CHAPTER 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

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

KEY CONCEPT 7.1

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

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.

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 desired significance level of the test. The theoretical underpinning of this procedure is that the OLS estimator has a large-sample normal distribution which, 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

CONFIDENCE INTERVALS FOR A SINGLE COEFFICIENT IN MULTIPLE REGRESSION

KEY CONCEPT 7.2

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

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

A 90% confidence interval is obtained by replacing 1.96 in Equation (7.4) with 1.645.

denoted t^{act} in this Key Concept. 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.

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:

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

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 help 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

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

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 ($|-0.60| < 1.645$). 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

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

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 2nd, 4th, and 5th 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 $|t_1| \le 1.96$ and $|t_2| \le 1.96$. Because the t-statistics are independent, $\Pr(|t_1| \le 1.96$ and $|t_2| \le 1.96) = \Pr(|t_1| \le 1.96) \times \Pr(|t_2| \le 1.96) = 0.95^2 = 0.9025 = 90.25\%$. 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

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

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. That is, 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 sum of squared residuals or by 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 rule-of-thumb F-statistic under the null hypothesis is, in large samples, $F_{q,\infty}$.

These rule-of-thumb formulas are easy to compute and have an intuitive interpretation in terms of how well the unrestricted and restricted regressions fit the data. Unfortunately, they are valid 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 *SSR* (or R^2) 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

$$F = [(0.4366 - 0.4149)/2]/[(1 - 0.4366)/(420 - 3 - 1)] = 8.01.$$

Because 8.01 exceeds the 1% critical value of 4.61, the hypothesis is rejected at the 1% level using this rule-of-thumb approach.

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 rule-of-thumb 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 one will be easiest 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 $\hat{\gamma}_1 \pm 1.96SE(\hat{\gamma}_1)$.

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.

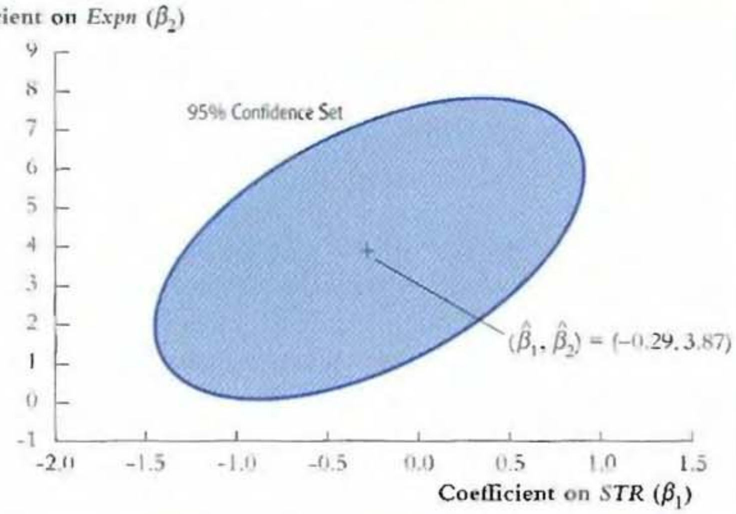
Recall that a 95% confidence interval is computed by finding the set of values of the coefficients that are not rejected using a t-statistic at the 5% significance level. This approach can be extended to the case of multiple coefficients. To make this concrete, suppose you are interested in constructing 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 construct 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 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.

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 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

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 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 affluence of the students 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 average district income, even after controlling for the percentage of English learners. In short, omitting 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 estimators will have omitted variable bias. As was discussed in Section 6.6, the OLS estimators are correlated, so in general the OLS estimators of all the coefficients will be biased. 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 that 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.

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.

Omitted Variable Bias in Multiple Regression

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.**

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 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.

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.

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 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 you have the most appropriate set of regressors, nor does a low R^2 or $\bar{R}^2$ necessarily mean 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.

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 the factors that could affect test scores are correlated with the student–teacher ratio, so omitting them from the regression will result in omitted variable bias. If data are available on these omitted variables, the solution to this problem is to include them as additional regressors in the multiple regression. When we do this, 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. One of these control variables is the one we have used previously, the fraction of students who are still learning English. The two other variables are new and control for the economic background of the students. There is no perfect measure of economic background in the data set, so instead we use two imperfect indicators of low income in the district. The first new variable is the percentage of students who are eligible for receiving a subsidized or free lunch at school. Students are eligible for this program if their family income is less than a certain threshold (approximately 150% of the poverty line). The second new variable is the percentage of students in the district whose families qualify for a California income assistance program. Families are eligible for this income assistance program depending in part on their family income, but the threshold is lower (stricter) than the threshold for the subsidized lunch program. These two variables thus measure the fraction of economically disadvantaged children in the district; although they are related, they are not perfectly correlated (their correlation coefficient is 0.74). Although theory suggests that economic background could be an important omitted factor, theory and expert judgment do not really help us decide which of these two variables (percentage eligible for a subsidized lunch or percentage eligible for income assistance) is a better measure of background. For our base specification, we choose the percentage eligible for a subsidized lunch as the economic background variable, but we consider an alternative specification that includes the other variable as well.

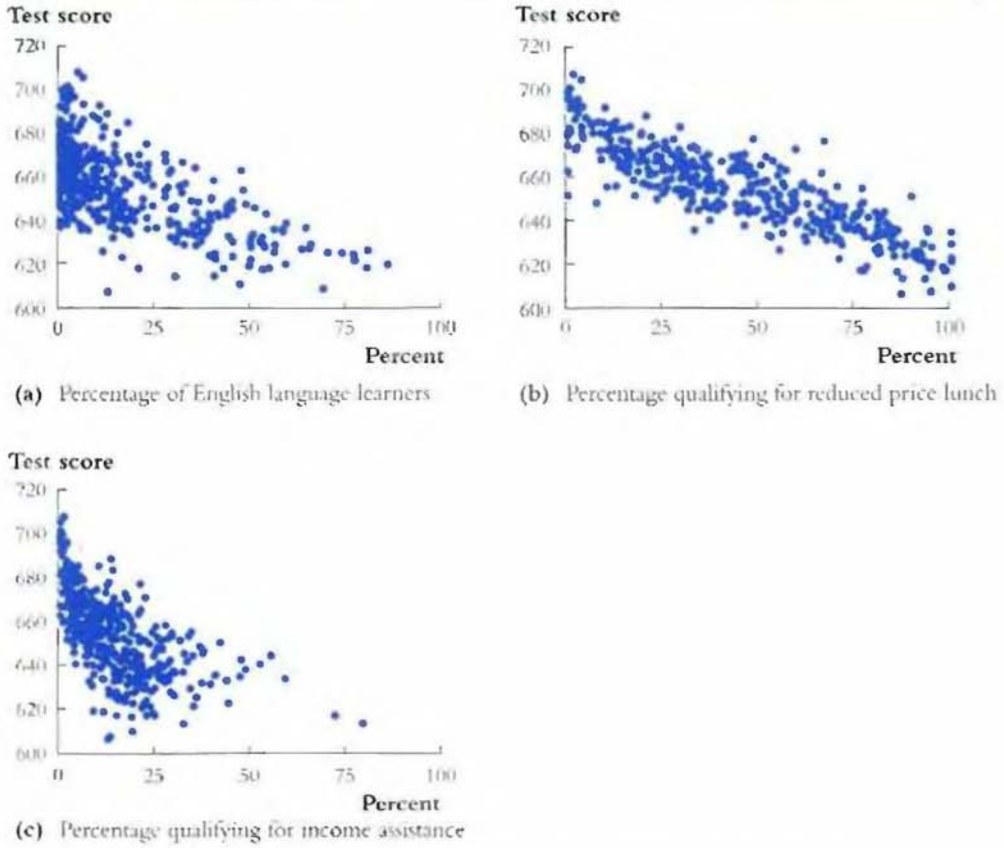
Scatterplots of tests scores and these variables are presented in Figure 7.2. Each of these variables exhibits a negative correlation with test scores. The correlation between test scores and the percentage of English learners is −0.64; between test scores and the percentage eligible for a subsidized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

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.

FIGURE 7.2 Scatterplots of Test Scores vs. Three Student Characteristics



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 subsidized lunch (correlation = −0.87); and (c) the percentage qualifying for income assistance (correlation = −0.63).

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 regression results by writing out the estimated regression equations, as in Equation (7.6). 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

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
$\bar{R}^2$	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. 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.

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, $\bar{R}^2$) 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

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR,\ \bar{R}^2 = 0.049, SER = 18.58, n = 420. \qquad (7.19)$$

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, these 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)–(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)–(5), reducing the student–teacher ratio by one student per teacher is estimated to increase average test scores by approximately one point, holding constant student characteristics.

2. The student characteristic variables are very useful 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 varible 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

Summary

1. Hypothesis tests and confidence intervals for a single regression coefficient are carried out using essentially the same procedures that were used in the one-variable linear regression model of Chapter 5. For example, a 95% confidence interval for β_1 is given by $\hat{\beta}_1 \pm 1.96SE(\hat{\beta}_1)$.
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 (226)
joint hypothesis (226)
F-statistic (227)
restricted regression (230)
unrestricted regression (230)
homoskedasticity-only F-statistic (231)
95% confidence set (234)
base specification (236)
alternative specifications (237)
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.

Exercises

The first six exercises refer to the table of estimated regressions on page 247, computed using data for 1998 from the CPS. The data set consists of information on 4000 full-time full-year workers. The highest educational achievement for each worker was either a high school diploma or a bachelor's degree. The worker's ages ranged from 25 to 34 years. The data set also contained 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 1998 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 1998 Data from the Current Population Survey

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

Regressor	(1)	(2)	(3)
College (X_1)	5.46 (0.21)	5.48 (0.21)	5.44 (0.21)
Female (X_2)	−2.64 (0.20)	−2.62 (0.20)	−2.62 (0.20)
Age (X_3)		0.29 (0.04)	0.29 (0.04)
Northeast (X_4)			0.69 (0.30)
Midwest (X_5)			0.60 (0.28)
South (X_6)			−0.27 (0.26)
Intercept	12.69 (0.14)	4.40 (1.05)	3.75 (1.06)
Summary Statistics and Joint Tests			
F-statistic for regional effects = 0			6.10
SER	6.27	6.22	6.21
$\bar{R}^2$	0.176	0.190	0.194
n	4000	4000	4000

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 1998 dollars using the consumer price index). The results are

$$\widehat{AHE} = \underset{(0.98)}{0.77} + \underset{(0.20)}{5.29}College - \underset{(0.18)}{2.59}Female + \underset{(0.03)}{0.40}Age,\ SER = 5.85,\ \bar{R}^2 = 0.21.$$

Comparing this regression to the regression for 1998 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):

$$\widehat{Price} = \underset{(23.9)}{119.2} + \underset{(2.61)}{0.485}BDR + \underset{(8.94)}{23.4}Bath + \underset{(0.011)}{0.156}Hsize + \underset{(0.00048)}{0.002}Lsize$$

$$+ \underset{(0.311)}{0.090}Age - \underset{(10.5)}{48.8}Poor,\ \bar{R}^2 = 0.72,\ SER = 41.5$$

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 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. $\beta_1 + a\beta_2 = 0$, where a is a constant;

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.

Empirical Exercises

E7.1 Use the data set **CPS04** described in Empirical Exercise 4.1 to answer the following questions.

a. Run a regression of average hourly earnings (*AHE*) on age (*Age*). What is the estimated intercept? What is the estimated slope?

b. Run a regression of *AHE* on *Age*, gender (*Female*), and education (*Bachelor*). What is the estimated effect of *Age* on earnings? Construct a 95% confidence interval for the coefficient on *Age* in the regression.

c. Are the results from the regression in (b) substantively different from the results in (a) regarding the effects of *Age* and *AHE*? Does the regression in (a) seem to suffer from omitted variable bias?

d. Bob is a 26-year-old male worker with a high school diploma. Predict Bob's earnings using the estimated regression in (b). Alexis is a 30-year-old female worker with a college degree. Predict Alexis's earnings using the regression.

e. Compare the fit of the regression in (a) and (b) using the regression standard errors, R^2 and $\bar{R}^2$. Why are the R^2 and $\bar{R}^2$ so similar in regression (b)?

f. Are gender and education determinants of earnings? Test the null hypothesis that *Female* can be deleted from the regression. Test the null hypothesis that *Bachelor* can be deleted from the regression. Test the null hypothesis that both *Female* and *Bachelor* can be deleted from the regression.

g. A regression will suffer from omitted variable bias when two conditions hold. What are these two conditions? Do these conditions seem to hold here?

E7.2 Using the data set **TeachingRatings** described in Empirical Exercise 4.2, carry out the following exercises.

a. Run a regression of *Course_Eval* on *Beauty*. Construct a 95% confidence interval for the effect of *Beauty* on *Course_Eval*.

b. Consider the various 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 that you constructed in (a). What is a reasonable 95% confidence interval for the effect of *Beauty* on *Course_Eval*?

E7.3 Use the data set **CollegeDistance** described in Empirical Exercise 4.3 to answer the following questions.

a. An education advocacy group argues that, on average, a person's educational attainment would increase by approximately 0.15 year if distance to the nearest college is decreased by 20 miles. Run a regression of years of completed education (*ED*) on distance to the nearest college (*Dist*). Is the advocacy groups' claim consistent with the estimated regression? Explain.

b. Other factors also affect how much college a person completes. Does controlling for these other factors change the estimated effect of distance on college years completed? To answer this question, construct a table like Table 7.1. Include a simple specification [constructed in (a)], a base specification (that includes a set of important control variables), and several modifications of the base specification. Discuss how the estimated effect of *Dist* on *ED* changes across the specifications.

c. It has been argued that, controlling for other factors, blacks and Hispanics complete more college than whites. Is this result consistent with the regressions that you constructed in part (b)?

E7.4 Using the data set **Growth** described in Empirical Exercise 4.4, but excluding the data for Malta, carry out the following exercises.

a. Run a regression of *Growth* on *TradeShare*, *YearsSchool*, *Rev_Coups*, *Assassinations* and *RGDP60*. Construct a 95% confidence interval for the coefficient on *TradeShare*. Is the coefficient statistically significant at the 5% level?

b. Test whether, taken as a group, *YearsSchool*, *Rev_Coups*, *Assassinations*, and *RGDP60* can be omitted from the regression. What is the p-value of the F-statistic?

APPENDIX 7.1 The Bonferroni Test of a Joint Hypotheses

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 you do not have the original data, then you will not be able to compute the F-statistic of 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 hypotheses 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:

$$\text{Accept if } |t_1| \le c \text{ and if } |t_2| \le c\text{; otherwise, reject} \quad \text{(Bonferroni one-at-a-time } t\text{-statistic test)}, \tag{7.20}$$

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 $\Pr(A \cap B) \ge 0$, it follows that $\Pr(A \cup B) \le \Pr(A) + \Pr(B)$. This inequality in turn implies that $1 - \Pr(A \cup B) \ge 1 - [\Pr(A) + \Pr(B)]$. 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, $\Pr(A^c \cap B^c) \geq 1 - [\Pr(A) + \Pr(B)]$.

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

$$\Pr(|t_1| > c \text{ or } |t_2| > c \text{ or both}) \leq \Pr(|t_1| > c) + \Pr(|t_2| > c). \qquad (7.21)$$

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 (7.21) provides a way to choose the critical value c so that the "one at a time" t-statistic has the desired significance level in large samples. Under the null hypothesis in large samples, $\Pr(|t_1| > c) = \Pr(|t_2| > c) = \Pr(|Z| > c)$. Thus Equation (7.21) implies that, in large samples, the probability that the one-at-a-time test rejects under the null is

$$\Pr_{H_0}(\text{one-at-a-time test rejects}) \leq 2\Pr(|Z| > c). \qquad (7.22)$$

The inequality in Equation (7.22) provides a way to choose 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 (7.22) is replaced by q.

Table 7.3 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 7.3, the critical value c is 2.241. This critical value is the 1.25% percentile of the standard normal distribution, so $\Pr(|Z| > 2.241) = 2.5\%$. Thus Equation (7.22) tells us that, in large samples, the one-at-a-time test in Equation (7.20) will reject at most 5% of the time under the null hypothesis.

The critical values in Table 7.3 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.

TABLE 7.3 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

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.