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 be primarily 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 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 – i.e., the observed sale prices of paintings. In order to proceed, the obvious approach for the analyst would be to fit the regression first, 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, i.e., 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), the broader statement is (5-2) while the narrower one 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, and 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 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,

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

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

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

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,

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

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 lnArea 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 of our work in this book is based on consistent estimators and save for the latter sections of this chapter, 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, i.e., 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 the Bayesian analyst never reaches 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 their 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, $\beta_j = 0$

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

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

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

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

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

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

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

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 **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 **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 **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 an 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,

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

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,

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

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, 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

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

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

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

(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 if W_k is "large."

Before we determine a benchmark for large, we note that the Wald measure suggested here is not useable 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

$$t_k = \frac{b_k - \beta_k^0}{\sqrt{s^2 S^{kk}}} \qquad \textbf{(5-13)}$$

Assuming that β_k does indeed equal β_k^0, ie, "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%. 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 $100(1-\alpha)\%$, where we choose α, typically 5%. 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% 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 below. 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;

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

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.)

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	−8.23237e−5	9.73928e−7		
−0.0092609	5.08549e−5	−4.96761e−7	0.00063729	
0.026749	−0.0026412	3.84102e−5	−5.46193e−5	0.021766

5-11

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

$$\boldsymbol{\beta}_1 = \mathbf{R}_1^{-1}[\mathbf{q} - \mathbf{R}_2\boldsymbol{\beta}_2]. \tag{5-4}$$

The implication is that although $\boldsymbol{\beta}_2$ is free to vary, once $\boldsymbol{\beta}_2$ is determined, $\boldsymbol{\beta}_1$ is determined by (5-4). Thus, only the $K - J$ elements of $\boldsymbol{\beta}_2$ are free parameters in the restricted model.

5.3 TWO APPROACHES TO TESTING HYPOTHESES

Hypothesis testing of the sort suggested in the preceding section can be approached from two viewpoints. First, having computed a set of parameter estimates, we can ask whether the estimates come reasonably close to satisfying the restrictions implied by the hypothesis. More formally, we can ascertain whether the failure of the estimates to satisfy the restrictions is simply the result of sampling error or is instead systematic. An alternative approach might proceed as follows. Suppose that we impose the restrictions implied by the theory. Since unrestricted least squares is, by definition, "least squares," this imposition must lead to a loss of fit. We can then ascertain whether this loss of fit results merely from sampling error or whether it is so large as to cast doubt on the validity of the restrictions. We will consider these two approaches in turn, then show that (as one might hope) within the framework of the linear regression model, the two approaches are equivalent.

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 the next section, we will consider the implications of relaxing this assumption and develop an alternative set of results that allows us to proceed without it.

5.4.2 ~~5.3.1~~ 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,**

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

Each row of $\mathbf{R}$ is the coefficients in a linear restriction on the coefficient vector. 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, $\beta_j = 0$

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

Ae; KT "null hypothesis" already a KT in chapter. Here also?

2. Two of the coefficients are equal, $\beta_k = \beta_j$,
$$\mathbf{R} = [0 \quad 0 \quad 1 \quad \cdots \quad -1 \quad \cdots \quad 0] \quad \text{and} \quad \mathbf{q} = 0.$$
3. A set of the coefficients sum to one, $\beta_2 + \beta_3 + \beta_4 = 1$,
$$\mathbf{R} = [0 \quad 1 \quad 1 \quad 1 \quad 0 \quad \cdots] \quad \text{and} \quad \mathbf{q} = 1.$$
4. A subset of the coefficients are all zero, $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_3 = 0$,
$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} = [\mathbf{I} : \mathbf{0}] \quad \text{and} \quad \mathbf{q} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$
5. Several linear restrictions, $\beta_2 + \beta_3 = 1$, $\beta_4 + \beta_6 = 0$, and $\beta_5 + \beta_6 = 0$,
$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{q} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$
6. All the coefficients in the model except the constant term are zero. [See (4-15) and Section 4.7.4.]
$$\mathbf{R} = [\mathbf{0} : \mathbf{I}_{K-1}] \quad \text{and} \quad \mathbf{q} = \mathbf{0}.$$

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 (4-8)] 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:

$$\begin{aligned} W &= \mathbf{m}'\{\text{Var}[\mathbf{m} \mid \mathbf{X}]\}^{-1}\mathbf{m}. \\ &= (\mathbf{Rb} - \mathbf{q})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q}) \\ &= \frac{(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})}{\sigma^2} \\ &\sim \chi^2[J]. \end{aligned} \tag{5-8}$$

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.

[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 $\mathbf{X}$.

The chi-squared statistic in (5-5) 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 (5-5), dividing by J, and multiplying and dividing by $n-K$, we obtain

$$F = \frac{W}{J}\frac{\sigma^2}{s^2}$$

$$= \left(\frac{(\mathbf{Rb}-\mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})}{\sigma^2}\right)\left(\frac{1}{J}\right)\left(\frac{\sigma^2}{s^2}\right)\left(\frac{(n-K)}{(n-K)}\right) \qquad (5\text{-}6)$$

$$= \frac{(\mathbf{Rb}-\mathbf{q})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})/J}{[(n-K)s^2/\sigma^2]/(n-K)}.$$

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 (5-5) and is distributed as $1/J$ times a chi-squared $[J]$, as we showed earlier. We found in (4-6) 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 $[n-K]$. [See (4-11).] 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 $\mathbf{T}(\boldsymbol{\varepsilon}/\sigma)$ 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 (5-6) leaves the F statistic for testing a linear hypothesis:

$$F[J, n-K|\mathbf{X}] = \frac{(\mathbf{Rb}-\mathbf{q})'\left\{\mathbf{R}[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}'\right\}^{-1}(\mathbf{Rb}-\mathbf{q})}{J}. \qquad (5\text{-}7)$$

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 rs 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

$$t = \frac{\hat{q} - q}{\text{se}(\hat{q})}. \tag{5-8}$$

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

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

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 (5-7) and (5-8). We can write the square of the t statistic as

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

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 5.1 Restricted Investment Equation

Section 5.2 suggested a theory about the behavior of investors: that 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 (5-1) and (5-3) 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 F5.1 are presented in Table 5.1. (One observation is lost in computing the change in the CPI.)

TABLE 5.1 Estimated Investment Equations (Estimated standard errors in parentheses)

	β_1	β_2	β_3	β_4	β_5
Model (5-1)	−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 (5-3)	−7.907	−0.00443	0.00443	1.764	−0.00440
	(1.201)	(0.00227)	(0.00227)	(0.161)	(0.00133)
	$s = 0.8670$, $R^2 = 0.979405$, $\mathbf{e}'\mathbf{e} = 1.49578$				

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 (5-2). 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

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

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 (5-8) 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.

5.3.2 THE RESTRICTED LEAST SQUARES ESTIMATOR

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 in Example 5.1, 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.

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 **b** was chosen to minimize the sum of squared deviations, **e**′**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 **b**, it follows that **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^2s 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

$$\text{Minimize}_{\mathbf{b}_0}\ S(\mathbf{b}_0) = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'(\mathbf{y} - \mathbf{X}\mathbf{b}_0) \quad \text{subject to } \mathbf{R}\mathbf{b}_0 = \mathbf{q}. \tag{5-19}$$

A Lagrangean function for this problem can be written

$$L^*(\mathbf{b}_0, \boldsymbol{\lambda}) = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'(\mathbf{y} - \mathbf{X}\mathbf{b}_0) + 2\boldsymbol{\lambda}'(\mathbf{R}\mathbf{b}_0 - \mathbf{q}). \tag{5-20}$$

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

$$\frac{\partial L^*}{\partial \mathbf{b}_*} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}_*) + 2\mathbf{R}'\boldsymbol{\lambda}_* = \mathbf{0}$$
$$\frac{\partial L^*}{\partial \boldsymbol{\lambda}_*} = 2(\mathbf{R}\mathbf{b}_* - \mathbf{q}) = \mathbf{0}. \tag{5-21}$$

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

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{R}' \\ \mathbf{R} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b}_* \\ \boldsymbol{\lambda}_* \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{q} \end{bmatrix} \tag{5-22}$$

or

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

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

$$\mathbf{d}_* = \mathbf{A}^{-1}\mathbf{v}. \tag{5-23}$$

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]

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

and

$$\boldsymbol{\lambda}_* = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}). \tag{5-24}$$

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:

$$\text{Var}[\mathbf{b}_* \mid \mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}. \tag{5-25}$$

Thus,

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

[2] Since λ is not restricted, we can formulate the constraints in terms of 2λ. The convenience of the scaling shows up in (5-11).

[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).

AU: Confirm "(5-11)" is correct x-ref. Or should it be "(5-20)"?

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 $\mathbf{Rb} - \mathbf{q}$. 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 $\mathbf{b}$ plus 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, 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

$$R^2_{\mathbf{Xz}} = R^2_{\mathbf{X}} + (1 - R^2_{\mathbf{X}})r^{*2}_{yz}, \tag{5-16}$$

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

$$r^{*2}_{yz} = \frac{t_z^2}{t_z^2 + (n - K)}, \tag{5-17}$$

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 (5-16) for r^{*2}_{yz} and (5-17) for t_z^2 and then insert the first solution in the second, then we obtain the result

$$t_z^2 = \frac{(R^2_{\mathbf{Xz}} - R^2_{\mathbf{X}})/1}{(1 - R^2_{\mathbf{Xz}})/(n - K)}. \tag{5-18}$$

5.4.2

We saw at the end of Section 5.3.1 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 (5-18), 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 (5-8). 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 $\mathbf{y} - \mathbf{Xb}_*$. Then, using a familiar device,

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

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

$$\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e} = (\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q}). \tag{5-19}$$

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

$$F[J, n-K] = \frac{(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(n-K)}. \tag{5-20}$$

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

$$F[J, n-K] = \frac{(R^2 - R_*^2)/J}{(1-R^2)/(n-K)}. \tag{5-21}$$

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 earlier joint test that all of 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 (5-19) 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 5.2 Production Function

The data in Appendix Table F5.2 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 5.2.3 We will construct several hypothesis tests based on these results. A generalization of the

[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.

TABLE 5.2 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	2.397				
	(−0.01984)	(0.01586)				
$\ln K$	−0.3313	−1.231	1.033			
	(0.001189)	(−0.00961)	(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

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 5.2. 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,

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

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

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 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 (5-20)], then no problem will arise. If (5-21) 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.

5.4 NONNORMAL DISTURBANCES AND LARGE SAMPLE TESTS

The distributions of the F, t, and chi-squared statistics that we used in the previous section rely on the assumption of normally distributed disturbances. Without this assumption,

[6] This case is not true when the restrictions are nonlinear. We consider this issue in Chapter 10.

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^2s 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 (5-29). The general result from this example is that one must be careful in using (5-30) 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 $H_0 : \beta_k = \beta_k^0$,

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

Conditional on $\mathbf{X}$, if $\beta_k = \beta_k^0$ (i.e., under H_0), then $t \mid \mathbf{X}$ 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 $(b_k - \beta_k)/\sqrt{s^2 (\mathbf{X}'\mathbf{X})^{-1}_{kk}} \mid \mathbf{X}$ and $[(n - K)s_2/\sigma^2] \mid \mathbf{X}$ 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 $f(t \mid \mathbf{X})$ 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,

5-25

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

$$\mathbf{b} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}^{-1}\right] \quad \text{where } \mathbf{Q} = \operatorname{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right). \tag{5-22}$$

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 finite sample approximation above. 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

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

(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, ~~Section 4.7.5, and (4-17)~~.] The exact distribution of this statistic is unknown, however, if $\boldsymbol{\varepsilon}$ is not normally distributed. From the results above, we find that the denominator of t_k converges to $\sqrt{\sigma^2\mathbf{Q}^{-1}_{kk}}$. 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. That is, the large-sample distribution of t_k is the same as that of

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

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

5-26

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 above, 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 (5-6) by σ^2 and rearrange the fraction slightly, so

$$F = \frac{(\mathbf{Rb}-\mathbf{q})'\{\mathbf{R}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}'\}^{-1}(\mathbf{Rb}-\mathbf{q})}{J(s^2/\sigma^2)}. \qquad (5\text{-}23)$$

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 $(\sigma^2/n)\mathbf{R}\mathbf{Q}^{-1}\mathbf{R}'$. 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*

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

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 (5-14) 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

$$z = \frac{c(\hat{\boldsymbol{\beta}}) - q}{\text{estimated standard error}} \tag{5-24}$$

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 4.9.4, 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

$$c(\hat{\boldsymbol{\beta}}) \approx c(\boldsymbol{\beta}) + \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{5-25}$$

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 $\operatorname{plim} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, 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,

$$\operatorname{Var}[c(\hat{\boldsymbol{\beta}})] \approx \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)' \operatorname{Var}[\hat{\boldsymbol{\beta}}] \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right). \tag{5-26}$$

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 5.3 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 F5.1. 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

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

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 $\gamma \geq 1$ 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

$$\hat{\mathbf{G}} = \frac{\partial \mathbf{c}(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}'}. \tag{5-27}$$

The estimate of the asymptotic covariance matrix of these functions is

$$\text{Est. Asy. Var}[\hat{\mathbf{c}}] = \hat{\mathbf{G}}\{\text{Est. Asy. Var}[\hat{\boldsymbol{\beta}}]\}\hat{\mathbf{G}}'. \tag{5-28}$$

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

$$W = (\hat{\mathbf{c}} - \mathbf{q})'\{\text{Est. Asy. Var}[\hat{\mathbf{c}}]\}^{-1}(\hat{\mathbf{c}} - \mathbf{q}). \tag{5-29}$$

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 (5-24).

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 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.

5.8 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:

$$H_0 : \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0 \qquad (7\text{-}12a)$$

and

$$H_1 : \mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_1. \qquad (7\text{-}12b)$$

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]

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

[8] Recent 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.

7.3.1 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 (7-12a, b), one of the models will be rejected and the other will be retained; the analysis will then proceed in 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.

7.3.2 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":

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

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

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