5
HYPOTHESIS TESTS AND MODEL SELECTION

5.1 INTRODUCTION

The linear regression model is used for three major purposes: estimation and prediction, which were the subjects of the previous chapter, and hypothesis testing. In this chapter,we will examine some applications of hypothesis tests using the linear regression model. We begin with the methodological and statistical theory. Some of this theory was developed in Chapter 4 (including the idea of a pivotal statistic in Section 4.5.1) and in Appendix C.7. In Section 5.2, we will extend the methodology to hypothesis testing based on the regression model. After the theory is developed, Sections 5.3–5.7 will examine some applications in regression modeling. This development will be concerned with the implications of restrictions on the parameters of the model, such as whether a variable is 'relevant' (i.e., has a nonzero coefficient) or whether the regression model itself is supported by the data (i.e., whether the data seem consistent with the hypothesis that all of the coefficients are zero). We will primarily be concerned with linear restrictions in this discussion. We will turn to nonlinear restrictions near the end of the development in Section 5.7. Section 5.8 considers some broader types of hypotheses, such as choosing between two competing models, such as whether a linear or a loglinear model is better suited to the data. In each of the cases so far, the testing procedure attempts to resolve a competition between two theories for the data; in Sections 5.2–5.7 between a narrow model and a broader one and in Section 5.8, between two arguably equal models. Section 5.9 illustrates a particular **specification test**, which is essentially a test of a proposition such as "the model is correct" vs. "the model is inadequate." This test pits the theory of the model against "some other unstated theory." Finally, Section 5.10 presents some general principles and elements of a strategy of model testing and selection.

5.2 HYPOTHESIS TESTING METHODOLOGY

We begin the analysis with the regression model as a statement of a proposition,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{5-1}$$

To consider a specific application, Example 4.6 depicted the auction prices of paintings

$$\ln Price = \beta_1 + \beta_2 \ln Size + \beta_3 AspectRatio + \varepsilon. \tag{5-2}$$

Some questions might be raised about the "model" in (5-2), fundamentally, about the variables. It seems natural that fine art enthusiasts would be concerned about aspect ratio, which is an element of the aesthetic quality of a painting. But, the idea that size should

be an element of the price is counterintuitive, particularly weighed against the surprisingly small sizes of some of the world's most iconic paintings such as the *Mona Lisa* (30″ high and 21″ wide) or Dali's *Persistence of Memory* (only 9.5″ high and 13″ wide). A skeptic might question the presence of ln*Size* in the equation, or, equivalently, the nonzero coefficient, β_2. To settle the issue, the relevant empirical question is whether the equation specified appears to be consistent with the data—that is, the observed sale prices of paintings. In order to proceed, the obvious approach for the analyst would be to fit the regression first and then examine the estimate of β_2. The "test" at this point, is whether b_2 in the least squares regression is zero or not. Recognizing that the least squares slope is a random variable that will never be exactly zero even if β_2 really is, we would soften the question to be whether the sample estimate seems to be close enough to zero for us to conclude that its population counterpart is actually zero, that is, that the nonzero value we observe is nothing more than noise that is due to sampling variability. Remaining to be answered are questions including; How close to zero is close enough to reach this conclusion? What metric is to be used? How certain can we be that we have reached the right conclusion? (Not absolutely, of course.) How likely is it that our decision rule, whatever we choose, will lead us to the wrong conclusion? This section will formalize these ideas. After developing the methodology in detail, we will construct a number of numerical examples.

5.2.1 RESTRICTIONS AND HYPOTHESES

The approach we will take is to formulate a hypothesis as a restriction on a model. Thus, in the classical methodology considered here, the model is a general statement and a hypothesis is a proposition that narrows that statement. In the art example in (5-2), while the narrower statement is (5-2) with the additional statement that $\beta_2 = 0$—without comment on β_1 or β_3. We define the **null hypothesis** as the statement that narrows the model and the **alternative hypothesis** as the broader one. In the example, the broader model allows the equation to contain both ln*Size* and *AspectRatio*—it admits the possibility that either coefficient might be zero but does not insist upon it. The null hypothesis insists that $\beta_2 = 0$ while it also makes no comment about β_1 or β_3. The formal notation used to frame this hypothesis would be

$$\begin{aligned} &\ln Price = \beta_1 + \beta_2 \ln Size + \beta_3 AspectRatio + \varepsilon, \\ &H_0: \beta_2 = 0, \\ &H_1: \beta_2 \neq 0. \end{aligned} \qquad \textbf{(5-3)}$$

Note that the null and alternative hypotheses, together, are exclusive and exhaustive. There is no third possibility; either one or the other of them is true, not both.

The analysis from this point on will be to measure the null hypothesis against the data. The data might persuade the econometrician to reject the null hypothesis. It would seem appropriate at that point to "accept" the alternative. However, in the interest of maintaining flexibility in the methodology, that is, an openness to new information, the appropriate conclusion here will be either to reject the null hypothesis or not to reject it. Not rejecting the null hypothesis is not equivalent to "accepting" it—though the language might suggest so. By accepting the null hypothesis, we would implicitly be closing off further investigation. Thus, the traditional, classical methodology leaves open the possibility that further evidence might still change the conclusion. Our testing

methodology will be constructed so as either to

Reject H_0: The data are inconsistent with the hypothesis with a reasonable degree of certainty.

Do not reject H_0: The data appear to be consistent with the null hypothesis.

5.2.2 NESTED MODELS

The general approach to testing a hypothesis is to formulate a statistical model that contains the hypothesis as a restriction on its parameters. A theory is said to have **testable implications** if it implies some testable restrictions on the model. Consider, for example, a model of investment, I_t,

$$\ln I_t = \beta_1 + \beta_2 i_t + \beta_3 \Delta p_t + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t, \quad \textbf{(5-4)}$$

which states that investors are sensitive to nominal interest rates, i_t, the rate of inflation, Δp_t, (the log of) real output, $\ln Y_t$, and other factors that trend upward through time, embodied in the time trend, t. An alternative theory states that "investors care about real interest rates." The alternative model is

$$\ln I_t = \beta_1 + \beta_2 (i_t - \Delta p_t) + \beta_3 \Delta p_t + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t. \quad \textbf{(5-5)}$$

Although this new model does embody the theory, the equation still contains both nominal interest and inflation. The theory has no testable implication for our model. But, consider the stronger hypothesis, "investors care *only* about real interest rates." The resulting equation,

$$\ln I_t = \beta_1 + \beta_2 (i_t - \Delta p_t) + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t, \quad \textbf{(5-6)}$$

is now restricted; in the context of (5-4), the implication is that $\beta_2 + \beta_3 = 0$. The stronger statement implies something specific about the parameters in the equation that may or may not be supported by the empirical evidence.

The description of testable implications in the preceding paragraph suggests (correctly) that testable restrictions will imply that only some of the possible models contained in the original specification will be "valid"; that is, consistent with the theory. In the example given earlier, (5-4) specifies a model in which there are five unrestricted parameters $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$. But, (5-6) shows that only some values are consistent with the theory, that is, those for which $\beta_3 = -\beta_2$. This subset of values is contained within the unrestricted set. In this way, the models are said to be **nested.** Consider a different hypothesis, "investors do not care about inflation." In this case, the smaller set of coefficients is $(\beta_1, \beta_2, 0, \beta_4, \beta_5)$. Once again, the restrictions imply a valid **parameter space** that is "smaller" (has fewer dimensions) than the unrestricted one. The general result is that the hypothesis specified by the restricted model is contained within the unrestricted model.

Now, consider an alternative pair of models: Model_0: "Investors care only about inflation"; Model_1: "Investors care only about the nominal interest rate." In this case, the two parameter vectors are $(\beta_1, 0, \beta_3, \beta_4, \beta_5)$ by Model_0 and $(\beta_1, \beta_2, 0, \beta_4, \beta_5)$ by Model_1. In this case, the two specifications are both subsets of the unrestricted model, but neither model is obtained as a restriction on the other.They have the same number of parameters; they just contain different variables. These two models are **nonnested.** For the present, we are concerned only with nested models. Nonnested models are considered in Section 5.8.

5.2.3 TESTING PROCEDURES—NEYMAN–PEARSON METHODOLOGY

In the example in (5-2), intuition suggests a testing approach based on measuring the data against the hypothesis. The essential methodology suggested by the work of Neyman and Pearson (1933) provides a reliable guide to testing hypotheses in the setting we are considering in this chapter. Broadly, the analyst follows the logic, "What type of data will lead me to reject the hypothesis?" Given the way the hypothesis is posed in Section 5.2.1, the question is equivalent to asking what sorts of data will support the model. The data that one can observe are divided into a **rejection region** and an **acceptance region**. The testing procedure will then be reduced to a simple up or down examination of the statistical evidence. Once it is determined what the rejection region is, if the observed data appear in that region, the null hypothesis is rejected. To see how this operates in practice, consider, once again, the hypothesis about size in the art price equation. Our test is of the hypothesis that β_2 equals zero. We will compute the least squares slope. We will decide in advance how far the estimate of β_2 must be from zero to lead to rejection of the null hypothesis. Once the rule is laid out, the test, itself, is mechanical. In particular, for this case, b_2 is "far" from zero if $b_2 > \beta_2^{0+}$ or $b_2 < \beta_2^{0-}$. If either case occurs, the hypothesis is rejected. The crucial element is that the rule is decided upon in advance.

5.2.4 SIZE, POWER, AND CONSISTENCY OF A TEST

Since the testing procedure is determined in advance and the estimated coefficient(s) in the regression are random, there are two ways the Neyman–Pearson method can make an error. To put this in a numerical context, the sample regression corresponding to (5-2) appears in Table 4.6. The estimate of the coefficient on ln*Area* is 1.33372 with an estimated standard error of 0.09072. Suppose the rule to be used to test is decided arbitrarily (at this point—we will formalize it shortly) to be: If b_2 is greater than +1.0 or less than −1.0, then we will reject the hypothesis that the coefficient is zero (and conclude that art buyers really do care about the sizes of paintings). So, based on this rule, we will, in fact, reject the hypothesis. However, since b_2 is a random variable, there are the following possible errors:

Type I error: $\beta_2 = 0$, but we reject the hypothesis.
The null hypothesis is incorrectly rejected.
Type II error: $\beta_2 \neq 0$, but we do not reject the hypothesis.
The null hypothesis is incorrectly retained.

The probability of a Type I error is called the **size of the test**. The size of a test is the probability that the test will incorrectly reject the null hypothesis. As will emerge later, the analyst determines this in advance. One minus the probability of a Type II error is called the **power of a test**. The power of a test is the probability that it will correctly reject a false null hypothesis. The power of a test depends on the alternative. It is not under the control of the analyst. To consider the example once again, we are going to reject the hypothesis if $|b_2| > 1$. If β_2 is actually 1.5, based on the results we've seen, we are quite likely to find a value of b_2 that is greater than 1.0. On the other hand, if β_2 is only 0.3, then it does not appear likely that we will observe a sample value greater than 1.0. Thus, again, the power of a test depends on the actual parameters that underlie the data. The idea of power of a test relates to its ability to find what it is looking for.

A test procedure is **consistent** if its power goes to 1.0 as the sample size grows to infinity. This quality is easy to see, again, in the context of a single parameter, such as the one being considered here. Since least squares is consistent, it follows that as the sample size grows, we will be able to learn the exact value of β_2, so we will know if it is zero or not. Thus, for this example, it is clear that as the sample size grows, we will know with certainty if we should reject the hypothesis. For most of our work in this text, we can use the following guide: A testing procedure about the parameters in a model is consistent if it is based on a consistent estimator of those parameters. Since nearly all our work in this book is based on consistent estimators and save for the latter sections of this chapter, where our tests will be about the parameters in nested models, our tests will be consistent.

5.2.5 A METHODOLOGICAL DILEMMA: BAYESIAN VS. CLASSICAL TESTING

As we noted earlier, the Neyman–Pearson testing methodology we will employ here is an all-or-nothing proposition. We will determine the testing rule(s) in advance, gather the data, and either reject or not reject the null hypothesis. There is no middle ground. This presents the researcher with two uncomfortable dilemmas. First, the testing outcome, that is, the sample data might be uncomfortably close to the boundary of the rejection region. Consider our example. If we have decided in advance to reject the null hypothesis if $b_2 > 1.00$, and the sample value is 0.9999, it will be difficult to resist the urge to reject the null hypothesis anyway, particularly if we entered the analysis with a strongly held belief that the null hypothesis is incorrect. (I.e., intuition notwithstanding, I am convinced that art buyers really do care about size.) Second, the methodology we have laid out here has no way of incorporating other studies. To continue our example, if I were the tenth analyst to study the art market, and the previous nine had decisively rejected the hypothesis that $\beta_2 = 0$, I will find it very difficult not to reject that hypothesis even if my evidence suggests, based on my testing procedure, that I should.

This dilemma is built into the classical testing methodology. There is a middle ground. The Bayesian methodology that we will discuss in Chapter 15 does not face this dilemma because Bayesian analysts never reache a firm conclusion. They merely update their priors. Thus, the first case noted, in which the observed data are close to the boundary of the rejection region, the analyst will merely be updating the prior with somethat slightly less persuasive evidence than might be hoped for. But, the methodology is comfortable with this. For the second instance, we have a case in which there is a wealth of prior evidence in favor of rejecting H_0. It will take a powerful tenth body of evidence to overturn the previous nine conclusions. The results of the tenth study (the posterior results) will incorporate not only the current evidence, but the wealth of prior data as well.

5.3 TWO APPROACHES TO TESTING HYPOTHESES

The **general linear hypothesis** is a set of J restrictions on the linear regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

The restrictions are written

$$\begin{aligned} r_{11}\beta_1 + r_{12}\beta_2 + \cdots + r_{1K}\beta_K &= q_1 \\ r_{21}\beta_1 + r_{22}\beta_2 + \cdots + r_{2K}\beta_K &= q_2 \\ &\cdots \\ r_{J1}\beta_1 + r_{J2}\beta_2 + \cdots + r_{JK}\beta_K &= q_J. \end{aligned} \tag{5-7}$$

The simplest case is a single restriction on one coefficient, such as

$$\beta_k = 0.$$

The more general case can be written in the matrix form,

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{q}. \tag{5-8}$$

Each row of $\mathbf{R}$ is the coefficients in one of the restrictions. Typically, $\mathbf{R}$ will have only a few rows and numerous zeros in each row. Some examples would be as follows:

1. One of the coefficients is zero, $\beta_j = 0$,

$$\mathbf{R} = [0 \quad 0 \quad \cdots \quad 1 \quad 0 \quad \cdots \quad 0] \text{ and } \mathbf{q} = 0.$$

2. Two of the coefficients are equal, $\beta_k = \beta_j$,

$$\mathbf{R} = [0 \quad 0 \quad 1 \quad \cdots \quad -1 \quad \cdots \quad 0] \text{ and } \mathbf{q} = 0.$$

3. A set of the coefficients sum to one, $\beta_2 + \beta_3 + \beta_4 = 1$,

$$\mathbf{R} = [0 \quad 1 \quad 1 \quad 1 \quad 0 \quad \cdots] \text{ and } \mathbf{q} = 1.$$

4. A subset of the coefficients are all zero, $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_3 = 0$,

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} = [\mathbf{I} \quad \mathbf{0}] \text{ and } \mathbf{q} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

5. Several linear restrictions, $\beta_2 + \beta_3 = 1$, $\beta_4 + \beta_6 = 0$, and $\beta_5 + \beta_6 = 0$,

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \text{ and } \mathbf{q} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

6. All the coefficients in the model except the constant term are zero,

$$\mathbf{R} = [\mathbf{0} : \mathbf{I}_{K-1}] \text{ and } \mathbf{q} = \mathbf{0}.$$

The matrix $\mathbf{R}$ has K columns to be conformable with $\boldsymbol{\beta}$, J rows for a total of J restrictions, and *full row rank*, so J must be less than or equal to K. The rows of $\mathbf{R}$ must be linearly independent. Although it does not violate the condition, the case of $J = K$ must also be ruled out. If the K coefficients satisfy $J = K$ restrictions, then $\mathbf{R}$ is square and nonsingular and $\boldsymbol{\beta} = \mathbf{R}^{-1}\mathbf{q}$. There is no estimation or inference problem. The restriction $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$ imposes J restrictions on K otherwise free parameters. Hence, with the restrictions imposed, there are, in principle, only $K - J$ free parameters remaining.

We will want to extend the methods to nonlinear restrictions. In a following example, below, the hypothesis takes the form H_0: $\beta_j/\beta_k = \beta_l/\beta_m$. The **general nonlinear**

hypothesis involves a set of J possibly nonlinear restrictions,

$$\mathbf{c}(\boldsymbol{\beta}) = \mathbf{q}, \tag{5-9}$$

where $\mathbf{c}(\boldsymbol{\beta})$ is a set of **J** nonlinear functions of $\boldsymbol{\beta}$. The linear hypothesis is a special case. The counterpart to our requirements for the linear case are that, once again, J be strictly less than K, and the matrix of derivatives,

$$\mathbf{G}(\boldsymbol{\beta}) = \partial \mathbf{c}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}', \tag{5-10}$$

have full row rank. This means that the restrictions are **functionally independent**. In the linear case, $\mathbf{G}(\boldsymbol{\beta})$ is the matrix of constants, **R** that we saw earlier and functional independence is equivalent to linear independence. We will consider nonlinear restrictions in detail in Section 5.7. For the present, we will restict attention to the general linear hypothesis.

The hypothesis implied by the restrictions is written

$$H_0: \mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0},$$

$$H_1: \mathbf{R}\boldsymbol{\beta} - \mathbf{q} \neq \mathbf{0}.$$

We will consider two approaches to testing the hypothesis, Wald tests and fit based tests. The hypothesis characterizes the population. If the hypothesis is correct, then the sample statistics should mimic that description. To continue our earlier example, the hypothesis states that a certain coefficient in a regression model equals zero. If the hypothesis is correct, then the least squares coefficient should be close to zero, at least within sampling variability. The tests will proceed as follows:

- Wald tests: The hypothesis states that $\mathbf{R}\boldsymbol{\beta} - \mathbf{q}$ equals $\mathbf{0}$. The least squares estimator, **b**, is an unbiased and consistent estimator of $\boldsymbol{\beta}$. If the hypothesis is correct, then the **sample discrepancy**, $\mathbf{Rb} - \mathbf{q}$ should be close to zero. For the example of a single coefficient, if the hypothesis that β_k equals zero is correct, then b_k should be close to zero. The Wald test measures how close $\mathbf{Rb} - \mathbf{q}$ is to zero.
- Fit based tests: We obtain the best possible fit—highest R^2—by using least squares without imposing the restrictions. We proved this in Chapter 3. We will show here that the sum of squares will never decrease when we impose the restrictions—except for an unlikely special case, it will increase. For example, when we impose $\beta_k = 0$ by leaving x_k out of the model, we should expect R^2 to fall. The empirical device to use for testing the hypothesis will be a measure of how much R^2 falls when we impose the restrictions.

AN IMPORTANT ASSUMPTION

To develop the test statistics in this section, we will assume normally distributed disturbances. As we saw in Chapter 4, with this assumption, we will be able to obtain the exact distributions of the test statistics. In Section 5.6, we will consider the implications of relaxing this assumption and develop an alternative set of results that allows us to proceed without it.

5.4 WALD TESTS BASED ON THE DISTANCE MEASURE

The **Wald test** is the most commonly used procedure. It is often called a "significance test." The operating principle of the procedure is to fit the regression without the restrictions, and then assess whether the results appear, within sampling variability, to agree with the hypothesis.

5.4.1 TESTING A HYPOTHESIS ABOUT A COEFFICIENT

The simplest case is a test of the value of a single coefficient. Consider, once again, our art market example in Section 5.2. The null hypothesis is

$$H_0: \beta_2 = \beta_2^0,$$

where β_2^0 is the hypothesized value of the coefficient, in this case, zero. The **Wald distance** of a coefficient estimate from a hypothesized value is the linear distance, measured in standard deviation units. Thus, for this case, the distance of b_k from β_k^0 would be

$$W_k = \frac{b_k - \beta_k^0}{\sqrt{\sigma^2 S^{kk}}}. \tag{5-11}$$

As we saw in (4-38), W_k (which we called z_k before) has a standard normal distribution assuming that $E[b_k] = \beta_k^0$. Note that if $E[b_k]$ is not equal to β_k^0, then W_k still has a normal distribution, but the mean is not zero. In particular, if $E[b_k]$ is β_k^1 which is different from β_k^0, then

$$E\{W_k | E[b_k] = \beta_k^1\} = \frac{\beta_k^1 - \beta_k^0}{\sqrt{\sigma^2 S^{kk}}}. \tag{5-12}$$

(E.g., if the hypothesis is that $\beta_k = \beta_k^0 = 0$, and β_k does not equal zero, then the expected of $W_k = b_k/\sqrt{\sigma^2 S^{kk}}$ will equal $\beta_k^1/\sqrt{\sigma^2 S^{kk}}$, which is not zero.) For purposes of using W_k to test the hypothesis, our interpretation is that if β_k does equal β_k^0, then b_k will be close to β_k^0, with the distance measured in standard error units. Therefore, the logic of the test, to this point, will be to conclude that H_0 is incorrect—should be rejected—if W_k is "large."

Before we determine a benchmark for large, we note that the Wald measure suggested here is not usable because σ^2 is not known. It was estimated by s^2. Once again, invoking our results from Chapter 4, if we compute W_k using the sample estimate of σ^2, we obtain

$$t_k = \frac{b_k - \beta_k^0}{\sqrt{s^2 S^{kk}}} \tag{5-13}$$

Assuming that β_k does indeed equal β_k^0, that is, "under the assumption of the null hypothesis," then t_k has a t distribution with $n - K$ degrees of freedom. [See (4-41).] We can now construct the testing procedure. The test is carried out by determining in advance the desired confidence with which we would like to draw the conclusion—the standard value is 95 percent. Based on (5-13), we can say that

$$\text{Prob}\{-t^*_{(1-\alpha/2),[n-K]} < t_k < +t^*_{(1-\alpha/2),[n-K]}\}$$

where $t^*_{(1-\alpha/2),[n-K]}$ is the appropriate value from the t table (in Appendix G of this book). By this construction, finding a sample value of t_k that falls outside this range is unlikely. Our test procedure states that it is so unlikely that we would conclude that it could not happen if the hypothesis were correct, so the hypothesis must be incorrect.

A common test is the hypothesis that a parameter equals zero—equivalently, this is a test of the relevance of a variable in the regression. To construct the test statistic, we set β_k^0 to zero in (5-13) to obtain the standard "t ratio,"

$$t_k = \frac{b_k}{s_{bk}}.$$

This statistic is reported in the regression results in several of our earlier examples, such as 4.10 where the regression results for the model in (5-2) appear. This statistic is usually labeled the ***t* ratio** for the estimator b_k. If $|b_k|/s_{bk} > t_{(1-\alpha/2),[n-K]}$, where $t_{(1-\alpha/2),[n-K]}$ is the $100(1-\alpha/2)$ percent critical value from the t distribution with $(n-K)$ degrees of freedom, then the null hypothesis that the coefficient is zero is rejected and the coefficient (actually, the associated variable) is said to be "statistically significant." The value of 1.96, which would apply for the 95 percent significance level in a large sample, is often used as a benchmark value when a table of critical values is not immediately available. The t ratio for the test of the hypothesis that a coefficient equals zero is a standard part of the regression output of most computer programs.

Another view of the testing procedure is useful. Also based on (4-39) and (5-13), we formed a confidence interval for β_k as $b_k \pm t^* s_k$. We may view this interval as the set of plausible values of β_k with a confidence level of $100(1-\alpha)$ percent, where we choose α, typically 5 percent. The confidence interval provides a convenient tool for testing a hypothesis about β_k, since we may simply ask whether the hypothesized value, β_k^0 is contained in this range of plausible values.

Example 5.1 Art Appreciation

Regression results for the model in (5-3) based on a sample of 430 sales of Monet paintings appear in Table 4.6 in Example 4.10. The estimated coefficient on ln*Area* is 1.33372 with an estimated standard error of 0.09072. The distance of the estimated coefficient from zero is $1.33372/0.09072 = 14.70$. Since this is far larger than the 95 percent critical value of 1.96, we reject the hypothesis that β_2 equals zero; evidently buyers of Monet paintings do care about size. In constrast, the coefficient on *AspectRatio* is -0.16537 with an estimated standard error of 0.12753, so the associated t ratio for the test of H_0:$\beta_3 = 0$ is only -1.30. Since this is well under 1.96, we conclude that art buyers (of Monet paintings) do not care about the aspect ratio of the paintings. As a final consideration, we examine another (equally bemusing) hypothesis, whether auction prices are inelastic H_0: $\beta_2 \leq 1$ or elastic H_1: $\beta_2 > 1$ with respect to area. This is a **one-sided test**. Using our Neyman–Pearson guideline for formulating the test, we will reject the null hypothesis if the estimated coefficient is sufficiently larger than 1.0 (and not if it is less than or equal to 1.0). To maintain a test of size 0.05, we will then place all of the area for the critical region (the rejection region) to the right of 1.0; the critical value from the table is 1.645. The test statistic is $(1.33372 - 1.0)/0.09072 = 3.679 > 1.645$. Thus, we will reject this null hypothesis as well.

Example 5.2 Earnings Equation

Appendix Table F5.1 contains 753 observations used in Mroz's (1987) study of the labor supply behavior of married women. We will use these data at several points in this example. Of the 753 individuals in the sample, 428 were participants in the formal labor market. For these individuals, we will fit a semilog earnings equation of the form suggested in Example 2.2;

$$\ln earnings = \beta_1 + \beta_2 age + \beta_3 age^2 + \beta_4 education + \beta_5 kids + \varepsilon,$$

TABLE 5.1 Regression Results for an Earnings Equation

Sum of squared residuals:	599.4582
Standard error of the regression:	1.19044
R^2 based on 428 observations	0.040995

Variable	*Coefficient*	*Standard Error*	*t Ratio*
Constant	3.24009	1.7674	1.833
Age	0.20056	0.08386	2.392
Age^2	−0.0023147	0.00098688	−2.345
Education	0.067472	0.025248	2.672
Kids	−0.35119	0.14753	−2.380

Estimated Covariance Matrix for b (e − n = times 10^{-n})

Constant	*Age*	Age^2	*Education*	*Kids*
3.12381				
−0.14409	0.0070325			
0.0016617	−8.23237*e*−5	9.73928*e*−7		
−0.0092609	5.08549*e*−5	−4.96761*e*−7	0.00063729	
0.026749	−0.0026412	3.84102*e*−5	−5.46193*e*−5	0.021766

where *earnings* is *hourly wage* times *hours worked*, *education* is measured in years of schooling, and *kids* is a binary variable which equals one if there are children under 18 in the household. (See the data description in Appendix F for details.) Regression results are shown in Table 5.1. There are 428 observations and 5 parameters, so the t statistics have $(428-5) = 423$ degrees of freedom. For 95 percent significance levels, the standard normal value of 1.96 is appropriate when the degrees of freedom are this large. By this measure, all variables are statistically significant and signs are consistent with expectations. It will be interesting to investigate whether the effect of *kids* is on the wage or hours, or both. We interpret the schooling variable to imply that an additional year of schooling is associated with a 6.7 percent increase in earnings. The quadratic age profile suggests that for a given education level and family size, earnings rise to the peak at $-b_2/(2b_3)$ which is about 43 years of age, at which point they begin to decline. Some points to note: (1) Our selection of only those individuals who had positive hours worked is not an innocent sample selection mechanism. Since individuals chose whether or not to be in the labor force, it is likely (almost certain) that earnings potential was a significant factor, along with some other aspects we will consider in Chapter 18.

(2) The earnings equation is a mixture of a labor supply equation—hours worked by the individual—and a labor demand outcome—the wage is, presumably, an accepted offer. As such, it is unclear what the precise nature of this equation is. Presumably it is a hash of the equations of an elaborate structural equation system. (See Example 1.1 for discussion.)

5.4.2 THE *F* STATISTIC AND THE LEAST SQUARES DISCREPANCY

We now consider testing a set of J linear restrictions stated in the **null hypothesis**

$$H_0 : \mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$$

against the **alternative hypothesis,**

$$H_1 : \mathbf{R}\boldsymbol{\beta} - \mathbf{q} \neq \mathbf{0}.$$

Given the least squares estimator $\mathbf{b}$, our interest centers on the **discrepancy vector** $\mathbf{Rb} - \mathbf{q} = \mathbf{m}$. It is unlikely that $\mathbf{m}$ will be exactly $\mathbf{0}$. The statistical question is whether

the deviation of $\mathbf{m}$ from $\mathbf{0}$ can be attributed to sampling error or whether it is significant. Since $\mathbf{b}$ is normally distributed [see (4-18)] and $\mathbf{m}$ is a linear function of $\mathbf{b}$, $\mathbf{m}$ is also normally distributed. If the null hypothesis is true, then $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$ and $\mathbf{m}$ has mean vector

$$E[\mathbf{m} \mid \mathbf{X}] = \mathbf{R}E[\mathbf{b} \mid \mathbf{X}] - \mathbf{q} = \mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}.$$

and covariance matrix

$$\text{Var}[\mathbf{m} \mid \mathbf{X}] = \text{Var}[\mathbf{Rb} - \mathbf{q} \mid \mathbf{X}] = \mathbf{R}\{\text{Var}[\mathbf{b} \mid \mathbf{X}]\}\mathbf{R}' = \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'.$$

We can base a test of H_0 on the **Wald criterion.** Conditioned on $\mathbf{X}$, we find:

$$\begin{aligned} W &= \mathbf{m}'\{\text{Var}[\mathbf{m} \mid \mathbf{X}]\}^{-1}\mathbf{m} \\ &= (\mathbf{Rb} - \mathbf{q})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q}) \\ &= \frac{(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})}{\sigma^2} \\ &\sim \chi^2[J]. \end{aligned} \tag{5-14}$$

The statistic W has a chi-squared distribution with J degrees of freedom if the hypothesis is correct.[1] Intuitively, the larger $\mathbf{m}$ is—that is, the worse the failure of least squares to satisfy the restrictions—the larger the chi-squared statistic. Therefore, a large chi-squared value will weigh against the hypothesis.

The chi-squared statistic in (5-14) is not usable because of the unknown σ^2. By using s^2 instead of σ^2 and dividing the result by J, we obtain a usable F statistic with J and $n - K$ degrees of freedom. Making the substitution in (5-14), dividing by J, and multiplying and dividing by $n - K$, we obtain

$$\begin{aligned} F &= \frac{W}{J}\frac{\sigma^2}{s^2} \\ &= \left(\frac{(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})}{\sigma^2}\right)\left(\frac{1}{J}\right)\left(\frac{\sigma^2}{s^2}\right)\left(\frac{(n-K)}{(n-K)}\right) \\ &= \frac{(\mathbf{Rb} - \mathbf{q})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})/J}{[(n-K)s^2/\sigma^2]/(n-K)}. \end{aligned} \tag{5-15}$$

If $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, that is, if the null hypothesis is true, then $\mathbf{Rb} - \mathbf{q} = \mathbf{Rb} - \mathbf{R}\boldsymbol{\beta} = \mathbf{R}(\mathbf{b} - \boldsymbol{\beta}) = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$. [See (4-4).] Let $\mathbf{C} = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']$ since

$$\frac{\mathbf{R}(\mathbf{b} - \boldsymbol{\beta})}{\sigma} = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) = \mathbf{D}\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right),$$

the numerator of F equals $[(\boldsymbol{\varepsilon}/\sigma)'\mathbf{T}(\boldsymbol{\varepsilon}/\sigma)]/J$ where $\mathbf{T} = \mathbf{D}'\mathbf{C}^{-1}\mathbf{D}$. The numerator is W/J from (5-14) and is distributed as $1/J$ times a chi-squared $[J]$, as we showed earlier. We found in (4-16) that $s^2 = \mathbf{e}'\mathbf{e}/(n-K) = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}/(n-K)$ where $\mathbf{M}$ is an idempotent matrix. Therefore, the denominator of F equals $[(\boldsymbol{\varepsilon}/\sigma)'\mathbf{M}(\boldsymbol{\varepsilon}/\sigma)]/(n-K)$. This statistic is distributed as $1/(n-K)$ times a chi-squared $[n-K]$. Therefore, the F statistic is the ratio of two chi-squared variables each divided by its degrees of freedom. Since $\mathbf{M}(\boldsymbol{\varepsilon}/\sigma)$ and

[1] This calculation is an application of the "full rank quadratic form" of Section B.11.6. Note that although the chi-squared distribution is conditioned on $\mathbf{X}$, it is also free of $\mathbf{X}$.

$\mathbf{T}(\boldsymbol{\varepsilon}/\sigma)$ are both normally distributed and their covariance $\mathbf{TM}$ is $\mathbf{0}$, the vectors of the quadratic forms are independent. The numerator and denominator of F are functions of independent random vectors and are therefore independent. This completes the proof of the F distribution. [See (B-35).] Canceling the two appearances of σ^2 in (5-15) leaves the F statistic for testing a linear hypothesis:

$$F[J, n-K|\mathbf{X}] = \frac{(\mathbf{Rb}-\mathbf{q})'\left\{\mathbf{R}[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}'\right\}^{-1}(\mathbf{Rb}-\mathbf{q})}{J}. \tag{5-16}$$

For testing one linear restriction of the form

$$H_0 : r_1\beta_1 + r_2\beta_2 + \cdots + r_K\beta_K = \mathbf{r}'\boldsymbol{\beta} = q$$

(usually, some of the r's will be zero), the F statistic is

$$F[1, n-K] = \frac{(\Sigma_j r_j b_j - q)^2}{\Sigma_j \Sigma_k r_j r_k \,\text{Est. Cov}[b_j, b_k]}.$$

If the hypothesis is that the jth coefficient is equal to a particular value, then $\mathbf{R}$ has a single row with a 1 in the jth position and 0s elsewhere, $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ is the jth diagonal element of the inverse matrix, and $\mathbf{Rb}-\mathbf{q}$ is $(b_j - q)$. The F statistic is then

$$F[1, n-K] = \frac{(b_j - q)^2}{\text{Est. Var}[b_j]}.$$

Consider an alternative approach. The sample estimate of $\mathbf{r}'\boldsymbol{\beta}$ is

$$r_1b_1 + r_2b_2 + \cdots + r_Kb_K = \mathbf{r}'\mathbf{b} = \hat{q}.$$

If $\hat{q}$ differs significantly from q, then we conclude that the sample data are not consistent with the hypothesis. It is natural to base the test on

$$t = \frac{\hat{q} - q}{\text{se}(\hat{q})}. \tag{5-17}$$

We require an estimate of the standard error of $\hat{q}$. Since $\hat{q}$ is a linear function of $\mathbf{b}$ and we have an estimate of the covariance matrix of $\mathbf{b}$, $s^2(\mathbf{X}'\mathbf{X})^{-1}$, we can estimate the variance of $\hat{q}$ with

$$\text{Est. Var}[\hat{q}\,|\,\mathbf{X}] = \mathbf{r}'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{r}.$$

The denominator of t is the square root of this quantity. In words, t is the distance in standard error units between the hypothesized function of the true coefficients and the same function of our estimates of them. If the hypothesis is true, then our estimates should reflect that, at least within the range of sampling variability. Thus, if the absolute value of the preceding t ratio is larger than the appropriate critical value, then doubt is cast on the hypothesis.

There is a useful relationship between the statistics in (5-16) and (5-17). We can write the square of the t statistic as

$$t^2 = \frac{(\hat{q}-q)^2}{\text{Var}(\hat{q}-q\,|\,\mathbf{X})} = \frac{(\mathbf{r}'\mathbf{b}-q)\left\{\mathbf{r}'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{r}\right\}^{-1}(\mathbf{r}'\mathbf{b}-q)}{1}.$$

It follows, therefore, that for testing a single restriction, the t statistic is the square root of the F statistic that would be used to test that hypothesis.

Example 5.3 Restricted Investment Equation

Section 5.2.2 suggested a theory about the behavior of investors: They care only about real interest rates. If investors were only interested in the real rate of interest, then equal increases in interest rates and the rate of inflation would have no independent effect on investment. The null hypothesis is

$$H_0 : \beta_2 + \beta_3 = 0.$$

Estimates of the parameters of equations (5-4) and (5-6) using 1950.1 to 2000.4 quarterly data on real investment, real GDP, an interest rate (the 90-day T-bill rate), and inflation measured by the change in the log of the CPI given in Appendix Table F5.2 are presented in Table 5.2. (One observation is lost in computing the change in the CPI.)

To form the appropriate test statistic, we require the standard error of $\hat{q} = b_2 + b_3$, which is

$$\text{se}(\hat{q}) = [0.00319^2 + 0.00234^2 + 2(-3.718 \times 10^{-6})]^{1/2} = 0.002866.$$

The t ratio for the test is therefore

$$t = \frac{-0.00860 + 0.00331}{0.002866} = -1.845.$$

Using the 95 percent critical value from t [203-5] $= 1.96$ (the standard normal value), we conclude that the sum of the two coefficients is not significantly different from zero, so the hypothesis should not be rejected.

There will usually be more than one way to formulate a restriction in a regression model. One convenient way to parameterize a constraint is to set it up in such a way that the standard test statistics produced by the regression can be used without further computation to test the hypothesis. In the preceding example, we could write the regression model as specified in (5-5). Then an equivalent way to test H_0 would be to fit the investment equation with both the real interest rate and the rate of inflation as regressors and to test our theory by simply testing the hypothesis that β_3 equals zero, using the standard t statistic that is routinely computed. When the regression is computed this way, $b_3 = -0.00529$ and the estimated standard error is 0.00287, resulting in a t ratio of -1.844(!). (Exercise: Suppose that the nominal interest rate, rather than the rate of inflation, were included as the extra regressor. What do you think the coefficient and its standard error would be?)

Finally, consider a test of the joint hypothesis

$$\begin{aligned} \beta_2 + \beta_3 &= 0 \quad \text{(investors consider the real interest rate),} \\ \beta_4 &= 1 \quad \text{(the marginal propensity to invest equals 1),} \\ \beta_5 &= 0 \quad \text{(there is no time trend).} \end{aligned}$$

TABLE 5.2 Estimated Investment Equations (Estimated standard errors in parentheses)

	β_1	β_2	β_3	β_4	β_5
Model (5-4)	−9.135	−0.00860	0.00331	1.930	−0.00566
	(1.366)	(0.00319)	(0.00234)	(0.183)	(0.00149)
	$s = 0.08618$, $R^2 = 0.979753$, $\mathbf{e}'\mathbf{e} = 1.47052$, Est. Cov$[b_2, b_3] = -3.718\text{e}{-6}$				
Model (5-6)	−7.907	−0.00443	0.00443	1.764	−0.00440
	(1.201)	(0.00227)	(0.00227)	(0.161)	(0.00133)
	$s = 0.8670$, $R^2 = 0.979405$, $\mathbf{e}'\mathbf{e} = 1.49578$				

Then,

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{q} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{Rb} - \mathbf{q} = \begin{bmatrix} -0.0053 \\ 0.9302 \\ -0.0057 \end{bmatrix}.$$

Inserting these values in F yields $F = 109.84$. The 5 percent critical value for $F[3, 198]$ is 2.65. We conclude, therefore, that these data are not consistent with the hypothesis. The result gives no indication as to which of the restrictions is most influential in the rejection of the hypothesis. If the three restrictions are tested one at a time, the t statistics in (5-17) are -1.844, 5.076, and -3.803. Based on the individual test statistics, therefore, we would expect both the second and third hypotheses to be rejected.

5.5 TESTING RESTRICTIONS USING THE FIT OF THE REGRESSION

A different approach to hypothesis testing focuses on the fit of the regression. Recall that the least squares vector $\mathbf{b}$ was chosen to minimize the sum of squared deviations, $\mathbf{e}'\mathbf{e}$. Since R^2 equals $1 - \mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$ and $\mathbf{y}'\mathbf{M}^0\mathbf{y}$ is a constant that does not involve $\mathbf{b}$, it follows that $\mathbf{b}$ is chosen to maximize R^2. One might ask whether choosing some other value for the slopes of the regression leads to a significant loss of fit. For example, in the investment equation (5-4), one might be interested in whether assuming the hypothesis (that investors care only about real interest rates) leads to a substantially worse fit than leaving the model unrestricted. To develop the test statistic, we first examine the computation of the least squares estimator subject to a set of restrictions. We will then construct a test statistic that is based on comparing the R^2's from the two regressions.

5.5.1 THE RESTRICTED LEAST SQUARES ESTIMATOR

Suppose that we explicitly impose the restrictions of the general linear hypothesis in the regression. The restricted least squares estimator is obtained as the solution to

$$\text{Minimize}_{\mathbf{b}_0} \; S(\mathbf{b}_0) = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'(\mathbf{y} - \mathbf{X}\mathbf{b}_0) \quad \text{subject to } \mathbf{R}\mathbf{b}_0 = \mathbf{q}. \tag{5-18}$$

A Lagrangean function for this problem can be written

$$L^*(\mathbf{b}_0, \boldsymbol{\lambda}) = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'(\mathbf{y} - \mathbf{X}\mathbf{b}_0) + 2\boldsymbol{\lambda}'(\mathbf{R}\mathbf{b}_0 - \mathbf{q}).^2 \tag{5-19}$$

The solutions $\mathbf{b}_*$ and $\boldsymbol{\lambda}_*$ will satisfy the necessary conditions

$$\begin{aligned} \frac{\partial L^*}{\partial \mathbf{b}_*} &= -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}_*) + 2\mathbf{R}'\boldsymbol{\lambda}_* = \mathbf{0} \\ \frac{\partial L^*}{\partial \boldsymbol{\lambda}_*} &= 2(\mathbf{R}\mathbf{b}_* - \mathbf{q}) = \mathbf{0}. \end{aligned} \tag{5-20}$$

Dividing through by 2 and expanding terms produces the partitioned matrix equation

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{R}' \\ \mathbf{R} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b}_* \\ \boldsymbol{\lambda}_* \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{q} \end{bmatrix} \tag{5-21}$$

or

$$\mathbf{A}\mathbf{d}_* = \mathbf{v}.$$

[2] Since $\boldsymbol{\lambda}$ is not restricted, we can formulate the constraints in terms of $2\boldsymbol{\lambda}$. The convenience of the scaling shows up in (5-20).

Assuming that the partitioned matrix in brackets is nonsingular, the restricted least squares estimator is the upper part of the solution

$$\mathbf{d}_* = \mathbf{A}^{-1}\mathbf{v}. \tag{5-22}$$

If, in addition, $\mathbf{X}'\mathbf{X}$ is nonsingular, then explicit solutions for $\mathbf{b}_*$ and $\boldsymbol{\lambda}_*$ may be obtained by using the formula for the partitioned inverse (A-74),[3]

$$\begin{aligned}\mathbf{b}_* &= \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}) \\ &= \mathbf{b} - \mathbf{C}\mathbf{m}\end{aligned}$$

and

$$\boldsymbol{\lambda}_* = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}). \tag{5-23}$$

Greene and Seaks (1991) show that the covariance matrix for $\mathbf{b}_*$ is simply σ^2 times the upper left block of $\mathbf{A}^{-1}$. Once again, in the usual case in which $\mathbf{X}'\mathbf{X}$ is nonsingular, an explicit formulation may be obtained:

$$\text{Var}[\mathbf{b}_* \mid \mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}. \tag{5-24}$$

Thus,

$$\text{Var}[\mathbf{b}_* \mid \mathbf{X}] = \text{Var}[\mathbf{b} \mid \mathbf{X}]\text{—a nonnegative definite matrix.}$$

One way to interpret this reduction in variance is as the value of the information contained in the restrictions.

Note that the explicit solution for $\boldsymbol{\lambda}_*$ involves the discrepancy vector $\mathbf{R}\mathbf{b} - \mathbf{q}$. If the unrestricted least squares estimator satisfies the restriction, the Lagrangean multipliers will equal zero and $\mathbf{b}_*$ will equal $\mathbf{b}$. Of course, this is unlikely. The constrained solution $\mathbf{b}_*$ is equal to the unconstrained solution $\mathbf{b}$ minus a term that accounts for the failure of the unrestricted solution to satisfy the constraints.

5.5.2 THE LOSS OF FIT FROM RESTRICTED LEAST SQUARES

To develop a test based on the restricted least squares estimator, we consider a single coefficient first and then turn to the general case of J linear restrictions. Consider the change in the fit of a multiple regression when a variable z is added to a model that already contains $K-1$ variables, $\mathbf{x}$. We showed in Section 3.5 (Theorem 3.6) (3-29) that the effect on the fit would be given by

$$R^2_{\mathbf{Xz}} = R^2_{\mathbf{X}} + (1 - R^2_{\mathbf{X}})r^{*2}_{yz}, \tag{5-25}$$

where $R^2_{\mathbf{Xz}}$ is the new R^2 after z is added, $R^2_{\mathbf{X}}$ is the original R^2 and r^*_{yz} is the partial correlation between y and z, controlling for $\mathbf{x}$. So, as we knew, the fit improves (or, at the least, does not deteriorate). In deriving the partial correlation coefficient between y and z in (3-22) we obtained the convenient result

$$r^{*2}_{yz} = \frac{t_z^2}{t_z^2 + (n-K)}, \tag{5-26}$$

[3]The general solution given for $\mathbf{d}_*$ may be usable even if $\mathbf{X}'\mathbf{X}$ is singular. Suppose, for example, that $\mathbf{X}'\mathbf{X}$ is 4×4 with rank 3. Then $\mathbf{X}'\mathbf{X}$ is singular. But if there is a parametric restriction on $\boldsymbol{\beta}$, then the 5×5 matrix in brackets may still have rank 5. This formulation and a number of related results are given in Greene and Seaks (1991).

where t_z^2 is the square of the t ratio for testing the hypothesis that the coefficient on z is zero in the *multiple* regression of $\mathbf{y}$ on $\mathbf{X}$ and $\mathbf{z}$. If we solve (5-25) for r_{yz}^{*2} and (5-26) for t_z^2 and then insert the first solution in the second, then we obtain the result

$$t_z^2 = \frac{\left(R_{\mathbf{Xz}}^2 - R_{\mathbf{X}}^2\right)/1}{\left(1 - R_{\mathbf{Xz}}^2\right)/(n-K)}. \tag{5-27}$$

We saw at the end of Section 5.4.2 that for a single restriction, such as $\beta_z = 0$,

$$F[1, n-K] = t^2[n-K],$$

which gives us our result. That is, in (5-27), we see that the squared t statistic (i.e., the F statistic) is computed using the change in the R^2. By interpreting the preceding as the result of *removing* z from the regression, we see that we have proved a result for the case of testing whether a single slope is zero. But the preceding result is general. The test statistic for a single linear restriction is the square of the t ratio in (5-17). By this construction, we see that for a single restriction, F is a measure of the loss of fit that results from imposing that restriction. To obtain this result, we will proceed to the general case of J linear restrictions, which will include one restriction as a special case.

The fit of the restricted least squares coefficients cannot be better than that of the unrestricted solution. Let $\mathbf{e}_*$ equal $\mathbf{y} - \mathbf{X}\mathbf{b}_*$. Then, using a familiar device,

$$\mathbf{e}_* = \mathbf{y} - \mathbf{X}\mathbf{b} - \mathbf{X}(\mathbf{b}_* - \mathbf{b}) = \mathbf{e} - \mathbf{X}(\mathbf{b}_* - \mathbf{b}).$$

The new sum of squared deviations is

$$\mathbf{e}_*'\mathbf{e}_* = \mathbf{e}'\mathbf{e} + (\mathbf{b}_* - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{b}_* - \mathbf{b}) \geq \mathbf{e}'\mathbf{e}.$$

(The middle term in the expression involves $\mathbf{X}'\mathbf{e}$, which is zero.) The loss of fit is

$$\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e} = (\mathbf{R}\mathbf{b} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}). \tag{5-28}$$

This expression appears in the numerator of the F statistic in (5-7). Inserting the remaining parts, we obtain

$$F[J, n-K] = \frac{(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(n-K)}. \tag{5-29}$$

Finally, by dividing both numerator and denominator of F by $\Sigma_i (y_i - \overline{y})^2$, we obtain the general result:

$$F[J, n-K] = \frac{(R^2 - R_*^2)/J}{(1-R^2)/(n-K)}. \tag{5-30}$$

This form has some intuitive appeal in that the difference in the fits of the two models is directly incorporated in the test statistic. As an example of this approach, consider the joint test that all the slopes in the model are zero. This is the overall F ratio that will be discussed in Section 5.5.3, where $R_*^2 = 0$.

For imposing a set of **exclusion restrictions** such as $\beta_k = 0$ for one or more coefficients, the obvious approach is simply to omit the variables from the regression and base the test on the sums of squared residuals for the restricted and unrestricted regressions. The F statistic for testing the hypothesis that a subset, say $\boldsymbol{\beta}_2$, of the coefficients are all zero is constructed using $\mathbf{R} = (\mathbf{0} : \mathbf{I})$, $\mathbf{q} = \mathbf{0}$, and $J = K_2 =$ the number of elements in $\boldsymbol{\beta}_2$. The matrix $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ is the $K_2 \times K_2$ lower right block of the full inverse matrix.

Using our earlier results for partitioned inverses and the results of Section 3.3, we have

$$\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}$$

and

$$\mathbf{Rb} - \mathbf{q} = \mathbf{b}_2.$$

Inserting these in (5-28) gives the loss of fit that results when we drop a subset of the variables from the regression:

$$\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e} = \mathbf{b}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2.$$

The procedure for computing the appropriate F statistic amounts simply to comparing the sums of squared deviations from the "short" and "long" regressions, which we saw earlier.

Example 5.4 Production Function

The data in Appendix Table F5.3 have been used in several studies of production functions.[4] Least squares regression of log output (value added) on a constant and the logs of labor and capital produce the estimates of a Cobb–Douglas production function shown in Table 5.3. We will construct several hypothesis tests based on these results. A generalization of the Cobb–Douglas model is the *translog* model,[5] which is

$$\ln Y = \beta_1 + \beta_2 \ln L + \beta_3 \ln K + \beta_4\left(\tfrac{1}{2}\ln^2 L\right) + \beta_5\left(\tfrac{1}{2}\ln^2 K\right) + \beta_6 \ln L \ln K + \varepsilon.$$

As we shall analyze further in Chapter 10, this model differs from the Cobb–Douglas model in that it relaxes the Cobb–Douglas's assumption of a unitary elasticity of substitution. The Cobb–Douglas model is obtained by the restriction $\beta_4 = \beta_5 = \beta_6 = 0$. The results for the two regressions are given in Table 5.3. The F statistic for the hypothesis of a Cobb–Douglas model is

$$F[3, 21] = \frac{(0.85163 - 0.67993)/3}{0.67993/21} = 1.768.$$

The critical value from the F table is 3.07, so we would not reject the hypothesis that a Cobb–Douglas model is appropriate.

The hypothesis of constant returns to scale is often tested in studies of production. This hypothesis is equivalent to a restriction that the two coefficients of the Cobb–Douglas production function sum to 1. For the preceding data,

$$F[1, 24] = \frac{(0.6030 + 0.3757 - 1)^2}{0.01586 + 0.00728 - 2(0.00961)} = 0.1157,$$

which is substantially less than the 95 percent critical value of 4.26. We would not reject the hypothesis; the data are consistent with the hypothesis of constant returns to scale. The equivalent test for the translog model would be $\beta_2 + \beta_3 = 1$ and $\beta_4 + \beta_5 + 2\beta_6 = 0$. The F statistic with 2 and 21 degrees of freedom is 1.8991, which is less than the critical value of 3.47. Once again, the hypothesis is not rejected.

In most cases encountered in practice, it is possible to incorporate the restrictions of a hypothesis directly on the regression and estimate a restricted model.[6] For example, to

[4]The data are statewide observations on SIC 33, the primary metals industry. They were originally constructed by Hildebrand and Liu (1957) and have subsequently been used by a number of authors, notably Aigner, Lovell, and Schmidt (1977). The 28th data point used in the original study is incomplete; we have used only the remaining 27.

[5]Berndt and Christensen (1973). See Example 2.4 and Section 10.4.2 for discussion.

[6]This case is not true when the restrictions are nonlinear. We consider this issue in Chapter 7.

TABLE 5.3 Estimated Production Functions

	Translog	*Cobb–Douglas*
Sum of squared residuals	0.67993	0.85163
Standard error of regression	0.17994	0.18837
R-squared	0.95486	0.94346
Adjusted R-squared	0.94411	0.93875
Number of observations	27	27

Variable	*Coefficient*	*Standard Error*	*t Ratio*	*Coefficient*	*Standard Error*	*t Ratio*
Constant	0.944196	2.911	0.324	1.171	0.3268	3.582
$\ln L$	3.61364	1.548	2.334	0.6030	0.1260	4.787
$\ln K$	−1.89311	1.016	−1.863	0.3757	0.0853	4.402
$\frac{1}{2}\ln^2 L$	−0.96405	0.7074	−1.363			
$\frac{1}{2}\ln^2 K$	0.08529	0.2926	0.291			
$\ln L \times \ln K$	0.31239	0.4389	0.712			

Estimated Covariance Matrix for Translog (Cobb–Douglas) Coefficient Estimates

	Constant	$\ln L$	$\ln K$	$\frac{1}{2}\ln^2 L$	$\frac{1}{2}\ln^2 K$	$\ln L \ln K$
Constant	8.472 (0.1068)					
$\ln L$	−2.388 (−0.01984)	2.397 (0.01586)				
$\ln K$	−0.3313 (0.001189)	−1.231 (−0.00961)	1.033 (0.00728)			
$\frac{1}{2}\ln^2 L$	−0.08760	−0.6658	0.5231	0.5004		
$\frac{1}{2}\ln^2 K$	−0.2332	0.03477	0.02637	0.1467	0.08562	
$\ln L \ln K$	0.3635	0.1831	−0.2255	−0.2880	−0.1160	0.1927

impose the constraint $\beta_2 = 1$ on the Cobb–Douglas model, we would write

$$\ln Y = \beta_1 + 1.0 \ln L + \beta_3 \ln K + \varepsilon$$

or

$$\ln Y - \ln L = \beta_1 + \beta_3 \ln K + \varepsilon.$$

Thus, the restricted model is estimated by regressing $\ln Y - \ln L$ on a constant and $\ln K$. Some care is needed if this regression is to be used to compute an F statistic. If the F statistic is computed using the sum of squared residuals [see (5-29)], then no problem will arise. If (5-30) is used instead, however, then it may be necessary to account for the restricted regression having a different dependent variable from the unrestricted one. In the preceding regression, the dependent variable in the unrestricted regression is $\ln Y$, whereas in the restricted regression, it is $\ln Y - \ln L$. The R^2 from the restricted regression is only 0.26979, which would imply an F statistic of 285.96, whereas the correct value is 9.935. If we compute the appropriate R_*^2 using the correct denominator, however, then its value is 0.92006 and the correct F value results.

Note that the coefficient on $\ln K$ is negative in the translog model. We might conclude that the estimated output elasticity with respect to capital now has the wrong sign. This conclusion would be incorrect, however; in the translog model, the capital elasticity of output is

$$\frac{\partial \ln Y}{\partial \ln K} = \beta_3 + \beta_5 \ln K + \beta_6 \ln L.$$

If we insert the coefficient estimates and the mean values for $\ln K$ and $\ln L$ (not the logs of the means) of 7.44592 and 5.7637, respectively, then the result is 0.5425, which is quite in line with our expectations and is fairly close to the value of 0.3757 obtained for the Cobb–Douglas model. The estimated standard error for this linear combination of the least squares estimates is computed as the square root of

$$\text{Est. Var}[b_3 + b_5 \overline{\ln K} + b_6 \overline{\ln L}] = \mathbf{w}'(\text{Est. Var}[\mathbf{b}])\mathbf{w},$$

where

$$\mathbf{w} = (0, 0, 1, 0, \overline{\ln K}, \overline{\ln L})'$$

and $\mathbf{b}$ is the full 6×1 least squares coefficient vector. This value is 0.1122, which is reasonably close to the earlier estimate of 0.0853.

5.5.3 TESTING THE SIGNIFICANCE OF THE REGRESSION

A question that is usually of interest is whether the regression equation as a whole is significant. This test is a joint test of the hypotheses that *all* the coefficients except the constant term are zero. If all the slopes are zero, then the multiple correlation coefficient, R^2, is zero as well, so we can base a test of this hypothesis on the value of R^2. The central result needed to carry out the test is given in (5-30). This is the special case with $R_*^2 = 0$, so the F statistic, which is usually reported with multiple regression results is

$$F[K-1, n-K] = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}.$$

If the hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$ (the part of $\boldsymbol{\beta}$ not including the constant) is true and the disturbances are normally distributed, then this statistic has an F distribution with K-1 and n- K degrees of freedom. Large values of F give evidence against the validity of the hypothesis. Note that a large F is induced by a large value of R^2. The logic of the test is that the F statistic is a measure of the loss of fit (namely, all of R^2) that results when we impose the restriction that all the slopes are zero. If F is large, then the hypothesis is rejected.

Example 5.5 F Test for the Earnings Equation

The F ratio for testing the hypothesis that the four slopes in the earnings equation in Example 5.2 are all zero is

$$F[4, 423] = \frac{0.040995/(5-1)}{(1-0.040995)/(428-5)} = 4.521,$$

which is far larger than the 95 percent critical value of 2.39. We conclude that the data are inconsistent with the hypothesis that all the slopes in the earnings equation are zero. We might have expected the preceding result, given the substantial t ratios presented earlier. But this case need not always be true. Examples can be constructed in which the individual coefficients are statistically significant, while jointly they are not. This case can be regarded as pathological, but the opposite one, in which none of the coefficients is significantly different from zero while R^2 is highly significant, is relatively common. The problem is that the interaction among the variables may serve to obscure their individual contribution to the fit of the regression, whereas their joint effect may still be significant.

5.5.4 SOLVING OUT THE RESTRICTIONS AND A CAUTION ABOUT USING R^2

In principle, one can usually solve out the restrictions imposed by a linear hypothesis. Algebraically, we would begin by partitioning $\mathbf{R}$ into two groups of columns, one with

J and one with $K - J$, so that the first set are linearly independent. (There are many ways to do so; any one will do for the present.) Then, with $\boldsymbol{\beta}$ likewise partitioned and its elements reordered in whatever way is needed, we may write

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{R}_1\boldsymbol{\beta}_1 + \mathbf{R}_2\boldsymbol{\beta}_2 = \mathbf{q}.$$

If the J columns of $\mathbf{R}_1$ are independent, then

$$\boldsymbol{\beta}_1 = \mathbf{R}_1^{-1}[\mathbf{q} - \mathbf{R}_2\boldsymbol{\beta}_2].$$

This suggests that one might estimate the restricted model directly using a transformed equation, rather than use the rather cumbersome restricted estimator shown in (5-23). A simple example illustrates. Consider imposing constant returns to scale on a two input production function,

$$\ln y = \beta_1 + \beta_2 \ln x_1 + \beta_3 \ln x_2 + \varepsilon.$$

The hypothesis of linear homogeneity is $\beta_2 + \beta_3 = 1$ or $\beta_3 = 1 - \beta_2$. Simply building the restriction into the model produces

$$\ln y = \beta_1 + \beta_2 \ln x_1 + (1 - \beta_2) \ln x_2 + \varepsilon$$

or

$$\ln y = \ln x_2 + \beta_1 + \beta_2(\ln x_1 - \ln x_2) + \varepsilon.$$

One can obtain the restricted least squares estimates by linear regression of $(\ln y - \ln x_2)$ on a constant and $(\ln x_1 - \ln x_2)$. However, the test statistic for the hypothesis cannot be tested using the familiar result in (5-30), because the denominators in the two R^2's are different. The statistic in (5-30) could even be negative. The appropriate approach would be to use the equivalent, but appropriate computation based on the sum of squared residuals in ( ). The general result from this example is that one must be careful in using (5-30) and that the dependent variable in the two regressions must be the same.

5.6 NONNORMAL DISTURBANCES AND LARGE-SAMPLE TESTS

We now consider the relation between the sample test statistics and the data in $\mathbf{X}$. First, consider the conventional t statistic in (4-41) for testing $H_0 : \beta_k = \beta_k^0$,

$$t|\mathbf{X} = \frac{b_k - \beta_k^0}{\sqrt{s^2(\mathbf{X}'\mathbf{X})_{kk}^{-1}}}.$$

Conditional on $\mathbf{X}$, if $\beta_k = \beta_k^0$ (i.e., under H_0), then $t|\mathbf{X}$ has a t distribution with $(n - K)$ degrees of freedom. What interests us, however, is the marginal, that is, the unconditional distribution of t. As we saw, $\mathbf{b}$ is only normally distributed conditionally on $\mathbf{X}$; the marginal distribution may not be normal because it depends on $\mathbf{X}$ (through the conditional variance). Similarly, because of the presence of $\mathbf{X}$, the denominator of the t statistic is not the square root of a chi-squared variable divided by its degrees of freedom, again, except conditional on this $\mathbf{X}$. But, because the distributions of $(b_k - \beta_k)/\sqrt{s^2(\mathbf{X}'\mathbf{X})_{kk}^{-1}}|\mathbf{X}$ and $[(n - K)s_2/\sigma^2]|\mathbf{X}$ are still independent $N[0, 1]$ and

$\chi^2[n-K]$, respectively, which do not involve $\mathbf{X}$, we have the surprising result that, regardless of the distribution of $\mathbf{X}$, or even of whether $\mathbf{X}$ is stochastic or nonstochastic, the marginal distributions of t is still t, even though the marginal distribution of b_k may be nonnormal. This intriguing result follows because $f(t\,|\mathbf{X})$ is not a function of $\mathbf{X}$. The same reasoning can be used to deduce that the usual F ratio used for testing linear restrictions, discussed in the previous section, is valid whether $\mathbf{X}$ is stochastic or not. This result is very powerful. The implication is that *if the disturbances are normally distributed, then we may carry out tests and construct confidence intervals for the parameters without making any changes in our procedures, regardless of whether the regressors are stochastic, nonstochastic, or some mix of the two.*

The distributions of these statistics do follow from the normality assumption for $\boldsymbol{\varepsilon}$, but they do not depend on $\mathbf{X}$. Without the normality assumption, however, the exact distributions of these statistics depend on the data and the parameters and are not F, t, and chi-squared. At least at first blush, it would seem that we need either a new set of critical values for the tests or perhaps a new set of test statistics. In this section, we will examine results that will generalize the familiar procedures. These large-sample results suggest that although the usual t and F statistics are still usable, in the more general case without the special assumption of normality, they are viewed as approximations whose quality improves as the sample size increases. By using the results of Section D.3 (on asymptotic distributions) and some large-sample results for the least squares estimator, we can construct a set of usable inference procedures based on already familiar computations.

Assuming the data are well behaved, the *asymptotic* distribution of the least squares coefficient estimator, $\mathbf{b}$, is given by

$$\mathbf{b} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}^{-1}\right] \quad \text{where } \mathbf{Q} = \operatorname{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right). \tag{5-31}$$

The interpretation is that, absent normality of $\boldsymbol{\varepsilon}$, as the sample size, n, grows, the normal distribution becomes an increasingly better approximation to the true, though at this point unknown, distribution of $\mathbf{b}$. As n increases, the distribution of $\sqrt{n}(\mathbf{b}-\boldsymbol{\beta})$ *converges* exactly to a normal distribution, which is how we obtain the preceding finite-sample approximation. This result is based on the central limit theorem and does not require normally distributed disturbances. The second result we will need concerns the estimator of σ^2:

$$\operatorname{plim} s^2 = \sigma^2, \quad \text{where } s^2 = \mathbf{e}'\mathbf{e}/(n-K).$$

With these in place, we can obtain some large-sample results for our test statistics that suggest how to proceed in a finite sample with nonnormal disturbances.

The sample statistic for testing the hypothesis that one of the coefficients, β_k equals a particular value, β_k^0 is

$$t_k = \frac{\sqrt{n}(b_k - \beta_k^0)}{\sqrt{s^2\left(\mathbf{X}'\mathbf{X}/n\right)^{-1}_{kk}}}.$$

(Note that two occurrences of $\sqrt{n}$ cancel to produce our familiar result.) Under the null hypothesis, with normally distributed disturbances, t_k is exactly distributed as t with $n-K$ degrees of freedom. [See Theorem 4.4 and the beginning of this section.] The

exact distribution of this statistic is unknown, however, if $\boldsymbol{\varepsilon}$ is not normally distributed. From the preceding results, we find that the denominator of t_k converges to $\sqrt{\sigma^2 \mathbf{Q}_{kk}^{-1}}$. Hence, if t_k has a limiting distribution, then it is the same as that of the statistic that has this latter quantity in the denominator. (See point 3 Theorem D.16.) That is, the large-sample distribution of t_k is the same as that of

$$\tau_k = \frac{\sqrt{n}\left(b_k - \beta_k^0\right)}{\sqrt{\sigma^2 \mathbf{Q}_{kk}^{-1}}}.$$

But $\tau_k = (b_k - E[b_k])/(\text{Asy. Var}[b_k])^{1/2}$ from the asymptotic normal distribution (under the hypothesis $\beta_k = \beta_k^0$), so it follows that τ_k has a standard normal asymptotic distribution, and this result is the large-sample distribution of our t statistic. Thus, as a large-sample approximation, we will use the standard normal distribution to approximate the true distribution of the test statistic t_k and use the critical values from the standard normal distribution for testing hypotheses.

The result in the preceding paragraph is valid only in large samples. For moderately sized samples, it provides only a suggestion that the t distribution may be a reasonable approximation. The appropriate critical values only *converge* to those from the standard normal, and generally *from above,* although we cannot be sure of this. In the interest of conservatism—that is, in controlling the probability of a Type I error—one should generally use the critical value from the t distribution even in the absence of normality. Consider, for example, using the standard normal critical value of 1.96 for a two-tailed test of a hypothesis based on 25 degrees of freedom. The nominal size of this test is 0.05. The actual size of the test, however, is the true, but unknown, probability that $|t_k| > 1.96$, which is 0.0612 if the $t[25]$ distribution is correct, and some other value if the disturbances are not normally distributed. The end result is that the standard t test retains a large sample validity. Little can be said about the true size of a test based on the t distribution unless one makes some other equally narrow assumption about $\boldsymbol{\varepsilon}$, but the t distribution is generally used as a reliable approximation.

We will use the same approach to analyze the F statistic for testing a set of J linear restrictions. Step 1 will be to show that with normally distributed disturbances, JF converges to a chi-squared variable as the sample size increases. We will then show that this result is actually independent of the normality of the disturbances; it relies on the central limit theorem. Finally, we consider, as before, the appropriate critical values to use for this test statistic, which only has large sample validity.

The F statistic for testing the validity of J linear restrictions, $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$, is given in (5-6). With normally distributed disturbances and under the null hypothesis, the exact distribution of this statistic is $F[J, n - K]$. To see how F behaves more generally, divide the numerator and denominator in (5-16) by σ^2 and rearrange the fraction slightly, so

$$F = \frac{(\mathbf{Rb} - \mathbf{q})' \left\{ \mathbf{R}[\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}' \right\}^{-1} (\mathbf{Rb} - \mathbf{q})}{J(s^2/\sigma^2)}. \tag{5-32}$$

Since $\text{plim}\, s^2 = \sigma^2$, and $\text{plim}(\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}$, the denominator of F converges to J and the bracketed term in the numerator will behave the same as $(\sigma^2/n)\mathbf{R}\mathbf{Q}^{-1}\mathbf{R}'$. (See Theorem D16.3.) Hence, regardless of what this distribution is, if F has a limiting distribution,

then it is the same as the limiting distribution of

$$W^* = \frac{1}{J}(\mathbf{Rb}-\mathbf{q})'[\mathbf{R}(\sigma^2/n)\mathbf{Q}^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})$$

$$= \frac{1}{J}(\mathbf{Rb}-\mathbf{q})'\{\text{Asy. Var}[\mathbf{Rb}-\mathbf{q}]\}^{-1}(\mathbf{Rb}-\mathbf{q}).$$

This expression is $(1/J)$ times a **Wald statistic,** based on the asymptotic distribution. The large-sample distribution of W^* will be that of $(1/J)$ times a chi-squared with J degrees of freedom. It follows that with normally distributed disturbances, JF converges to a chi-squared variate with J degrees of freedom. The proof is instructive. [See White (2001, 9.76).]

THEOREM 5.1 Limiting Distribution of the Wald Statistic

If $\sqrt{n}(\mathbf{b}-\boldsymbol{\beta}) \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}^{-1}]$ *and if* $H_0 : \mathbf{R}\boldsymbol{\beta}-\mathbf{q}=\mathbf{0}$ *is true, then*

$$W = (\mathbf{Rb}-\mathbf{q})'\{\mathbf{R}s^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\}^{-1}(\mathbf{Rb}-\mathbf{q}) = JF \xrightarrow{d} \chi^2[\mathrm{J}].$$

Proof: *Since* $\mathbf{R}$ *is a matrix of constants and* $\mathbf{R}\boldsymbol{\beta}=\mathbf{q}$,

$$\sqrt{n}\mathbf{R}(\mathbf{b}-\boldsymbol{\beta}) = \sqrt{n}(\mathbf{Rb}-\mathbf{q}) \xrightarrow{d} N[\mathbf{0}, \mathbf{R}(\sigma^2\mathbf{Q}^{-1})\mathbf{R}']. \tag{1}$$

For convenience, write this equation as

$$\mathbf{z} \xrightarrow{d} N[\mathbf{0}, \mathbf{P}]. \tag{2}$$

In Section A.6.11, we define the inverse square root of a positive definite matrix $\mathbf{P}$ *as another matrix, say* $\mathbf{T}$, *such that* $\mathbf{T}^2=\mathbf{P}^{-1}$, *and denote* $\mathbf{T}$ *as* $\mathbf{P}^{-1/2}$. *Then, by the same reasoning as in (1) and (2),*

$$\text{if } \mathbf{z} \xrightarrow{d} N[\mathbf{0}, \mathbf{P}], \quad \text{then } \mathbf{P}^{-1/2}\mathbf{z} \xrightarrow{d} N[\mathbf{0}, \mathbf{P}^{-1/2}\mathbf{P}\mathbf{P}^{-1/2}] = N[\mathbf{0}, \mathbf{I}]. \tag{3}$$

We now invoke Theorem D.21 for the limiting distribution of a function of a random variable. The sum of squares of uncorrelated (i.e., independent) standard normal variables is distributed as chi-squared. Thus, the limiting distribution of

$$(\mathbf{P}^{-1/2}\mathbf{z})'(\mathbf{P}^{-1/2}\mathbf{z}) = \mathbf{z}'\mathbf{P}^{-1}\mathbf{z} \xrightarrow{d} \chi^2(J). \tag{4}$$

Reassembling the parts from before, we have shown that the limiting distribution of

$$n(\mathbf{Rb}-\mathbf{q})'[\mathbf{R}(\sigma^2\mathbf{Q}^{-1})\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q}) \tag{5}$$

is chi-squared, with J *degrees of freedom. Note the similarity of this result to the results of Section B.11.6. Finally, if*

$$\text{plim } s^2\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} = \sigma^2\mathbf{Q}^{-1}, \tag{6}$$

then the statistic obtained by replacing $\sigma^2\mathbf{Q}^{-1}$ *by* $s^2(\mathbf{X}'\mathbf{X}/n)^{-1}$ *in (5) has the same limiting distribution. The n's cancel, and we are left with the same Wald statistic we looked at before. This step completes the proof.*

The appropriate critical values for the F test of the restrictions $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$ converge from above to $1/J$ times those for a chi-squared test based on the Wald statistic (see the Appendix tables). For example, for testing $J = 5$ restrictions, the critical value from the chi-squared table (Appendix Table G.4) for 95 percent significance is 11.07. The critical values from the F table (Appendix Table G.5) are $3.33 = 16.65/5$ for $n - K = 10$, $2.60 = 13.00/5$ for $n - K = 25$, $2.40 = 12.00/5$ for $n - K = 50$, $2.31 = 11.55/5$ for $n - K = 100$, and $2.214 = 11.07/5$ for large $n - K$. Thus, with normally distributed disturbances, as n gets large, the F test can be carried out by referring JF to the critical values from the chi-squared table.

The crucial result for our purposes here is that the distribution of the Wald statistic is built up from the distribution of $\mathbf{b}$, which is asymptotically normal even without normally distributed disturbances. The implication is that an appropriate large sample test statistic is chi-squared $= JF$. Once again, this implication relies on the central limit theorem, not on normally distributed disturbances. Now, what is the appropriate approach for a small or moderately sized sample? As we saw earlier, the critical values for the F distribution converge from above to $(1/J)$ times those for the preceding chi-squared distribution. As before, one cannot say that this will always be true in every case for every possible configuration of the data and parameters. Without some special configuration of the data and parameters, however, one, can expect it to occur generally. The implication is that absent some additional firm characterization of the model, the F statistic, with the critical values from the F table, remains a conservative approach that becomes more accurate as the sample size increases.

Exercise 7 at the end of this chapter suggests another approach to testing that has validity in large samples, a **Lagrange multiplier test.** The vector of Lagrange multipliers in (5-23) is $[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})$, that is, a multiple of the least squares discrepancy vector. In principle, a test of the hypothesis that $\boldsymbol{\lambda}_*$ equals zero should be equivalent to a test of the null hypothesis. Since the leading matrix has full rank, this can only equal zero if the discrepancy equals zero. A Wald test of the hypothesis that $\boldsymbol{\lambda}_* = \mathbf{0}$ is indeed a valid way to proceed. The large sample distribution of the Wald statistic would be chi-squared with J degrees of freedom. (The procedure is considered in Exercise 7.) For a set of exclusion restrictions, $\boldsymbol{\beta}_2 = \mathbf{0}$, there is a simple way to carry out this test. The chi-squared statistic, in this case with K_2 degrees of freedom can be computed as nR^2 in the regression of $\mathbf{e}_*$ (the residuals in the short regression) on the full set of independent variables.

5.7 TESTING NONLINEAR RESTRICTIONS

The preceding discussion has relied heavily on the linearity of the regression model. When we analyze nonlinear functions of the parameters and nonlinear regression models, most of these exact distributional results no longer hold.

The general problem is that of testing a hypothesis that involves a nonlinear function of the regression coefficients:

$$H_0 : c(\boldsymbol{\beta}) = q.$$

We shall look first at the case of a single restriction. The more general one, in which $\mathbf{c}(\boldsymbol{\beta}) = \mathbf{q}$ is a set of restrictions, is a simple extension. The counterpart to the test statistic

we used earlier would be

$$z = \frac{c(\hat{\boldsymbol{\beta}}) - q}{\text{estimated standard error}} \tag{5-33}$$

or its square, which in the preceding were distributed as $t[n - K]$ and $F[1, n - K]$, respectively. The discrepancy in the numerator presents no difficulty. Obtaining an estimate of the sampling variance of $c(\hat{\boldsymbol{\beta}}) - q$, however, involves the variance of a nonlinear function of $\hat{\boldsymbol{\beta}}$.

The results we need for this computation are presented in Sections 4.4.4, B.10.3, and D.3.1. A linear Taylor series approximation to $c(\hat{\boldsymbol{\beta}})$ around the true parameter vector $\boldsymbol{\beta}$ is

$$c(\hat{\boldsymbol{\beta}}) \approx c(\boldsymbol{\beta}) + \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{5-34}$$

We must rely on consistency rather than unbiasedness here, since, in general, the expected value of a nonlinear function is not equal to the function of the expected value. If plim $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, then we are justified in using $c(\hat{\boldsymbol{\beta}})$ as an estimate of $c(\boldsymbol{\beta})$. (The relevant result is the Slutsky theorem.) Assuming that our use of this approximation is appropriate, the variance of the nonlinear function is approximately equal to the variance of the right-hand side, which is, then,

$$\text{Var}[c(\hat{\boldsymbol{\beta}})] \approx \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)' \text{Var}[\hat{\boldsymbol{\beta}}] \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right). \tag{5-35}$$

The derivatives in the expression for the variance are functions of the unknown parameters. Since these are being estimated, we use our sample estimates in computing the derivatives. To estimate the variance of the estimator, we can use $s^2(\mathbf{X}'\mathbf{X})^{-1}$. Finally, we rely on Theorem D.22 in Section D.3.1 and use the standard normal distribution instead of the t distribution for the test statistic. Using $\mathbf{g}(\hat{\boldsymbol{\beta}})$ to estimate $\mathbf{g}(\boldsymbol{\beta}) = \partial c(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$, we can now test a hypothesis in the same fashion we did earlier.

Example 5.6 A Long-Run Marginal Propensity to Consume

A consumption function that has different short- and long-run marginal propensities to consume can be written in the form

$$\ln C_t = \alpha + \beta \ln Y_t + \gamma \ln C_{t-1} + \varepsilon_t,$$

which is a **distributed lag** model. In this model, the short-run marginal propensity to consume (MPC) (elasticity, since the variables are in logs) is β, and the long-run MPC is $\delta = \beta/(1 - \gamma)$. Consider testing the hypothesis that $\delta = 1$.

Quarterly data on aggregate U.S. consumption and disposable personal income for the years 1950 to 2000 are given in Appendix Table F5.2. The estimated equation based on these data is

$$\ln C_t = \underset{(0.01055)}{0.003142} + \underset{(0.02873)}{0.07495} \ln Y_t + \underset{(0.02859)}{0.9246} \ln C_{t-1} + e_t, \quad R^2 = 0.999712, \quad s = 0.00874$$

Estimated standard errors are shown in parentheses. We will also require Est. Asy. Cov[b, c] $= -0.0008207$. The estimate of the long-run MPC is $d = b/(1 - c) = 0.07495/(1 - 0.9246) = 0.99403$. To compute the estimated variance of d, we will require

$$g_b = \frac{\partial d}{\partial b} = \frac{1}{1 - c} = 13.2626, \quad g_c = \frac{\partial d}{\partial c} = \frac{b}{(1 - c)^2} = 13.1834.$$

The estimated asymptotic variance of d is

$$\begin{aligned}\text{Est. Asy. Var}[d] &= g_b^2\text{Est. Asy. Var}[b] + g_c^2\text{Est. Asy. Var}[c] + 2g_bg_c\text{Est. Asy. Cov}[b, c]\\ &= 13.2626^2 \times 0.02873^2 + 13.1834^2 \times 0.02859^2\\ &\quad + 2(13.2626)(13.1834)(-0.0008207) = 0.0002585.\end{aligned}$$

The square root is 0.016078. To test the hypothesis that the long-run MPC is greater than or equal to 1, we would use

$$z = \frac{0.99403 - 1}{0.016078} = -0.37131.$$

Because we are using a large sample approximation, we refer to a standard normal table instead of the t distribution. The hypothesis that $\gamma = 1$ is not rejected.

You may have noticed that we could have tested this hypothesis with a linear restriction instead; if $\delta = 1$, then $\beta = 1 - \gamma$, or $\beta + \gamma = 1$. The estimate is $q = b + c - 1 = -0.00045$. The estimated standard error of this linear function is $[0.02873^2 + 0.02859^2 - 2(0.0008207)]^{1/2} = 0.00118$. The t ratio for this test is -0.38135, which is almost the same as before. Since the sample used here is fairly large, this is to be expected. However, there is nothing in the computations that ensures this outcome. In a smaller sample, we might have obtained a different answer. For example, using the last 11 years of the data, the t statistics for the two hypotheses are 7.652 and 5.681. The Wald test is not invariant to how the hypothesis is formulated. In a borderline case, we could have reached a different conclusion. This **lack of invariance** does not occur with the likelihood ratio or Lagrange multiplier tests discussed in Chapter 16. On the other hand, both of these tests require an assumption of normality, whereas the Wald statistic does not. This illustrates one of the trade-offs between a more detailed specification and the power of the test procedures that are implied.

The generalization to more than one function of the parameters proceeds along similar lines. Let $\mathbf{c}(\hat{\boldsymbol{\beta}})$ be a set of J functions of the estimated parameter vector and let the $J \times K$ matrix of derivatives of $\mathbf{c}(\hat{\boldsymbol{\beta}})$ be

$$\hat{\mathbf{G}} = \frac{\partial \mathbf{c}(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}'}. \tag{5-36}$$

The estimate of the asymptotic covariance matrix of these functions is

$$\text{Est. Asy. Var}[\hat{\mathbf{c}}] = \hat{\mathbf{G}}\{\text{Est. Asy. Var}[\hat{\boldsymbol{\beta}}]\}\hat{\mathbf{G}}'. \tag{5-37}$$

The jth row of $\hat{\mathbf{G}}$ is K derivatives of c_j with respect to the K elements of $\hat{\boldsymbol{\beta}}$. For example, the covariance matrix for estimates of the short- and long-run marginal propensities to consume would be obtained using

$$\mathbf{G} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1/(1-\gamma) & \beta/(1-\gamma)^2 \end{bmatrix}.$$

The statistic for testing the J hypotheses $\mathbf{c}(\boldsymbol{\beta}) = \mathbf{q}$ is

$$W = (\hat{\mathbf{c}} - \mathbf{q})'\{\text{Est. Asy. Var}[\hat{\mathbf{c}}]\}^{-1}(\hat{\mathbf{c}} - \mathbf{q}). \tag{5-38}$$

In large samples, W has a chi-squared distribution with degrees of freedom equal to the number of restrictions. Note that for a single restriction, this value is the square of the statistic in (5-33).

5.8 CHOOSING BETWEEN NONNESTED MODELS

The classical testing procedures that we have been using have been shown to be most powerful for the types of hypotheses we have considered.[7] Although use of these procedures is clearly desirable, the requirement that we express the hypotheses in the form of restrictions on the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{q}$$

versus

$$H_1 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{q},$$

can be limiting. Two common exceptions are the general problem of determining which of two possible sets of regressors is more appropriate and whether a linear or loglinear model is more appropriate for a given analysis. For the present, we are interested in comparing two competing linear models:

$$H_0 : \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0 \qquad \textbf{(5-39a)}$$

and

$$H_1 : \mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_1. \qquad \textbf{(5-39b)}$$

The classical procedures we have considered thus far provide no means of forming a preference for one model or the other. The general problem of testing nonnested hypotheses such as these has attracted an impressive amount of attention in the theoretical literature and has appeared in a wide variety of empirical applications.[8]

5.8.1 TESTING NONNESTED HYPOTHESES

A useful distinction between hypothesis testing as discussed in the preceding chapters and model selection as considered here will turn on the asymmetry between the null and alternative hypotheses that is a part of the classical testing procedure.[9] Because, by construction, the classical procedures seek evidence in the sample to refute the "null" hypothesis, how one frames the null can be crucial to the outcome. Fortunately, the Neyman–Pearson methodology provides a prescription; the null is usually cast as the narrowest model in the set under consideration. On the other hand, the classical procedures never reach a sharp conclusion. Unless the significance level of the testing procedure is made so high as to exclude all alternatives, there will always remain the possibility of a Type 1 error. As such, the null hypothesis is never rejected with certainty, but only with a prespecified degree of confidence. Model selection tests, in contrast, give the competing hypotheses equal standing. There is no natural null hypothesis. However, the end of the process is a firm decision—in testing (5-39a, b), one of the models will be rejected and the other will be retained; the analysis will then proceed in

[7] See, for example, Stuart and Ord (1989, Chap. 27).

[8] Surveys on this subject are White (1982a, 1983), Gourieroux and Monfort (1994), McAleer (1995), and Pesaran and Weeks (2001). McAleer's survey tabulates an array of applications, while Gourieroux and Monfort focus on the underlying theory.

[9] See Granger and Pesaran (2000) for discussion.

the framework of that one model and not the other. Indeed, it cannot proceed until one of the models is discarded. It is common, for example, in this new setting for the analyst first to test with one model cast as the null, then with the other. Unfortunately, given the way the tests are constructed, it can happen that both or neither model is rejected; in either case, further analysis is clearly warranted. As we shall see, the science is a bit inexact.

The earliest work on nonnested hypothesis testing, notably Cox (1961, 1962), was done in the framework of sample likelihoods and maximum likelihood procedures. Recent developments have been structured around a common pillar labeled the **encompassing principle** [Mizon and Richard (1986)]. In the large, the principle directs attention to the question of whether a maintained model can explain the features of its competitors, that is, whether the maintained model encompasses the alternative. Yet a third approach is based on forming a **comprehensive model** that contains both competitors as special cases. When possible, the test between models can be based, essentially, on classical (-like) testing procedures. We will examine tests that exemplify all three approaches.

5.8.2 AN ENCOMPASSING MODEL

The encompassing approach is one in which the ability of one model to explain features of another is tested. Model 0 "encompasses" Model 1 if the features of Model 1 can be explained by Model 0, but the reverse is not true.[10] Because H_0 cannot be written as a restriction on H_1, none of the procedures we have considered thus far is appropriate. One possibility is an artificial nesting of the two models. Let $\overline{\mathbf{X}}$ be the set of variables in $\mathbf{X}$ that are not in $\mathbf{Z}$, define $\overline{\mathbf{Z}}$ likewise with respect to $\mathbf{X}$, and let $\mathbf{W}$ be the variables that the models have in common. Then H_0 and H_1 could be combined in a "supermodel":

$$\mathbf{y} = \overline{\mathbf{X}}\,\overline{\boldsymbol{\beta}} + \overline{\mathbf{Z}}\,\overline{\boldsymbol{\gamma}} + \mathbf{W}\boldsymbol{\delta} + \boldsymbol{\varepsilon}.$$

In principle, H_1 is rejected if it is found that $\overline{\boldsymbol{\gamma}} = \mathbf{0}$ by a conventional F test, whereas H_0 is rejected if it is found that $\overline{\boldsymbol{\beta}} = \mathbf{0}$. There are two problems with this approach. First, $\boldsymbol{\delta}$ remains a mixture of parts of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and it is not established by the F test that either of these parts is zero. Hence, this test does not really distinguish between H_0 and H_1; it distinguishes between H_1 and a hybrid model. Second, this compound model may have an extremely large number of regressors. In a time-series setting, the problem of collinearity may be severe.

Consider an alternative approach. If H_0 is correct, then $\mathbf{y}$ will, apart from the random disturbance $\boldsymbol{\varepsilon}$, be fully explained by $\mathbf{X}$. Suppose we then attempt to estimate $\boldsymbol{\gamma}$ by regression of $\mathbf{y}$ on $\mathbf{Z}$. Whatever set of parameters is estimated by this regression, say, $\mathbf{c}$, if H_0 is correct, then we should estimate exactly the same coefficient vector if we were to regress $\mathbf{X}\boldsymbol{\beta}$ on $\mathbf{Z}$, since $\boldsymbol{\varepsilon}_0$ is random noise under H_0. Because $\boldsymbol{\beta}$ must be estimated, suppose that we use $\mathbf{Xb}$ instead and compute $\mathbf{c}_0$. A test of the proposition that Model 0 "encompasses" Model 1 would be a test of the hypothesis that $E[\mathbf{c} - \mathbf{c}_0] = \mathbf{0}$. It is straightforward to show [see Davidson and MacKinnon (2004, pp. 671–672)] that the test can be carried out by using a standard F test to test the hypothesis that $\boldsymbol{\gamma}_1 = \mathbf{0}$

[10] See Deaton (1982), Dastoor (1983), Gourieroux et al. (1983, 1995) and, especially, Mizon and Richard (1986).

in the augmented regression,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}_1,$$

where $\mathbf{Z}_1$ is the variables in $\mathbf{Z}$ that are not in $\mathbf{X}$. (Of course, a line of manipulation reveals that $\overline{\mathbf{Z}}$ and $\mathbf{Z}_1$ are the same, so the tests are also.)

5.8.3 COMPREHENSIVE APPROACH—THE *J* TEST

The underpinnings of the comprehensive approach are tied to the density function as the characterization of the data generating process. Let $f_0(y_i \mid data, \boldsymbol{\beta}_0)$ be the assumed density under Model 0 and define the alternative likewise as $f_1(y_i \mid data, \boldsymbol{\beta}_1)$. Then, a comprehensive model which subsumes both of these is

$$f_c(y_i \mid data, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1) = \frac{[f_0(y_i \mid data, \boldsymbol{\beta}_0)]^{1-\lambda}[f_1(y_i \mid data, \boldsymbol{\beta}_1)]^{\lambda}}{\int_{\text{range of } y_i}[f_0(y_i \mid data, \boldsymbol{\beta}_0)]^{1-\lambda}[f_1(y_i \mid data, \boldsymbol{\beta}_1)]^{\lambda}\, dy_i}.$$

Estimation of the comprehensive model followed by a test of $\lambda = 0$ or 1 is used to assess the validity of Model 0 or 1, respectively.[11]

The ***J* test** proposed by Davidson and MacKinnon (1981) can be shown [see Pesaran and Weeks (2001)] to be an application of this principle to the linear regression model. Their suggested alternative to the preceding compound model is

$$\mathbf{y} = (1-\lambda)\mathbf{X}\boldsymbol{\beta} + \lambda(\mathbf{Z}\boldsymbol{\gamma}) + \boldsymbol{\varepsilon}.$$

In this model, a test of $\lambda = 0$ would be a test against H_1. The problem is that λ cannot be separately estimated in this model; it would amount to a redundant scaling of the regression coefficients. Davidson and MacKinnon's J test consists of estimating $\boldsymbol{\gamma}$ by a least squares regression of $\mathbf{y}$ on $\mathbf{Z}$ followed by a least squares regression of $\mathbf{y}$ on $\mathbf{X}$ and $\mathbf{Z}\hat{\boldsymbol{\gamma}}$, the fitted values in the first regression. A valid test, at least asymptotically, of H_1 is to test $H_0 : \lambda = 0$. If H_0 is true, then plim $\hat{\lambda} = 0$. Asymptotically, the ratio $\hat{\lambda}/\text{se}(\hat{\lambda})$ (i.e., the usual t ratio) is distributed as standard normal and may be referred to the standard table to carry out the test. Unfortunately, in testing H_0 versus H_1 and vice versa, all four possibilities (reject both, neither, or either one of the two hypotheses) could occur. This issue, however, is a finite sample problem. Davidson and MacKinnon show that as $n \to \infty$, if H_1 is true, then the probability that $\hat{\lambda}$ will differ significantly from 0 approaches 1.

Example 5.7* J *Test for a Consumption Function

Gaver and Geisel (1974) propose two forms of a consumption function:

$$H_0 : C_t = \beta_1 + \beta_2 Y_t + \beta_3 Y_{t-1} + \varepsilon_{0t}$$

and

$$H_1 : C_t = \gamma_1 + \gamma_2 Y_t + \gamma_3 C_{t-1} + \varepsilon_{1t}.$$

The first model states that consumption responds to changes in income over two periods, whereas the second states that the effects of changes in income on consumption persist for many periods. Quarterly data on aggregate U.S. real consumption and real disposable income are given in Appendix Table F5.2. Here we apply the J test to these data and the two proposed specifications. First, the two models are estimated separately (using observations

[11] Silva (2001) presents an application to the choice of probit or logit model for binary choice.

1950.2 through 2000.4). The least squares regression of C on a constant, Y, lagged Y, and the fitted values from the second model produces an estimate of λ of 1.0145 with a t ratio of 62.861. Thus, H_0 should be rejected in favor of H_1. But reversing the roles of H_0 and H_1, we obtain an estimate of λ of -10.677 with a t ratio of -7.188. Thus, H_1 is rejected as well.[12]

5.9 A SPECIFICATION TEST

The tests considered so far have evaluated nested models. The presumption is that one of the two models is correct. In Section 5.8, we broadened the range of models considered to allow two nonnested models. It is not assumed that either model is necessarily the true data generating process; the test attempts to ascertain which of two competing models is closer to the truth. Specification tests fall between these two approaches. The idea of a **specification test** is to consider a particular null model and alternatives that are not explicitly given in the form of restrictions on the regression equation. A useful way to consider some specification tests is as if the core model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is the null hypothesis and the alternative is a possibly unstated generalization of that model. Ramsey's (1969) **RESET test** is one such test which seeks to uncover nonlinearities in the functional form. One (admittedly ambiguous) way to frame the analysis is

H_0: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,
H_1: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} +$ higher order powers of x_k and other terms $+ \boldsymbol{\varepsilon}$.

A straightforward approach would be to add squares, cubes, and cross products of the regressors to the equation and test down to H_0 as a restriction on the larger model. Two complications are that this approach might be too specific about the form of the alternative hypothesis and, second, with a large number of variables in $\mathbf{X}$, it could become unwieldy. Ramsey's proposed solution is to add powers of $\mathbf{x}_i'\boldsymbol{\beta}$ to the regression using the least squares predictions—typically, one would add the square and, perhaps the cube. This would require a two-step estimation procedure, since in order to add $(\mathbf{x}_i'\mathbf{b})^2$ and $(\mathbf{x}_i'\mathbf{b})^3$, one needs the coefficients. The suggestion, then, is to fit the null model first, using least squares. Then, for the second step, the squares (and cubes) of the predicted values from this first-step regression are added to the equation and it is refit with the additional variables. A (large-sample) Wald test is then used to test the hypothesis of the null model.

As a general strategy, this sort of specification is designed to detect failures of the assumptions of the null model. The obvious virtue of such a test is that it provides much greater generality than a simple test of restrictions such as whether a coefficient is zero. But, that generality comes at considerable cost:

1. The test is nonconstructive. It gives no indication what the researcher should do next if the null model is rejected. This is a general feature of specification tests. Rejection of the null model does not imply any particular alternative.
2. Since the alternative hypothesis is unstated, it is unclear what the power of this test is against any specific alternative.
3. For this specific test (perhaps not for some other specification tests we will examine later), because $\mathbf{x}_i'\mathbf{b}$ uses the same $\mathbf{b}$ for every observation, the observations are

[12] For related discussion of this possibility, see McAleer, Fisher, and Volker (1982).

correlated, while they are assumed to be uncorrelated in the original model. Because of the two-step nature of the estimator, it is not clear what is the appropriate covariance matrix to use for the Wald test. Two other complications emerge for this test. First, it is unclear what γ converges to, assuming it converges to anything. Second, variance of the difference between $\mathbf{x}_i'\mathbf{b}$ and $\mathbf{x}_i'\boldsymbol{\beta}$ is a function of x, so the second-step regression might be heteroscedastic. The implication is that neither the size nor the power of this test is necessarily what might be expected.

Example 5.8 Size of a RESET Test

To investigate the true size of the RESET test in a particular application, we carried out a Monte Carlo experiment. The results in Table 4.6 give the following estimates of equation (5-2):

$$\ln Price = -8.42653 + 1.33372 \ln Area - 0.16537 Aspect\ Ratio + e \text{ where } sd(e) = 1.10266.$$

We take the estimated right-hand side to be our population. We generated 5,000 samples of 430 (the original sample size), by reusing the regression coefficients and generating a new sample of disturbances for each replication. Thus, with each replication, r, we have a new sample of observations on $\ln Price_{ir}$ where the regression part is as above reused and a new set of disturbances is generated each time. With each sample, we computed the least squares coefficient, then the predictions. We then recomputed the least squares regression while adding the square and cube of the prediction to the regression. Finally, with each sample, we computed the chi-squared statistic, and rejected the null model if the chi-squared statistic is larger than 5.99, the 95th percentile of the chi-squared distribution with two degrees of freedom. The **nominal size** of this test is 0.05. Thus, in samples of 100, 500, 1,000, and 5,000, we should reject the null nodel 5, 25, 50, and 250 times. In our experiment, the computed chi-squared exceeded 5.99 8, 31, 65, and 259 times, respectively, which suggests that at least with sufficient replications, the test performs as might be expected. We then investigated the power of the test by adding 0.1 times the square of ln *Area* to the predictions. It is not possible to deduce the exact power of the RESET test to detect this failure of the null model. In our experiment, with 1,000 replications, the null hypothesis is rejected 321 times. We conclude that the procedure does appear have power to detect this failure of the model assumptions.

5.10 MODEL BUILDING—A GENERAL TO SIMPLE STRATEGY

There has been a shift in the general approach to model building in the past 20 years or so, partly based on the results in the previous two sections. With an eye toward maintaining simplicity, model builders would generally begin with a small specification and gradually build up the model ultimately of interest by adding variables. But, based on the preceding results, we can surmise that just about any criterion that would be used to decide whether to add a variable to a current specification would be tainted by the biases caused by the incomplete specification at the early steps. Omitting variables from the equation seems generally to be the worse of the two errors. Thus, the **simple-to-general** approach to model building has little to recommend it. Building on the work of Hendry [e.g., (1995)] and aided by advances in estimation hardware and software, researchers are now more comfortable beginning their specification searches with large elaborate models involving many variables and perhaps long and complex lag structures. The attractive strategy is then to adopt a **general-to-simple,** downward reduction of the

model to the preferred specification. [This approach has been completely automated in Hendry's *PCGets*(c) computer program. See, e.g., Hendry and Kotzis (2001).] Of course, this must be tempered by two related considerations. In the "kitchen sink" regression, which contains every variable that might conceivably be relevant, the adoption of a fixed probability for the Type I error, say, 5 percent, ensures that in a big enough model, some variables will appear to be significant, even if "by accident." Second, the problems of pretest estimation and **stepwise model building** also pose some risk of ultimately misspecifying the model. To cite one unfortunately common example, the statistics involved often produce unexplainable lag structures in dynamic models with many lags of the dependent or independent variables.

5.10.1 MODEL SELECTION CRITERIA

The preceding discussion suggested some approaches to model selection based on nonnested hypothesis tests. Fit measures and testing procedures based on the sum of squared residuals, such as R^2 and the Cox test, are useful when interest centers on the within-sample fit or within-sample prediction of the dependent variable. When the model building is directed toward forecasting, within-sample measures are not necessarily optimal. As we have seen, R^2 cannot fall when variables are added to a model, so there is a built-in tendency to overfit the model. This criterion may point us away from the best forecasting model, because adding variables to a model may increase the variance of the forecast error (see Section 4.6) despite the improved fit to the data. With this thought in mind, the **adjusted R^2**,

$$\overline{R}^2 = 1 - \frac{n-1}{n-K}(1-R^2) = 1 - \frac{n-1}{n-K}\left(\frac{\mathbf{e}'\mathbf{e}}{\sum_{i=1}^{n}(y_i-\overline{y})^2}\right), \tag{5-40}$$

has been suggested as a fit measure that appropriately penalizes the loss of degrees of freedom that result from adding variables to the model. Note that $\overline{R}^2$ may fall when a variable is added to a model if the sum of squares does not fall fast enough. (The applicable result appears in Theorem 3.7; $\overline{R}^2$ does not rise when a variable is added to a model unless the t ratio associated with that variable exceeds one in absolute value.) The adjusted R^2 has been found to be a preferable fit measure for assessing the fit of forecasting models. [See Diebold (2003), who argues that the simple R^2 has a downward bias as a measure of the out-of-sample, one-step-ahead prediction error variance.]

The adjusted R^2 penalizes the loss of degrees of freedom that occurs when a model is expanded. There is, however, some question about whether the penalty is sufficiently large to ensure that the criterion will necessarily lead the analyst to the correct model (assuming that it is among the ones considered) as the sample size increases. Two alternative fit measures that have seen suggested are the **Akaike Information Criterion,**

$$\text{AIC}(K) = s_y^2(1-R^2)e^{2K/n} \tag{5-41}$$

and the Schwarz or **Bayesian Information Criterion,**

$$\text{BIC}(K) = s_y^2(1-R^2)n^{K/n}. \tag{5-42}$$

(There is no degrees of freedom correction in s_y^2.) Both measures improve (decline) as R^2 increases (decreases), but, everything else constant, degrade as the model size increases. Like $\overline{R}^2$, these measures place a premium on achieving a given fit with a smaller

number of parameters per observation, K/n. Logs are usually more convenient; the measures reported by most software are

$$\text{AIC}(K) = \ln\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{2K}{n} \quad \textbf{(5-43)}$$

$$\text{BIC}(K) = \ln\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{K \ln n}{n}. \quad \textbf{(5-44)}$$

Both **prediction criteria** have their virtues, and neither has an obvious advantage over the other. [See Diebold (2003).] The **Schwarz criterion,** with its heavier penalty for degrees of freedom lost, will lean toward a simpler model. All else given, simplicity does have some appeal.

5.10.2 MODEL SELECTION

The preceding has laid out a number of choices for **model selection,** but, at the same time, has posed some uncomfortable propositions. The pretest estimation aspects of specification search are based on the model builder's knowledge of "the truth" and the consequences of failing to use that knowledge. While the cautions about blind search for statistical significance are well taken, it does seem optimistic to assume that the correct model is likely to be known with hard certainty at the outset of the analysis. The bias documented in (4-10) is well worth the modeler's attention. But, in practical terms, knowing anything about the magnitude presumes that we know what variables are in $\mathbf{X}_2$, which need not be the case. While we can agree that the model builder will omit income from a demand equation at their peril, we could also have some sympathy for the analyst faced with finding the right specification for their forecasting model among dozens of choices. The tests for nonnested models would seem to free the modeler from having to claim that the specified set of models contain "the truth." But, a moment's thought should suggest that the cost of this is the possibly deflated power of these procedures to point toward that truth, The J test may provide a sharp choice between two alternatives, but it neglects the third possibility, that both models are wrong. Vuong's test does but, of course, it suffers from the fairly large inconclusive region, which is a symptom of its relatively low power against many alternatives. The upshot of all of this is that there remains much to be accomplished in the area of model selection. Recent commentary has provided suggestions from two perspective, classical and Bayesian.

5.10.3 CLASSICAL MODEL SELECTION

Hansen (2005) lists four shortcomings of the methodology we have considered here:

1. parametric vision
2. assuming a true data generating process
3. evaluation based on fit
4. ignoring model uncertainty

All four of these aspects have framed the analysis of the preceding sections. Hansen's view is that the analysis considered here is too narrow and stands in the way of progress in model discovery.

All the model selection procedures considered here are based on the likelihood function, which requires a specific distributional assumption. Hansen argues for a focus, instead, on semiparametric structures. For regression analysis, this points toward generalized method of moments estimators. Casualties of this reorientation will be distributionally based test statistics such as the Cox and Vuong statistics, and even the AIC and BIC measures, which are transformations of the likelihood function. However, alternatives have been proposed [e.g, by Hong, Preston, and Shum (2000)]. The second criticism is one we have addressed. The assumed "true" model can be a straightjacket. Rather (he argues), we should view our specifications as approximations to the underlying true data generating process—this greatly widens the specification search, to one for a model which provides the best approximation. Of course, that now forces the question of what is "best." So far, we have focused on the likelihood function, which in the classical regression can be viewed as an increasing function of R^2. The author argues for a more "focused" information criterion (FIC) that examines directly the parameters of interest, rather than the fit of the model to the data. Each of these suggestions seeks to improve the process of model selection based on familiar criteria, such as test statistics based on fit measures and on characteristics of the model.

A (perhaps *the*) crucial issue remaining is uncertainty about the model itself. The search for the correct model is likely to have the same kinds of impacts on statistical inference as the search for a specification given the form of the model (see Sections 4.3.2 and 4.3.3). Unfortunately, incorporation of this kind of uncertainty in statistical inference procedures remains an unsolved problem. Hansen suggests one potential route would be the Bayesian model averaging methods discussed next although he does express some skepticism about Bayesian methods in general.

5.10.4 BAYESIAN MODEL AVERAGING

If we have doubts as to which of two models is appropriate, then we might well be convinced to concede that possibly neither one is really "the truth." We have painted ourselves into a corner with our "left or right" approach to testing. The Bayesian approach to this question would treat it as a problem of comparing the two hypotheses rather than testing for the validity of one over the other. We enter our sampling experiment with a set of prior probabilities about the relative merits of the two hypotheses, which is summarized in a "prior odds ratio," $P_{01} = \text{Prob}[H_0]/\text{Prob}[H_1]$. After gathering our data, we construct the Bayes factor, which summarizes the weight of the sample evidence in favor of one model or the other. After the data have been analyzed, we have our "posterior odds ratio," $P_{01} \mid \text{data} = \text{Bayes factor} \times P_{01}$. The upshot is that ex post, neither model is discarded; we have merely revised our assessment of the comparative likelihood of the two in the face of the sample data. Of course, this still leaves the specification question open. Faced with a choice among models, how can we best use the information we have? Recent work on **Bayesian model averaging** [Hoeting et al. (1999)] has suggested an answer.

An application by Wright (2003) provides an interesting illustration. Recent advances such as Bayesian VARs have improved the forecasting performance of econometric models. Stock and Watson (2001, 2004) report that striking improvements in predictive performance of international inflation can be obtained by averaging a large

number of forecasts from different models and sources. The result is remarkably consistent across subperiods and countries. Two ideas are suggested by this outcome. First, the idea of blending different models is very much in the spirit of Hansen's fourth point. Second, note that the focus of the improvement is not on the fit of the model (point 3), but its predictive ability. Stock and Watson suggested that simple equal-weighted averaging, while one could not readily explain why, seems to bring large improvements. Wright proposed Bayesian model averaging as a means of making the choice of the weights for the average more systematic and of gaining even greater predictive performance.

Leamer (1978) appears to be the first to propose Bayesian model averaging as a means of combining models. The idea has been studied more recently by Min and Zellner (1993) for output growth forecasting, Doppelhofer et al. (2000) for cross-country growth regressions, Koop and Potter (2004) for macroeconomic forecasts, and others. Assume that there are M models to be considered, indexed by $m = 1, \ldots, M$. For simplicity, we will write the mth model in a simple form, $f_m(\mathbf{y} \mid \mathbf{Z}, \boldsymbol{\theta}_m)$ where $f(.)$ is the density, $\mathbf{y}$ and $\mathbf{Z}$ are the data, and $\boldsymbol{\theta}_m$ is the parameter vector for model m. Assume, as well, that model m^* is the true model, unknown to the analyst. The analyst has priors π_m over the probabilities that model m is the correct model, so π_m is the prior probability that $m = m^*$. The posterior probabilities for the models are

$$\Pi_m = \text{Prob}(m = m^* \mid \mathbf{y}, \mathbf{Z}) = \frac{P(\mathbf{y}, \mathbf{Z} \mid m)\pi_m}{\sum_{r=1}^{M} P(\mathbf{y}, \mathbf{Z} \mid r)\pi_r}, \tag{5-45}$$

where $P(\mathbf{y}, \mathbf{Z} \mid m)$ is the marginal likelihood for the mth model,

$$P(\mathbf{y}, \mathbf{Z} \mid m) = \int_{\theta_m} P(\mathbf{y}, \mathbf{Z} \mid \theta_m, m)\, P(\theta_m) d\theta_m, \tag{5-46}$$

while $P(\mathbf{y}, \mathbf{Z} \mid \theta_m, m)$ is the conditional (on θ_m) likelihood for the mth model and $P(\theta_m)$ is the analyst's prior over the parameters of the mth model. This provides an alternative set of weights to the $\Pi_m = 1/M$ suggested by Stock and Watson. Let $\hat{\theta}_m$ denote the Bayesian estimate (posterior mean) of the parameters of model m. (See Chapter 16.) Each model provides an appropriate posterior forecast density, $f^*(\mathbf{y} \mid \mathbf{Z}, \hat{\theta}_m, m)$. The Bayesian model averaged forecast density would then be

$$\overline{f}^* = \sum_{m=1}^{M} f^*(\mathbf{y} \mid \mathbf{Z}, \hat{\theta}, m)\Pi_m. \tag{5-47}$$

A point forecast would be a similarly weighted average of the forecasts from the individual models.

Example 5.9 Bayesian Averaging of Classical Estimates

Many researchers have expressed skepticism of Bayesian methods because of the apparent arbitrariness of the specifications of prior densities over unknown parameters. In the Bayesian model averaging setting, the analyst requires prior densities over not only the model probabilities, π_m, but also the model specific parameters, θ_m. In their application, Doppelhofer, Miller, and Sala-i-Martin (2000) were interested in the appropriate set of regressors to include in a long-term macroeconomic (income) growth equation. With 32 candidates, M for their application was 2^{32} (minus one if the zero regressors model is ignored), or roughly four billion. Forming this many priors would be optimistic in the extreme. The authors proposed a novel method of weighting a large subset (roughly 21 million) of the 2^M possible (classical) least squares regressions. The weights are formed using a Bayesian procedure; however,

the estimates that are weighted are the classical least squares estimates. While this saves considerable computational effort, it still requires the computation of millions of least squares coefficient vectors. [See Sala-i-Martin (1997).] The end result is a model with 12 independent variables.

5.11 SUMMARY AND CONCLUSIONS

This chapter has focused on two uses of the linear regression model, hypothesis testing, and basic prediction. The central result for testing hypotheses is the F statistic. The F ratio can be produced in two equivalent ways; first, by measuring the extent to which the unrestricted least squares estimate differs from what a hypothesis would predict, and second, by measuring the loss of fit that results from assuming that a hypothesis is correct. We then extended the F statistic to more general settings by examining its large-sample properties, which allow us to discard the assumption of normally distributed disturbances and by extending it to nonlinear restrictions.

This is the last of five chapters that we have devoted specifically to the methodology surrounding the most heavily used tool in econometrics, the classical linear regression model. We began in Chapter 2 with a statement of the regression model. Chapter 3 then described computation of the parameters by least squares—a purely algebraic exercise. Chapter 4 reinterpreted least squares as an estimator of an unknown parameter vector and described the finite sample and large-sample characteristics of the sampling distribution of the estimator. Chapter 5 was devoted to building and sharpening the regression model, with statistical results for testing hypotheses about the underlying population. In this chapter, we have examined some broad issues related to model specification and selection of a model among a set of competing alternatives. The concepts considered here are tied very closely to one of the pillars of the paradigm of econometrics; Underlying the model is a theoretical construction, a set of true behavioral relationships that constitute *the model*. It is only on this notion that the concepts of bias and biased estimation and model selection make any sense—"bias" as a concept can only be described with respect to some underlying "model" against which an estimator can be said to be biased. That is, there must be a yardstick. This concept is a central result in the analysis of specification, where we considered the implications of underfitting (omitting variables) and overfitting (including superfluous variables) the model. We concluded this chapter (and our discussion of the classical linear regression model) with an examination of procedures that are used to choose among competing model specifications.

Key Terms and Concepts

- Acceptance region
- Adjusted R-squared
- Akaike Information Criterion
- Alternative hypothesis
- Bayesian model averaging
- Bayesian Information Criterion
- Biased estimator
- Comprehensive model
- Consistent
- Distributed lag
- Discrepancy vector
- Encompassing principle
- Exclusion restrictions
- Ex post forecast
- Functionally independent
- General nonlinear hypothesis
- General-to-simple strategy
- Inclusion of superfluous variables
- J test
- Lack of invariance

- Lagrange multiplier test
- Linear restrictions
- Mean squared error
- Model selection
- Nested
- Nested models
- Nominal size
- Nonnested
- Nonnested models
- Nonnormality
- Null hypothesis
- One-sided test
- Parameter space
- Power of a test
- Prediction criterion
- Prediction interval
- Prediction variance
- Rejection region
- Restricted least squares
- Root mean squared error
- Sample discrepancy
- Schwarz criterion
- Simple-to-general
- Size of the test
- Specification test
- Stepwise model building
- t ratio
- Testable implications
- Theil U statistic
- Wald criterion
- Wald distance
- Wald statistic
- Wald test

Exercises

1. A multiple regression of y on a constant x_1 and x_2 produces the following results: $\hat{y} = 4 + 0.4x_1 + 0.9x_2$, $R^2 = 8/60$, $\mathbf{e}'\mathbf{e} = 520$, $n = 29$,
$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 29 & 0 & 0 \\ 0 & 50 & 10 \\ 0 & 10 & 80 \end{bmatrix}.$$
Test the hypothesis that the two slopes sum to 1.
2. Using the results in Exercise 1, test the hypothesis that the slope on x_1 is 0 by running the restricted regression and comparing the two sums of squared deviations.
3. The regression model to be analyzed is $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, where $\mathbf{X}_1$ and $\mathbf{X}_2$ have K_1 and K_2 columns, respectively. The restriction is $\boldsymbol{\beta}_2 = \mathbf{0}$.
   a. Using (5-23), prove that the restricted estimator is simply $[\mathbf{b}_{1*}, \mathbf{0}]$, where $\mathbf{b}_{1*}$ is the least squares coefficient vector in the regression of $\mathbf{y}$ on $\mathbf{X}_1$.
   b. Prove that if the restriction is $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0$ for a nonzero $\boldsymbol{\beta}_2^0$, then the restricted estimator of $\boldsymbol{\beta}_1$ is $\mathbf{b}_{1*} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_2^0)$.
4. The expression for the restricted coefficient vector in (5-23) may be written in the form $\mathbf{b}_* = [\mathbf{I} - \mathbf{CR}]\mathbf{b} + \mathbf{w}$, where $\mathbf{w}$ does not involve $\mathbf{b}$. What is $\mathbf{C}$? Show that the covariance matrix of the restricted least squares estimator is
$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$$
and that this matrix may be written as
$$\text{Var}[\mathbf{b} \mid \mathbf{X}]\{[\text{Var}(\mathbf{b} \mid \mathbf{X})]^{-1} - \mathbf{R}'[\text{Var}(\mathbf{Rb}) \mid \mathbf{X}]^{-1}\mathbf{R}\}\text{Var}[\mathbf{b} \mid \mathbf{X}].$$
5. Prove the result that the restricted least squares estimator never has a larger covariance matrix than the unrestricted least squares estimator.
6. Prove the result that the R^2 associated with a restricted least squares estimator is never larger than that associated with the unrestricted least squares estimator. Conclude that imposing restrictions never improves the fit of the regression.
7. An alternative way to test the hypothesis $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$ is to use a Wald test of the hypothesis that $\boldsymbol{\lambda}_* = \mathbf{0}$, where $\boldsymbol{\lambda}_*$ is defined in (5-23). Prove that
$$\chi^2 = \boldsymbol{\lambda}_*'\{\text{Est. Var}[\boldsymbol{\lambda}_*]\}^{-1}\boldsymbol{\lambda}_* = (n - K)\left[\frac{\mathbf{e}_*'\mathbf{e}_*}{\mathbf{e}'\mathbf{e}} - 1\right].$$

Note that the fraction in brackets is the ratio of two estimators of σ^2. By virtue of (5-28) and the preceding discussion, we know that this ratio is greater than 1. Finally, prove that this test statistic is equivalent to JF, where J is the number of restrictions being tested and F is the conventional F statistic given in (5-16). Formally, the Lagrange multiplier test requires that the variance estimator be based on the restricted sum of squares, not the unrestricted. Then, the test statistic would be $\text{LM} = nJ/[(n-K)/F + J]$. See Godfrey (1988).

8. Use the test statistic defined in Exercise 7 to test the hypothesis in Exercise 1.
9. Prove that under the hypothesis that $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, the estimator

$$s_*^2 = \frac{(\mathbf{y} - \mathbf{X}\mathbf{b}_*)'(\mathbf{y} - \mathbf{X}\mathbf{b}_*)}{n - K + J},$$

where J is the number of restrictions, is unbiased for σ^2.

10. Show that in the multiple regression of $\mathbf{y}$ on a constant, $\mathbf{x}_1$ and $\mathbf{x}_2$ while imposing the restriction $\beta_1 + \beta_2 = 1$ leads to the regression of $\mathbf{y} - \mathbf{x}_1$ on a constant and $\mathbf{x}_2 - \mathbf{x}_1$.
11. Suppose the true regression model is given by (4-8). The result in (4-10) shows that if either $\mathbf{P}_{1.2}$ is nonzero or $\boldsymbol{\beta}_2$ is nonzero, then regression of $\mathbf{y}$ on $\mathbf{X}_1$ alone produces a biased and inconsistent estimator of $\boldsymbol{\beta}_1$. Suppose the objective is to forecast $\mathbf{y}$, not to estimate the parameters. Consider regression of $\mathbf{y}$ on $\mathbf{X}_1$ alone to estimate $\boldsymbol{\beta}_1$ with $\mathbf{b}_1$ (which is biased). Is the forecast of $\mathbf{y}$ computed using $\mathbf{X}_1\mathbf{b}_1$ also biased? Assume that $E[\mathbf{X}_2 \mid \mathbf{X}_1]$ is a linear function of $\mathbf{X}_1$. Discuss your findings generally. What are the implications for prediction when variables are omitted from a regression?
12. Compare the mean squared errors of b_1 and $b_{1.2}$ in Section 4.7.2. (*Hint*: The comparison depends on the data and the model parameters, but you can devise a compact expression for the two quantities.)
13. The log likelihood function for the linear regression model with normally distributed disturbances is shown in Example 4.6. Show that at the maximum likelihood estimators of $\mathbf{b}$ for $\boldsymbol{\beta}$ and $\mathbf{e}'\mathbf{e}/n$ for σ^2, the log likelihood is an increasing function of R^2 for the model.
14. Show that the model of the alternative hypothesis in Example 5.7 can be written

$$H_1\text{: } C_t = \theta_1 + \theta_2 Y_t + \theta_3 Y_{t-1} + \sum_{s=2}^{\infty} \theta_{s+2} Y_{t-s} + \varepsilon_{it} + \sum_{s=1}^{\infty} \lambda_s \varepsilon_{t-s}.$$

As such, it does appear that H_0 is a restriction on H_1. However, because there are an infinite number of constraints, this does not reduce the test to a standard test of restrictions. It does suggest the connections between the two formulations. (We will revisit models of this sort in Chapter 21.)

Applications

1. The application in Chapter 3 used 15 of the 17,919 observations in Koop and Tobias's (2004) study of the relationship between wages and education, ability, and family characteristics. (See Appendix Table F3.2.) We will use the full data set for this exercise. The data may be downloaded from the *Journal of Applied Econometrics* data archive at http://www.econ.queensu.ca/jae/12004-vl9.7/koop-tobias/. The

data file is in two parts. The first file contains the panel of 17,919 observations on variables:

Column 1; *Person id* (ranging from 1 to 2,178),
Column 2; *Education,*
Column 3; *Log of hourly wage,*
Column 4; *Potential experience,*
Column 5; *Time trend*.

Columns 2–5 contain time varying variables. The second part of the data set contains time invariant variables for the 2,178 households. These are

Column 1; *Ability,*
Column 2; *Mother's education,*
Column 3; *Father's education,*
Column 4; *Dummy variable for residence in a broken home,*
Column 5; *Number of siblings*.

To create the data set for this exercise, it is necessary to merge these two data files. The ith observation in the second file will be replicated T_i times for the set of T_i observations in the first file. The *person id* variable indicates which rows must contain the data from the second file. (How this preparation is carried out will vary from one computer package to another.) (*Note*: We are not attempting to replicate Koop and Tobias's results here—we are only employing their interesting data set.) Let $\mathbf{X}_1 = [$*constant, education, experience, ability*$]$ and let $\mathbf{X}_2 = [$*mother's education, father's education, broken home, number of siblings*$]$.

a. Compute the full regression of log *wage* on $\mathbf{X}_1$ and $\mathbf{X}_2$ and report all results.
b. Use an F test to test the hypothesis that all coefficients except the constant term are zero.
c. Use an F statistic to test the joint hypothesis that the coefficients on the four household variables in $\mathbf{X}_2$ are zero.
d. Use a Wald test to carry out the test in part c.

2. The generalized Cobb–Douglas cost function examined in Application 2 in Chapter 4 is a special case of the **translog cost function,**

$$\begin{aligned}\ln C = \alpha &+ \beta \ln Q + \delta_k \ln P_k + \delta_l \ln P_l + \delta_f \ln P_f \\ &+ \phi_{kk}[\tfrac{1}{2}(\ln P_k)^2] + \phi_{ll}[\tfrac{1}{2}(\ln P_l)^2] + \phi_{ff}[\tfrac{1}{2}(\ln P_f)^2] \\ &+ \phi_{kl}[\ln P_k][\ln P_l] + \phi_{kf}[\ln P_k][\ln P_f] + \phi_{lf}[\ln P_l][\ln P_f] \\ &+ \gamma[\tfrac{1}{2}(\ln Q)^2] \\ &+ \theta_{Qk}[\ln Q][\ln P_k] + \theta_{Ql}[\ln Q][\ln P_l] + \theta_{Qf}[\ln Q][\ln P_k] + \varepsilon.\end{aligned}$$

The theoretical requirement of linear homogeneity in the factor prices imposes the following restrictions:

$$\delta_k + \delta_l + \delta_f = 1 \qquad \phi_{kk} + \phi_{kl} + \phi_{kf} = 0 \qquad \phi_{kl} + \phi_{ll} + \phi_{lf} = 0$$
$$\phi_{kf} + \phi_{lf} + \phi_{ff} = 0 \qquad \theta_{QK} + \theta_{Ql} + \theta_{Qf} = 0$$

Note that although the underlying theory requires it, the model can be estimated (by least squares) without imposing the linear homogeneity restrictions. [Thus, one

could "test" the underlying theory by testing the validity of these restrictions. See Christensen, Jorgenson, and Lau (1975).] We will repeat this exercise in part b.

A number of additional restrictions were explored in Christensen and Greene's (1976) study. The hypothesis of homotheticity of the production structure would add the additional restrictions

$$\theta_{Qk} = 0, \quad \theta_{Ql} = 0, \quad \theta_{Qf} = 0.$$

Homogeneity of the production structure adds the restriction $\gamma = 0$. The hypothesis that all elasticities of substitution in the production structure are equal to -1 is imposed by the six restrictions $\phi_{ij} = 0$ for all i and j.

We will use the data from the earlier application to test these restrictions. For the purposes of this exercise, denote by $\beta_1, \ldots, \beta_{15}$ the 15 parameters in the cost function above in the order that they appear in the model, starting in the first line and moving left to right and downward.

a. Write out the **R** matrix and **q** vector in (5-8) that are needed to impose the restriction of linear homogeneity in prices.
b. "Test" the theory of production using all 158 observations. Use an F test to test the restrictions of linear homogeneity. Note, you can use the general form of the F statistic in (5-16) to carry out the test. Christensen and Greene enforced the linear homogeneity restrictions by building them into the model. You can do this by dividing cost and the prices of capital and labor by the price of fuel. Terms with f subscripts fall out of the model, leaving an equation with 10 parameters. Compare the sums of squares for the two models to carry out the test. Of course, the test may be carried out either way and will produce the same result.
c. Test the hypothesis homotheticity of the production structure under the assumption of linear homogeneity in prices.
d. Test the hypothesis of the generalized Cobb–Douglas cost function in Chapter 4 against the more general translog model suggested here, once again (and henceforth) assuming linear homogeneity in the prices.
e. The simple Cobb–Douglas function appears in the first line of the model above. Test the hypothesis of the Cobb–Douglas model against the alternative of the full translog model.
f. Test the hypothesis of the generalized Cobb–Douglas model against the homothetic translog model.
g. Which of the several functional forms suggested here to you conclude is the most appropriate for these data?

3. The gasoline consumption model suggested in part d of Application 1 in Chapter 4 may be written as

$$\ln(G/Pop) = \alpha + \beta_P \ln P_g + \beta_l \ln (Income/Pop) + \gamma_{nc} \ln P_{nc} + \gamma_{uc} \ln P_{uc} + \gamma_{pt} \ln P_{pt} + \tau \text{year} + \delta_d \ln P_d + \delta_n \ln P_n + \delta_s \ln P_s + \varepsilon.$$

a. Carry out a test of the hypothesis that the three aggregate price indices are not significant determinants of the demand for gasoline.
b. Consider the hypothesis that the microelasticities are a constant proportion of the elasticity with respect to their corresponding aggregate. Thus, for some positive θ (presumably between 0 and 1), $\gamma_{nc} = \theta\delta_d$, $\gamma_{uc} = \theta\delta_d$, $\gamma_{pt} = \theta\delta_s$. The first

two imply the simple linear restriction $\gamma_{nc} = \gamma_{uc}$. By taking ratios, the first (or second) and third imply the nonlinear restriction

$$\frac{\gamma_{nc}}{\gamma_{pt}} = \frac{\delta_d}{\delta_s} \qquad \text{or} \qquad \gamma_{nc}\delta_s - \gamma_{pt}\delta_d = 0.$$

Describe in detail how you would test the validity of the restriction.

c. Using the gasoline market data in Table F2.2, test the two restrictions suggested here, separately and jointly.

4. The J test in Example 5.7 is carried out using more than 50 years of data. It is optimistic to hope that the underlying structure of the economy did not change in 50 years. Does the result of the test carried out in Example 5.7 persist if it is based on data only from 1980 to 2000? Repeat the computation with this subset of the data.