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

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

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

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

5.8.3 COMPREHENSIVE APPROACH—THE J TEST

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

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

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

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

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

In this model, a test of $\lambda = 0$ would be a test against H_1. The problem is that λ cannot be separately estimated in this model; it would amount to a redundant scaling of the regression coefficients. Davidson and MacKinnon's J test consists of estimating $\boldsymbol{\gamma}$ by a least squares regression of $\mathbf{y}$ on $\mathbf{Z}$ followed by a least squares regression of $\mathbf{y}$ on $\mathbf{X}$ and $\mathbf{Z}\hat{\boldsymbol{\gamma}}$, the fitted values in the first regression. A valid test, at least asymptotically, of H_1 is to test $H_0 : \lambda = 0$. If H_0 is true, then plim $\hat{\lambda} = 0$. Asymptotically, the ratio $\hat{\lambda}/\text{se}(\hat{\lambda})$ (i.e., the usual t ratio) is distributed as standard normal and may be referred to

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

the standard table to carry out the test. Unfortunately, in testing H_0 versus H_1 and vice versa, all four possibilities (reject both, neither, or either one of the two hypotheses) could occur. This issue, however, is a finite sample problem. Davidson and MacKinnon show that as $n \to \infty$, if H_1 is true, then the probability that $\hat{\lambda}$ will differ significantly from 0 approaches 1.

Example 7.2 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 Appendix 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 through 2000.4). The least squares regression of C on a constant, Y, lagged Y, and the fitted values from the second model produces an estimate of λ of 1.0145 with a t ratio of 62.861. Thus, H_0 should be rejected in favor of H_1. But reversing the roles of H_0 and H_1, we obtain an estimate of λ of −10.677 with a t ratio of −7.188. Thus, H_1 is rejected as well.[7]

7.3.4 VUONG'S TEST AND THE KULLBACK–LEIBLER INFORMATION CRITERION

Vuong's (1989) approach to testing **nonnested models** is also based on the likelihood ratio statistic.[8] The logic of the test is similar to that which motivates the likelihood ratio test in general. Suppose that $f(y_i \mid \mathbf{Z}_i, \theta)$ and $g(y_i \mid \mathbf{Z}_i, \gamma)$ are two competing models for the density of the random variable y_i, with f being the null model, H_0, and g being the alternative, H_1. For instance, in Example 7.2, both densities are (by assumption now) normal, y_i is consumption, C_t, $\mathbf{Z}_i$ is $[1, Y_t, Y_{t-1}, C_{t-1}]$, θ is $(\beta_1, \beta_2, \beta_3, 0, \sigma^2)$, γ is $(\gamma_1, \gamma_2, 0, \gamma_3, \omega^2)$, and σ^2 and ω^2 are the respective conditional variances of the disturbances, ε_{0t} and ε_{1t}. The crucial element of Vuong's analysis is that it need not be the case that either competing model is "true"; they may both be incorrect. What we want to do is attempt to use the data to determine which competitor is closer to the truth, that is, closer to the correct (unknown) model.

We assume that observations in the sample (disturbances) are conditionally independent. Let $L_{i,0}$ denote the ith contribution to the likelihood function under the null hypothesis. Thus, the log likelihood function under the null hypothesis is $\Sigma_i \ln L_{i,0}$. Define $L_{i,1}$ likewise for the alternative model. Now, let m_i equal $\ln L_{i,1} - \ln L_{i,0}$. If we were using the familiar likelihood ratio test, then, the likelihood ratio statistic would be simply $LR = 2\Sigma_i m_i = 2n\overline{m}$ when $L_{i,0}$ and $L_{i,1}$ are computed at the respective maximum likelihood estimators. When the competing models are nested—H_0 is a restriction on H_1—we know that $\Sigma_i m_i \geq 0$. The restrictions of the null hypothesis will never increase

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

[8] Once again, it is necessary to rely on results that we will develop more fully in Chapter 16. But, this discussion of nonnested models is a convenient point at which to introduce Vuong's useful statistic, and we will not be returning to the topic of nonnested models save for a short application in Chapter 24.

5.9 A Specification Test

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

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

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

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

1. The test is nonconstructive. It gives no indication what the researcher should do next if the null model is rejected. This is a general feature of specification tests. Rejection of the null model does not imply any particular alternative.

2. Since the alternative hypothesis is unstated, it is unclear what the power of this test is against any specific alternative.

3. For this specific test (perhaps not for some other specification tests we will examine later), because $\mathbf{x}_i'\mathbf{b}$ uses the same $\mathbf{b}$ for every observation, the observations are correlated, while they are assumed to be uncorrelated in the original model. Because of the two step nature of the estimator, it is not clear what is the appropriate covariance matrix to use for the Wald test. Two other complications emerge for this test. First, it is unclear what γ converges to, assuming it converges to anything. Second, variance of the difference between $\mathbf{x}_i'\mathbf{b}$ and $\mathbf{x}_i'\boldsymbol{\beta}$ is a function of $\mathbf{x}_i$, so the second step regression might be heteroscedastic. The implication is that neither the size nor the power of this test are necessarily what might be expected.

Example 5.9. Size of a RESET Test

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

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

We take the estimated right hand side to be our population. We generated 5,000 samples of 430 (the original sample size), by reusing the regression coefficients and generating a new sample of disturbances for each replication. Thus, with each replication, r, we have a new sample of observations on $\ln Price_{ir}$ where the regression part is as aboye (reused) and a new set of disturbances is generated each time. With each sample, we computed the least squares coefficient, then the predictions. We then recomputed the least squares regression while adding the square and cube of the prediction to the regression. Finally, with each sample, we computed the chi-squared statistic, and rejected the null model if the chi-squared statistic is larger than 5.99, the 95th percentile of the chi squared distribution with two degrees of freedom. The **nominal size** of this test is 0.05. Thus, in samples of 100, 500, 1000 and 5000, we should reject the null nodel 5, 25, 50 and 250 times. In our experiment, the computed chi squared exceeded 5.99 8, 31, 65 and 259 times, respectively, which suggests that at least with sufficient replications, the test performs as might be expected. We then investigated the power of the test by adding 0.1 times the square of ln Area to the predictions. It is not possible to deduce the exact power of the RESET test to detect this failure of the null model. In our experiment, with 1,000 replications, the null hypothesis is rejected 321 times. We conclude that the procedure does appear have power to detect this failure of the model assumptions.

AU! KT "nominal size" not in chap. list

5.10 MODEL BUILDING—A GENERAL TO SIMPLE STRATEGY

There has been a shift in the general approach to model building in the past 20 years or so, partly based on the results in the previous two sections. With an eye toward maintaining simplicity, model builders would generally begin with a small specification and gradually build up the model ultimately of interest by adding variables. But, based on the preceding results, we can surmise that just about any criterion that would be used to decide whether to add a variable to a current specification would be tainted by the biases caused by the incomplete specification at the early steps. Omitting variables from the equation seems generally to be the worse of the two errors. Thus, the **simple-to-general** approach to model building has little to recommend it. Building on the work of Hendry [e.g., (1995)] and aided by advances in estimation hardware and software, researchers are now more comfortable beginning their specification searches with large elaborate models involving many variables and perhaps long and complex lag structures. The attractive strategy is then to adopt a **general-to-simple,** downward reduction of the model to the preferred specification. (This approach has been completely automated in Hendry's *PCGets*(c) computer program. [See, e.g., Hendry and Kotzis (2001).]). Of course, this must be tempered by two related considerations. In the "kitchen sink" regression, which contains every variable that might conceivably be relevant, the adoption of a fixed probability for the type I error, say, 5 percent, ensures that in a big enough model, some variables will appear to be significant, even if "by accident." Second, the problems of pretest estimation and **stepwise model building** also pose some risk of ultimately misspecifying the model. To cite one unfortunately common example, the statistics involved often produce unexplainable lag structures in dynamic models with many lags of the dependent or independent variables.

Vuong's general result for nonnested models (his Theorem 5.1) describes the behavior of the statistic

$$V = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} m_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \overline{m})^2}} = \sqrt{n}(\overline{m}/s_m), \quad m_i = \ln L_{i,0} - \ln L_{i,1}. \tag{7-14}$$

He finds:

(1) Under the hypothesis that the models are "equivalent", $V \xrightarrow{D} N[0,1]$

(2) Under the hypothesis that $f(y_i \mid \mathbf{Z}_i, \theta)$ is "better", $V \xrightarrow{A.S.} +\infty$

(3) Under the hypothesis that $g(y_i \mid \mathbf{Z}_i, \gamma)$ is "better", $V \xrightarrow{A.S.} -\infty$.

This test is directional. Large positive values favor the null model while large negative values favor the alternative. The intermediate values (e.g., between −1.96 and +1.96 for 95 percent significance) are an inconclusive region.

Example 7.3 Vuong Test for a Consumption Function

We conclude Example 7.2 by applying the **Vuong test** to the consumption data. For the linear model, $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ with normally distributed disturbances,

$$\ln L_i = -1/2\,[\ln \sigma^2 + \ln 2\pi + (y_i - \mathbf{x}_i'\beta)^2/\sigma^2] \tag{7-15}$$

and the maximum likelihood estimators of β and σ^2 are b and $\mathbf{e}'\mathbf{e}/n$. For the time-series data in Example 7.2, define for H_0, $e_{t0} = C_t - b_1 - b_2 Y_t - b_3 Y_{t-1}$ and $\mathbf{e}_0'\mathbf{e}_0 = \Sigma_t e_{t0}^2$. Define e_{t1} and $\mathbf{e}_0'\mathbf{e}_1$ likewise for H_1. Then, based on (7-14), we will have

$$\hat{m}_t = -1/2[\ln(\mathbf{e}_0'\mathbf{e}_0/\mathbf{e}_1'\mathbf{e}_1) + (e_{t0}^2/(\mathbf{e}_0'\mathbf{e}_0/T) - e_{t1}^2/(\mathbf{e}_0'\mathbf{e}_1/T))], \quad t = 1950.2, \ldots, 2000.4$$

(where $T = 203$). The Vuong statistic is −13.604, which once again strongly favors the alternative, H_1.

5.10.1 7.4 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 5.6) despite the improved fit to the data. With this thought in mind, the **adjusted R^2**,

$$\overline{R}^2 = 1 - \frac{n-1}{n-K}(1 - R^2) = 1 - \frac{n-1}{n-K}\left(\frac{\mathbf{e}'\mathbf{e}}{\sum_{i=1}^{n}(y_i - \overline{y})^2}\right), \tag{7-16}$$

has been suggested as a fit measure that appropriately penalizes the loss of degrees of freedom that result from adding variables to the model. Note that $\overline{R}^2$ may fall when a variable is added to a model if the sum of squares does not fall fast enough. (The applicable result appears in Theorem 3.7; $\overline{R}^2$ does not rise when a variable is added to a model unless the t ratio associated with that variable exceeds one in absolute value.)

The adjusted R^2 has been found to be a preferable fit measure for assessing the fit of forecasting models. [See Diebold (2003), who argues that the simple R^2 has a downward bias as a measure of the out-of-sample, one-step-ahead prediction error variance.]

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

$$\text{AIC}(K) = s_y^2(1 - R^2)e^{2K/n} \tag{7-17}$$

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

$$\text{BIC}(K) = s_y^2(1 - R^2)n^{K/n}. \tag{7-18}$$

(There is no degrees of freedom correction in s_y^2.) Both measures improve (decline) as R^2 increases (decreases), but, everything else constant, degrade as the model size increases. Like $\bar{R}^2$, 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

$$\text{AIC}(K) = \ln\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{2K}{n} \tag{7-19}$$

$$\text{BIC}(K) = \ln\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{K \ln n}{n}. \tag{7-20}$$

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

7.5 MODEL SELECTION

The preceding has laid out a number of choices for **model selection,** but, at the same time, has posed some uncomfortable propositions. The pretest estimation aspects of specification search are based on the model builder's knowledge of "the truth" and the consequences of failing to use that knowledge. While the cautions about blind search for statistical significance are well taken, it does seem optimistic to assume that the correct model is likely to be known with hard certainty at the outset of the analysis. The bias documented in (7-4) is well worth the modeler's attention. But, in practical terms, knowing anything about the magnitude presumes that we know what variables are in $\mathbf{X}_2$, which need not be the case. While we can agree that the model builder will omit income from a demand equation at their peril, we could also have some sympathy for the analyst faced with finding the right specification for their forecasting model among dozens of choices. The tests for nonnested models would seem to free the modeler from having to claim that the specified set of models contain "the truth." But, a moment's thought should suggest that the cost of this is the possibly deflated power of these procedures to

point toward that truth, The J test may provide a sharp choice between two alternatives, but it neglects the third possibility, that both models are wrong. Vuong's test does but, of course, it suffers from the fairly large inconclusive region, which is a symptom of its relatively low power against many alternatives. The upshot of all of this is that there remains much to be accomplished in the area of model selection. Recent commentary has provided suggestions from two perspective, classical and Bayesian.

5. Ref 10.3

~~7.5.1~~ CLASSICAL MODEL SELECTION

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

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

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

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

A (perhaps *the*) crucial issue remaining is uncertainty about the model itself. The search for the correct model is likely to have the same kinds of impacts on statistical inference as the search for a specification given the form of the model (see Section ~~7.2~~). Unfortunately, incorporation of this kind of uncertainty in statistical inference procedures remains an unsolved problem. Hansen suggests one potential route would be the Bayesian model averaging methods discussed next although he does express some skepticism about Bayesian methods in general.

4.3.2 and 4.3.3)

5.10.4 Ref 10.2

~~7.5.2~~ BAYESIAN MODEL AVERAGING

If we have doubts as to which of two models is appropriate, then we might well be convinced to concede that possibly neither one is really "the truth." We have painted ourselves into a corner with our "left or right" approach to testing. The Bayesian

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

An application by Wright (2003) provides an interesting illustration. Recent advances such as Bayesian VARs have improved the forecasting performance of econometric models. Stock and Watson (2001, 2004) report that striking improvements in predictive performance of international inflation can be obtained by averaging a large number of forecasts from different models and sources. The result is remarkably consistent across subperiods and countries. Two ideas are suggested by this outcome. First, the idea of blending different models is very much in the spirit of Hansen's fourth point. Second, note that the focus of the improvement is not on the fit of the model (point 3), but its predictive ability. Stock and Watson suggested that simple equal-weighted averaging, while one could not readily explain why, seems to bring large improvements. Wright proposed Bayesian model averaging as a means of making the choice of the weights for the average more systematic and of gaining even greater predictive performance.

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

$$\Pi_m = \text{Prob}(m = m^* \mid \mathbf{y}, \mathbf{Z}) = \frac{P(\mathbf{y}, \mathbf{Z} \mid m)\pi_m}{\sum_{r=1}^{M} P(\mathbf{y}, \mathbf{Z} \mid r)\pi_r}, \qquad \text{5-45 (7-21)}$$

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

$$P(\mathbf{y}, \mathbf{Z} \mid m) = \int_{\theta_m} P(\mathbf{y}, \mathbf{Z} \mid \theta_m, m) P(\theta_m) d\theta_m, \qquad \text{5-46 (7-22)}$$

while $P(\mathbf{y}, \mathbf{Z} \mid \theta_m, m)$ is the conditional (on θ_m) likelihood for the mth model and $P(\theta_m)$ is the analyst's prior over the parameters of the mth model. This provides an alternative set of weights to the $\Pi_m = 1/M$ suggested by Stock and Watson. Let $\hat{\theta}_m$ denote the

Bayesian estimate (posterior mean) of the parameters of model m. (See Chapter 18.) Each model provides an appropriate posterior forecast density, $f^*(\mathbf{y} \mid \mathbf{Z}, \hat{\theta}_m, m)$. The Bayesian model averaged forecast density would then be

$$\bar{f}^* = \sum_{m=1}^{M} f^*(\mathbf{y} \mid \mathbf{Z}, \hat{\theta}, m)\Pi_m. \tag{7-23}$$

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

Example 7.4 Bayesian Averaging of Classical Estimates

Many researchers have expressed skepticism of Bayesian methods because of the apparent arbitrariness of the specifications of prior densities over unknown parameters. In the Bayesian model averaging setting, the analyst requires prior densities over not only the model probabilities, π_m, but also the model specific parameters, θ_m. In their application, Doppelhofer, Miller, and Sala-i-Martin (2000) were interested in the appropriate set of regressors to include in a long-term macroeconomic (income) growth equation. With 32 candidates, M for their application was 2^{32} (minus one if the zero regressors model is ignored), or roughly four billion. Forming this many priors would be optimistic in the extreme. The authors proposed a novel method of weighting a large subset (roughly 21 million) of the 2^M possible (classical) least squares regressions. The weights are formed using a Bayesian procedure; however, the estimates that are weighted are the classical least squares estimates. While this saves considerable computational effort, it still requires the computation of millions of least squares coefficient vectors. [See Sala-i-Martin (1997).] The end result is a model with 12 independent variables.

7.6 SUMMARY AND CONCLUSIONS

This is the last of six 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. Chapter 4 reinterpreted least squares as an estimator of an unknown parameter vector, and described the finite sample and large sample characteristics of the sampling distribution of the estimator. Chapters 5 and 6 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.

5.11 SUMMARY AND CONCLUSIONS

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

This is the last of five chapters that we have devoted specifically to the methodology surrounding the most heavily used tool in econometrics, the classical linear regression model. We began in Chapter 2 with a statement of the regression model. Chapter 3 then described computation of the parameters by least squares—a purely algebraic exercise. Chapter 4 reinterpreted least squares as an estimator of an unknown parameter vector, and described the finite sample and large sample characteristics of the sampling distribution of the estimator. Chapter 5 was devoted to building and sharpening the regression model, with statistical results for testing hypotheses about the underlying population. In this chapter, we have examined some broad issues related to model specification and selection of a model among a set of competing alternatives. The concepts considered here are tied very closely to one of the pillars of the paradigm of econometrics, 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.

5-44

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.

5.7 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
- Lack of invariance
- Lagrange multiplier test
- Linear restrictions
- Nested models
- Nonnested models
- Nonnormality
- Null hypothesis
- Parameter space
- Prediction interval
- Prediction variance
- Restricted least squares
- Root mean squared error
- Testable implications
- Theil U statistic
- Wald statistic

add (A) from page 147

Exercises

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

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

   Test the hypothesis that the two slopes sum to 1.
2. Using the results in Exercise 1, test the hypothesis that the slope on x_1 is 0 by running the restricted regression and comparing the two sums of squared deviations.
3. The regression model to be analyzed is $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, where $\mathbf{X}_1$ and $\mathbf{X}_2$ have K_1 and K_2 columns, respectively. The restriction is $\boldsymbol{\beta}_2 = \mathbf{0}$.
   a. Using (5-14), 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 $\mathbf{b}_{1*} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_2^0)$.

5-23

4. The expression for the restricted coefficient vector in (5-14) 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

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

TU: The following terms were not bold KTs in text: ~~Ex post forecast~~, ~~Linear restrictions~~, ~~nonnormality~~, ~~Prediction interval~~, ~~Prediction variance~~, Restricted least squares, ~~Root mean squared error~~, ~~Theil U statistic~~, Biased estimator, ~~Inclusion of superfluous variables~~, ~~mean squared error~~, Omission of relevant variables, Omitted variable formula, ~~Pretest estimator~~, ~~Specification analysis~~. Mark bold in text or delete from list.

and that this matrix may be written as

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

5. Prove the result that the restricted least squares estimator never has a larger covariance matrix than the unrestricted least squares estimator.
6. Prove the result that the R^2 associated with a restricted least squares estimator is never larger than that associated with the unrestricted least squares estimator. Conclude that imposing restrictions never improves the fit of the regression.
7. An alternative way to test the hypothesis $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$ is to use a Wald test of the hypothesis that $\boldsymbol{\lambda}_* = \mathbf{0}$, where $\boldsymbol{\lambda}_*$ is defined in (5-14). 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 (5-19) and the preceding discussion, we know that this ratio is greater than 1. Finally, prove that this test statistic is equivalent to JF, where J is the number of restrictions being tested and F is the conventional F statistic given in (5-6). Formally, the Lagrange multiplier test requires that the variance estimator be based on the restricted sum of squares, not the unrestricted. Then, the test statistic would be $\text{LM} = nJ/[(n-K)/F + J]$. See Godfrey (1988).
8. Use the test statistic defined in Exercise 7 to test the hypothesis in Exercise 1.
9. Prove that under the hypothesis that $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, the estimator

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

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

Applications

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

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

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

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

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

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

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

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

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

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

Note that although the underlying theory requires it, the model can be estimated (by least squares) without imposing the linear homogeneity restrictions. [Thus, one could "test" the underlying theory by testing the validity of these restrictions. See Christensen, Jorgenson, and Lau (1975).] We will repeat this exercise in part b.

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

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

AU: KT "translog cost function" not in chap. list.

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

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

5-8

a. Write out the **R** matrix and **q** vector in (5-3) that are needed to impose the restriction of linear homogeneity in prices.

b. "Test" the theory of production using all 158 observations. Use an F test to test the restrictions of linear homogeneity. Note, you can use the general form of the F statistic in (5-7) to carry out the test. Christensen and Greene enforced the linear homogeneity restrictions by building them into the model. You can do this by dividing cost and the prices of capital and labor by the price of fuel. Terms with f subscripts fall out of the model, leaving an equation with ten parameters. Compare the sums of squares for the two models to carry out the test. Of course, the test may be carried out either way and will produce the same result.

5-16

10

c. Test the hypothesis homotheticity of the production structure under the assumption of linear homogeneity in prices.

d. Test the hypothesis of the generalized Cobb–Douglas cost function in Chapter 4 against the more general translog model suggested here, once again (and henceforth) assuming linear homogeneity in the prices.

e. The simple Cobb–Douglas function appears in the first line of the model above. Test the hypothesis of the Cobb–Douglas model against the alternative of the full translog model.

f. Test the hypothesis of the generalized Cobb–Douglas model against the homothetic translog model.

g. Which of the several functional forms suggested here to you conclude is the most appropriate for these data?

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

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

a. Carry out a test of the hypothesis that the three aggregate price indices are not significant determinants of the demand for gasoline.

b. Consider the hypothesis that the microelasticities are a constant proportion of the elasticity with respect to their corresponding aggregate. Thus, for some positive θ (presumably between 0 and 1), $\gamma_{nc} = \theta\delta_d$, $\gamma_{uc} = \theta\delta_d$, $\gamma_{pt} = \theta\delta_s$. The first two imply the simple linear restriction $\gamma_{nc} = \gamma_{uc}$. By taking ratios, the first (or second) and third imply the nonlinear restriction

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

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

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

insert (c) from page 5-48 msp 147

5-48 / End 5

Insert on msp 5-44 where indicated

Key Terms and Concepts

- Adjusted R-squared
- Akaike Information Criterion
- Bayesian model averaging
- Bayesian Information Criterion
- Biased estimator
- Comprehensive model
- Encompassing principle
- General-to-simple strategy
- Inclusion of superfluous variables
- J test
- ~~Kullback–Leibler Information Criterion~~
- 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
- ~~Vuong's test~~

Insert terms by number in alphabetical order. Do not set numbers

A

end of insert A

Insert on msp 5-45

Exercises

4-8 9-10

1. Suppose the true regression model is given by ~~(7-2)~~. The result in ~~(7-4)~~ shows that if either $\mathbf{P}_{1.2}$ is nonzero or β_2 is nonzero, then regression of $\mathbf{y}$ on $\mathbf{X}_1$ alone produces a biased and inconsistent estimator of β_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 β_1 with $\mathbf{b}_1$ (which is biased). Is the forecast of $\mathbf{y}$ computed using $\mathbf{X}_1\mathbf{b}_1$ also biased? Assume that $E[\mathbf{X}_2 \mid \mathbf{X}_1]$ is a linear function of $\mathbf{X}_1$. Discuss your findings generally. What are the implications for prediction when variables are omitted from a regression?
2. Compare the mean squared errors of b_1 and $b_{1.2}$ in Section ~~7.2.2~~ 4.7.2 (~~Hint: the comparison~~ Example 4.6. depends on the data and the model parameters, but you can devise a compact expression for the two quantities.)
3. ~~An individual certain~~ The log likelihood function for the linear regression model with normally distributed disturbances is shown in ~~(7-15)~~. Show that at the maximum likelihood estimators of $\mathbf{b}$ for $\boldsymbol{\beta}$ and $\mathbf{e}'\mathbf{e}/n$ for σ^2, the log likelihood is an increasing function of R^2 for the model.
4. Show that the model of the alternative hypothesis in Example ~~7.2~~ 5.7 can be written

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

As such, it does appear that H_0 is a restriction on H_1. However, because there are an infinite number of constraints, this does not reduce the test to a standard test of restrictions. It does suggest the connections between the two formulations. (We will revisit models of this sort in Chapter ~~20~~ 21.)

B

end of insert B

Insert on msp 5-47

Applications

4 ~~X~~ The J test in Example ~~7.2~~ 5.7 is carried out using more than 50 years of data. It is optimistic to hope that the underlying structure of the economy did not change in 50 years. Does the result of the test carried out in Example 7.2 persist if it is based on data only from 1980 to 2000? Repeat the computation with this subset of the data.

5.7

C

end of insert C