6
INFERENCE AND PREDICTION

6.1 INTRODUCTION

The linear regression model is used for three major functions: estimation, which was the subject of the previous three chapters (and most of the rest of this book), hypothesis testing, and prediction or forecasting. In this chapter, we will examine some applications of hypothesis tests using the classical model. The basic statistical theory was developed in Chapters 4, 5, and Appendix C, so the methods discussed here will use tools that are already familiar. After the theory is developed in Sections 6.2–6.4, we will examine some applications in Sections 6.4 and 6.5. We will be primarily concerned with linear restrictions in this chapter, and will turn to nonlinear restrictions near the end of the chapter, in Section 6.5. Section 6.6 discusses the third major use of the regression model, prediction.

6.2 RESTRICTIONS AND NESTED MODELS

One common 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 simple model of investment, I_t, suggested by Section 3.3.2,

$$\ln I_t = \beta_1 + \beta_2 i_t + \beta_3 \Delta p_t + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t, \qquad \textbf{(6-1)}$$

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 which 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

$$\ln I_t = \beta_1 + \beta_2 (i_t - \Delta p_t) + \beta_3 \Delta p_t + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t. \qquad \textbf{(6-2)}$$

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,

$$\ln I_t = \beta_1 + \beta_2 (i_t - \Delta p_t) + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t, \qquad \textbf{(6-3)}$$

is now restricted; in the context of the first model, 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, equation (6-1) specifies a model in which there are five unrestricted parameters $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$. But, equation (6-3) 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.** We are concerned only with nested models in this chapter. Nonnested models are considered in Section 8.3.

Beginning with the linear regression model

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

we consider a set of **linear restrictions** of the form

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

These can be combined into the single equation

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{q}.$$

Each row of $\mathbf{R}$ is the coefficients in one of the restrictions. The matrix $\mathbf{R}$ has K columns to be conformable with $\boldsymbol{\beta}$, J rows for a total of J restrictions, and full row rank, so J must be less than or equal to K. The rows of $\mathbf{R}$ must be linearly independent. Although it does not violate the condition, the case of $J = K$ must also be ruled out.[1] 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. One way to view this situation is to partition $\mathbf{R}$ 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

[1] If the K slopes satisfy $J = K$ restriction, then $\mathbf{R}$ is square and nonsingular and $\boldsymbol{\beta} = \mathbf{R}^{-1}\mathbf{q}$. There is no estimation or inference problem.

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

$$\boldsymbol{\beta}_1 = \mathbf{R}_1^{-1}[\mathbf{q} - \mathbf{R}_2\boldsymbol{\beta}_2]. \tag{6-4}$$

The implication is that although $\boldsymbol{\beta}_2$ is free to vary, once $\boldsymbol{\beta}_2$ is determined, $\boldsymbol{\beta}_1$ is determined by (6-4). Thus, only the $K-J$ elements of $\boldsymbol{\beta}_2$ are free parameters in the restricted model.

6.3 TWO APPROACHES TO TESTING HYPOTHESES

Hypothesis testing of the sort suggested above 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.

6.3.1 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, $\beta_j = 0$

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

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

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

We can base a test of H_0 on the **Wald criterion:**

$$\begin{aligned} W &= \mathbf{m}'\big\{\text{Var}[\mathbf{m} \mid \mathbf{X}]\big\}^{-1}\mathbf{m}. \\ &= (\mathbf{Rb} - \mathbf{q})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q}) \qquad \textbf{(6-5)} \\ &= \frac{(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})}{\sigma^2} \\ &\sim \chi^2[J]. \end{aligned}$$

The statistic W has a chi-squared distribution with J degrees of freedom if the hypothesis is correct.[2] 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.

[2]This calculation is an application of the "full rank quadratic form" of Section B.10.5.

The chi-squared statistic in (6-5) 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 (6-5), dividing by J, and multiplying and dividing by $n-K$, we obtain

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

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

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

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 $\mathbf{C}=[\mathbf{R}(\mathbf{X}'\mathbf{X}^{-1}R']$ 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 (6-5) 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 (6-6) leaves the F statistic for testing a linear hypothesis:

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

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

$$F[1,n-K] = \frac{(\Sigma_j r_j b_j - q)^2}{\Sigma_j \Sigma_k r_j r_k \,\text{Est. Cov}[b_j,b_k]}. \qquad \textbf{(6-7)}$$

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, $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ 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

$$t = \frac{\hat{q} - q}{\text{se}(\hat{q})}. \tag{6-8}$$

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 (6-7) and (6-8). We can write the square of the t statistic as

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

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

Example 6.1 Restricted Investment Equation

Section 6.2 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 (6-1) and (6-3) 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 (see Appendix Table F5.1) are given in Table 6.1. (One observation is lost in computing the change in the CPI.)

TABLE 6.1 Estimated Investment Equations (Estimated standard errors in parentheses)

	β_1	β_2	β_3	β_4	β_5
Model (6-1)	−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 (6-3)	−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 (6-2). 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, 199] from the table is 2.60. 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 (6-8) 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.

6.3.2 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 6.1, 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.

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

$$\text{Minimize}_{\mathbf{b}_0}\ S(\mathbf{b}_0) = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'(\mathbf{y} - \mathbf{X}\mathbf{b}_0) \quad \text{subject to } \mathbf{R}\mathbf{b}_0 = \mathbf{q}. \tag{6-9}$$

A Lagrangean function for this problem can be written

$$L^*(\mathbf{b}_0, \boldsymbol{\lambda}) = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'(\mathbf{y} - \mathbf{X}\mathbf{b}_0) + 2\boldsymbol{\lambda}'(\mathbf{R}\mathbf{b}_0 - \mathbf{q}).^3 \tag{6-10}$$

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

$$\begin{aligned} \frac{\partial L^*}{\partial \mathbf{b}_*} &= -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}_*) + 2\mathbf{R}'\boldsymbol{\lambda}_* = \mathbf{0} \\ \frac{\partial L^*}{\partial \boldsymbol{\lambda}_*} &= 2(\mathbf{R}\mathbf{b}_* - \mathbf{q}) = \mathbf{0}. \end{aligned} \tag{6-11}$$

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

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{R}' \\ \mathbf{R} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b}_* \\ \boldsymbol{\lambda}_* \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{q} \end{bmatrix} \tag{6-12}$$

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

$$\mathbf{d}_* = \mathbf{A}^{-1}\mathbf{v}. \tag{6-13}$$

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),[4]

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

and

$$\boldsymbol{\lambda}_* = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}). \tag{6-14}$$

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:

$$\text{Var}[\mathbf{b}_* \mid \mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}. \tag{6-15}$$

Thus,

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

[3] Since $\boldsymbol{\lambda}$ is not restricted, we can formulate the constraints in terms of $2\boldsymbol{\lambda}$. Why this scaling is convenient will be clear shortly.

[4] 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).

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.

6.3.3 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

$$R^2_{\mathbf{Xz}} = R^2_{\mathbf{X}} + \left(1 - R^2_{\mathbf{X}}\right) r^{*2}_{yz}, \tag{6-16}$$

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 (3-23) we obtained the convenient result

$$r^{*2}_{yz} = \frac{t_z^2}{t_z^2 + (n - K)}, \tag{6-17}$$

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 (6-16) for r^{*2}_{yz} and (6-17) for t_z^2 and then insert the first solution in the second, then we obtain the result

$$t_z^2 = \frac{\left(R^2_{\mathbf{Xz}} - R^2_{\mathbf{X}}\right)/1}{\left(1 - R^2_{\mathbf{Xz}}\right)/(n - K)}. \tag{6-18}$$

We saw at the end of Section 6.3.1 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 (6-18), 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 (6-8). 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

$$\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e} = (\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q}). \tag{6-19}$$

This expression appears in the numerator of the F statistic in (6-7). Inserting the remaining parts, we obtain

$$F[J, n-K] = \frac{(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(n-K)}. \tag{6-20}$$

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

$$F[J, n-K] = \frac{(R^2 - R_*^2)/J}{(1-R^2)/(n-K)}. \tag{6-21}$$

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 discussed in Section 4.7.4 (4-15), 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 (6-19) 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 6.2 Production Function

The data in Appendix Table F6.1 have been used in several studies of production functions.[5] 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 6.2. We will construct several hypothesis tests based on these results. A generalization of the

[5]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 6.2 Estimated Production Functions

	Translog	*Cobb–Douglas*
Sum of squared residuals	0.67993	0.85163
Standard error of regression	0.17994	0.18840
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	3.583
$\ln L$	3.61363	1.548	2.334	0.6030	0.1260	4.787
$\ln K$	−1.89311	1.016	−1.863	0.3757	0.0853	4.402
$\frac{1}{2}\ln^2 L$	−0.96406	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*	$\frac{1}{2}$ *ln² L*	$\frac{1}{2}$ *ln² K*	*ln L ln K*
Constant	8.472					
	(0.1068)					
lnL	−2.388	2.397				
	(−0.01984)	(0.01586)				
lnK	−0.3313	−1.231	1.033			
	(0.00189)	(−.00961)	(0.00728)			
$\frac{1}{2}$***ln²L***	−0.08760	−0.6658	0.5231	0.5004		
$\frac{1}{2}$***ln²K***	0.2332	0.03477	0.02637	0.1467	0.08562	
lnL lnK	0.3635	0.1831	−0.2255	−0.2880	−0.1160	0.1927

Cobb–Douglas model is the *translog* model,[6] 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 14, 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 6.2. 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,

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

[6] Berndt and Christensen (1973). See Example 2.5 for discussion.

which is substantially less than the critical value given earlier. 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 1.8891, which is less than the critical value of 3.49. 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.[7] 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 (6-20)], then no problem will arise. If (6-21) 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 9.375. If we compute the appropriate R_*^2 using the correct denominator, however, then its value is 0.94339 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.

6.4 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,

[7] This case is not true when the restrictions are nonlinear. We consider this issue in Chapter 9.

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

$$\mathbf{b} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}^{-1}\right] \quad \text{where } \mathbf{Q} = \operatorname{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right). \tag{6-22}$$

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

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

(Note that two occurrences of $\sqrt{n}$ cancel to produce our familiar result.) Under the null hypothesis, with normally distributed disturbances, t_k is exactly distributed as t with $n-K$ degrees of freedom. [See Theorem 4.4 and (4-13).] 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 $\tau_k = \left(b_k - E[b_k]\right)/\left(\text{Asy. Var}[b_k]\right)^{1/2}$ from the asymptotic normal distribution (under the hypothesis $\beta_k = \beta_k^0$), so it follows that τ_k has a standard normal asymptotic distribution, and this result is the large-sample distribution of our t statistic. Thus, as a large-sample approximation, we will use the standard normal distribution to approximate

the true distribution of the test statistic t_k and use the critical values from the standard normal distribution for testing hypotheses.

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

We will use the same approach to analyze the F statistic for testing a set of J linear restrictions. Step 1 will be to show that with normally distributed disturbances, JF converges to a chi-squared variable as the sample size increases. We will then show that this result is actually independent of the normality of the disturbances; it relies on the central limit theorem. Finally, we consider, as 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 (6-6). 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 (6-6) by σ^2 and rearrange the fraction slightly, so

$$F = \frac{(\mathbf{Rb} - \mathbf{q})' \left\{ \mathbf{R}[\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}] \mathbf{R}' \right\}^{-1} (\mathbf{Rb} - \mathbf{q})}{J (s^2/\sigma^2)}. \tag{6-23}$$

Since plim $s^2 = \sigma^2$, and 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

$$\begin{aligned} W^* &= \frac{1}{J} (\mathbf{Rb} - \mathbf{q})' [\mathbf{R}(\sigma^2/n)\mathbf{Q}^{-1}\mathbf{R}']^{-1} (\mathbf{Rb} - \mathbf{q}) \\ &= \frac{1}{J} (\mathbf{Rb} - \mathbf{q})' \left\{ \text{Asy. Var}[\mathbf{Rb} - \mathbf{q}] \right\}^{-1} (\mathbf{Rb} - \mathbf{q}). \end{aligned}$$

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 6.1 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* $\mathbf{P}^{-1/2}$. *Let* $\mathbf{T}$ *be the inverse square root of* $\mathbf{P}$. *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 ns 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 (6-14) 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.

6.5 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

$$z = \frac{c(\hat{\boldsymbol{\beta}}) - q}{\text{estimated standard error}} \qquad \textbf{(6-24)}$$

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 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

$$c(\hat{\boldsymbol{\beta}}) \approx c(\boldsymbol{\beta}) + \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{6-25}$$

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,

$$\operatorname{Var}[c(\hat{\boldsymbol{\beta}})] \approx \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)' \operatorname{Var}[\hat{\boldsymbol{\beta}}] \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right). \tag{6-26}$$

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.2.2 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 6.3 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

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

Estimated standard errors are shown in parentheses. We will also require Est.Asy. Cov[b, c] = -0.0003298. 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

$$\begin{aligned} \text{Est.Asy. Var}[d] &= g_b^2 \text{Est.Asy. Var}[b] + g_c^2 \text{Est.Asy. Var}[c] + 2 g_b g_c \text{Est.Asy. Cov}[b, c] \\ &= 13.2626^2 \times 0.02873^2 + 13.1834^2 \times 0.02859^2 \\ &\quad + 2(13.2626)(13.1834)(-0.0003298) = 0.17192. \end{aligned}$$

The square root is 0.41464. To test the hypothesis that the long-run MPC is greater than or equal to 1, we would use

$$z = \frac{0.99403 - 1}{0.41464} = -0.0144.$$

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

You may have noticed that we could have tested this hypothesis with a linear restriction instead; if $\delta = 1$, then $\beta = 1 - \gamma$, or $\beta + \gamma = 1$. The estimate is $q = b + c - 1 = -0.00045$. The estimated standard error of this linear function is $[0.02873^2 + 0.02859^2 - 2(0.0003298)]^{1/2} = 0.03136$. The t ratio for this test is -0.01435 which is 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 assures 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 17. 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

$$\hat{\mathbf{G}} = \frac{\partial \mathbf{c}(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}'}. \tag{6-27}$$

The estimate of the asymptotic covariance matrix of these functions is

$$\text{Est.Asy. Var}[\hat{\mathbf{c}}] = \hat{\mathbf{G}}\big\{\text{Est.Asy. Var}[\hat{\boldsymbol{\beta}}]\big\}\hat{\mathbf{G}}'. \tag{6-28}$$

The jth row of $\mathbf{G}$ 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

$$W = (\hat{\mathbf{c}} - \mathbf{q})'\big\{\text{Est. Asy. Var}[\hat{\mathbf{c}}]\big\}^{-1}(\hat{\mathbf{c}} - \mathbf{q}). \tag{6-29}$$

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 (6-24).

6.6 PREDICTION

After the estimation of parameters, a common use of regression is for prediction.[8] Suppose that we wish to predict the value of y^0 associated with a regressor vector $\mathbf{x}^0$. This value would be

$$y^0 = \mathbf{x}^{0\prime}\boldsymbol{\beta} + \varepsilon^0.$$

It follows from the Gauss–Markov theorem that

$$\hat{y}^0 = \mathbf{x}^{0\prime}\mathbf{b} \tag{6-30}$$

is the minimum variance linear unbiased estimator of $E[y^0|\mathbf{x}^0]$. The forecast error is

$$e^0 = y^0 - \hat{y}^0 = (\boldsymbol{\beta} - \mathbf{b})'\mathbf{x}^0 + \varepsilon^0.$$

The **prediction variance** to be applied to this estimate is

$$\text{Var}[e^0|\mathbf{X}, \mathbf{x}^0] = \sigma^2 + \text{Var}[(\boldsymbol{\beta} - \mathbf{b})'\mathbf{x}^0|\mathbf{X}, \mathbf{x}^0] = \sigma^2 + \mathbf{x}^{0\prime}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{x}^0. \tag{6-31}$$

If the regression contains a constant term, then an equivalent expression is

$$\text{Var}[e^0] = \sigma^2\left[1 + \frac{1}{n} + \sum_{j=1}^{K-1}\sum_{k=1}^{K-1}\left(x_j^0 - \bar{x}_j\right)\left(x_k^0 - \bar{x}_k\right)(\mathbf{Z}'\mathbf{M}^0\mathbf{Z})^{jk}\right]$$

where $\mathbf{Z}$ is the $K - 1$ columns of $\mathbf{X}$ not including the constant. This result shows that the width of the interval depends on the distance of the elements of $\mathbf{x}^0$ from the center of the data. Intuitively, this idea makes sense; the farther the forecasted point is from the center of our experience, the greater is the degree of uncertainty.

The prediction variance can be estimated by using s^2 in place of σ^2. A confidence interval for y^0 would be formed using a

$$\textbf{prediction interval} = \hat{y}^0 \pm t_{\lambda/2}\,\text{se}(e^0).$$

Figure 6.1 shows the effect for the bivariate case. Note that the prediction variance is composed of three parts. The second and third become progressively smaller as we accumulate more data (i.e., as n increases). But the first term σ^2 is constant, which implies that no matter how much data we have, we can never predict perfectly.

Example 6.4 Prediction for Investment

Suppose that we wish to "predict" the first quarter 2001 value of real investment. The average rate (secondary market) for the 90 day T-bill was 4.48% (down from 6.03 at the end of 2000); real GDP was 9316.8; the CPI_U was 528.0 and the time trend would equal 204. (We dropped one observation to compute the rate of inflation. Data were obtained from www.economagic.com.) The rate of inflation on a yearly basis would be

[8] It is necessary at this point to make a largely semantic distinction between "prediction" and "forecasting." We will use the term "prediction" to mean using the regression model to compute fitted values of the dependent variable, either within the sample or for observations outside the sample. The same set of results will apply to cross sections, time series, or panels. These are the methods considered in this section. It is helpful at this point to reserve the term "forecasting" for usage of the time series models discussed in Chapter 20. One of the distinguishing features of the models in that setting will be the explicit role of "time" and the presence of lagged variables and disturbances in the equations and correlation of variables with past values.

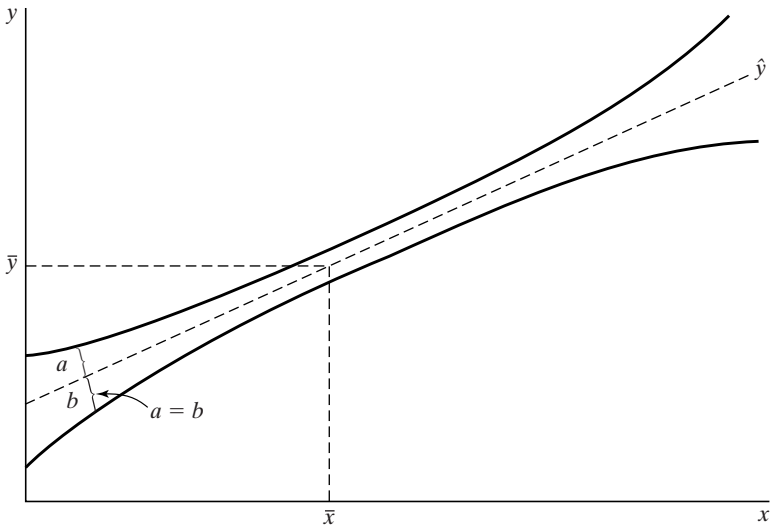

FIGURE 6.1 Prediction Intervals.

100% × 4 × ln(528.0/521.1) = 5.26%. The data vector for predicting $\ln I_{2001.1}$ would be $\mathbf{x}^0 = [1, 4.48, 5.26, 9.1396, 204]'$. Using the regression results in Example 6.1,

$$\mathbf{x}^{0\prime}\mathbf{b} = [1, 4.48, 5.26, 9.1396, 204] \times [-9.1345, -0.008601, 0.003308, 1.9302, -0.005659]'$$
$$= 7.3312.$$

The estimated variance of this prediction is

$$s^2[1 + \mathbf{x}^{0\prime}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}^0] = 0.0076912. \quad \textbf{(6-32)}$$

The square root, 0.087699, gives the prediction standard deviation. Using this value, we obtain the prediction interval:

$$7.3312 \pm 1.96(0.087699) = \langle 7.1593, 7.5031 \rangle.$$

The yearly rate of real investment in the first quarter of 2001 was 1721. The log is 7.4507, so our forecast interval contains the actual value.

We have forecasted the log of real investment with our regression model. If it is desired to forecast the level, the natural estimator would be $\hat{I} = \exp(\ln I)$. Assuming that the estimator, itself, is at least asymptotically normally distributed, this should systematically underestimate the level by a factor of $\exp(\hat{\sigma}^2/2)$ based on the mean of the lognormal distribution. [See Wooldridge (2000, p. 203) and Section B.4.4.] It remains to determine what to use for $\hat{\sigma}^2$. In (6-32), the second part of the expression will vanish in large samples, leaving (as Wooldridge suggests) $s^2 = 0.007427$.[9] Using this scaling, we obtain a prediction of 1532.9, which is still 11 percent below the actual value. Evidently, this model based on an extremely long time series does not do a very good job of predicting at the end of the sample period. One might surmise various reasons, including some related to the model specification that we will address in Chapter 20, but as a first guess, it seems optimistic to apply an equation this simple to more than 50 years of data while expecting the underlying structure to be unchanging

[9] Wooldridge suggests an alternative not necessarily based on an assumption of normality. Use as the scale factor the single coefficient in a within sample regression of y_i on the exponents of the fitted logs.

through the entire period. To investigate this possibility, we redid all the preceding calculations using only the data from 1990 to 2000 for the estimation. The prediction for the level of investment in 2001.1 is now 1885.2 (using the suggested scaling), which is an overestimate of 9.54 percent. But, this is more easily explained. The first quarter of 2001 began the first recession in the U.S. economy in nearly 10 years, and one of the early symptoms of a recession is a rapid decline in business investment.

All the preceding assumes that $\mathbf{x}^0$ is either known with certainty, ex post, or forecasted perfectly. If $\mathbf{x}^0$ must, itself, be forecasted (an ex ante forecast), then the formula for the forecast variance in (6-31) would have to be modified to include the variation in $\mathbf{x}^0$, which greatly complicates the computation. Most authors view it as simply intractable. Beginning with Feldstein (1971), derivation of firm analytical results for the correct forecast variance for this case remain to be derived except for simple special cases. The one qualitative result that seems certain is that (6-31) will understate the true variance. McCullough (1996) presents an alternative approach to computing appropriate forecast standard errors based on the method of bootstrapping. (See the end of Section 16.3.2.)

Various measures have been proposed for assessing the predictive accuracy of forecasting models.[10] Most of these measures are designed to evaluate **ex post forecasts,** that is, forecasts for which the independent variables do not themselves have to be forecasted. Two measures that are based on the residuals from the forecasts are the **root mean squared error**

$$\text{RMSE} = \sqrt{\frac{1}{n^0}\sum_i (y_i - \hat{y}_i)^2}$$

and the mean absolute error

$$\text{MAE} = \frac{1}{n^0}\sum_i |y_i - \hat{y}_i|,$$

where n^0 is the number of periods being forecasted. (Note that both of these as well as the measures below, are backward looking in that they are computed using the observed data on the independent variable.) These statistics have an obvious scaling problem—multiplying values of the dependent variable by any scalar multiplies the measure by that scalar as well. Several measures that are scale free are based on the **Theil *U* statistic:**[11]

$$U = \sqrt{\frac{(1/n^0)\sum_i (y_i - \hat{y}_i)^2}{(1/n^0)\sum_i y_i^2}}.$$

This measure is related to R^2 but is not bounded by zero and one. Large values indicate a poor forecasting performance. An alternative is to compute the measure in terms of the changes in y:

$$U_\Delta = \sqrt{\frac{(1/n^0)\sum_i (\Delta y_i - \Delta \hat{y}_i)^2}{(1/n^0)\sum_i (\Delta y_i)^2}},$$

[10] See Theil (1961) and Fair (1984).

[11] Theil (1961).

where $\Delta y_i = y_i - y_{i-1}$ and $\Delta \hat{y}_i = \hat{y}_i - y_{i-1}$, or, in percentage changes, $\Delta y_i = (y_i - y_{i-1})/y_{i-1}$ and $\Delta \hat{y}_i = (\hat{y}_i - y_{i-1})/y_{i-1}$. These measures will reflect the model's ability to track turning points in the data.

6.7 SUMMARY AND CONCLUSIONS

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

Key Terms and Concepts

- Alternative hypothesis
- Distributed lag
- Discrepancy vector
- Exclusion restrictions
- Ex post forecast
- Lagrange multiplier test
- Limiting distribution
- Linear restrictions
- Nested models
- Nonlinear restriction
- Nonnested models
- Noninvariance of Wald test
- Nonnormality
- Null hypothesis
- Parameter space
- Prediction interval
- Prediction variance
- Restricted least squares
- Root mean squared error
- Testable implications
- Theil U statistic
- Wald criterion

Exercises

1. A multiple regression of y on a constant x_1 and x_2 produces the following results: $\hat{y} = 4 + 0.4x_1 + 0.9x_2$, $R^2 = 8/60$, $\mathbf{e}'\mathbf{e} = 520$, $n = 29$,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 29 & 0 & 0 \\ 0 & 50 & 10 \\ 0 & 10 & 80 \end{bmatrix}.$$

   Test the hypothesis that the two slopes sum to 1.
2. Using the results in Exercise 1, test the hypothesis that the slope on x_1 is 0 by running the restricted regression and comparing the two sums of squared deviations.
3. The regression model to be analyzed is $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, where $\mathbf{X}_1$ and $\mathbf{X}_2$ have K_1 and K_2 columns, respectively. The restriction is $\boldsymbol{\beta}_2 = \mathbf{0}$.
   a. Using (6-14), prove that the restricted estimator is simply $[\mathbf{b}_{1*}, \mathbf{0}]$, where $\mathbf{b}_{1*}$ is the least squares coefficient vector in the regression of $\mathbf{y}$ on $\mathbf{X}_1$.
   b. Prove that if the restriction is $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0$ for a nonzero $\boldsymbol{\beta}_2^0$, then the restricted estimator of $\boldsymbol{\beta}_1$ is $\mathbf{b}_{1*} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_2^0)$.
4. The expression for the restricted coefficient vector in (6-14) may be written in the form $\mathbf{b}_* = [\mathbf{I} - \mathbf{CR}]\mathbf{b} + \mathbf{w}$, where $\mathbf{w}$ does not involve $\mathbf{b}$. What is $\mathbf{C}$? Show that the

covariance matrix of the restricted least squares estimator is

$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$$

and that this matrix may be written as

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

5. Prove the result that the restricted least squares estimator never has a larger covariance matrix than the unrestricted least squares estimator.
6. Prove the result that the R^2 associated with a restricted least squares estimator is never larger than that associated with the unrestricted least squares estimator. Conclude that imposing restrictions never improves the fit of the regression.
7. The **Lagrange multiplier test** of the hypothesis $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$ is equivalent to a Wald test of the hypothesis that $\boldsymbol{\lambda} = \mathbf{0}$, where $\boldsymbol{\lambda}$ is defined in (6-14). Prove that

$$\chi^2 = \boldsymbol{\lambda}'\{\text{Est. Var}[\boldsymbol{\lambda}]\}^{-1}\boldsymbol{\lambda} = (n-K)\left[\frac{\mathbf{e}_*'\mathbf{e}_*}{\mathbf{e}'\mathbf{e}} - 1\right].$$

   Note that the fraction in brackets is the ratio of two estimators of σ^2. By virtue of (6-19) and the preceding discussion, we know that this ratio is greater than 1. Finally, prove that the Lagrange multiplier statistic is equivalent to JF, where J is the number of restrictions being tested and F is the conventional F statistic given in (6-6).
8. Use the Lagrange multiplier test to test the hypothesis in Exercise 1.
9. Using the data and model of Example 2.3, carry out a test of the hypothesis that the three aggregate price indices are not significant determinants of the demand for gasoline.
10. The full model of Example 2.3 may be written in logarithmic terms as

$$\ln G/pop = \alpha + \beta_p \ln P_g + \beta_y \ln Y + \gamma_{nc} \ln P_{nc} + \gamma_{uc} \ln P_{uc} + \gamma_{pt} \ln P_{pt}$$
$$+ \beta \text{ year} + \delta_d \ln P_d + \delta_n \ln P_n + \delta_s \ln P_s + \varepsilon.$$

    Consider the hypothesis that the microelasticities are a constant proportion of the elasticity with respect to their corresponding aggregate. Thus, for some positive θ (presumably between 0 and 1), $\gamma_{nc} = \theta\delta_d$, $\gamma_{uc} = \theta\delta_d$, $\gamma_{pt} = \theta\delta_s$.

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

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

    a. Describe in detail how you would test the validity of the restriction.
    b. Using the gasoline market data in Table F2.2 , test the restrictions separately and jointly.
11. Prove that under the hypothesis that $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, the estimator

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

    where J is the number of restrictions, is unbiased for σ^2.
12. Show that in the multiple regression of $\mathbf{y}$ on a constant, $\mathbf{x}_1$ and $\mathbf{x}_2$ while imposing the restriction $\beta_1 + \beta_2 = 1$ leads to the regression of $\mathbf{y} - \mathbf{x}_1$ on a constant and $\mathbf{x}_2 - \mathbf{x}_1$.