Significance Level		
Number of Restrictions (q)	**10%**	**5%**	**1%**
2	1.960	2.241	2.807
3	2.128	2.394	2.935
4	2.241	2.498	3.023

[1]This inequality can be used to derive other interesting inequalities. For example, it implies that $1 - \Pr(A \cup B) \geq 1 - [\Pr(A) + \Pr(B)]$. Let A^c and B^c be the complements of A and B—that is, the events "not A" and "not B." Because the complement of $A \cup B$ is $A^c \cap B^c$, $1 - \Pr(A \cup B) = \Pr(A^c \cap B^c)$, which yields Bonferroni's inequality, $\Pr(A^c \cap B^c) \geq 1 - [\Pr(A) + \Pr(B)]$.

the 1.25% percentile of the standard normal distribution, so $\Pr(|Z| > 2.241) = 2.5\%$. Thus Equation (7.24) tells us that, in large samples, the one-at-a-time test in Equation (7.22) will reject at most 5% of the time under the null hypothesis.

The critical values in Table 7.2 are larger than the critical values for testing a single restriction. For example, with $q = 2$, the one-at-a-time test rejects if at least one t-statistic exceeds 2.241 in absolute value. This critical value is greater than 1.96 because it properly corrects for the fact that, by looking at two t-statistics, you get a second chance to reject the joint null hypothesis, as discussed in Section 7.2.

If the individual t-statistics are based on heteroskedasticity-robust standard errors, then the Bonferroni test is valid whether or not there is heteroskedasticity, but if the t-statistics are based on homoskedasticity-only standard errors, the Bonferroni test is valid only under homoskedasticity.

Application to Test Scores

The t-statistics testing the joint null hypothesis that the true coefficients on test scores and expenditures per pupil in Equation (7.6) are, respectively, $t_1 = -0.60$ and $t_2 = 2.43$. Although $|t_1| < 2.241$, because $|t_2| > 2.241$, we can reject the joint null hypothesis at the 5% significance level using the Bonferroni test. However, both t_1 and t_2 are less than 2.807 in absolute value, so we cannot reject the joint null hypothesis at the 1% significance level using the Bonferroni test. In contrast, using the F-statistic in Section 7.2, we were able to reject this hypothesis at the 1% significance level.

APPENDIX

7.2 Conditional Mean Independence

This appendix shows that, under the assumption of conditional mean independence introduced in Section 7.5 [Equation (7.20)], the OLS coefficient estimator is unbiased for the variable of interest but not for the control variable.

Consider a regression with two regressors, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. If $E(u_i | X_{1i}, X_{2i}) = 0$, as would be true if X_{1i} and X_{2i} are randomly assigned in an experiment, then the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased estimators of the causal effects β_1 and β_2.

Now suppose that X_{1i} is the variable of interest and X_{2i} is a control variable that is correlated with omitted factors in the error term. Although the conditional mean zero assumption does not hold, suppose that conditional mean independence does so that $E(u_i | X_{1i}, X_{2i}) = E(u_i | X_{2i})$. For convenience, further suppose that $E(u_i | X_{2i})$ is linear in X_{2i} so that $E(u_i | X_{2i}) = \gamma_0 + \gamma_2 X_{2i}$, where γ_0 and γ_1 are constants. (This linearity assumption is discussed below.) Define v_i to be the difference between u_i and the conditional expectation of

u_i given X_{1i} and X_{2i}—that is, $v_i = u_i - E(u_i|X_{1i}, X_{2i})$—so that v_i has conditional mean zero: $E(v_i|X_{1i}, X_{2i}) = E[u_i - E(u_i|X_{1i}, X_{2i})|X_{1i}, X_{2i}] = E(u_i|X_{1i}, X_{2i}) - E(u_i|X_{1i}, X_{2i}) = 0$. Thus,

$$\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \\
&= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + E(u_i \mid X_{1i}, X_{2i}) + v_i \quad \text{(using the definition of } v_i\text{)} \\
&= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + E(u_i \mid X_{2i}) + v_i \quad \text{(using conditional mean independence)} \\
&= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + (\gamma_0 + \gamma_2 X_{2i}) + v_i \quad [\text{using linearity of } E(u_i \mid X_{2i})] \\
&= (\beta_0 + \gamma_0) + \beta_1 X_{1i} + (\beta_2 + \gamma_2) X_{2i} + v_i \quad \text{(collecting terms)} \\
&= \delta_0 + \beta_1 X_{1i} + \delta_2 X_{2i} + v_i,
\end{aligned} \tag{7.25}$$

where $\delta_0 = \beta_0 + \gamma_0$ and $\delta_2 = \beta_2 + \gamma_2$.

The error v_i in Equation (7.25) has conditional mean zero; that is, $E(v_i|X_{1i}, X_{2i}) = 0$. Therefore, the first least squares assumption for multiple regression applies to the final line of Equation (7.25), and if the other three least squares assumptions for multiple regression also hold, then the OLS regression of Y_i on a constant, X_{1i}, and X_{2i} will yield unbiased and consistent estimators of δ_0, β_1, and δ_2. Thus the OLS estimator of the coefficient on X_{1i} is unbiased for the causal effect β_1. However, the OLS estimator of the coefficient on X_{2i} is *not* unbiased for β_2 and instead estimates the sum of the causal effect β_2 and the coefficient γ_2 arising from the correlation of the control variable X_{2i} with the original error term u_i.

The derivation in Equation (7.25) works for any value of β_2, including zero. A variable X_{2i} is a useful control variable if conditional mean independence holds; it need not have a direct causal effect on Y_i.

The fourth line in Equation (7.25) uses the assumption that $E(u_i|X_{2i})$ is linear in X_{2i}. As discussed in Section 2.4, this will be true if u_i and X_{2i} are jointly normally distributed. The assumption of linearity can be relaxed using methods discussed in Chapter 8. Exercise 18.9 works through the steps in Equation (7.25) for multiple variables of interest and multiple control variables.

In terms of the example in Section 7.5 [the regression in Equation (7.19)], if X_{2i} is *LchPct*, then β_2 is the causal effect of the subsidized lunch program (β_2 is positive if the program's nutritional benefits improve test scores), γ_2 is negative because *LchPct* is negatively correlated with (controls for) omitted learning advantages that improve test scores, and $\delta_2 = \beta_2 + \gamma_2$ would be negative if the omitted variable bias contribution through γ_2 outweights the positive causal effect β_2.

To better understand the conditional mean independence assumption, return to the concept of an ideal randomized controlled experiment. As discussed in Section 4.4, if X_{1i} is randomly assigned, then in a regression of Y_i on X_{1i}, the conditional mean zero assumption holds. If, however, X_{1i} is randomly assigned, conditional on another variable X_{2i}, then the conditional mean independence assumption holds, but if X_{2i} is correlated with u_i, the conditional mean zero assumption does not. For example, consider an experiment to study the effect on grades in econometrics of mandatory versus optional homework. Among economics majors ($X_{2i} = 1$), 75% are assigned to the treatment group (mandatory homework: $X_{1i} = 1$), while among noneconomics majors ($X_{2i} = 0$), only 25% are assigned to the treatment group. Because treatment is randomly assigned within majors and within nonmajors, u_i is independent of X_{1i}, given X_{2i}, so in particular, $E(u_i | X_{1i}, X_{2i}) = E(u_i | X_{2i})$. If choice of major is correlated with other characteristics (like prior math) that determine performance in an econometrics course, then $E(u_i | X_{2i}) \neq 0$, and the regression of the final exam grade (Y_i) on X_{1i} alone will be subject to omitted variable bias (X_{1i} is correlated with major and thus with other omitted determinants of grade). Including major (X_{2i}) in the regression eliminates this omitted variable bias (treatment is randomly assigned, given major), making the OLS estimator of the coefficient on X_{1i} an unbiased estimator of the causal effect on econometrics grades of requiring homework. However, the OLS estimator of the coefficient on major is not unbiased for the causal effect of switching into economics because major is not randomly assigned and is correlated with other omitted factors that would not change (like prior math) were a student to switch majors.