8
SPECIFICATION ANALYSIS AND MODEL SELECTION

8.1 INTRODUCTION

Chapter 7 presented results which were primarily focused on sharpening the functional form of the model. Functional form and hypothesis testing are directed toward improving the specification of the model or using that model to draw generally narrow inferences about the population. In this chapter we turn to some broader techniques that relate to choosing a specific model when there is more than one competing candidate. Section 8.2 describes some larger issues related to the use of the multiple regression model—specifically the impacts of an incomplete or excessive specification on estimation and inference. Sections 8.3 and 8.4 turn to the broad question of statistical methods for choosing among alternative models.

8.2 SPECIFICATION ANALYSIS AND MODEL BUILDING

Our analysis has been based on the assumption that the correct specification of the regression model is known to be

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{8-1}$$

There are numerous types of errors that one might make in the specification of the estimated equation. Perhaps the most common ones are the **omission of relevant variables** and the **inclusion of superfluous variables.**

8.2.1 BIAS CAUSED BY OMISSION OF RELEVANT VARIABLES

Suppose that a correctly specified regression model would be

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}, \tag{8-2}$$

where the two parts of $\mathbf{X}$ have K_1 and K_2 columns, respectively. If we regress $\mathbf{y}$ on $\mathbf{X}_1$ without including $\mathbf{X}_2$, then the estimator is

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} = \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\boldsymbol{\varepsilon}. \tag{8-3}$$

Taking the expectation, we see that unless $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$ or $\boldsymbol{\beta}_2 = \mathbf{0}$, $\mathbf{b}_1$ is biased. The well-known result is **the omitted variable formula:**

$$E[\mathbf{b}_1 \mid \mathbf{X}] = \boldsymbol{\beta}_1 + \mathbf{P}_{1.2}\boldsymbol{\beta}_2, \tag{8-4}$$

where

$$\mathbf{P}_{1.2} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2. \tag{8-5}$$

Each column of the $K_1 \times K_2$ matrix $\mathbf{P}_{1.2}$ is the column of slopes in the least squares regression of the corresponding column of $\mathbf{X}_2$ on the columns of $\mathbf{X}_1$.

Example 8.1 Omitted Variables

If a demand equation is estimated without the relevant income variable, then (8-4) shows how the estimated price elasticity will be biased. Letting b be the estimator, we obtain

$$E[b \mid \text{price, income}] = \beta + \frac{\text{Cov[price, income]}}{\text{Var[price]}}\gamma,$$

where γ is the income coefficient. In aggregate data, it is unclear whether the missing covariance would be positive or negative. The sign of the bias in b would be the same as this covariance, however, because Var[price] and γ would be positive.

The gasoline market data we have examined in Examples 2.3 and 7.6 provide a striking example. Figure 7.5 showed a simple plot of per capita gasoline consumption, G/pop against the price index P_G. The plot is considerably at odds with what one might expect. But a look at the data in Appendix Table F2.2 shows clearly what is at work. Holding per capita income, I/pop and other prices constant, these data might well conform to expectations. In these data, however, income is persistently growing, and the simple correlations between G/pop and I/pop and between P_G and I/pop are 0.86 and 0.58, respectively, which are quite large. To see if the expected relationship between price and consumption shows up, we will have to purge our data of the intervening effect of I/pop. To do so, we rely on the Frisch–Waugh result in Theorem 3.3. The regression results appear in Table 7.6. The first column shows the full regression model, with ln PG, ln Income, and several other variables. The estimated demand elasticity is −0.11553, which conforms with expectations. If income is omitted from this equation, the estimated price elasticity is +0.074499 which has the wrong sign, but is what we would expect given the theoretical results above.

In this development, it is straightforward to deduce the directions of bias when there is a single included variable and one omitted variable. It is important to note, however, that if more than one variable is included, then the terms in the omitted variable formula involve multiple regression coefficients, which themselves have the signs of partial, not simple, correlations. For example, in the demand equation of the previous example, if the price of a closely related product had been included as well, then the simple correlation between price and income would be insufficient to determine the direction of the bias in the price elasticity. What would be required is the sign of the correlation between price and income net of the effect of the other price. This requirement might not be obvious, and it would become even less so as more regressors were added to the equation.

8.2.2 PRETEST ESTIMATION

The variance of $\mathbf{b}_1$ is that of the third term in (8-3), which is

$$\text{Var}[\mathbf{b}_1 \mid \mathbf{X}] = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}. \tag{8-6}$$

If we had computed the correct regression, including $\mathbf{X}_2$, then the slopes on $\mathbf{X}_1$ would have been unbiased and would have had a covariance matrix equal to the upper left block of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. This matrix is

$$\text{Var}[\mathbf{b}_{1.2} \mid \mathbf{X}] = \sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}, \tag{8-7}$$

where

$$\mathbf{M}_2 = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2',$$

or

$$\text{Var}[\mathbf{b}_{1.2} \mid \mathbf{X}] = \sigma^2[\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1]^{-1}.$$

We can compare the covariance matrices of $\mathbf{b}_1$ and $\mathbf{b}_{1.2}$ more easily by comparing their inverses [see result (A-120)];

$$\text{Var}[\mathbf{b}_1 \mid \mathbf{X}]^{-1} - \text{Var}[\mathbf{b}_{1.2} \mid \mathbf{X}]^{-1} = (1/\sigma^2)\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1,$$

which is nonnegative definite. We conclude that although $\mathbf{b}_1$ is **biased,** its variance is never larger than that of $\mathbf{b}_{1.2}$ (since the inverse of its variance is at least as large).

Suppose, for instance, that $\mathbf{X}_1$ and $\mathbf{X}_2$ are each a single column and that the variables are measured as deviations from their respective means. Then

$$\text{Var}[b_1 \mid \mathbf{X}] = \frac{\sigma^2}{s_{11}}, \quad \text{where } s_{11} = \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2,$$

whereas

$$\text{Var}[b_{1.2} \mid \mathbf{X}] = \sigma^2[\mathbf{x}_1'\mathbf{x}_1 - \mathbf{x}_1'\mathbf{x}_2(\mathbf{x}_2'\mathbf{x}_2)^{-1}\mathbf{x}_2'\mathbf{x}_1]^{-1} = \frac{\sigma^2}{s_{11}(1 - r_{12}^2)}, \tag{8-8}$$

where

$$r_{12}^2 = \frac{(\mathbf{x}_1'\mathbf{x}_2)^2}{\mathbf{x}_1'\mathbf{x}_1\mathbf{x}_2'\mathbf{x}_2}$$

is the squared sample correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$. The more highly correlated $\mathbf{x}_1$ and $\mathbf{x}_2$ are, the larger is the variance of $b_{1.2}$ compared with that of b_1. Therefore, it is possible that b_1 is a more precise estimator based on the **mean-squared error** criterion.

The result in the preceding paragraph poses a bit of a dilemma for applied researchers. The situation arises frequently in the search for a model specification. Faced with a variable that a researcher suspects should be in their model, but which is causing a problem of collinearity, the analyst faces a choice of omitting the relevant variable or including it and estimating its (and all the other variables') coefficient imprecisely. This presents a choice between two estimators, b_1 and $b_{1.2}$. In fact, what researchers usually do actually creates a third estimator. It is common to include the problem variable provisionally. If its t ratio is sufficiently large, it is retained; otherwise it is discarded. This third estimator is called a **pretest estimator.** What is known about pretest estimators is not encouraging. Certainly they are biased. How badly depends on the unknown parameters. Analytical results suggest that the pretest estimator is the least precise of the three when the researcher is most likely to use it. [See Judge et al. (1985).]

8.2.3 INCLUSION OF IRRELEVANT VARIABLES

If the regression model is correctly given by

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \tag{8-9}$$

and we estimate it as if (8-2) were correct (i.e., we include some extra variables), then it might seem that the same sorts of problems considered earlier would arise. In fact, this case is not true. We can view the omission of a set of relevant variables as equivalent to imposing an incorrect restriction on (8-2). In particular, omitting $\mathbf{X}_2$ is equivalent to *incorrectly* estimating (8-2) subject to the restriction $\boldsymbol{\beta}_2 = \mathbf{0}$. As we discovered, incorrectly imposing a restriction produces a biased estimator. Another way to view this error is to note that it amounts to incorporating incorrect information in our estimation. Suppose, however, that our error is simply a failure to use some information that is *correct.*

The inclusion of the irrelevant variables $\mathbf{X}_2$ in the regression is equivalent to failing to impose $\boldsymbol{\beta}_2 = \mathbf{0}$ on (8-2) in estimation. But (8-2) is not incorrect; it simply fails to incorporate $\boldsymbol{\beta}_2 = \mathbf{0}$. Therefore, we do not need to prove formally that the least squares estimator of $\boldsymbol{\beta}$ in (8-2) is unbiased *even given* the restriction; we have already proved it. We can assert on the basis of all our earlier results that

$$E[\mathbf{b} \,|\, \mathbf{X}] = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{bmatrix}. \tag{8-10}$$

By the same reasoning, s^2 is also unbiased:

$$E\left[\frac{\mathbf{e}'\mathbf{e}}{n - K_1 - K_2} \,\middle|\, \mathbf{X}\right] = \sigma^2. \tag{8-11}$$

Then where is the problem? It would seem that one would generally want to “overfit” the model. From a theoretical standpoint, the difficulty with this view is that the failure to use correct information is always costly. In this instance, the cost is the reduced precision of the estimates. As we have shown, the covariance matrix in the short regression (omitting $\mathbf{X}_2$) is never larger than the covariance matrix for the estimator obtained in the presence of the superfluous variables.[1] Consider again the single-variable comparison given earlier. If $\mathbf{x}_2$ is highly correlated with $\mathbf{x}_1$, then incorrectly including it in the regression will greatly inflate the variance of the estimator.

8.2.4 MODEL BUILDING—A GENERAL TO SIMPLE STRATEGY

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

[1] There is no loss if $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$, which makes sense in terms of the information about $\mathbf{X}_1$ contained in $\mathbf{X}_2$ (here, none). This situation is not likely to occur in practice, however.

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. 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 assures 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.

8.3 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.[2] 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:

$$H_0 : \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0 \tag{8-12a}$$

and

$$H_1 : \mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_1. \tag{8-12b}$$

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.[3]

Before turning to classical- (frequentist-) based tests in this setting, we should note that the Bayesian approach to this question might be more intellectually appealing. Our procedures will continue to be directed toward an objective of rejecting one model in favor of the other. Yet, in fact, if we have doubts as to which of two models is appropriate, then we might well be convinced to concede that possibly neither one is really "the truth." We have rather painted ourselves into a corner with our "left or right"

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

[3] 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.

approach. The Bayesian approach to this question treats it as a problem of comparing the two hypotheses rather than testing for the validity of one over the other. We enter our sampling experiment with a set of prior probabilities about the relative merits of the two hypotheses, which is summarized in a "prior odds ratio," $P_{01} = \text{Prob}[H_0]/\text{Prob}[H_1]$. After gathering our data, we construct the Bayes factor, which summarizes the weight of the sample evidence in favor of one model or the other. After the data have been analyzed, we have our "posterior odds ratio,"

$$P_{01} \mid \text{data} = \text{Bayes factor} \times P_{01}.$$

The upshot is that ex post, neither model is discarded; we have merely revised our assessment of the comparative likelihood of the two in the face of the sample data. Some of the formalities of this approach are discussed in Chapter 16.

8.3.1 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.[4] Since, 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 one error. As such, the null 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 (8-12a, 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** which 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.

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

8.3.2 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.[5] Since 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 $\bar{\mathbf{X}}$ be the set of variables in $\mathbf{X}$ that are not in $\mathbf{Z}$, define $\bar{\mathbf{Z}}$ 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":

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

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

Consider an alternative approach. If H_0 is correct, then $\mathbf{y}$ will, apart from the random disturbance $\boldsymbol{\varepsilon}$, be fully explained by $\mathbf{X}$. Suppose we then attempt to estimate $\boldsymbol{\gamma}$ by regression of $\mathbf{y}$ on $\mathbf{Z}$. Whatever set of parameters is estimated by this regression, say $\mathbf{c}$, if H_0 is correct, then we should estimate exactly the same coefficient vector if we were to regress $\mathbf{X}\boldsymbol{\beta}$ on $\mathbf{Z}$, since $\boldsymbol{\varepsilon}_0$ is random noise under H_0. Since $\boldsymbol{\beta}$ must be estimated, suppose that we use $\mathbf{Xb}$ instead and compute $\mathbf{c}_0$. A test of the proposition that Model 0 "encompasses" Model 1 would be a test of the hypothesis that $E[\mathbf{c} - \mathbf{c}_0] = \mathbf{0}$. It is straightforward to show [see Davidson and MacKinnon (1993, pp. 384–387)] that the test can be carried out by using a standard F test to test the hypothesis that $\boldsymbol{\gamma}_1 = \mathbf{0}$ in the augmented regression,

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

where $\mathbf{Z}_1$ is the variables in $\mathbf{Z}$ that are not in $\mathbf{X}$.

8.3.3 COMPREHENSIVE APPROACH—THE *J* TEST

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

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

Estimation of the comprehensive model followed by a test of $\lambda = 0$ or 1 is used to assess the validity of Model 0 or 1, respectively.[6]

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

[6] See Section 21.4.4c for an application to the choice of probit or logit model for binary choice suggested by Silva (2001).

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

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

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

Example 8.2 ***J Test for a Consumption Function***

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

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

and

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

The first model states that consumption responds to changes in income over two periods, whereas the second states that the effects of changes in income on consumption persist for many periods. Quarterly data on aggregate U.S. real consumption and real disposable income are given in Table F5.1. Here we apply the J test to these data and the two proposed specifications. First, the two models are estimated separately (using observations 1950.2–2000.4). The least squares regression of C on a constant, Y, lagged Y, and the fitted values from the second model produces an estimate of λ of 1.0145 with a t ratio of 62.861. Thus, H_0 should be rejected in favor of H_1. But reversing the roles of H_0 and H_1, we obtain an estimate of λ of -10.677 with a t ratio of -7.188. Thus, H_1 is rejected as well.[7]

8.3.4 THE COX TEST[8]

Likelihood ratio tests rely on three features of the density of the random variable of interest. First, under the null hypothesis, the average log density of the null hypothesis will be less than under the alternative—this is a consequence of the fact that the null model is nested within the alternative. Second, the degrees of freedom for the chi-squared statistic is the reduction in the dimension of the parameter space that is specified by the null hypothesis, compared to the alternative. Third, in order to carry out the test, under the null hypothesis, the test statistic must have a known distribution which is free of the model parameters under the alternative hypothesis. When the models are

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

[8] The Cox test is based upon the likelihood ratio statistic, which will be developed in Chapter 17. The results for the linear regression model, however, are based on sums of squared residuals, and therefore, rely on nothing more than least squares, which is already familiar.

nonnested, none of these requirements will be met. The first need not hold at all. With regard to the second, the parameter space under the null model may well be larger than (or, at least the same size) as under the alternative. (Merely reversing the two models does not solve this problem. The test must be able to work in both directions.) Finally, because of the symmetry of the null and alternative hypotheses, the distributions of likelihood based test statistics will generally be functions of the parameters of the alternative model. Cox's (1961, 1962) analysis of this problem produced a reformulated test statistic that is based on the standard normal distribution and is centered at zero.[9]

Versions of the Cox test appropriate for the linear and nonlinear regression models have been derived by Pesaran (1974) and Pesaran and Deaton (1978). The latter present a test statistic for testing linear versus loglinear models that is extended in Aneuryn-Evans and Deaton (1980). Since in the classical regression model the least squares estimator is also the maximum likelihood estimator, it is perhaps not surprising that Davidson and MacKinnon (1981, p. 789) find that their test statistic is asymptotically equal to the negative of the Cox–Pesaran and Deaton statistic.

The Cox statistic for testing the hypothesis that **X** is the correct set of regressors and that **Z** is not is

$$c_{01} = \frac{n}{2}\ln\left[\frac{s_{\mathbf{Z}}^2}{s_{\mathbf{X}}^2 + (1/n)\mathbf{b}'\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{X}\mathbf{b}}\right] = \frac{n}{2}\ln\left[\frac{s_{\mathbf{Z}}^2}{s_{\mathbf{ZX}}^2}\right], \tag{8-13}$$

where

$$\begin{aligned}
\mathbf{M}_{\mathbf{Z}} &= \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}',\\
\mathbf{M}_{\mathbf{X}} &= \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',\\
\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.\\
s_{\mathbf{Z}}^2 &= \mathbf{e}_{\mathbf{Z}}'\mathbf{e}_{\mathbf{Z}}/n = \text{mean-squared residual in the regression of } \mathbf{y} \text{ on } \mathbf{Z},\\
s_{\mathbf{X}}^2 &= \mathbf{e}_{\mathbf{X}}'\mathbf{e}_{\mathbf{X}}/n = \text{mean-squared residual in the regression of } \mathbf{y} \text{ on } \mathbf{X},\\
s_{\mathbf{ZX}}^2 &= s_{\mathbf{X}}^2 + \mathbf{b}'\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{X}\mathbf{b}/n.
\end{aligned}$$

The hypothesis is tested by comparing

$$q = \frac{c_{01}}{\{\text{Est. Var}[c_{01}]\}^{1/2}} = \frac{c_{01}}{\sqrt{\dfrac{s_{\mathbf{X}}^2}{s_{\mathbf{ZX}}^4}\mathbf{b}'\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{M}_{\mathbf{X}}\mathbf{M}_{\mathbf{Z}}\mathbf{X}\mathbf{b}}} \tag{8-14}$$

to the critical value from the standard normal table. A large value of q is evidence against the null hypothesis (H_0).

The Cox test appears to involve an impressive amount of matrix algebra. But the algebraic results are deceptive. One needs only to compute linear regressions and retrieve fitted values and sums of squared residuals. The following does the first test. The roles of **X** and **Z** are reversed for the second.

1. Regress **y** on **X** to obtain **b** and $\hat{\mathbf{y}}_{\mathbf{X}} = \mathbf{Xb}$, $\mathbf{e}_{\mathbf{X}} = \mathbf{y} - \mathbf{Xb}$, $s_{\mathbf{X}}^2 = \mathbf{e}_{\mathbf{X}}'\mathbf{e}_{\mathbf{X}}/n$.
2. Regress **y** on **Z** to obtain **d** and $\hat{\mathbf{y}}_{\mathbf{Z}} = \mathbf{Zd}$, $\mathbf{e}_{\mathbf{Z}} = \mathbf{y} - \mathbf{Zd}$, $s_{\mathbf{Z}}^2 = \mathbf{e}_{\mathbf{Z}}'\mathbf{e}_{\mathbf{Z}}/n$.

[9] See Pesaran and Weeks (2001) for some of the formalities of these results.

3. Regress $\hat{\mathbf{y}}_{\mathbf{X}}$ on $\mathbf{Z}$ to obtain $\mathbf{d}_{\mathbf{X}}$ and $\mathbf{e}_{\mathbf{Z}.\mathbf{X}} = \hat{\mathbf{y}}_{\mathbf{X}} - \mathbf{Z}\mathbf{d}_{\mathbf{X}} = \mathbf{M}_{\mathbf{Z}}\mathbf{X}\mathbf{b}$, $\mathbf{e}'_{\mathbf{Z}.\mathbf{X}}\mathbf{e}_{\mathbf{Z}.\mathbf{X}} = \mathbf{b}'\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{X}\mathbf{b}$.
4. Regress $\mathbf{e}_{\mathbf{Z}.\mathbf{X}}$ on $\mathbf{X}$ and compute residuals $\mathbf{e}_{\mathbf{X}.\mathbf{Z}\mathbf{X}}$, $\mathbf{e}'_{\mathbf{X}.\mathbf{Z}\mathbf{X}}\mathbf{e}_{\mathbf{X}.\mathbf{Z}\mathbf{X}} = \mathbf{b}'\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{M}_{\mathbf{X}}\mathbf{M}_{\mathbf{Z}}\mathbf{X}\mathbf{b}$.
5. Compute $s^2_{\mathbf{Z}\mathbf{X}} = s^2_{\mathbf{X}} + \mathbf{e}'_{\mathbf{Z}.\mathbf{X}}\mathbf{e}_{\mathbf{Z}.\mathbf{X}}/n$.
6. Compute $c_{01} = \frac{n}{2}\log\frac{s^2_{\mathbf{Z}}}{s^2_{\mathbf{Z}\mathbf{X}}}$, $v_{01} = \frac{s^2_{\mathbf{X}}(\mathbf{e}'_{\mathbf{X}.\mathbf{Z}\mathbf{X}}\mathbf{e}_{\mathbf{X}.\mathbf{Z}\mathbf{X}})}{s^4_{\mathbf{Z}\mathbf{X}}}$, $q = \frac{c_{01}}{\sqrt{v_{01}}}$.

Therefore, the Cox statistic can be computed simply by computing a series of least squares regressions.

Example 8.3 Cox Test for a Consumption Function

We continue the previous example by applying the Cox test to the data of Example 8.2. For purposes of the test, let $\mathbf{X} = [\mathbf{i}\ \mathbf{y}\ \mathbf{y}_{-1}]$ and $\mathbf{Z} = [\mathbf{i}\ \mathbf{y}\ \mathbf{c}_{-1}]$. Using the notation of (8-13) and (8-14), we find that

$$\begin{aligned} s^2_{\mathbf{X}} &= 7{,}556.657,\\ s^2_{\mathbf{Z}} &= 456.3751,\\ \mathbf{b}'\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{X}\mathbf{b} &= 167.50707,\\ \mathbf{b}'\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{M}_{\mathbf{X}}\mathbf{M}_{\mathbf{Z}}\mathbf{X}\mathbf{b} &= 2.61944,\\ s^2_{\mathbf{Z}\mathbf{X}} &= 7556.657 + 167.50707/203 = 7{,}557.483. \end{aligned}$$

Thus,

$$c_{01} = \frac{203}{2}\ln\left(\frac{456.3751}{7{,}557.483}\right) = -284.908$$

and

$$\text{Est. Var}[c_{01}] = \frac{7{,}556.657(2.61944)}{7{,}557.483^2} = 0.00034656.$$

Thus, $q = -15{,}304.281$. On this basis, we reject the hypothesis that $\mathbf{X}$ is the correct set of regressors. Note in the previous example that we reached the same conclusion based on a t ratio of 62.861. As expected, the result has the opposite sign from the corresponding J statistic in the previous example. Now we reverse the roles of $\mathbf{X}$ and $\mathbf{Z}$ in our calculations. Letting $\mathbf{d}$ denote the least squares coefficients in the regression of consumption on $\mathbf{Z}$, we find that

$$\begin{aligned} \mathbf{d}'\mathbf{Z}'\mathbf{M}_{\mathbf{X}}\mathbf{Z}\mathbf{d} &= 1{,}418{,}985.185,\\ \mathbf{d}'\mathbf{Z}'\mathbf{M}_{\mathbf{X}}\mathbf{M}_{\mathbf{Z}}\mathbf{M}_{\mathbf{X}}\mathbf{Z}\mathbf{d} &= 22{,}189.811,\\ s^2_{\mathbf{X}\mathbf{Z}} &= 456.3751 + 1{,}418{,}985.185/203 = 7446.4499. \end{aligned}$$

Thus,

$$c_{10} = \frac{203}{2}\ln\left(\frac{7{,}556.657}{7{,}446.4499}\right) = 1.491$$

and

$$\text{Est. Var}[c_{10}] = \frac{456.3751(22{,}189.811)}{7{,}446.4499^2} = 0.18263.$$

This computation produces a value of $q = 3.489$, which is roughly equal (in absolute value) to its counterpart in Example 8.2, -7.188. Since -2.595 is less than the 5 percent critical value of to -1.96, we once again reject the hypothesis that $\mathbf{Z}$ is the preferred set of regressors though the results do strongly favor $\mathbf{Z}$ in qualitative terms.

Pesaran and Hall (1988) have extended the Cox test to testing which of two nonnested restricted regressions is preferred. The modeling framework is

$$H_0\colon\quad \mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_0, \quad \text{Var}[\boldsymbol{\varepsilon}_0 \,|\, \mathbf{X}_0] = \sigma_0^2\mathbf{I}, \quad \text{subject to } \mathbf{R}_0\boldsymbol{\beta}_0 = \mathbf{q}_0$$

$$H_0\colon\quad \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1, \quad \text{Var}[\boldsymbol{\varepsilon}_1 \,|\, \mathbf{X}_1] = \sigma_1^2\mathbf{I}, \quad \text{subject to } \mathbf{R}_1\boldsymbol{\beta}_1 = \mathbf{q}_1.$$

Like its counterpart for unrestricted regressions, this Cox test requires a large amount of matrix algebra. However, once again, it reduces to a sequence of regressions, though this time with some unavoidable matrix manipulation remaining. Let

$$\mathbf{G}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1} - (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{R}_i'[\mathbf{R}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{R}_i']^{-1}\mathbf{R}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}, \quad i = 0, 1,$$

and $\mathbf{T}_i = \mathbf{X}_i\mathbf{G}_i\mathbf{X}_i'$, $m_i = \text{rank}(\mathbf{R}_i)$, $k_i = \text{rank}(\mathbf{X}_i)$, $h_i = k_i - m_i$ and $d_i = n - h_i$ where n is the sample size. The following steps produce the needed statistics:

1. Compute $\mathbf{e}_i =$ the residuals from the restricted regression, $i = 0, 1$.
2. Compute $\mathbf{e}_{10}$ by computing the residuals from the restricted regression of $\mathbf{y} - \mathbf{e}_0$ on $\mathbf{X}_1$. Compute $\mathbf{e}_{01}$ likewise by reversing the subscripts.
3. Compute $\mathbf{e}_{100}$ as the residuals from the restricted regression of $\mathbf{y} - \mathbf{e}_{10}$ on $\mathbf{X}_0$ and $\mathbf{e}_{110}$ likewise by reversing the subscripts.

   Let v_i, v_{ij} and v_{ijk} denote the sums of squared residuals in Steps 1, 2, and 3 and let $s_i^2 = \mathbf{e}_i'\mathbf{e}_i/d_i$.
4. Compute trace $(\mathbf{B}_0^2) = h_1 - \text{trace}[(\mathbf{T}_0\mathbf{T}_1)^2] - \left\{h_1 - \text{trace}[(\mathbf{T}_0\mathbf{T}_1)^2]\right\}^2/(n - h_0)$ and trace $(\mathbf{B}_1^2)$ likewise by reversing subscripts.
5. Compute $s_{10}^2 = \left(v_{10} + s_0^2 \,\text{trace}[\mathbf{I} - \mathbf{T}_0 - \mathbf{T}_1 + \mathbf{T}_0\mathbf{T}_1]\right)$ and s_{01}^2 likewise.

The authors propose several statistics. A Wald test based on Godfrey and Pesaran (1983) is based on the difference between an estimator of σ_1^2 and the probability limit of this estimator assuming that H_0 is true

$$W_0 = \sqrt{n}(v_1 - v_0 - v_{10})\big/\sqrt{4v_0v_{100}}.$$

Under the null hypothesis of Model 0, the limiting distribution of W_0 is standard normal. An alternative statistic based on Cox's likelihood approach is

$$N_0 = (n/2)\ln\left(s_1^2/s_{10}^2\right)\Big/\sqrt{4v_{100}s_0^2/\left(s_{10}^2\right)^2}.$$

Example 8.4 Cox Test for Restricted Regressions

The example they suggest is two competing models for expected inflation, P_t^e, based on commonly used lag structures involving lags of P_t^e and current lagged values of actual inflation, P_t;

$$\text{(Regressive):}\quad P_t^e = P_t + \theta_1(P_t - P_{t-1}) + \theta_2(P_{t-1} - P_{t-2}) + \varepsilon_{0t}$$

$$\text{(Adaptive)}\quad P_t^e = P_{t-1}^e + \lambda_1\left(P_t - P_{t-1}^e\right) + \lambda_2\left(P_{t-1} - P_{t-2}^e\right) + \varepsilon_{1t}.$$

By formulating these models as

$$y_t = \beta_1 P_{t-1}^e + \beta_2 P_{t-2}^e + \beta_3 P_t + \beta_4 P_{t-1} + \beta_5 P_{t-2} + \varepsilon_t,$$

K They show that the hypotheses are

$$H_0:\ \beta_1 = \beta_2 = 0,\quad \beta_3 + \beta_4 + \beta_5 = 1$$

$$H_1:\ \beta_1 + \beta_3 = 1,\quad \beta_2 + \beta_4 = 0, \beta_5 = 0.$$

Pesaran and Hall's analysis was based on quarterly data for British manufacturing from 1972 to 1981. The data appear in the Appendix to Pesaran (1987) and are reproduced in Table F8.1. Using their data, the computations listed before produce the following results:

$$W_0:\ \text{Null is } H_0; -3.887,\quad \text{Null is } H_1; -0.134$$

$$N_0:\ \text{Null is } H_0; -2.437,\quad \text{Null is } H_1; -0.032.$$

These results fairly strongly support Model 1 and lead to rejection of Model 0.[10]

8.4 MODEL SELECTION CRITERIA

The preceding discussion suggested some approaches to model selection based on nonnested hypothesis tests. Fit measures and testing procedures based on the sum of squared residuals, such as R^2 and the Cox test, are useful when interest centers on the within-sample fit or within-sample prediction of the dependent variable. When the model building is directed toward forecasting, within-sample measures are not necessarily optimal. As we have seen, R^2 cannot fall when variables are added to a model, so there is a built-in tendency to overfit the model. This criterion may point us away from the best forecasting model, because adding variables to a model may increase the variance of the forecast error (see Section 6.6) despite the improved fit to the data. With this thought in mind, the **adjusted R^2**,

$$\bar{R}^2 = 1 - \frac{n-1}{n-K}(1-R^2) = 1 - \frac{n-1}{n-K}\left(\frac{\mathbf{e}'\mathbf{e}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right), \tag{8-15}$$

has been suggested as a fit measure that appropriately penalizes the loss of degrees of freedom that result from adding variables to the model. Note that $\bar{R}^2$ may fall when a variable is added to a model if the sum of squares does not fall fast enough. (The applicable result appears in Theorem 3.7; $\bar{R}^2$ does not rise when a variable is added to a model unless the t ratio associated with that variable exceeds one in absolute value.) The adjusted R^2 has been found to be a preferable fit measure for assessing the fit of forecasting models. [See Diebold (1998b, p. 87), who argues that the simple R^2 has a downward bias as a measure of the out-of-sample, one-step-ahead prediction error variance.]

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

$$\text{AIC}(K) = s_y^2(1-R^2)e^{2K/n} \tag{8-16}$$

[10] Our results differ somewhat from Pesaran and Hall's. For the first row of the table, they reported (−2.180, −1.690) and for the second, (−2.456, −1.907). They reach the same conclusion, but the numbers do differ substantively. We have been unable to resolve the difference.

and the Schwartz or Bayesian information criterion,

$$\text{BIC}(K) = s_y^2(1 - R^2)n^{K/n}. \tag{8-17}$$

(There is no degrees of freedom correction in s_y^2.) Both measures improve (decline) as R^2 increases, but, everything else constant, degrade as the model size increases. Like $\bar{R}^2$, these measures place a premium on achieving a given fit with a smaller number of parameters per observation, K/n. Logs are usually more convenient; the measures reported by most software are

$$\text{AIC}(K) = \log\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{2K}{n} \tag{8-18}$$

$$\text{BIC}(K) = \log\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{K\log n}{n}. \tag{8-19}$$

Both **prediction criteria** have their virtues, and neither has an obvious advantage over the other. [See Diebold (1998b, p. 90).] The **Schwarz criterion,** with its heavier penalty for degrees of freedom lost, will lean toward a simpler model. All else given, simplicity does have some appeal.

8.5 SUMMARY AND CONCLUSIONS

This is the last of seven chapters that we have devoted specifically to the most heavily used tool in econometrics, the classical linear regression model. We began in Chapter 2 with a statement of the regression model. Chapter 3 then described computation of the parameters by least squares—a purely algebraic exercise. Chapters 4 and 5 reinterpreted least squares as an estimator of an unknown parameter vector, and described the finite sample and large sample characteristics of the sampling distribution of the estimator. Chapters 6 and 7 were devoted to building and sharpening the regression model, with tools for developing the functional form and statistical results for testing hypotheses about the underlying population. In this chapter, we have examined some broad issues related to model specification and selection of a model among a set of competing alternatives. The concepts considered here are tied very closely to one of the pillars of the paradigm of econometrics, that underlying the model is a theoretical construction, a set of true behavioral relationships that constitute *the model.* It is only on this notion that the concepts of bias and biased estimation and model selection make any sense—"bias" as a concept can only be described with respect to some underlying "model" against which an estimator can be said to be biased. That is, there must be a yardstick. This concept is a central result in the analysis of specification, where we considered the implications of underfitting (omitting variables) and overfitting (including superfluous variables) the model. We concluded this chapter (and our discussion of the classical linear regression model) with an examination of procedures that are used to choose among competing model specifications.

Key Terms and Concepts

- Adjusted R-squared
- Akaike criterion
- Biased estimator
- Comprehensive model
- Cox test
- Encompassing principle
- General-to-simple strategy
- Inclusion of superfluous variables
- *J* test
- Mean squared error
- Model selection
- Nonnested models
- Omission of relevant variables
- Omitted variable formula
- Prediction criterion
- Pretest estimator
- Schwarz criterion
- Simple-to-general
- Specification analysis
- Stepwise model building

Exercises

1. Suppose the true regression model is given by (8-2). The result in (8-4) shows that if either $\mathbf{P}_{1.2}$ is nonzero or $\boldsymbol{\beta}_2$ is nonzero, then regression of $\mathbf{y}$ on $\mathbf{X}_1$ alone produces a biased and inconsistent estimator of $\boldsymbol{\beta}_1$. Suppose the objective is to forecast $\mathbf{y}$, not to estimate the parameters. Consider regression of $\mathbf{y}$ on $\mathbf{X}_1$ alone to estimate $\boldsymbol{\beta}_1$ with $\mathbf{b}_1$ (which is biased). Is the forecast of y computed using $\mathbf{X}_1\mathbf{b}_1$ also biased? Assume that $E[\mathbf{X}_2 \mid \mathbf{X}_1]$ is a linear function of $\mathbf{X}_1$. Discuss your findings generally. What are the implications for prediction when variables are omitted from a regression?
2. Compare the mean squared errors of b_1 and $b_{1.2}$ in Section 8.2.2. (Hint: The comparison depends on the data and the model parameters, but you can devise a compact expression for the two quantities.)
3. The *J* test in Example 8.2 is carried out using over 50 years of data. It is optimistic to hope that the underlying structure of the economy did not change in 50 years. Does the result of the test carried out in Example 8.2 persist if it is based on data only from 1980 to 2000? Repeat the computation with this subset of the data.
4. The Cox test in Example 8.3 has the same difficulty as the *J* test in Example 8.2. The sample period might be too long for the test not to have been affected by underlying structural change. Repeat the computations using the 1980 to 2000 data.