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

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

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

5.3 Regression When X Is a Binary Variable

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

Interpretation of the Regression Coefficients

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

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

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

The population regression model with D_i as the regressor is

$$Y_i = \beta_0 + \beta_1 D_i + u_i, i = 1, \ldots, n. \quad (5.15)$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.4 Heteroskedasticity and Homoskedasticity

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

CHAPTER 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

KEY CONCEPT 6.1

Omitted Variable Bias in Regression with a Single Regressor

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.

term, and if it is correlated with X_i, then the error term is correlated with X_i. 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.

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 $\Delta Y = \beta_1 \Delta X_1$. That is,

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

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

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

KEY CONCEPT 6.3

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

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

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

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

$$\widehat{TestScore} = 698.9 - 2.28 \times STR. \tag{6.11}$$

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 to 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.