5
Specification

5.1 Introduction

At one time econometricians tended to assume that the model provided by economic theory represented accurately the real-world mechanism generating the data, and viewed their role as one of providing "good" estimates for the key parameters of that model. If any uncertainty was expressed about the model specification, there was a tendency to think in terms of using econometrics to "find" the real-world data-generating mechanism. Both these views of econometrics are obsolete. It is now generally acknowledged that econometric models are "false" and that there is no hope, or pretense, that through them "truth" will be found. Feldstein's (1982, p. 829) remarks are typical of this view: "in practice all econometric specifications are necessarily 'false' models. . . The applied econometrician, like the theorist, soon discovers from experience that a useful model is not one that is 'true' or 'realistic' but one that is parsimonious, plausible and informative." This is echoed by an oft-quoted remark attributed to George Box - "All models are wrong, but some are useful" - and another from Theil (1971, p. vi): "Models are to be used, but not to be believed."

In light of this recognition, econometricians have been forced to articulate more clearly what econometric models are. There is some consensus that models are metaphors, or windows, through which researchers view the observable world, and that their acceptance and use depends not upon whether they can be deemed "true" but rather upon whether they can be said to correspond to the facts. Econometric specification analysis is a means of formalizing what is meant by "corresponding to the facts," thereby defining what is meant by a "correctly specified model." From this perspective econometric analysis becomes much more than estimation and inference in the context of a given model; in conjunction with economic theory, it plays a crucial, preliminary role of searching for and evaluating a model, leading ultimately to its acceptance or rejection.

Econometrics textbooks are mainly devoted to the exposition of econometrics for estimation and inference in the context of a given model for the data-generating process. The more important problem of specification of this model is not given much attention, for three main reasons. First, specification is not easy. In

the words of Hendry and Richard (1983, p. 112), "the data generation process is complicated, data are scarce and of uncertain relevance, experimentation is uncontrolled and available theories are highly abstract and rarely uncontroversial." Second, most econometricians would agree that specification is an innovative/imaginative process that cannot be taught: "Even with a vast arsenal of diagnostics, it is very hard to write down rules that can be used to guide a data analysis. So much is really subjective and subtle. . . A great deal of what we teach in applied statistics is *not* written down, let alone in a form suitable for formal encoding. It is just simply 'lore'" (Welsch, 1986, p. 405). And third, there is no accepted "best" way of going about finding a correct specification.

There is little that can be done about items one and two above; they must be lived with. Item three, however, is worthy of further discussion: regardless of how difficult a specification problem, or how limited a researcher's powers of innovation/imagination, an appropriate methodology should be employed when undertaking empirical work. The purpose of this chapter is to discuss this issue; it should be viewed as a prelude to the examination in chapter 6 of specific violations of the first assumption of the CLR model.

5.2 Three Methodologies

Until about the mid-1970s, econometricians were too busy doing econometrics to worry about the principles that were or should be guiding empirical research. Sparked by the predictive failure of large-scale econometric models, and fueled by dissatisfaction with the gap between how econometrics was taught and how it was applied by practitioners, the profession began to examine with a critical eye the way in which econometric models were specified. This chapter is in part a summary of the state of this ongoing methodological debate. At considerable risk of oversimplification, three main approaches to the

specification problem are described below in stylized form.

(I) Average Economic Regression (AER)

This approach describes what is thought to be the usual way in which empirical work in economics is undertaken. The researcher begins with a specification that is viewed as being known to be correct, with data being used primarily to determine the orders of magnitude of a small number of unknown parameters. Significant values of diagnostic test statistics, such as the Durbin-Watson statistic, are initially interpreted as suggesting estimation problems that should be dealt with by adopting more sophisticated estimation methods, rather than as pointing to a misspecification of the chosen model. If these more sophisticated methods fail to "solve" the problem, the researcher then conducts "specification" tests, hunting for an alternative specification that is "better", using age-old criteria such as correct signs, high R2s, and significant *t* values on coefficients

"known" to be nonzero. Thus in the AER approach the data ultimately do play a role in the specification, despite the researcher's initial attitude regarding the validity of the theoretical specification. This role may be characterized as proceeding from a simple model and "testing up" to a specific more general model.

(2) Test, Test, Test (TTT)

This approach uses econometrics to discover which models of the economy are tenable, and to test rival views. To begin, the initial specification is made more general than the researcher expects the specification ultimately chosen to be, and testing of various restrictions, such as sets of coefficients equal to the zero vector, is undertaken to simplify this general specification; this testing can be characterized as "testing down" from a general to a more specific model. Following this, the model is subjected to a battery of diagnostic, or misspecification, tests, hunting for signs that the model is misspecified. (Note the contrast with AER "specification" tests, which hunt for specific alternative specifications.) A significant diagnostic, such as a small DW value, is interpreted as pointing to a model misspecification rather than as pointing to a need for more sophisticated estimation methods. The model is continually respecified until a battery of diagnostic tests allows

a researcher to conclude that the model is satisfactory on several specific criteria (discussed in the general notes), in which case it is said to be "congruent" with the evidence.

(3) Fragility Analysis

The specification ultimately arrived at by the typical AER or TTT search may be inappropriate because its choice is sensitive to the initial specification investigated, the order in which tests were undertaken, type I and type II errors, and innumerable prior beliefs of researchers concerning the parameters that subtly influence decisions taken (through the exercise of innovation/imagination) throughout the specification process. It may, however, be the case that the different possible specifications that could have arisen from the AER or the TTT approaches would all lead to the same conclusion with respect to the purpose for which the study was undertaken, in which case why worry about the specification? This is the attitude towards specification adopted by the third approach. Suppose that the purpose of the study is to estimate the coefficients of some "key" variables. The first step of this approach, after identifying a general family of models, is to undertake an "extreme bounds analysis," in which the coefficients of the key variables are estimated using all combinations of included/excluded "doubtful" variables. If the resulting range of estimates is too wide for comfort, an attempt is made to narrow this range by conducting a "fragility analysis." A Bayesian method (see chapter 13) is used to incorporate non-sample information into the estimation, but in such a way as to allow for a range of this Bayesian information, corresponding to the range of such informa-

tion that will surely characterize the many researchers interested in this estimation. This range of information will produce a range of estimates of the parameters of interest; a narrow range ("sturdy" estimates) implies that the data at hand yield useful information, but if this is not the case ("fragile" estimates), it must be concluded that inferences from these data are too fragile to be believed.

Which is the best of these three general approaches? There is no agreement that one of these methodologies is unequivocally the best to employ; each has faced criticism, a general summary of which is

provided below.

(1) The AER is the most heavily criticized, perhaps because it reflects most accurately what researchers actually do. It is accused of using econometrics merely to illustrate assumed-known theories. The attitude that significant diagnostics reflect estimation problems rather than specification errors is viewed in an especially negative light, even by those defending this approach. "Testing up" is recognized as inappropriate, inviting type I errors through loss of control over the probability of a type I error. The *ad hoc* use of extraneous information, such as the "right" signs on coefficient estimates, is deplored, especially by those with a Bayesian bent. The use of statistics such as R2, popular with those following this methodology, is frowned upon. Perhaps most frustrating to critics is the lack of a well-defined structure and set of criteria for this approach; there is never an adequate description of the path taken to the ultimate specification.

(2) The TTT methodology is also criticized for failing in practice to provide an adequate description of the path taken to the ultimate specification, reflecting an underlying suspicion that practitioners using this methodology find it necessary to use many of the *ad hoc* rules of thumb followed in the AER approach. This could in part be a reflection of the role played in specification by innovation/imagination, which cannot adequately be explained or defended, but is nonetheless unsettling. The heavy reliance on testing in this methodology raises fears of a proliferation of type I errors (creating pretest bias, discussed in section 12.4 of chapter 12), exacerbated by the small degrees of freedom due to the very general initial specification and by the fact that many of these tests have only asymptotic justification. When "testing up" the probability of a type I error is neither known nor controlled; using the "testing down" approach can allay these fears by the adoption of a lower a value for the tests, but this is not routinely done.

(3) Objections to fragility analysis usually come from those not comfortable with the Bayesian approach, even though care has been taken to make it palatable to non-Bayesians. Such objections are theological in nature and not likely to be resolved. There is vagueness regarding how large a range of parameter estimates has to be to conclude that it is fragile; attempts to formalize this lead to measures comparable to the test statistics this approach seeks to avoid. The methodology never does lead to the adoption of a specific specification, something that researchers find unsatisfactory. There is no scope for the general fami-

ly of models initially chosen to be changed in the light of what the data has to say. Many researchers find Bayesian prior formulation both difficult and alien. Some object that this analysis too often concludes that results are fragile.

5.3 General Principles for Specification

Although the controversy over econometric methodology has not yet been resolved, the debate has been fruitful in that some general principles have emerged to guide model specification.

(1) Although "letting the data speak for themselves" through econometric estimation and testing is an important part of model specification, economic theory should be the foundation of and guiding force in a specification search.

(2) Models whose residuals do not test as insignificantly different from white noise (random errors) should be initially viewed as containing a misspecification, not as needing a special estimation procedure, as too many researchers are prone to do.

(3) "Testing down" is more suitable than "testing up"; one should begin with a general, unrestricted model and then systematically simplify it in light of the sample evidence. In doing this a researcher should control the overall probability of a type I error by adjusting the a value used at each stage of the testing (as explained in the technical notes), something which too many researchers neglect to do. This approach, deliberate overfitting, involves a loss of efficiency (and thus loss of power) when compared to a search beginning with a correct simple model. But this simple model may not be correct, in which case the approach of beginning with a simple model and expanding as the data permit runs the danger of biased inference resulting from underspecification.

(4) Tests of misspecification are better undertaken by testing simultaneously for several misspecifications rather than testing one-by-one for these misspecifications. By such an "overtesting" technique one avoids the problem of one type of misspecification. This approach helps to deflect the common criticism that such tests rely for their power on aspects of the maintained hypothesis about which little is

known.

(5) Regardless of whether or not it is possible to test simultaneously for misspecifications, models should routinely be exposed to a battery of misspecification diagnostic tests before being accepted. A subset of the data should be set aside before model specification and estimation, so that these tests can include tests for predicting extra-sample observations.

(6) Researchers should be obliged to show that their model encompasses rival models, in the sense that it can predict what results would be obtained were one to run the regression suggested by a rival model. The chosen model

should be capable of explaining the data and of explaining the successes and failures of rival models in accounting for the same data.

(7) Bounds on the range of results corresponding to different reasonable specifications should be reported, rather than just providing the results of the specification eventually adopted, and the path taken to the selection of this specification should be fully reported.

5.4 Misspecification Tests/Diagnostics

Despite the protestations of fragility analysis advocates, testing has come to play a more and more prominent role in econometric work. Thanks to the ingenuity of econometric theorists, and the power of asymptotic algebra, an extremely large number of tests have been developed, seemingly catering to practitioners' every possible need, but at the same time courting confusion because of unknown small-sample properties, suspicions of low power, and often-conflicting prescriptions. It is not possible in this book to discuss all or even a majority of these tests. The more prominent among them are discussed briefly in later chapters when it is relevant to do so; before moving on to these chapters, however, it may be useful to have an overview of tests used for specification purposes. They fall into several categories.

(1) *Omitted variable (OV) tests* F and t tests for zero restrictions on (or, more generally, linear combinations of) the parameters, as discussed in chapter 4, are commonly used for specification purposes. Several more

complicated tests, such as Hausman tests, can be reformulated as OV tests in an artificial regression, greatly simplifying testing.

(2) *RESET tests* RESET tests, discussed in chapter 6, are used to test for whether unknown variables have been omitted from a regression specification, and are not to be confused with OV tests that test for zero coefficients on known variables. They can also be used to detect a misspecified functional form.

(3) *Tests for functional form* Two types of tests for functional form are available, as discussed in chapter 6. The first type, such as tests based on recursive residuals and the rainbow test, does not specify a specific alternative functional form. For the second type, functional form is tested by testing a restriction on a more general functional form, such as a Box-Cox transformation.

(4) *Tests for structural change* In this category fall tests for parameter constancy, discussed in chapter 6, such as Chow tests, cusum and cusum-of-squares tests, and predictive failure (or post-sample prediction) tests.

(5) *Tests for outliers* These tests, among which are included tests for normality, are sometimes used as general tests for misspecification. Examples

are the Jarque-Bera test, the Shapiro-Wilk test, the Cook outlier test, and the use of the DFITS measure (discussed in chapter 18).

(6) *Tests for non-spherical errors* These are tests for various types of serial correlation and heteroskedasticity, discussed in chapter 8. Examples are the Durbin-Watson test, the Breusch-Godfrey test, Durbin's *h* and *m* tests, the Goldfeld-Quandt test, the Breusch-Pagan test and the White test.

(7) *Tests for exogeneity* These tests, often referred to as Hausman tests, test for contemporaneous correlation between regressors and the error. They are discussed in chapter 9 (testing for measurement error) and chapter 10 (testing for simultaneous equation bias).

(8) *Data transformation tests* These tests, which do not have any specific alternative hypothesis, are considered variants of the Hausman test. Examples are the grouping test and the differencing test.

(9) *Non-nested tests* When testing rival models that are not nested, as might arise when testing for encompassing, non-nested tests must be employed. Examples are the non-nested *F* test and the *J* test.

(10) *Conditional moment tests* These tests are based on a very general testing methodology which in special cases gives rise to most of the tests listed above. Beyond serving as a unifying framework for existing tests, the value of this testing methodology is that it suggests how specification tests can be undertaken in circumstances in which alternative tests are difficult to construct.

Categorizing tests in this way is awkward, for several reasons.

(1) Such a list will inevitably be incomplete. For example, it could be expanded to incorporate tests for specification encountered in more advanced work. Should there be categories for unit root and cointegration tests (see chapter 17), identification tests (see chapter 10), and selection bias tests (see chapter 16), for example? What about Bayesian "tests"?

(2) It is common for practitioners to use a selection criterion, such as the Akaike information criterion, or adjusted R2, to aid in model specification, particularly for determining things like the number of lags to include. Should this methodology be classified as a test?

(3) These categories are not mutually exclusive. There are non-nested variants of tests for non-spherical errors and of functional form tests, some tests for functional form are just variants of tests for structural break, and the RESET test is a special case of an OV test, for example.

(4) Tests take different forms. Some are LM tests, some are LR tests, and some are W tests. Some use *F*-tables, some use *t*-tables, some use *X*2-tables, and some require their own special tables. Some are exact tests, whereas some rest on an asymptotic justification.

(5) Some tests are "specification" tests, involving a specific alternative, whereas others are "misspecification" tests, with no specific alternative.

This last distinction is particularly relevant for this chapter. A prominent feature of the list of general principles given earlier is the use of misspecification tests, the more common of which are often referred to as diagnostics. These tests are designed to detect an inadequate specification (as opposed to "specification" tests, which examine the validity of a specific alternative). There have been calls for researchers to submit their models to misspecification tests as a matter of course, and it is becoming common for computer packages automatically to print out selected diagnostics.

Of the tests listed above, several fall into the misspecification category. Possibly the most prominent are the non-spherical error tests. As stressed in chapter 8, a significant value for the DW statistic could be due to several misspecifications (an omitted variable, a dynamic misspecification, or an incorrect functional form), not just to autocorrelated errors, the usual conclusion drawn by those following the AER methodology. The same is true of tests for heteroskedasticity. As noted in chapter 6, significant values of RESET could be due to an incorrect functional form, and tests for structural break and the first type of functional form test statistic could be significant because of a structural break, an omitted variable or an incorrect functional form. So these tests should be viewed as misspecification tests. Outliers could arise from a variety of specification errors, so they also can be classified as misspecification tests.

It could be argued that the misspecification tests mentioned in the preceding paragraph are to some extent specification tests because they can be associated with one or more specific classes of alternatives that have inspired their construction. Because of this they are discussed in later chapters when that class of alternative is addressed. Three of the tests listed above, however, are sufficiently general in nature that there is no obvious alternative specification determining where they should appear in later chapters. These are data transformation tests, non-nested tests, and conditional moment tests.

Data transformation tests The idea behind data transformation tests is that if the null hypothesis of a linear functional form with a set of specific explanatory variables is correct, then estimating with raw data should yield coefficient estimates very similar to those obtained from using linearly transformed data. If the two sets of estimated coefficients are not similar, one can conclude that the null hypothesis is not correct, but one cannot draw any conclusion about what dimension of that null

hypothesis is incorrect, since many different misspecifications could have given rise to this discrepancy. Choosing a specific transformation, and formalizing what is meant by "very similar," produces a test statistic. Fortunately, as explained in the technical notes, data transformation tests have been shown to be equivalent to OV tests, greatly simplifying their application.

Non-nested tests Two models are non-nested (or "separate") if one cannot be obtained from the other by imposing a restriction. The importance of this distinction is that in this circumstance it is not possible to follow the usual testing methodology, namely to employ a test of the restriction as a specification test. Non-nested hypothesis tests provide a means of testing the specification of one

model by exploiting the supposed "falsity" of other models. A model chosen to play the role of the "other" model need not be an alternative model under consideration, but this is usually the case. If the null model is the "correct" model, then the "other" model should not be able to explain anything beyond that explained by the null model. Formalizing this, as explained in the technical notes, produces a non-nested hypothesis test, on the basis of which the null can be either rejected or not rejected/accepted. If the former is the case, then one cannot conclude that the "other" model should be accepted - the role of the "other" model in this exercise is simply to act as a standard against which to measure the performance of the null. (This is what makes this test a misspecification test, rather than a specification test.) If one wants to say something about the "other" model, then the roles of the two hypotheses must be reversed, with the "other" model becoming the null, and the test repeated. Note that in this testing procedure it is possible to reject both models or to accept both models.

Conditional moment tests These tests are undertaken by selecting a function of the data and parameters that under a correct specification should be zero, computing this function for each observation (evaluated at the MLEs), taking the average over all the observations and testing this average against zero. The function used for this purpose is usually a moment or a conditional moment (such as the product of an exogenous variable and the residual), explaining why these tests are called *moment* (M) or *conditional moment* (CM) tests. The test would be formed by creating an estimate of the variance-covariance matrix of this average

and using a Wald test formula. Its main appeal is that in some circumstances it is easier to formulate appropriate moment conditions than to derive alternative tests.

5.5 R2 Again

The coefficient of determination, R2, is often used in specification searches and in undertaking hypothesis tests. Because it is so frequently abused by practitioners, an extension of our earlier (section 2.4) discussion of this statistic is warranted.

It is noted in the general notes to section 4.3 that the *F* test could be interpreted in terms of R2 and changes in R2. Whether or not a set of extra independent variables belongs in a relationship depends on whether or not, by adding the extra regressors, the R2 statistic increases significantly. This suggests that, when one is trying to determine which independent variable should be included in a relationship, one should search for the highest R2.

This rule would lead to the choice of a relationship with too many regressors (independent variables) in it, because the addition of a regressor cannot cause the R2 statistic to fall (for the same reason that the addition of a regressor cannot cause the minimized sum of squared residuals to become larger - minimizing without the restriction that the extra regressor must be ignored gives at least as

low a minimand as when the restriction is imposed). Correcting the R2 statistic for degrees of freedom solves this problem. The R2 statistic adjusted to account for degrees of freedom is called the "adjusted R2" or "R2" and is now reported by most packaged computer regression programs, and by practically all researchers, in place of the unadjusted R2.

Adding another regressor changes the degrees of freedom associated with the measures that make up the R2 statistic. If an additional regressor accounts for very little of the unexplained variation in the dependent variable, R2 falls (whereas R2 rises). Thus, only if R2 rises should an extra variable be seriously considered for inclusion in the set

of independent variables. This suggests that econometricians should search for the "best" set of independent variables by determining which potential set of independent variables produces the highest R2. This procedure is valid only in the sense that the "correct set" of independent variables will produce, on average in repeated samples, a higher R2 than will any "incorrect" set of independent variables.

Another common use of the R2 statistic is in the context of measuring the relative importance of different independent variables in determining the dependent variable. Textbooks present several ways of decomposing the R2 statistic into component parts, each component being identified with one independent variable and used as a measure of the relative importance of that independent variable in the regression. Unfortunately, none of these partitions of R2 is meaningful unless it happens that the independent variables are uncorrelated with one another in the sample at hand. (This happens only by experimental design or by extraordinary luck, economists almost never being in a position to effect either.) In the typical case in which the independent variables are correlated in the sample, these suggested partitionings are not meaningful because: (a) they can no longer be legitimately allocated to the independent variables; (b) they no longer add up to R2; or (c) they do add up to R2 but contain negative as well as positive terms.

The main reason for this can be explained as follows. Suppose there are only two independent variables, and they are correlated in the sample. Two correlated variables can be thought of as having, between them, three sorts of variation: variation unique to the first variable, variation unique to the second variable and variation common to both variables. (When the variables are uncorrelated, this third type of variation does not exists.) Each of the three types of variation in this set of two variables "explains" some of the variation in the dependent variable. The basic problem is that no one can agree how to divide the explanatory power of the common variation between the two independent variables. If the dependent variable is regressed on both independent variables, the resulting R2 reflects the explanatory power of all three types of independent variable variation. If the dependent variable is regressed on only one independent variable, variation unique to the other variable is removed and the resulting R2 reflects the explanatory power of the other two types of independent variable variation. Thus, if one

independent variable is removed, the remaining variable gets credit for *all* of the common variation. If the second independent variable were reinstated and the resulting increase in R2 were used to measure the influence of this second variable, this variable would get credit for *none* of the common variation. Thus it would be illegitimate to measure the influence of an independent variable either by its R2 in a regression of the dependent variable on only that independent variable, or by the addition to R2 when that independent variable is added to a set of regressors. This latter measure clearly depends on the order in which the independent variables are added. Such procedures, and others like them, can only be used when the independent variables are uncorrelated in the sample. The use of breakdowns of the R2 statistic in this context should be avoided.

5.1 General Notes

5.1 Introduction

Economists' search for "truth" has over the years given rise to the view that economists are people searching in a dark room for a non-existent black cat; econometricians are regularly accused of finding one.

The consensus reported in the second paragraph of this chapter may or may not exist. Some quotations reflecting views consistent with this interpretation are Pesaran (1988, p. 339), "econometric models are metaphors providing different windows on a complex and bewildering reality," and Poirier (1988, p. 139), "'Truth' is really nothing more than a 'correspondence' with the facts, and an important role for econometrics is to articulate what constitutes a satisfactory degree of correspondence."

That model specification requires creativity and cannot be taught is widely acknowledged. Consider, for example, the remarks of Pagan (1987, p. 20): "Constructing 'systematic theologies' for econometrics can well stifle creativity, and some evidence of this has already become apparent. Few would deny that in the hands of the masters the methodologies perform impressively, but in the hands of their disciples it is all much less convincing."

Economists are often accused of never looking at their data - they seldom dirty their hands with primary data collection, using instead secondary data sources available in electronic form. Indeed, as noted by Reuter (1982, p. 137) "Economists are unique among social scientists in that they are trained only to analyses, not to collect, data.. . . One consequence is a lack of scepticism about the quality of data." Aigner (1988) stresses how dependent we are on data of unknown quality, generated by others for purposes that do not necessarily correspond with our own, and notes (p. 323) that "data generation is a dirty, time-consuming, expensive and non-glorious job." All this leads to an inexcusable lack of familiarity with the data, a source of many errors in econometric specification and analysis. This suggests that a possible route to finding better specifications is to focus on getting more and better data, and looking more carefully at these data, rather than on fancier techniques for dealing with existing data. Breuer and Wohar (1996) is an example in which knowing the institutional details of how the data were produced can aid an econometric analysis. Chatfield (1991) has some good examples of how empirical work can be greatly enhanced by being sensi-

tive to the context of the problem (the data-generating process) and knowing a lot about one's data.

EDA, exploratory data analysis, is an approach to statistics which emphasizes that a researcher should always begin by looking carefully at the data in a variety of imaginative ways. Exploratory data analysts use the *interocular trauma test:* keep looking at the data until the answer hits you between the eyes! For an exposition of EDA, see Hartwig and Dearing (1979), and for examples see L. S. Mayer (1980) and Denby and Pregibon (1987). Maddala (1988, pp. 557) presents a nice example from Anscombe (1973) in which four sets of data give rise to almost identical regression coefficients, but very different graphs. Leamer (1994, p. xiii) has an amusing graph in which the data spell out HELP. Unwin (1992) discusses how interactive graphics should revolutionize statistical practice. Perhaps econometric software should have built into it some means of preventing a user from running a regression until the data have been examined!

5.2 Three Methodologies

Pagan (1987) has a good account of the wakening of the profession's interest in econometric methodology. Pagan (1995) is an update; Granger (1990) contains a selection of articles prominent in this controversy. Hendry et al. (1990) is an instructive informal discussion of these issues. In this context the word "methodology" refers to the principles of the procedures adopted in the testing and quantification of economic theories, in contrast to its more popular use as a synonym for econometric "technique" or "method." Nakamura et al. (1990) is a useful survey of methods of model specification. Readers should be warned that many econometricians do not view this debate over econometric methodology with favor; they prefer not to worry about such issues. This invariably means that they continue to use the approach to specification they have always used, the AER approach, albeit with more testing than in the past. Dharmapala and McAleer (1996) discuss econometric methodology in the context of the philosophy of science.

The nomenclature AER (average economic regression) is taken from Gilbert (1986), which contains an extraordinarily clear exposition of the TTT (test, test, test) approach. Pagan (1987) has an excellent presentation of TTT and of fragility analysis, along with critiques of both. Pagan also identifies a fourth approach, the VAR methodology (discussed in chapter 10); it has not been included here because it cannot be interpreted as a general specification methodology (it applies only to time series data) and because it in effect makes no effort to seek or evaluate a traditional specification.

The AER approach is defended by Darnell and Evans (1990), who refer to it as the "traditional" approach. They argue that if the traditional approach were modified to focus on finding specifications that exhibit non-spherical errors before undertaking tests, then it would be more palatable than TTT and fragility analysis, both of which they criticize.

Johnston (1984, pp. 498510) has a good description of how the AER approach ought to be implemented. He stresses the need for the researcher to talk with experts in the area being modeled, become familiar with the relevant institutions, actually look at the data, recognize the data limitations, avoid data mining, use economic theory, and, of utmost importance, exploit the judgement of an experienced critic. He gives an amus-

ing account of his experience on an energy research project; his specification search did not end until his experienced critic, Alex Cairncross, stated that he "wouldn't mind getting on a plane and taking this to Riyadh."

The TTT approach is identified with David Hendry, the most prominent of its advocates. Hendry (1993) is a selection of papers tracing the evolution of this econometric methodology, of which Hansen (1996) is an interesting review and critique. Particularly useful are Hendry's Introduction (pp. 17), the introductions to each section, the preambles associated with each article, and the chapter 19 summary which also describes the PC-GIVE software designed for this type of specification work in the time series context. A well-known application is Davidson et al. (1978); a more recent application is Hendry and Ericsson (1991), with a critique by Friedman and Schwartz (1991). The nomenclature TTT was chosen with reference to an oft-cited quote from Hendry (1980, p. 403): "The three golden rules of econometrics are test, test, and test." This methodology has been developed in the context of a specific type of time series modeling, called autoregressive distributed lag models (discussed in chapter 17 under the heading "error-correction models"), but the general principles apply to other contexts.

What does it mean to say, following TTT, that the model is "congruent" with the evidence? There are five main criteria.

(1) *Data-admissible* The model must not be capable of producing predictions that are not logically possible. For example, if the data to be explained are proportions, then the model should force all outcomes into the zero to one range.

(2) *Theory-consistent* The model must be consistent with the economic theory from which it is derived; it must make good economic sense. For example, if economic theory suggests that a certain long-run equilibrium should characterize a relationship, then the dynamic formulation of that relationship should be such that its equilibrium solution yields that long-run equilibrium.

(3) *Predictive validity* The model should be capable of adequately predicting observations not used in its estimation/specification. This is sometimes referred to as parameter constancy. This test is particularly important because it addresses the concern that exploring the data to develop a specification implies that those data

cannot be used to test that specification.

(4) *Data coherency* The residuals from the model should be white noise (i.e., random), since otherwise some regularity has not yet been included in the specification. Many econometricians consider this requirement too strong because it rules out genuine autocorrelation or heteroskedasticity. A more realistic interpretation of this requirement is that if the errors are not white noise the researcher's first reaction should be to check the specification very carefully, not to adopt GLS.

(5) *Encompassing* The model should be able to encompass its rivals in the sense that it can explain other models' results, implying that these other models contain no information capable of improving the current model.

The "fragility analysis" approach to specification is identified with Ed Leamer, its foremost proponent; the standard references are Leamer (1983a) and Leamer and Leonard (1983). An instructive critique is McAleer et al. (1985). For applications see Cooley and LeRoy (1981) and Leamer (1986). Fragility analysis can be undertaken using software called SEARCH, developed by Leamer. Caudill (1988) suggests that fragility analysis be reported by presenting a histogram reflecting the confidence inter-

vals produced by running the range of regressions associated with that analysis. Leamer's view of the AER and TTT methodologies is reflected in the comments of Leamer and Leonard (1983, p. 306):

> Empirical results reported in economic journals are selected from a large set of estimated models. Journals, through their editorial policies, engage in some selection, which in turn stimulates extensive model searching and prescreening by prospective authors. Since this process is well known to professional readers, the reported results are widely regarded to overstate the precision of the estimates, and probably to distort them as well. As a consequence, statistical analyses are either greatly discounted or completely ignored.

Leamer (1978, p. vi) is refreshingly frank in describing the wide gap between econometric theory and econometric practice:

> We comfortably divide ourselves into a celibate priesthood of statistical theorists, on the one hand, and a legion of inveterate sinner-data analysts, on the other. The priests are empowered to draw up lists of sins and are revered for the special talents they display. Sinners are not expected to avoid sins; they need only confess their errors openly.

His description (1978, p. vi) of how he was first led to this view, as a graduate student, is widely quoted:

> As it happens, the econometric modeling was done in the basement of the building and the econometric theory courses were taught on the top floor (the third). I was perplexed by the fact that the same language was used in both places. Even more amazing was the transmogrification of particular individuals who wantonly sinned in the basement and metamorphosed into the highest of high priests as they ascended to the third floor.

Leamer (1978, pp. 513) contains an instructive taxonomy of specification searches, summarized in Darnell and Evans (1990).

Using techniques that adopt specifications on the basis of searches for high R2 or high *t* values, as practitioners of the AER approach are often accused of doing, is called data mining, fishing, grubbing or number-crunching. This methodology is described eloquently by Case: "if you torture the data long enough, Nature will confess." Karni and Shapiro (1980) is an amusing account of data torturing. In reference to this unjustified (but unfortunately typical) means of specifying relationships, Leamer (1983a) is moved to comment: "There are two things you are better off not watching in the making: sausages and econometric estimates."

Both searching for high R2 and searching for high *t* values are known to be poor mechanisms for model choice; convincing arguments can be found in T. Mayer (1975, 1980), Peach and Webb (1983) and Lovell (1983). Mayer focuses on adjusted R2, showing that it does a poor job of picking out the correct specification, mainly because it capitalizes on chance, choosing a specification because it is able to explain better the peculiarities of that particular data set. This underlines the importance of setting aside some data to use for extra-sample prediction testing after a tentative specification has been chosen and estimated (as urged

by TTT). Peach and Webb fabricated 50 macroeconomic models at random and discovered that the majority of these models exhibited very high R2 and *t* statistics. Lovell focuses on the search for significant *t* values, branding it data mining, and concludes that such searches will lead to

inappropriate specifications, mainly owing to a high probability of type I errors because of the many tests performed. Denton (1985) suggests that this phenomenon is not confined to individual researchers - that many independent researchers, working with the same data, will collectively be performing these many tests, ensuring that journals will tend to be full of type I errors. All this is summed up nicely by Lovell (1983, p. 10): "It is ironic that the data mining procedure that is most likely to produce regression results that appear impressive in terms of the customary criteria is also likely to be the most misleading in terms of what it asserts about the underlying process generating the data under study."

It must be noted that the data mining methodology has one positive feature: sometimes such experimentation uncovers empirical regularities that point to errors in theoretical specifications. For example, through data mining one of my colleagues stumbled across a result that led him to re-examine the details of the British Columbia stumpage fee system. He discovered that he had overlooked some features of this tax that had an important bearing on the behavior of the forest industry. Because of this, he was able to develop a much more satisfactory theoretical specification, and thereby to produce better empirical results. I give John Maynard Keynes (1940, p. 155) the last word on the subject:

> It will be remembered that the seventy translators of the Septuagint were shut up in seventy separate rooms with the Hebrew text and brought out with them, when they emerged, seventy identical translations. Would the same miracle be vouchsafed if seventy multiple correlators were shut up with the same statistical material?

One important dimension of TTT is that the data should be allowed to help to determine the specification, especially for model features such as lag lengths, about which economic theory offers little guidance. The

earlier comments on data mining suggest, however, that letting the data speak for themselves can be dangerous. It may be necessary to have certain features in a model for logical consistency, even if a particular sample of data fails to reflect them, to avoid the common experience of an apparently well-fitting model performing poorly out-of-sample. Belsley (1986a) argues for the use of prior information in specification analysis; discussants of the Belsley paper wonder whether adoption of an incorrect model based on poor prior information is more dangerous than letting the data speak for themselves. Belsley (1988a) has a good general discussion of this issue in the context of forecasting. A balance must be found between letting the data help with the specification and not letting the data dominate the specification, which unfortunately returns use to the "specification is an art that cannot be taught" phenomenon.

5.3 General Principles for Specification

That economic theory should be at the heart of a specification search is too often forgotten by practitioners. Belsley and Welsch (1988) provide a cogent example of the use of such *a priori* information and note (p. 447): "Don't try to model without understanding the nonstatistical aspects of the real-life system you are trying to subject to statistical analysis. Statistical analysis done in ignorance of the subject matter is just that - ignorant statistical analysis."

Pagan (1987) calls for a greater integration of competing methodologies, in much the same spirit as that in which the general principles for guiding model specification were presented earlier. Since these principles may not be endorsed by all econometricians, some references may be warranted. On the issue of requiring white noise residuals see Darnell and Evans (1990, chapter 4), who defend the traditional (AER) approach providing it adopts this view. On "testing down" see Harvey (1990, pp. 1857). On "overtesting" see Bera and Jarque (1982). On diagnostics and extra sample prediction see Harvey (1990, pp. 1879). On encompassing see Mizon (1984). On the reporting of bounds and specification paths see Pagan (1987).

Unusual observations can often be of particular value in specification, as they prompt researchers to develop their theoretical models more carefully to explain those observations. For discussion and examples see

Zellner (1981). It should be noted that some robust estimation procedures (discussed in chapter 18) have a tendency to throw such "outliers" away, something that should not be done until they have been carefully examined.

Koenker (1988) suggests that specification is affected by sample size, noting that as the sample size increases the number of explanatory variables in published studies tends to increase at a rate proportional to the sample size raised to the power one-quarter. Larger samples tempt researchers to ask new questions and refine old ones; implicitly, they are less and less willing to accept bias in the face of the extra precision brought by the larger sample size. Koenker notes (p. 139) and interesting implication for asymptotic theory, claiming that it rests on the following "willing suspension of disbelief": "Daily an extremely diligent research assistant arrives with buckets of (independent) new observations, but our imaginary colleague is so uninspired by curiosity and convinced of the validity of his original model, that each day he simply re-estimates his initial model - without alteration - employing ever-larger samples."

Hogg (1988) suggests a useful rule of thumb for specification: compare the estimates from OLS and a robust method; if they disagree, take another hard look at both the data and the model. Note that this could be viewed as a (casual) variant of the Hausman specification testing method.

5.4 Misspecification Tests/Diagnostics

Kramer and Sonnberger (1986) have a good exposition of many misspecification tests, along with examples of their application. Pagan (1984a) notes that most tests can be written in the form of an OV test, which he refers to as a variable addition test. McAleer (1994) tabulates (pp. 330, 331) possible causes of diagnostic failures. Beggs (1988) and McGuirk, Driscoll and Alway (1993) have good discussions for practitioners, with examples. MacKinnon (1992) is a very informative survey of the use of artificial regressions for calculating a wide variety of specification tests.

Extensive use of diagnostic tests/checks is not universally applauded. Goldberger (1986) claims a recent empirical study reported more diagnostic test statistics than the number of observations in the data set; Oxley (1996, p. 229) opens that "we probably have more papers creating new test statistics than papers using them." Several complaints and warnings have been issued:

(1) their use may decrease the intensity with which researchers investigate their data and theoretical specifications;

(2) it may be replacing one kind of data mining with another;

(3) many tests are only valid in large samples, something often forgotten;

(4) inexperienced researchers frequently apply tests in contexts in which they are inappropriate;

(5) many tests are only valid if the model is "correctly specified";

(6) sequences of tests distort things like the probability of a type I error;

(7) most of the tests used are not independent of one another;

(8) the properties of pre-test estimators are not well understood.

These points suggest that some care should be taken in applying diagnostic tests, and that results should be viewed with a healthy degree of scepticism.

That most researchers do not bother to subject their models to misspecification tests is illustrated convincingly by Kramer et al. (1985), who apply a battery of such tests to several empirical studies and find that these tests are failed with high frequency.

Doran (1993) is a good exposition of non-nested testing. McAleer (1987) and MacKinnon (1983) are good surveys of the non-nested test literature; the commentary on the MacKinnon paper provides an interesting view of controversies in this area. The feature of non-nested tests that all models under consideration may be rejected (or accepted) is discussed by Dastoor (1981). Kennedy (1989) uses the Ballentine to exposit some of the non-nested tests and their common features.

The non-nested F test is regarded as one of the best non-nested testing procedures, because of its computational ease and its relatively good performance in Monte Carlo studies. Suppose there are two theories, $H0$

and *H*1. According to H0, the independent variables are *X* and *Z;* according to *H*1, they are *X* and *W*. A general model with *X, Z,* and *W* as explanatory variables is formed (without any economic rationale!), called an artificial nesting model. To test *H*0 the coefficients of *W* are tested against zero, using an *F* test, and to test *H*1 the coefficients of *Z* are tested against zero, using an *F* test. Note that if neither *H*0 nor *H*1 is correct it is possible for both hypotheses to be rejected, and if one of *H*0 and *H*1 is correct, but *X* and *Z* happen to be highly collinear, it is possible for both to be accepted. It is often the case that degrees-of-freedom problems (the artificial nesting model could contain a lot of variables), collinearity problems or nonlinear functional forms make this test unattractive. There most popular alternatives are the *J* test and its variants.

As is made clearer in the technical notes to this section, the *J* test is akin to the *F* test in that it stems from an artificial nesting model. To conduct this test, the dependent variable *y* is regressed on the explanatory variables of hypothesis *H*0, together with $\hat{y}$ the estimated *y* from the regression associated with *H*1. If $\hat{y}$ has some explanatory power beyond that contributed by the explanatory variables of *H*0, then *H*0 cannot be the "true" model. This question is addressed by using a *t* test to test if the coefficient of $\hat{y}$ is significantly different from zero: if it is, *H*0 is rejected; otherwise, *H*0 is accepted. The roles of *H*0 and *H*1 are reversed and the procedure is repeated to allow *H*1 to be either accepted or rejected.

Mizon and Richard (1986) exposit the encompassing principle and use it to unify several testing procedures. They show that the different non-nested tests all have different implicit null hypotheses. For example, the *J* test is a "variance" encompassing test - they test if a hypothesis can predict the estimated variance obtained by running the

regression suggested by the other hypothesis. In contrast, the non-nested *F* test is a "mean" encompassing test - it tests if a hypothesis can predict the coefficient estimate obtained by running the regression suggested by the other hypothesis. This explains the different degrees of

freedom of the *J* and non-nested *F* tests. A third type of encompassing test is a "forecast" encompassing test. Model 1 forecast encompasses model 2 if model 2 forecasts can be explained by model 1. The one-step-ahead forecast errors from model 2 are regressed on the difference between the one-step-ahead forecasts from models 1 and 2; a *t* test on the slope coefficient from this regression is used for the forecast encompassing test.

Data transformation tests are said to be Hausman-type tests, because they are based on a principle popularized by Hausman (1978) in the context of testing for contemporaneous correlation between the regressor(s) and the error (discussed further in chapters 9 and 10). This principle is as follows: if the model specification is correct, estimates by any two consistent methods should be close to one another; if they are not close to one another, doubt is cast on the model.

Several variants of the data transformation test exist, the more popular of which are Farebrother (1979), where the transformation groups the data; Davidson et al. (1985), where the transformation is first differencing; and Boothe and MacKinnon (1986), where the transformation is that usually employed for doing GLS. Breusch and Godfrey (1985) have a good discussion, as do Kramer and Sonnberger (1986, pp. 1115). See the technical notes for discussion of how such tests can be operationalized as OV tests.

For examples of situations in which conditional moment tests are easier to construct than alternatives, see Pagan and Vella (1989). Newey (1985) and Tauchen (1985) have developed a computationally attractive way of calculating CM tests by running an artificial regression. (Regress a column of ones on the moments and the first derivatives of the log-likelihood with respect to each parameter, and test the slopes of the moments against zero.) Unfortunately, this method relies on OPG (outer product of the gradient - see appendix B) estimates of variance-covariance matrices which cause the type I error (size) of tests to be far too large. For discussion of CM tests see Godfrey (1988, pp. 37ene (1997, pp. 534Davidson and MacKinnon (1993, pp. 5718). Although most tests are such that their asymptotic distributions are not sensitive to the assumption of normal errors, in small samples this may be of some concern. Rank tests are robust in this respect; McCabe (1989) suggests several rank tests for use as misspecification tests and claims that they have good power.

5.5 R2 Again

R2, the adjusted R2, is derived from an interpretation of R2 as I minus the ratio of the variance of the disturbance term to the variance of the dependent variable (i.e., it is concerned with variances rather than variation). Estimation of these variances involves corrections for degrees of freedom, yielding (after manipulation) the expression

$$\bar{R}^2 = R^2 - \frac{K-1}{T-K}(1-R^2) \text{ or } 1 - \frac{T-1}{T-K}(1-R^2)$$

where K is the number of independent variables and T is the number of observations. Armstrong (1978, p. 324) discusses some alternative adjustments to R2. It is interesting to note that, if the true R2 is zero (i.e., if there is no relationship between the dependent and independent variables), then the expected value of the unadjusted R2 is K/T, a value that could be quite large. See Montgomery and Morrison (1973) for the general formula when the true R2 is not zero.

Both R2 and R2 are biased but consistent estimators of the "true" or "population" coefficient of determination. R2 has a smaller bias than R2, though. An unbiased estimator of the population coefficient of determination has not been developed because the distributions of R2 and R2 are intractable when this population coefficient is nonzero.

The result that the "correct" set of independent variables produces a higher R2 on average in repeated samples was derived by Theil (1957).

If adding an independent variable increases R2, its t value is greater than unity. See Edwards (1969). Thus the rule of maximizing R2 is quite different from the rule of keeping variables only if their t values are significant at the 5% level.

It is worth reiterating that searching for a high R2 or a high $R2$ runs the real danger of finding, through perseverance, an equation that fits the data well but is incorrect because it captures accidental features of the particular data set at hand (called "capitalizing on chance") rather than the true underlying relationship. This is illustrated in convincing fashion by Mayer (1975) and Bacon (1977).

Aigner (1971, pp. 1017) presents a good critical summary of measures used to capture the relative importance of independent variables in determining the dependent variable. He stresses the point that the relative strength of individual regressors should be discussed in a policy context, so that, for example, the impact on the dependent variable per dollar of policy action is what is relevant.

Anderson-Sprecher (1994) offers an interpretation of the R2 measure that clarifies many of the problems with its use.

Technical Notes

5.2 Three Methodologies

TTT was developed in the context of autoregressive distributed lag models, where the initial "more general" specification takes the form of a very generous lag length on all explanatory variables, as well as on the lagged dependent variable. This is done to reflect the fact that economic theory typically has very little to say about the nature of the dynamics of a relationship. Common sense is used to choose the initial lag lengths. For example, if quarterly data are being used, five lags might be initially specified, allowing for fourth-differences and first-differences of the fourth-differenced data. One of the problems this creates is a lack of degrees of freedom. There is a tendency to solve this problem by "cheating" a little on the general-to-specific methodology - by not including at first all explanatory variables under consideration (adding them in later after the initial over-parameterized model has been simplified).

The main input to a fragility analysis is a Bayesian prior distribution, with its vari-

ance-covariance matrix indexed by a scale parameter. By varying this scale parameter to reflect different degrees of confidence in this prior held by different researchers, a range of parameter estimates is produced, the output of a fragility analysis. An alternative approach, suggested by Granger and Uhlig (1990), is to modify the extreme-bounds analysis by considering only specifications that produce R2 values within 10 or 15% of the highest R2.

In "testing down," the size of the overall test (the overall probability of a type I error), a, can be determined/controlled from the result that (1 - a) is equal to the product over *i* of (1 - a*i*), where a*i* is the size of the *i*th individual test. For example, suppose we are conducting n tests during a testing down process and we want the overall type I error to be 5%. What common type I error a* of the individual n tests will accomplish this? This is calculated from 0.95 = (1 - a*)n, yielding a* = 1- 0.951/n which becomes smaller and smaller as n grows.

A sixth criterion is often found in the list of criteria used to determine data congruency, namely that the explanatory variables should be at least weakly exogenous (i.e., it is valid to condition on these regressors), since otherwise it will be necessary to model the regress and and the regressor jointly. This criterion is out of place in general application of the TTT methodology. What is meant is that exogeneity should be tested for, not that a model must be such that all its explanatory variables are exogenous, however convenient that may be. If an explanatory variable is found not to be exogenous, an alternative specification may be required, but not necessarily one in which that variable must be exogenous.

There are three types of exogeneity. Suppose *y* is thought to be explained by *x*. The variable *x* is said to be weakly exogenous if current *y* does not also explain *x*. This implies that estimation and testing can be undertaken by conditioning on *x*. It is strongly exogenous if also the lagged value of *y* does not explain *x* (i.e., there is no "feedback" from *y* to *x*); strong exogeneity has implications mainly for using *x* to forecast *y*. The variable *x* is "super exogenous" if the *x* coefficients in the relationship determining *y* are not affected by changes in the *x* values or by the process generating the *x* values. This has relevance for policy; it reflects the "Lucas critique" (Lucas, 1976), which claims that a policy change will cause rational economic agents to change their behavior, and questions what meaning one can attach to the assumed-constant parameters estimated by econometrics. Maddala (1988, pp. 32531) has a good textbook exposition of exogeneity.

5.4 Misspecification Tests/Diagnostics

The rationale behind the *J* test is easily seen by structuring the artificial nesting model on which it rests. Suppose there are two competing linear hypotheses:

H_0: $y = X\beta + \varepsilon_0$, and
H_1: $y = Z\delta + \varepsilon_1$.

The artificial nesting model

$$y = (1 - \lambda)X\beta + \lambda Z\delta + \varepsilon_2$$

is formed, combining *H*0 and *H*1 with weights (1 - l) and l, respectively. Under the null hypothesis that *H*0 is the correct specification, l is zero, so a specification test of *H*0 can be formed by testing l = 0. Regressing *y* on *X Z* will permit estimation of (1 - l)b and ld, but not l. Even this cannot be done if *X* and *Z* have a common variable. This dilemma is resolved by the following two-step procedure:

(1) regress *y* on *Z,* obtain dOLS and calculate $\hat{y}1 = Z\text{eOLS}$, the estimated *y* from this regression;

(2) regress *y* on *X* and $\hat{y}1$ one and test the (single) slope coefficient estimate $\hat{\lambda}$ of $\hat{y}1$ one against zero by a *t* test.

This permits *H*0 to be either accepted or rejected. The roles of *H*0 and *H*1 are then reversed and the procedure is repeated to allow *H*1 to be either accepted or rejected. (Why not just test l = 1 from the regression in (2)? The logic of the test described above is based on *H*0 being the null; when *H*1 is the null $\hat{\lambda}$ - 1 divided by its standard error turns out not to be distributed as a *t.)*

In small samples the type I error of the *J* test tends to be too large; Fan and Li (1995) find that bootstrapping eliminates this problem. Bera et al. (1992) show how non-nested testing can be undertaken simultaneously with testing for other features of the specification. In nonlinear contexts Xb and/or Ze above would be replaced with the relevant nonlinear function. If this creates computational difficulties the P test is employed. For an exposition see Davidson and MacKinnon (1993, pp. 3823); Smith and Smyth (1990) is a good example.

Suppose we have specified $y = X\beta + e$ and have suggested the transformation matrix P for the purpose of constructing a data transformation test. Transforming the data produces $Py = PX\beta + Pe$ to which OLS is applied to obtain $\beta^* = (X'P'PX)^{-1}X'P'Py$. This must be compared to $\beta OLS = (X'X)^{-1}X'y$. Now write y as $X\beta OLS + eOLS$ where eOLS is the OLS residual vector, and substitute this in the expression for β^* to get $\beta^* = \beta OLS + (X'P'PX)^{-1}X'P'PeOLS$. For this to be insignificantly different from zero, $P'PX$ must be uncorrelated (or nearly so) with eOLS. It turns out that this can be tested by using a familiar F test to test if the coefficient vector on $P'PX$ is zero when y is regressed on X and $P'PX$. (For an intuitive explanation of this, see the technical notes to section 9.2, where the Hausman test for measurement errors is explained.) Thus a data transformation test can be performed as an OV test, where the omitted variables are defined by $P'PX$. Any redundancies (a column of $P'PX$ equal to a column of X, for example) created in this way are handled by omitting the offending column of $P'PX$ and changing the degrees of freedom of the F test accordingly.

An unusual variant of a Hausman-type misspecification test is White's information matrix test, in which two different estimates of the information matrix (the inverse of the variance-covariance matrix) are compared. If the model is correctly specified, these estimates are asymptotically equivalent. One estimate is based on the matrix of second derivatives of the log-likelihood (the Hessian form), while the other is obtained by adding up the outer products of the vector of first derivatives of the log-likelihood (the OPG, or outer product of the gradient form). Hall (1989) provides a computationally feasible way of calculating this test statistic.

6
Violating Assumption One: Wrong Regressors, Nonlinearities, and Parameter Inconstancy

6.1 Introduction

The first assumption of the CLR model states that the conditional expectation of the dependent variable is an unchanging linear function of known independent variables. It is usually referred to as the "model specification." Chapter 5 discussed in general terms the question of how to go about finding a model specification that is in accord, or "congruent," with the data. The purpose of this chapter is to be more specific on this issue, examining the three major ways in which this first assumption can be violated. First is the case in which the specified set of independent variables omits relevant variables or includes irrelevant variables. Second is the case of a nonlinear functional form. And third is the case in which the parameters do not remain constant.

6.2 Incorrect Set of Independent Variables

The consequences of using an incorrect set of independent variables fall into two categories. Intuitive explanations for these results are given in the general notes to this section.

(1) *Omission of a relevant independent variable*

(a) In general, the OLS estimator of the coefficients of the remaining variables is biased. If by luck (or experimental design, should the researcher be fortunate enough to have control over the data) the observations on the omitted variables(s) are uncorrelated in the sample with the observations on the other independent variables (i.e., if the

omitted variable is orthogonal to the included variables), the slope coefficient estimator will be unbiased; the intercept estimator will retain its bias unless the mean of the observations on the omitted variable is zero.

(b) The variance-covariance matrix of bOLS becomes smaller (unless the omitted variable is orthogonal to the included variables, in which case it remains unchanged). This result, in conjunction with the bias noted in (a) above, implies that omitting a relevant variable can either raise or lower an estimator's MSE, depending on the relative magnitudes of the variance reduction and the bias.

(c) The estimator of the (now smaller) variance-covariance matrix of bOLS is biased upward, because the estimator of s2, the variance of the error term, is biased upward. This causes inferences concerning these parameters to be inaccurate. This is the case even if the omitted variable is orthogonal to the others.

(2) *Inclusion of an irrelevant variable*

(a) bOLS and the estimator of its variance-covariance matrix remain unbiased.

(b) Unless the irrelevant variable is orthogonal to the other independent variables, the variance-covariance matrix bOLS becomes larger; the OLS estimator is not as efficient. Thus is this case the MSE of the estimator is unequivocally raised.

At first glance a strategy of "throwing in everything but the kitchen sink as regressors" seems to be a good way of avoiding bias. This creates what is sometimes referred to as the "kitchen sink" dilemma - omitted variables, and the bias they cause, will be avoided, but the irrelevant variables that will inevitably be present will cause high variances.

There is no easy way out of this dilemma. The first and foremost ingredient in a search for the correct set of explanatory variables is economic theory. If economic theory cannot defend the use of a variable as an explanatory variable, it should not be included in the set of potential independent variables. Such theorizing should take place *before* any empirical testing of the appropriateness of potential independent variables; this guards against the adoption of an independent variable just because it happens to "explain" a significant portion of the variation in the dependent variable in the particular sample at hand. Unfortunately, there is a limit to the information that economic theory can provide in this respect. For example, economic theory can suggest that lagged values of an explanatory variable should be included, but will seldom suggest how many such variables should be included. Because of this, economic theory must be supplemented by some additional mechanism for determining the correct set of explanatory variables.

According to the TTT methodology discussed in chapter 5, this should be done by including more variables than thought necessary and then "testing

down" to obtain a final specification. If this approach is followed, the question arises as to what critical value of the relevant *t* or *F* statistic should be employed to operationalize the testing procedure. An obvious possibility is the traditional 5% value, perhaps adjusted downwards if several tests are to be performed. An alternative, as mentioned in the general notes to section 4.5, is to use a critical value of unity, implying maximization of adjusted R2. Several other suggestions for critical values correspond to maximization of alternative adjusted forms of R2, with slightly different trade-offs between goodness of fit and parsimony (number of explanatory variables). These are usually formalized in terms of finding the set of explanatory variables that minimizes a specific function of the sum of squared errors and the number of explanatory variables. The more popular of these model selection criteria are the Akaike information criterion (AIC), Amemiya's prediction criterion (PC), and the Schwarz criterion (SC), which are discussed in the general notes.

Unfortunately there are no unequivocal means of testing for whether an unknown explanatory variable has been omitted, mainly because other misspecifications, such as incorrect functional form, affect available tests. Many of the misspecification tests discussed in chapter 5 are used to check for an omitted explanatory variable. Particularly popular in this regard are tests for serial correlation in the errors (discussed in chapter 8), since any cyclical movement in an omitted variable will be transmitted to the OLS residuals.

Also popular is the RESET test. When a relevant variable is omitted from a model, the "disturbance" term of the false model incorporates the influence of the omitted variable. If some variable or set of variables *Z* can be used as a proxy for the (unknown) omitted variable(s), a specification error test can be formed by examining *Z*'s relationship to the false model's error term. The RESET (regression specification error test) does this by adding *Z* to the set of regressors and then testing *Z*'s set of coefficient estimates against the zero vector by means of a traditional *F* test. There are two popular choices of *Z:* the squares, cubes and fourth powers of the predicted dependent variable, and the squares, cubes and fourth powers of the explanatory variables.

6.3 Nonlinearity

The first assumption of the CLR model specifies that the functional form of the relationship to be estimated is linear. Running an OLS regression when this is not true is clearly unsatisfactory, since parameter estimates not only are biased but also are without meaning except in so far as the linear functional form can be interpreted as an approximation to a nonlinear functional form. Functional forms popular in applied econometric work are summarized in the technical notes to this section.

The OLS procedure must be revised to handle a nonlinear functional form. These revisions fall into two categories.

(1) Transformations

If by transforming one or more variables a nonlinear function can be translated into a linear function in the transformed variables, OLS estimation procedures can be applied to transformed data. These transformations are of two types.

(a) *Transforming only independent variables* If, for example, the nonlinear functional form is

$$y = a + bx + cx^2 + \varepsilon$$

a linear function

$$y = a + bx + cz + \varepsilon$$

can be created by structuring a new independent variable z whose observations are the squares of the observations on x. This is an example of an equation nonlinear in variables but linear in parameters. The dependent variable y can be regressed on the independent variables x and z using bOLS to estimate the parameters. The OLS estimator has its CLR model properties, the R2 statistic retains its traditional properties, and the standard hypothesis tests are valid.

(b) *Transforming the entire equation* When transforming only independent variables cannot create a linear functional form, it is sometimes possible to create a linear function in transformed variables by transforming the entire equation. If, for example, the nonlinear

function is the Cobb-Douglas production function (with a multiplicative disturbance)

$$Y = AK^{\alpha}L^{\gamma}\varepsilon$$

then transforming the entire equation by taking natural logarithms of both sides creates

$$\ln Y = \ln A + \alpha \ln K + \gamma \ln L + \ln \varepsilon$$

or

$$Y^* = A^* + \alpha K^* + \gamma L^* + \varepsilon^*,$$

a linear function in the transformed variables Y^*, K^* and L^*. If this new relationship meets the CLR model assumptions, which econometricians usually assume is the case, the OLS estimates from a regression using these transformed variables have their traditional desirable properties.

(2) Computer-Assisted Numerical Techniques

Some nonlinear functions cannot be transformed into a linear form. The CES production function is an example of this, as is the Cobb-Douglas function with an additive, rather than a multiplicative, disturbance. In these cases econometricians turn to either nonlinear least squares or maximum likelihood methods, both of which require computer search procedures. In nonlinear least squares the computer uses an iterative technique to find those values of the parameters in the relationship that cause the sum of squared residuals to be minimized. It starts with approximate guesses of the parameter values and computes the residuals and then the sum of squared residuals; next, it changes one of the parameter values slightly, recomputes the residuals and sees if the sum of squared residuals becomes larger or smaller. It keeps changing parameter values in directions that lead to smaller sums of squared residuals until it finds the set of parameter values that, when changed slightly in any direction, causes the sum of squared residuals to rise. These parameter values are the least squares estimates in the nonlinear context. A good initial guess of the parameter values is necessary to ensure that the procedure reaches a global and not a local minimum for

the sum of squared residuals. For maximum likelihood estimation a similar computer search technique is used to find parameter values that maximize the likelihood function. See the technical notes for a discussion of the way in which computer searches are structured, some of which have led to the development of new estimators.

In general, the desirable properties of the OLS estimator in the CLR model do not carry over to the nonlinear least squares estimator. For this reason the maximum likelihood estimator is usually chosen in preference to the nonlinear least squares estimator. The two techniques are identical whenever the dependent variable is determined by a nonlinear function of the independent variables plus a normally distributed, additive disturbance.

There are five main methods of testing for nonlinearity.

(1) *RESET* Although the Regression Equation Specification Error Test was designed to be used to test for missing regressors, it turns out to be powerful for detecting nonlinearities. This weakens its overall attractiveness, since rejection of a model could be due to either a nonlinearity or an omitted explanatory variable. (No test can discriminate between unknown omitted variables and unknown functional form; a strong case can be made that the RESET test can only test for functional form.)

(2) *Recursive residuals* The *n*th recursive residual is the error in predicting the *n*th observation using parameters estimated from a linear regression employing the first *n* - 1 observations. If the true functional form is non-linear, then, if the data are ordered according to the variable entering non-linearly, these residuals could become either all positive or all negative, a result that can be exploited to test for nonlinearity.

(3) *General functional forms* Some functional forms contain particular forms, such as linearity or log-linearity, as special cases corresponding to specific values of a parameter. These particular functional forms can then be tested by testing the estimate of this parameter against these specific values.

(4) *Non-nested tests* Variants of the non-nested testing methodology discussed in chapter 5 can be used to test functional form.

(5) *Structural change tests* Because a nonlinear function can be approximated by two or more linear segments, the structural change tests discussed in the next section can be interpreted as tests for nonlinearity.

6.4 Changing Parameter Values

A common criticism of econometricians concerns their assumption that the parameters are constants. In time series estimation, changing institutions and social mores have surely caused the parameter values characterizing the economy to change over time, and in cross-section estimation it is surely unrealistic to assume that the parameters for every individual or every region are exactly the same. Although most econometricians usually ignore these criticisms, maintaining that with small sample sizes they are forced to make these simplifying assumptions to obtain estimates of any sort, several techniques are available for addressing this problem.

(1) Switching Regimes

It may be known that at a particular point in time the economic structure changed. For example, the date of the Canada-USA auto pact might mark a change in parameter values associated with the Canadian or US auto industries. In such a case we need run only two regressions, one for each "regime." More often than not, however, the point in time at which the parameter values changed is unknown and must be estimated. If the error variances are the same for both regimes, this can be done by selecting several likely points of change, running pairs of regressions for each and then choosing among these points of change by determining which corresponds to the smallest total sum of squared residuals. (If the error variances cannot be assumed equal, a maximum likelihood technique must be used.) This approach has been extended in several directions:

(a) to accommodate more than two regimes;

(b) to permit continuous switching back and forth, either randomly or according to a critical value of an unknown function of some additional variables;

(c) to eliminate discontinuities, so that the function describing one regime blends into the function describing the next regime over an adjustment period.