Journal of Economic Behavior and Organization 14 (1990) 97–125. North-Holland

# ALTERNATIVE APPROACHES TO MODEL CHOICE*

Alice Orcutt NAKAMURA and Masao NAKAMURA
*University of Alberta, Edmonton, Alta., Canada T6G 2R6*

Harriet Orcutt DULEEP
*U.S. Commission on Civil Rights, Washington, DC 20425, USA*

(Received October 1988, final version received March 1990)

Accepted model choice procedures are characterized. Problems concerning the determination of significance levels for multiple tests, Type II errors for specification error tests, and the use of point null hypotheses are reviewed. The potential gains from greater reliance on sensitivity analysis, incorporation of uncertain prior information, and experimentation are discussed. Attention is turned then to models which are explicitly approximations, and to the need, in this context, for choice procedures that provide a basis for ranking models in terms of the relative goodness of the approximations they provide.

## 1. Introduction

Twenty years ago, Guy Orcutt (1968, p. 94) wrote:

> Economic research has become quantitative in nature; it does avail itself of the best that statisticians and econometricians have had to offer, it does make extensive use of mathematics and computers, and it has begun to seriously use sample survey data. In other words economic research has adopted most of the trappings of modern science and technology, but it still fails to achieve adequate testing of basic economic hypotheses. Theories come and go but not on the basis of sound and convincing evidence.

In the intervening years there have been spectacular improvements in the

*The authors are particularly grateful to Noxy Dastoor and Quang Vuong for detailed comments that helped us to greatly improve the paper. Thanks are also due to Marcel Dagenais, Erwin Diewert, Arthur Goldberger, Zvi Griliches, and Arnold Zellner; and to participants in seminars at the Universities of Illinois, Ohio State and Wisconsin for comments on earlier versions of this paper. Earlier conversations and correspondence about problems of model choice and specification error testing with Takeshi Amemiya, Don Andrews, Jerry Hausman, Jim Heckman, Jan Kmenta, Hiroki Tsurumi and Kenneth Wallis stimulated and helped us to develop the ideas discussed in this paper. Guy Orcutt's influence is evident throughout. Any errors are, of course, our own.

0167–2681/90/$03.50 © 1990—Elsevier Science Publishers B.V. (North-Holland)

computational, econometric, data resource and other tools of applied economic research. Yet the problem of how to choose among competing economic hypotheses, represented in terms of competing models, is still largely unresolved.

We begin by reviewing key aspects of currently accepted model choice procedures. In section 3 we consider three sorts of statistical problems associated with current model choice practices. Accepted strategies for dealing with situations where economic theory does not provide an adequate basis for model specification are reviewed in section 4. In section 5 we consider reasons for interest in explicitly approximate models. The issue of how to evaluate approximate models is the topic of section 6. Our recommendations are summarized in section 7.

## 2. Accepted model choice procedures

### *2.1. A simple textbook approach*

In the ideal textbook case, economic theory suggests a single set of explanatory variables and the functional form [Kmenta (1986, p. 516)]. Economic theory provides sign predictions for the marginal impacts of the explanatory variables on the dependent variable, and may provide general magnitude predictions as well. The initial choice of an estimation method is based on the theoretically prescribed form of the model and a guess as to the properties of the error term. Specification error tests are used to check on these assumptions.[1] Based on the outcomes of these tests, the manner in which the model is estimated may be altered or alterations may be made in the formulas used to compute the relevant standard errors for the model. For instance, if the error term is found to be heteroskedastic, alternative variance estimators may be used [see, e.g., Harvey (1976), Cragg (1983), and White (1980a)].

Having obtained an estimated model which passes a battery of specification error tests, the next step is to check the statistical significance of the estimated model as a whole, and the significance and signs of key coefficients. For example, Johnston (1984, p. 505) recommends that 'one looks for correctly signed coefficients which have reasonable statistical significance'. A finding that the model as a whole is insignificant or that theoretically crucial coefficients are insignificantly different from zero or have the 'wrong' signs may lead the econometrician to try to expand the data base or to seek a new one. It may also touch off a search for a more efficient estimation method, or for previously unnoticed estimation problems or logical flaws in the theoreti-

[1] References and surveys of available specification error tests can be found, for instance in Krämer and Sonnberger (1986), Ruud (1984), Engle (1982), Bera and Jarque (1982), and Thursby (1979).

cal specification of the model. The reformulated and reestimated model can then be subjected to specification error tests, the significance checked, and the extent of the agreement between the new estimated model and a priori theoretical expectations examined. This iterative model specification process is judged to be satisfactorily completed when all of the above steps are satisfactorily completed.[2]

Key elements of this model choice procedure include the following:

(1) Statistical methods are called upon only to provide the precise magnitudes of the model's coefficients, to aid in the determination of the error term's properties, to check on logical errors in the model's theoretical derivation or possible problems with the data collection process, and to provide the basis for the construction of confidence intervals and tests of significance for the model's coefficients.
(2) Considerable weight is placed on demonstrating that the estimated behavioral responses of interest are stronger than what might be expected by chance. Yet it is recognized that the statistical tests used for this purpose are sensitive to the assumed properties of the unobserved error term. Thus Johnston (1984, p. 505) cautions that 'it is essential to examine the properties of the disturbance term in order to assess the validity of the statistical tests being applied'.
(3) Economic theory is the final arbiter of when a satisfactory empirical model has been achieved, subject to the aforementioned specification error tests (some of which also presume that the systematic part of the model is correctly specified).
(4) There is an unstated presumption that the underlying behavioral process can be fully represented in terms of a 'true' model consisting of an equation (or possibly a system of equations) which is simple enough in its form, which has few enough observable variables, and which has a sufficiently well-behaved error term that it can be consistently estimated using available econometric methods and available or obtainable data. Data problems such as errors in variables are usually ignored [see, e.g., Griliches (1985), Carter and Fuller (1980), and Feldstein (1974)].

### *2.2. Minor complications*

Suppose now that economic theory leaves some limited doubt as to the choice of explanatory variables or the functional form relating the dependent

[2]In its more extreme forms, this model choice process is disparagingly referred to sometimes as data mining. See, for instance, Lovell (1983). Leamer (1978, p. 130) suggests that in actual practice 'few researchers are willing to accept "peculiar" estimates and the standard operating procedure is to search for constraints that yield "acceptable" estimates. The fact that the resulting estimator is neither unbiased, linear, nor "best" is no large deterrent to a person whose research project would be dubbed "fruitless" if it were summarized in a nonsensical estimate.'

variable to the explanatory variables. For instance, two competing theories may suggest two alternative specifications which both yield satisfactory empirical models according to the textbook criteria laid out above.

If the competing models can be expressed so that one is a special case of the other, then nested model selection statistical tests can be used to differentiate them. These techniques result in one or the other of the alternative models being selected as the 'true' model. If neither model can be formulated as a special case of the other, then statistical tests suitable for non-nested models must be used.[3] Non-nested tests can potentially lead to the rejection of both models. This might prompt the econometrician to search for logical flaws in the theoretical derivations of the models, or to broaden the search for problems concerning a priori assumptions. Either or both of the alternative models might then be reformulated and reestimated and the preferred non-nested tests might be applied again to see if either model could be accepted as the 'true' model.

This more involved model selection procedure subsumes all of the steps, and hence all of the properties, of the simplest case with the exception that economic theory is no longer the final arbiter of when a satisfactory empirical model has been achieved. Nevertheless, economic theory is still used to restrict the model specification as much as possible prior to estimation, and a satisfactory empirical model must still agree with theoretical sign and magnitude expectations. With important exceptions that are discussed in section 6, there is also little demonstrated interest in (and no accepted way of) evaluating 'approximate' empirical models – models whch cannot be regarded as consistently estimated representations of the 'true' underlying behavioral process.

## 3. Statistical problems

There are a number of statistical problems with accepted model choice procedures. Most of these have to do with the specification error testing or the significance testing phases of the choice process.

### *3.1. Determination of significance levels for multiple tests*

As econometric knowledge has deepened, so has the list of specification tests. Hendry (1979, p. 403) is widely quoted in econometrics texts and monographs for his commandment: 'The three golden rules of econometrics are test, test and test.'

[3]For an introduction to and references on nested model selection methods see, e.g., Kmenta (1986, pp. 593–595) and Judge et al. (1985, pp. 855–880). For an introduction to basic issues and references on tests for nonnested model selection see Kmenta (1986, pp. 595–598), Judge et al. (1985, pp. 881–884), McAleer (1987), MacKinnon (1983), and Dastoor (1983).

Kramer and Sonnberger (1986, p. 147) succinctly state a fundamental inference problem which must be dealt with if Hendry's commandment is to be appropriately implemented:

> Consider first the following rule (which is often called an 'induced test'): 'Reject the model when at least one individual test is significant.' What is the probability of a Type I error for this induced test?

This question is particularly troublesome when the individual specification error tests are not independent. Despite considerable research focused on this problem area, Kramer and Sonnberger (1986, p. 148) acknowledge that 'multiple testing is a research activity with many unsolved problems and few solutions' (p. 155). [For recent research in this area, see, e.g., Dufour (1989).] Yet they conclude 'it is better than not to test at all...' (p. 155). Others have expressed concerns about the soundness of this approach. For example, Leamer (1985, p. 308) writes:

> And what inferences are allowable after a model passes a battery of 'specification error' tests that are sometimes more numerous than even the set of observations? This recommendation ... merits the retort: 'There are two things you are better off not seeing in the making: sausages and econometric estimates," to which they might reply: 'It must be right, I've been doing it since my youth.'

A related, but conceptually distinct, problem has to do with how specification error testing relates to the ultimate goals of empirical research. In 1969 before the current preoccupation with specification error testing was in full swing, Edwards and Orcutt (1969, p. 1) wrote:

> ... the estimated standard errors of the coefficients and the Durbin–Watson statistic are familiar concomitants of published results of model estimations. To what extent do these statistics measure what we want them to in real applications? The bulk of econometric literature has ignored the relationship between statistical indicators and model performance, has assumed that data could be found that would satisfy the conditions under which an estimation technique was derived, and has limited the choice of models to those with suitable properties for deducing desirable estimators. This is all well and good if the primary interest lies with theoretical developments. However, the course of applied economic modeling would be strengthened by more empirical evidence about how well our models perform in reality.

One aspect of the problem to which Edwards and Orcutt are alluding in the last sentence of the above quote are the poorly understood relationships between the outcomes of specification error tests and the related distortions of estimation results and hypothesis tests for the parameters of final interest.

[Further discussion of this point can be found, for instance, in Nakamura, Nakamura and Orcutt (1976) and Nakamura and Nakamura (1978, 1981, 1985c).]

### 3.2. *Type II errors for specification error tests*

One of the key differences between specification error tests and ordinary significance tests is the relative importance, in terms of research objectives, of Type I and Type II errors. In the case of the test of significance for a slope parameter of a model, the researcher is typically hoping that the null hypothesis of insignificance (that is the null hypothesis that the true value of the parameter is zero) will be rejected. Control of the probability of a Type I error thus represents a precaution against an overly optimistic interpretaton of the empirical results. The range of parameter values under the alternative hypothesis that is viewed as interesting usually lies a considerable distance from the null hypothesis, where the probabilities of a Type II error are low.

On the other hand, in the case of a specification error test, a researcher often hopes the null hypothesis will be *accepted*; rejection of the null hypothesis usually implies the necessity of searching for an alternative (and less efficient) estimation method which will compensate for the identified specification problem. The probability of a Type II error thus takes on special importance as a check against an overly optimistic interpretation of the empirical results.

Unfortunately the probability of a Type II error cannot be controlled in the manner that the probability of a Type I error can be [however, Andrews (1989) has some useful related suggestions.] Also, there is often concern about values of the parameter which is the object of the specification error test (usually some parameter having to do with the distribution of the error term) that are close to the value specified under the null hypothesis. This is not necessarily because of the concern about the resulting quality of the estimates of the model's response coefficients; standard estimation methods such as ordinary least squares are known to be robust against a range of small to moderate misspecifications. However, the *tests of significance* for these response coefficients are sensitive to small departures from some of the assumed properties of the distribution of the error term.[4]

### 3.3. *The use of point null hypotheses*

Reminding us of the Lindley paradox, McCloskey (1985, p. 202) writes:

[4]See, for instance, Orcutt and James (1948), Orcutt and Cochrane (1949), Orcutt and Winokur (1969), Nakamura and Nakamura (1973), Nakamura, Nakamura and Orcutt (1976), and Nakamura and Nakamura (1978 and 1985c).

> Except in the limiting case of literally zero correlation, if the sample were large enough all the coefficients would be significant....

To clarify the substantive, as opposed to the statistical, difficulty with point null hypotheses, let us consider McCloskey's example concerning the usual test of purchasing power parity. According to this test, the hypothesis of purchasing power parity is rejected if the estimate of the slope coefficient $B$ is statistically significantly different from 1.0. McCloskey (1985, pp. 201–202) points out:

> But 'exactly' true is not relevant for most economic purposes. What is relevant is merely that $B$ is in the neighborhood of 1.0, where 'the neighborhood' is defined by *why* it is relevant...

We regard McCloskey's 'neighborhood' observation as more relevant than concerns about the inevitable significance of even a small departure from a point null hypothesis if the sample size is large enough.

A similar argument can be made with respect to the usual tests of significance for whether each of the slope coefficients of a model equals zero. Yet how is an appropriate neighborhood about zero to be chosen for an interval null hypothesis of insignificance? There are also inherent difficulties in accommodating an interval hypothesis within the framework of conventional hypothesis testing. The only explicit consideration this isue receives is couched in terms of the desirability of adjusting the size of a test as the sample size increases to reflect the increasing power of the test, and arguments (that are rarely put into practice) concerning the need to consider some sort of a loss function in setting the size of a test. However, the problem of accounting for the seriousness of the 'losses' associated with departures of various magnitudes from the null hypothesis of interest is distinct from the problem of the appropriate specification of the null hypothesis itself.

## 4. Relaxing the central role of economic theory

The model choice procedures discussed in section 2 presume the existence of a body of economic theory rich enough and of sufficiently established veracity to serve as the basis for determining most details of the specification of a model. Yet economic theory is not adequately developed to always be relied on in this manner. In some circumstances theoretical exclusions are sufficiently weak that the econometrician is faced with a large number of potential models differing in terms of included variables, functional form and the properties of the error term. We will briefly discuss three accepted ways of proceeding in such circumstances.

### 4.1. Sensitivity analysis

The first of these approaches is sensitivity analysis. In its simplest form the approach consists of estimating all conceivable alternative models and then examining the inferences provided by these different models. If it turns out that, regardless of the model, the inferences are essentially the same, then there is no need to determine which is the 'true' one [see, e.g., Leamer (1978, 1985]. There is still a problem, however, when the inferences provided by the different empirical models differ in crucial ways. In this situation Leamer (1985, p. 311) argues:

> ... a sensible and general characterization of the problem of inference begins with a broad family of alternative models and a representative, but hypothetical, prior distribution over that family. Because no prior distribution can be taken to be an exact representation of opinion, a global sensitivity analysis is carried out to determine which are sturdy. A neighborhood of prior distributions around the representative distribution is selected and inferences that depend in a significant way on the choice of prior from this neighborhood are judged to be fragile. Ideally, the neighborhood of distributions is credibly wide, and the corresponding interval of inferences is usefully narrow. But if it is discovered that an incredibly narrow neighborhood of prior distributions is required to produce a usefully narrow set of inferences, then inferences from the given data set are suspended, and pronounced too fragile to serve as a basis for action.

### 4.2. Incorporation of uncertain prior information

Theoretically implied properties are often imposed on a model prior to estimation. If these properties are not subjected to any sort of a direct empirical test, this amounts to treating these properties as prior information known with certainty. When the imposition of all available prior information viewed as certain still leaves a plethora of alternative models yielding inferences that differ in crucial ways, one possiblity is to use *uncertain* prior information as a basis for discriminating among the models. This uncertain prior information may even include 'informed guesses' about the values of some of the parameters of a model. Zellner (1979, p. 635) quotes Tukey as stating:

> It is my impression that ... it is considered decent to use judgment in choosing a functional form, but indecent to use judgment in choosing a coefficient. If judgment about important things is quite all right, why should it not be used for less important ones as well?

In the present context, we are not concerned with the issue of *which* prior

information should be regarded as uncertain. We simply wish to draw attention to the fact that Bayesian econometrics explicitly deals with the issue of incorporating uncertain prior information.[5] Moreover, in the Bayesian approach an attempt is made to formally take account of the *degree* of uncertainty associated with uncertain prior information.

### *4.3. An experimental approach*

We often do not know whether, or how, a particular explanatory variable should be incorporated into a model. Also many relevant explanatory variables in socioeconomic models are correlated. In discussing the determination of income inequality, Duleep (1968a, p. 137) argues:

> Since all the explanatory variables are normally correlated with each other, the variation which does occur among these variables (which allows the estimation of their separate effects) may occur only as the result of extraordinary personal traits or circumstances. For instance, the likelihood of observing persons with a low socioeconomic family background but high educational attainment is low, and may be due to extraordinary circumstances or traits.... Hence, analysis of variation in any given sample may not shed much light on the question – what will be the effect on earnings of policies which increase the education of persons who would not otherwise receive this education?

These and related concerns suggest an experimental approach in which 'treatments' and their outcomes (or lack of outcomes) are analyzed.

Laboratory experimentation is attractive because of the possibilities it offers for achieving control over both the explanatory variable(s) of key interest and other variables which are nuisance factors in the context of the experiment. Yet laboratory experiments have their limitations. Orcutt and Orcutt (1968, pp. 766–767) single out the treatment of nuisance factors as an important disadvantage:

> The primary drawback of placing much reliance on laboratory experimentation is that the conditions under which research results need to be applied cannot be satisfactorily duplicated in the laboratory. Many variables besides potential policy variables influence behavior.... The mere fact that the values of these other variables might be held constant in the laboratory would be of little help, since the basic difficulty is that they would be held constant at the wrong levels.... To guard against

[5]See Zellner (1971, 1984, 1985), and Tiao and Zellner (1964). For applications see, e.g., Zellner and Rossi (1984), H. Tsurumi (1976), and H. Tsurumi and Y. Tsurumi (1983).

> such surprises it is important for experimentation to be carried out in as realistic settings as possible, and with samples of experimental units which are representative of the population with which policy makers must deal.

For different reasons, nuisance factors have also clouded the interpretation of findings from field experiments, though field experimentation continues to be viewed as a promising research tool [see Hausman and Wise (1985) and Nakamura and Nakamura (1986)].

By more carefully considering the advantages of experimental data, econometricians may develop better ways of collecting and analyzing non-experimental data. Orcutt, Nakamura and Nakamura (1980, p. 62) suggest, for example:

> A central focus of econometric research using nonexperimental data has been on accounting for and predicting the variation of one or more variables given current and lagged values of other variables.... In addressing similar sorts of problems in areas of science for which controlled experimentation is possible, relatively little attention is focused on variance reduction. Rather experiments are designed so as to facilitate observation of the impacts of well-defined treatments. This sort of treatment-response approach to research could also be applied more widely in analyzing non-experimental data. The key elements of this approach are the identification of well-defined treatments, and the collection of data that allow observation of the impacts of these treatments.

Orcutt, Nakamura and Nakamura (1980, pp. 64–65) also argue that econometricians should consider using data more selectively:

> Econometricians accustomed to working with macro time series have learned to use every possible observation in estimating any relation.... But what researcher tries to pool data, even before estimating and testing, from quite different experiments? The fact that the same or an, not any overlapping set of variables happens to be involved is not sufficient.... In planned experimentation an attempt is made to apply an action of interest on at least three or more widely separated levels of application. If implications of two or more actions are being explored, then an attempt is made to avoid or minimize covariation between assigned treatment levels. In an effort to avoid mistakenly attributing outcomes to treatments, experimentalists make use of observations on carefully selected control groups.... The experimentalist is thus extremely selective with respect to sample points.... The researcher who wishes to learn from naturally occurring applications of treatments of interest has every reason to be equally selective of sample points.

In many situations it may be useful to combine insights derived from experimental evidence and full model estimation: Insights from the estimation of complete models may guide selection of experimental evidence while experimental evidence may be used to test causal assumptions. For instance, in an examination of income effects on mortality experience, Duleep (1986a) used natural experiments to test whether income affects mortality, as opposed to a third variable (or variables) determining both socioeconomic status and health. Results from the estimation of an income-mortality model were used to select experimental evidence. In particular, the finding that an inverse relationship between mortality and income occurs only at levels of income focused attention on experimental evidence relating reductions in poverty to mortality outcomes.

## 5. Models as approximations

Regardless of the model selection methodology, in some cases, the isolation of a unique 'true' model may be an unrealistic or even an undesirable goal. There are at least four reasons for interest in explicitly approximate models. The first is ignorance of aspects of the true model such as the functional form or properties of the error term(s). A second is lack of data for some of the variables in the model. Many variables such as human capital, user and opportunity costs, and 'real' monetary variables cannot usually be directly observed. A third reason is that the true model may be too complex to be estimated accurately. There may be a trade-off between the realism of a model and the accuracy with which the parameters of the model can be estimated [see, e.g., Sawa (1978) and Zellner (1984, pp. 31–32)]. Obviously the smaller the quantity of data that is available, and the poorer the quality of the data (due, for example, to autocorrelation or multicollinearity), the greater the limitations are on the accuracy with which models of increasing complexity can be estimated.

A fourth reason for interest in approximate models is that a simpler model may be preferable for the intended purpose for which the model will be used. This is the reason Leamer (1978, p. 114) has in mind when he writes that 'what often appears to be choice among potentially true models is, in fact, the choice of a simple model that works well for some decisions.' Leamer (1978, p. 205) gives an example about maps that clarifies what he means by 'a simple model that works well for some decisions':

> We may take as a theory of the world an enormously detailed globe which identifies every object down to the smallest grain of sand. The complexity of this theory effectively prevents us from using it for any purpose whatsoever. Instead, we simplify it in the form of a set of maps. I use one map to find my way to the subway station, another to select

> the station at which to depart. The pilot of the airplane uses yet another to navigate from Boston to Washington. Each map is a greatly simplified version of the theory of the world; each is designed for some class of decisions and works relatively poorly for others.

In the remainder of this section, we briefly review three (non-exhaustive) types of approaches for developing approximate empirical models.

### 5.1. *Model building when the functional form is unknown*

Kmenta (1986, pp. 516–517) writes:

> The choice of the functional form of a regression equation should be made on theoretical or [a priori] empirical grounds. Because such grounds are frequently difficult to establish, efforts have been made to solve the problem of choice in various ways. One approach has involved the use of a functional form that is flexible enough to approximate a variety of special forms.

The simplest and most widely used functional approximations are step functions.[6] When a step function is used to represent the impact of a continuous explanatory variable, the goodness of the approximation depends on the number of steps that are defined. Since a degree of freedom is lost for each additional 'step', there is a trade-off between the goodness of the approximation and the efficiency with which the parameters of the step function can be estimated. Multicollinearity is also a problem when many dummy variables are used. This trade-off is more serious when a step function representation is used for more than one of the explanatory variables, and particularly when interaction as well as direct effects of these variables are allowed for. We mention these well-known approximation tradeoffs because tradeoffs are an essential feature of all approximations.

A conceptually distinct problem with approximations concerns the problem of relating properties of an approximate model to properties of a hypothesized theoretical model. For example, Goldberger (1964, p. 222) notes that the fact that step functions remain flat over ranges of the explanatory variable(s) makes it difficult to define partial derivatives that play prominent roles in many theoretical models. There are ways of dealing with this problem. For instance, Orcutt et al. (1961, pp. 229–231 and 241–250) note that the estimated response coefficients of a step function are essentially the sample means for the dependent variable for those observations with values for the designated explanatory variable falling into the range for each step

[6]See Goldberger (1964, pp. 218–231), and Kmenta (1986, pp. 461–473). Step function models can be viewed as a simple case of varying coefficient models. Splines are another related type of approximate model. See Poirier (1974), Buse and Lim (1977), and Johnston (1984, pp. 392–396).

(holding any other explanatory variables fixed). The shape of the resulting estimated mean response function can be approximated by a continuous curvilinear functional form. In essence, the suggestion is that a further aproximation can be used to translate findings based on the approximate empirical model back into the framework of the theoretical model which motivated the specification of the empirical model.[7] Some might ask why we want to make this translation. Leamer (1978, p. 228) offers an interesting perspective on this question in the context of relating the parameters of approximate and theoretical models:

> I think the answer to this question has to do with the problem of pooling information from different sources. 'Pure' prior information may apply to the theoretical parameters, and even if interest centers on the other parameters it is necessary to know their relationships in order to make use of prior information.

The translation problem on which Leamer is commenting is less evident – but still there – with continuous functional approximations such as the Box–Cox transformations.[8] The first explicit attempts we are aware of to develop approximations that facilitate the theoretical-approximate model translation problem use Taylor expansions as the approximating mechanism [see Diewert (1969, 1971, 1973, 1974)]. Diewert (1986, p. 79) argues:

> We should attempt to choose functional forms that are *flexible*; i.e., they can provide a second order approximation to an arbitrary twice continuously differentiable function with the appropriate theoretical homogeneity and curvature properties.... If the researcher does not use flexible functional forms, then unwarranted a priori restrictions on elasticities of substitution will be imposed. For example, the use of the Cobb–Douglas, Leontief or C.E.S. functional forms does not allow any pair of goods to be complements....

Many others have stressed the importance of not imposing unwarranted a priori restrictions prior to estimation. Diewert brought additional ways of accomplishing this purpose to the attention of economists. But what truly distinguishes the approximation approach of Diewert from earlier attempts to develop functional forms that, in Kmenta's words, are 'flexible enough to approximate a variety of special forms' is the attention to the problem of how specific properties of a theoretical model can be allowed for and represented in a functional approximation of this model. Functional forms

[7]Actually Orcutt et al. propose making this additional functional approximation for the purpose of improving and facilitating the way in which an estimated step function is incorporated into a microsimulation model.

[8]Material on Box–Cox transformations can be found, e.g., in Kmenta (1986, pp. 518–520), Judge et al. (1985, p. 257 and pp. 839–842), Amemiya (1985, pp. 249–252) and Box and Cox (1964, 1982).

that are flexible, as this term is defined by Diewert, include the translog function and Diewert's biquadratic function.[9] Pursuing the same objectives enunciated in Diewert's work, Barnett (1983, 1985) develops functional approximations that are Diewert-flexible, but are based on Laurent series rather than Taylor series expansions.

A related, but technically different, problem of interpreting empirical results based on approximations of theoretical models has to do with understanding the stastistical behavior of estimators of the parameters of an approximate model. Gallant (1981, p. 212) contends that, viewed from this perspective, 'Taylor's theorem fails rather miserably' as a basis of developing functional approximations. He explains that this is because 'statistical regression methods essentially expand the true function in a (general) Fourier series – not in a Taylor's series.' Gallant explores the properties of a flexible functional form based on a Fourier series approximation. Approaching similar research questions in a different way, White (1980b) delves into the statistical properties of the parameter estimates and the predictions of the dependent variable when least squares regression is applied to a 'mispecified' model that is a functional approximation of some 'true' model. [See also White (1982).]

### 5.2. *Living with multicollinearity*

Approximate modelling strategies are pursued sometimes because it is thought that the available data would not permit satisfactory estimation of the 'true' model even if its form could be theoretically specified. Such a situation may arise, for instance, because of collinearity among two or more of the explanatory variables. The ridge regression and Stein-like estimators are some of the procedures that have been proposed for estimating approximations to underlying behavioral models in the face of multicollinearity problems.[10]

### 5.3. *The use of proxies for unobservable variables*

A typical empirical model of female labor supply incorporates variables for child status, a woman's age, the income of the husband (for married women), the woman's years of schooling, and so forth. All these variables are really proxies for underlying (and often unmeasured) hypothesized effects. For example, it is the time and dollar costs of caring for children, not their mere presence, that theoretically affect maternal labor supply. Likewise, labor

[9]See Diewert (1986, pp. 80–96), Diewert (1969, 1973, 1974), Christensen, Jorgenson and Lau (1971, 1975), Jorgenson (1984), and Deaton and Muellbauer (1980, pp. 73–78).

[10]See Hoerl and Kennard (1970a, 1970b), Kmenta (1986, pp. 430–442), Judge et al. (1985, pp. 912–926), and Amemiya (1985, pp. 55–69).

supply behavior may vary with age because of factors such as the accumulation of work experience, the reluctance of employers to hire older women, changing family responsibilities, and increasing health problems as a woman ages. Except for age-based rules for aspects of work behavior such as retirement, it is hard to think of any reason why simply getting a year older should alter a woman's labor supply.

One problem with the use of proxy variables is that they may capture not only the effects of the unobservable factors they are intended to capture, but also the effects of other unobservable (or omitted) factors. For example, Schultz (1978) has raised the possibility that women may differ in terms of background factors and tastes which affect both their child-status characteristics and their current labor supply. If Schultz's arguments are correct, then when child status variables are directly introduced into standard models of female labor supply these variables will capture not only the direct impacts of children, but will also serve as proxies for the persistent preferences and preconditions that have affected both fertility and (past as well as current) labor supply.

Schultz's argument have stimulated the development of models treating both the fertility and labor supply of women as endogenous. In empirical implementations of these models, the child status variables appearing in the labor supply equation(s) are replaced by estimated linear combinations of 'exogenous' variables. That is, the child status variables are split into 'explained' and 'unexplained' portions, and the unexplained portions of the original child status variables are relegated to the disturbance term(s) of the labor supply equation(s). The hope is that the explained portions of the child status variables (the instrumental child status variables, which are linear combinations of exogenous variables) will be uncorrelated with the omitted, persistent tastes and preconditions affecting both child status and labor supply. The distinction sometimes drawn between use of an instrumental variable in a case like this and the use of a proxy variable is circumstantial, not statistical, in nature. Kmenta (1986, p. 579) admonishes:

> A *proxy variable* is not to be confused with an *instrumental variable*.... Instrumental variables are used when $X$ [the explanatory variable thought to belong in the 'true' model] *is* observable but correlated with the regression disturbance.

Unfortunately it is difficult to find suitable exogenous variables to use in forming instrumental child status variables. A reconsideration of the issues raised by Schultz led Nakamura and Nakamura (1985b) to instead suggest the introduction into the labor supply model of additional proxy variables to control for the background factors and tastes that might otherwise be picked up by the original child status variables. They argue that if there are important unobservable factors which affect the labor supply of individual

women year after year, then the effects of these unobservable factors will be embedded in the observable past work behavior of these women. In this case, it should be possible to at least partially remove these unobservable factors from the error term(s) of the labor supply equation(s) by introducing some measure of past labor supply into the relationship(s) to be estimated.

There are a number of other areas of economic analysis where the problems of dealing with unobservable nuisance variables are severe and where the development of panel data bases has opened up the possibility of using a proxy approach of the sort implemented by Nakamura and Nakamura in their more recent studies of female labor supply. The study of income effects on mortality experience[11] and the analysis of firm behavior[12] are two such areas. Models incorporating explanatory variables that are included as proxies for other variables hypothesized to affect the behavior of the dependent variable are inherently approximate in nature.[13]

## 6. Ranking alternative approximations

The model choice procedures discussed in section 2 cannot be readily used for choosing among models that are explicitly regarded as approximations. One reason is that, even when a correspondence can be established between the parameters of an approximate model and the associated 'true' model, the resulting estimated coefficients cannot usually be regarded as consistent estimates of the parameters of the 'true' model [as demonstrated, e.g., by White (1980b) and Gallant (1981)]. Thus it is difficult to use economic theory in a convincing way as a criterion of model choice. In fact, it is difficult to use any sort of decision rule focusing primarily on the estimated coefficients, since the coefficients may have differing interpretations in alternative approximate models. A second problem is that most of the existing model choice procedures have been designed to select the 'true' model rather than to *rank* alternative imperfect models.[14] However, procedures for evaluating empirical

[11]See Duleep (1986a, 1986b, 1986c).

[12]See Day (1967), Crain, Shughart and Tollison (1984), Nelson and Winter (1982), and Nakamura and Nakamura (1985d, 1988).

[13]See McCallum (1972), Wickens (1972), Aigner (1974), Frost (1979), Ohtani (1985), Kmenta (1986, pp. 579–581), and Judge et al. (1985, pp. 709–711) on the use of proxy variables and the associated problems of coefficient bias and inconsistency.

[14]Consider a specification error test, such as the Wu–Hausman test for correlations between included explanatory variables and the error term. When the null hypothesis of no correlation is rejected, no information is provided about the seriousness or consequences of the implied correlation problem [See Nakamura and Nakamura (1985c)]. The researcher has no way of judging whether a better approximation to the underlying behavioral relationship could be obtained by directly estimating the original model despite the apparent correlation problem, or by adopting an instrumental variables approach which might reduce or eliminate coefficient inconsistencies resulting from the correlation problem but which will also result in a loss of efficiency. For differing points of view on appropriate objectives in ranking alternative imperfect models see, e.g.., Chin and Kennedy (1987) Dastoor (1990), and Pollak and Wales, forthcoming.

models based on the correspondence between the actual and predicted values of the dependent variable(s) are suitable for ranking or choosing among approximate models. This point is explicitly recognized by Sawa (1978), Leamer (1978), and Vuong (1989).

### *6.1. Output space model evaluation: single statistic methods*

In the case of output space evaluation methods which focus on the predicted versus the actual values of the dependent variable, a better fit, in some specified sense, is taken to be evidence of a better model. The more recent literature on output space evaluation methods has pursued the objective of identifying some single statistic or index that can be used to gauge the 'goodness' or 'relative goodness' of an estimated model.

The standard $R^2$, which is also the square of the simple correlation coefficient between the predicted and actual values for the dependent variable, is one possible index of goodness. The model with the highest in-sample $R^2$ is the one for which the in-sample sum of the squared residuals is minimized. In this sense it is the best fitting model. There is the obvious problem, however, that the in-sample $R^2$ can be increased simply by increasing the number of regressors. To deal with this problem, both Theil's adjusted $R^2$ [Theil (1961 and 1971, pp. 178–179)] and Amemiya's modified $R^2$ [Amemiya (1966 and 1985, p. 51)] take account of the in-sample loss of degrees of freedom as additional variables are added into a model. The modified $R^2$ imposes a higher penalty on increasing the number of regressors than Theil's adjusted $R^2$ does. Another conceptually important difference between Amemiya's and Theil's $R^2$s is that the modified $R^2$ is adjusted for degrees of freedom taking explicit account of a loss function. In particular, maximizing the modified $R^2$ is equivalent to minimizing Amemiya's Prediction Criterion, PC, which is essentially the unconditional mean square prediction error [see Amemiya (1980)].

Various measures of the mean square prediction error can be thought of as making up a second category of output space indices of model goodness. Amemiya's PC (which is minimized when the modified $R^2$ is maximized) belongs in this category. Several other members of this category are derived by minimizing the expected mean square prediction error conditional on the matrix of observations for the possible regressors [see Judge et al. (1985, pp. 875–879), Mallows (1964, 1973), Allen (1971), and Leamer (1983)].

A third category of indices of model goodness are based on the Kullback–Leibler Information Criterion (KLIC). Members of this category include Akaike's Information Criterion (AIC), Sawa's BIC Criterion, another criterion derived by Akaike using a Bayesian framework (AIC), yet another variant of AIC ($AIC_2$) developed by Akaike, and Vuong's likelihood ratio

test statistics [see Kullback (1959), Akaike (1973, 1978, 1981), Sawa (1978) and Vuong (1989)]. The Cox test statistic for non-nested hypotheses is also a variant of the Neyman–Pearson likelihood ratio [see Cox (1961, 1962), Pesaran (1974), Dastoor (1985), and McAleer (1987)], and the $J$ test suggested by Davidson and MacKinnon (1981) and the JA test proposed by Fisher and McAleer (1981) are closely related to the Cox test.

The fourth category consists of Bayesian measures of model goodness. The most important member of this category is probably the posterior odds ratio [see Zellner (1971, pp. 291–317)]. The posterior odds can be expressed as the prior odds times the ratio of the averaged likelihoods, with the prior densities serving as the weighting functions. (This contrasts with the usual likelihood ratio which is a ratio of maximized likelihood functions.) In sufficiently large samples the posterior odds are simply the conventional likelihood ratio multiplied by the prior odds. Zellner (1978) shows that, in a mathematical sense, the posterior odds ratio is closely related to Akaike's AIC statistic.

Pearson's chi-square statistic provides the basis for a fifth category of output space indices of model goodness. Chi-square type statistics are not mathematically related to the $R^2$, the unconditional or conditional mean square prediction error, the KLIC distance measure or the likelihood ratio statistic, or the posterior odds ratio.

Within the classical hypothesis testing framework, chi-square statistics can be used to test whether the specified model can be accepted as the 'true' model.[15] Chi-square type statistics can be used for diagnostic purposes in the spirit of conventional specification error tests [see Andrews (1988a, 1988b)]. Chi-square type statistics can also be used for *ranking* alternative models in terms of the degree of congruence between the predicted and actual distributions of the dependent variable(s) of interest. In fact, Massy, Montgomery and Morrison (1970, p. 36) assert: 'The chi-square statistic may be more useful for comparing the fit of two different models than it is in evaluating the correctness of either model.' The first application of this sort that we are aware of in the economics literature is due to Heckman (1981). Heckman's application involves ranking alternative models for the simple binary choice of working versus not working each year, including one model which is explicitly approximate in that it incorporates a proxy explanatory variable. Nakamura and Nakamura (1983, 1985a, 1985b) extend this approach to accommodate models involving multiple discrete choices, continuous dependent variables, and outcomes which reflect the joint outputs of multiple behavioral relationships. Heckman and Walker (1988) use this and other approaches in examining alternative models of fertility.

[15]See McFadden (1974), Heckman (1984) Horowitz (1985), and Andrews (1933a, 1988b). For marketing applications see Bagozzi and Yi (1988).

### *6.2. An older tradition*

Despite continuing efforts to develop a single index of goodness that can be used for ranking or choosing among alternative models, the limitations of *any* single measure are acknowledged by even those contributing to this literature. In the 'Conclusions' of a paper comparing several simple indices of goodness, including the modified $R^2$ and the PC index, Amemiya (1980, p. 352) writes:

> It is not my intention to recommend any single criterion as a definitely superior criterion or as a panacea of the problem of selection. On the contrary, the general picture that has emerged from this paper is that all of the criteria considered are based on a somewhat arbitrary assumption which cannot be fully justified, and that by slightly varying the loss function and the decision strategy one can indefinitely go on inventing new criteria. That is what one would expect, for there is no simple solution to a complex problem.... However, if I must rely on a single index, I would rather use....

And he goes on to sum up his conclusions on which of the indices he considers are the best ones. Likewise, in discussing the meaning and proper usage of the posterior odds ratio, Zellner (1971, pp. 296–297) writes:

> In the Bayesian approach *explicit* consideration is given to the loss structure.... It is extremely interesting to investigate the implications of various loss structures.... Obviously, with other loss structures the specific prescription for action will be different but the principles will be the same.

There is an older tradition in econometrics of output space evaluation procedures that are more descriptive and exploratory in nature. Moreover the spirit of these procedures is to develop a better understanding of which models might be more or less appropriate to use for particular applications, and to determine where research should be directed in trying to improve the usefulness of particular models. These are different, though related, objectives from general investigations of possible specification problems or attempts to determine the range of alternative models that yield essentially the same inferences (a Leamer-style investigation of fragility).

An example of this older tradition are graphical methods for analyzing residuals. Residual analyses are output space evaluation methods, since residuals are just the differences between the predicted and actual values of the dependent variable. At one point, a great variety of graphical methods for analyzing residuals were part of standard econometric practice [see Ezekiel and Fox (1959)]. For example, in his econometrics text, Christ (1966) writes:

> Graphical methods are excellent for getting a quick impression. The graph of calculated residuals against $t$ [time] is often helpful, for if its peaks and troughs coincide with (or lead or lag by a constant amount) those of some other economic variable that has not been used in the estimation process, it may be that this variable belongs in the equation being estimated (or at least in the model). Also, if residuals are plotted on one axis and a possibly important omitted variable on the other, we can very quickly see about how close a relationship exists between them and also whether it is approximately linear. If an omitted variable can be found that is correlated with the residuals and has good economic reason for being considered as related to the phenomenon described by the equation, it will be helpful to include that variable in the equation.

As the efforts of (particularly younger) economists have become increasingly focused on satisfying journal referees – as opposed to meeting the needs of firm managers, government policy makers, or others who might make substantive use of the research – there has been waning interest in models tailored for particular uses. This is surely one reason for the decline in interest in model choice methodologies, such as graphical analyses of residuals, providing evidence about *how* the predictions of an estimated model fit in a *particular* setting. A second factor has been the push for scientific rigor, and hence for selection methodologies that are well grounded in statistical, or some other accepted methodological body of, choice theory. A third factor has been the presumption that the 'true' model is typically among the alternatives being considered, and that the 'true' model would always be preferred to any alternative. The 'true' model is thought of as being unique. Any *one* method that results in the selection of this 'true' model is implicitly viewed as accomplishing the premier goal of specification analysis.

Two further reasons why once popular, more descriptive types of output space model choice and validation methods have fallen into disrepute are associated with the macroeconometric environment within which these methods evolved. The first of these reasons has to do with the paucity of the available data at the macro level. In a macro data environment, only time series comparisons between the actual and predicted values of the dependent variable are possible. All available observations, and most of the information contained in these observations, are usually used in estimating macro models. As a result, in-sample predictive evaluations of the model are of limited usefulness. Out-of-sample evaluations would be better. However, out-of-sample model evaluation cannot be carried out until more data become available. Moreover, due to autocorrelation, in most cases the out-of-sample observations on the variable of interest are largely an extrapolation of the in-sample values [see, e.g., Orcutt (1948) and Ames and Reiter (1961)]. Thus

even out-of-sample predictive comparisons will not provide a rigorous basis for model evaluation.

This criticism is less relevant in a micro data environment. Often there are sufficient data that some portion can be used for estimation and the rest can be reserved for immediate out-of-sample testing. The out-of-sample data are not usually autoregressively related to the in-sample data. Thus the out-of-sample tests that can be carried out are more meaningful. Also in a micro data (and particularly in a panel data) environment it is often the case that not all of the information in the in-sample data is used in estimating the model [see Nakamura and Nakamura (1985b, p. 198)]. For example, an equation for the probability of work in a year cannot incorporate all of the available information in a panel data set on the year-to-year employment status of each individual. Rather standard practice is to summarize this information in terms, say, of a variable for the number of years each individual has worked or a variable (or variables) characterizing the work behavior of each individual in the previous year [see Heckman (1981), and Nakamura and Nakamura (1985a, 1985b)]. In this situation, even appropriate in-sample predictive tests may provide useful insights as to the quality of an estimated model.

Finally, output space model evaluation methods have been criticized on the grounds that they focus exclusively on the predictive abilities of models. Applied economists are often more concerned about the extent to which their estimated models embody appropriate behavioral responses. Orcutt (1952, pp. 195–196) reminds us:

> ... it is necessary to be able to predict something about the way in which the use of the control instruments may be expected to modify an unknown future course. As a bare minimum, this means knowing with some confidence whether or not a given course of action will raise or lower the position of the variable we wish to control relative to its future path in the absence of the action.

It is true that output space model evaluation methods were first developed in a macroeconometric forecasting environment where predictive ability is of paramount importance. It is also true that the model which provides the most accurate forecasts for some dependent variable in, say, a mean square error sense, will not necessarily provide the most accurate estimates of how the dependent variables will change on average in response to a specified change in some explanatory variable. This point is often made in the context of discussing the relative merits of a structural versus a reduced form representation of a variable. Yet there is a relationship between predictive ability and the ability of a model to properly capture responses to the included explanatory variables. Moreover an estimated model, including the assumed and estimated properties of the error term, should be able to

reproduce the observed distribution of the dependent variable conditional on the observed values for the explanatory variables. If it cannot, then at the very least there is probably some difficulty with the specification of the properties of the error term, which may mean that the standard tests of significance are inappropriate.

## 7. Recommendations

In our view further research on model choice should be guided by the following recommendations:

(1) *It should be recognized that virtually all economic models are approximations.*

The rhetoric of econometric practice should be brought into line with White's (1980, p. 162) assertion that 'most econometric estimating relationships are intended as approximations, rather than as the "truth"'.

(2) *Less attention should be devoted to model choice methods focused on the acceptance or rejection of models without reference to intended uses of the models.*

Nor should papers be rejected by journals simply because they do not measure up to the standards of these methodologies. For instance, evidence of a specification problem, such as sample selection bias, for which no econometric adjustment has been made should not be a sufficient reason for rejection. Most specification error tests provide no obvious basis for judging the costs, in terms of the goodness of the behavioral approximation, of the specification problems that are detected by these tests. *Rather model specification and choice should be practiced with due regard for Leamer's (1978, p. 205) analogy about different maps for different purposes: 'Each map is a greatly simplified version of the theory of the world; each is designed for some class of decisions and works relatively poorly for others.'*

(3) *Accepted practices for choosing among competing models should exhibit a courtroom-style approach rather than a mechanical application of statistical tests.*

In the concluding section of *The Second Paycheck: A Socioeconomic Analysis of Earnings*, Nakamura and Nakamura (1985b, p. 365) explain:

> Our behavioral conclusions all rest on the accumulation of circumstantial evidence.... Moreover, the issue of when the weight of accumulated evidence is sufficient to warrant a particular conclusion is treated as a matter of judgment. In a courtroom proceeding, eyewitness reports, expert testimony and various sorts of circumstantial evidence may all be brought before the court, but it is the ultimate responsibility of a judge or jury to weigh this evidence and reach a verdict. In a study ... in which there is uncertainly about the proper specification of the func-

> tional forms of the behavioral relationships, about the distributions of the disturbance terms, and so forth, we do not believe that better conclusions will necessarily be reached by avoiding the degree of arbitrariness inherent in judgmental decision making by appeals to mathematical statistics predicated on assumptions that cannot be checked.

In the practice of a courtroom-style approach to model choice, we concur with Leamer (1978, p. 123) that 'arguments concerning the use of prior information should ... address the question of how rather than whether prior information should be used'. We also feel strongly that prior information about the aspects of an approximation that are most crucial for a particular application, or concerning the interpretation of or the qualification of the empirical findings, should enter into the 'court-room' consideration of the evidence even if it is not obvious how this information can be summarized in the form of prior distributions or a mathematically specified loss function.

(4) *In line with the spirit of a courtroom style approach to model choice, different indices of model goodness should be examined from the perspective of the extent to which they provide different sorts of information about the strengths and weaknesses of particular models.*

This is why we have organized our discussion of various output space indices of model goodness in terms of the basic statistics on which these indices are based, and have commented on or provided references concerning how these basic statistics are mathematically related.

(5) *In a cross-sectional or panel micro data environment, the potential for exploring the strengths and weaknesses of alternative models in different dimensions by examining the original data and model outputs from different perspectives should be exploited as fully as possible.*

Indices and other evidence of model fit should be examined for interesting subsets of the data as well as for the entire data set, out-of-sample as well as in-sample, at different levels of aggregation, and for meaningful combinations of model outputs (such as annual earnings defined as the product of the hourly wage rate times annual hours of work) as well as for individual dependent variables (such as annual hours of work), with the intended uses of the model guiding this investigation. This is the approach adopted, for example, in Nakamura and Nakamura (1985b) and in Heckman and Walker (1988). In his foreword to *The Second Paycheck* Heckman (1985) explains his views on a courtroom approach to model choice versus more standard methods in a labor economics context:

> The approach pursued in many recent studies of labor supply has been to arrive at final, empirical specifications for a single demographic group by means of a battery of '*t*' and '*F*' tests on the coefficients of candidate variables. The problem of pretest bias [the multiple tests problem] is

> conveniently ignored. Only rarely ... do analysts ask how well fitted micro relationships explain other aspects of labor supply such as the aggregate time series movement....
>
> This book does not adopt the conventional '*t*' ratio methodology. The authors estimate the same models for a variety of age, marital status, and sex groups and look for commonalities in findings across groups. They look for consistency in the impact of explanatory variables on different dimensions of labor supply. Models are simulated both within samples and out of samples.... The simulation format has the additional feature of spelling out the implications of complex models that are not obvious from reported coefficient estimates. The rigorous body of tests proposed and implemented by the authors of this book ... sets a new, high standard that will be followed by all serious scholars of the subject.

In summing up, the first of the above recommendations represents a return to reality. The second recommendation reaffirms the fundamental American principle that usefulness is more important than style and the elegance of generality. And recommendations 3 through 5 can be thought of loosely as a generalization of the important British 'Encompassing Principle' which originated from Hendry's work.[16]

[16]See Hendry (1983) and Mizon and Richard (1986). Mizon and Richard (1986, p. 657) state that the Encompassing Principle 'requires a model $M$ to be able to explain characteristics of rival models'. Here we are extending the principle to data encompassing: that is, a model should be able to explain the results of viewing the data from rival vantage points (such as in cross-sections, over time, grouped according to the values of certain explanatory variables, and so forth).

## References

Aigner, D.J., 1974, MSE dominance of least squares with errors of observations, Journal of Econometrics 2, 365–372.

Akaike, H., 1973, Information theory and the extension of the maximum likelihood principle, in: B.N. Petrov and F. Csaki, eds., 2nd international symposium on information theory (Akailseoniai–Kiudo, Budapest) 267–281.

Akaike, H., 1978, On the likelihood of a time series model, presented at the Institute of Statisticians 1978 Conference on Time Series Analysis (Cambridge University, Cambridge, England) July.

Akaike, H., 1981, Likelihood of a model and information criteria, Journal of Econometrics 16, 3–14.

Allen, D.M., 1971, Mean square error of predictions as a criterion for selecting variables, Technometrics 13, 469–475.

Amemiya, T., 1966, On the use of principal components of independent variables in two-stage least-squares estimation, International Economic Review 7, 283–303.

Amemiya, T., 1980, Selection of regressors, International Economic Review 21, no. 2, 331–354.

Amemiya, T., 1985, Advanced econometrics (Harvard University Press, Cambridge, MA).

Ames, E. and S. Reiter, 1961, Distribution of correlation coefficients in economic time series, Journal of the American Statistical Association 56, 736–656.

Andrews, D.W.K., 1988a, Chi-square diagnostic tests for econometric models: Introduction and applications, Journal of Econometrics 37, no. 1, 135–156.
Andrews, D.W.K., 1988b, Chi-square diagnostic tests for econometric models: Theory, Econometrica 56, no. 6, 1419–1453.
Andrews, D.W.K., 1989, Power in econometric applications, Econometrica 57 no. 5, 1059–1090.
Bagozzi, R.P. and Y. Yi, 1988, On the evaluation of structural equation models, Journal of the Academy of Marketing Science 16, no. 1, 74–94.
Barnett, W.A., 1983, New indices of money supply and the flexible Laurent demand system, Journal of Business and Economic Statistics 1, 7–23.
Barnett, W.A., 1985, The Minflex–Laurent translog flexible functional form, Journal of Econometrics 30, 33–44.
Bera, A.K.. and C.M. Jarque, 1982, Model specification tests: A simultaneous approach, Journal of Econometrics 20, no. 1, 59–82.
Box, G.E.P. and D.R. Cox, 1964, An analysis of transformations, Journal of the Royal Statistical Society, Series B 26, 211–243.
Box, G.E.P. and D.R. Cox, 1982, An analysis of transformations revisited, rebutted, Journal of the American Statistical Association 77, 209–210.
Buse, A. and L. Lim, 1977, Cubic splines as a special case of restricted least squares, Journal of the American Statistical Association 72, 64–68.
Carter, R.L. and W.A. Fuller, 1980, Instrumental variable estimation of the simple errors-in-variables model, Journal of the American Statistical Association 75, 687–692.
Chin, C.F. and P.E. Kennedy, 1987, On inferring the true model's direction, Canadian Journal of Economics 20, 876–879.
Christ, C.F., 1966, Econometric models and methods (Wiley, Inc., New York).
Christensen, L.R., D.W. Jorgenson and L.J. Lau, 1971, Conjugate duality and the transcendental logarithmic production function, Econometrica 39, 255–256.
Christensen, L.R., D.W. Jorgenson and L.J. Lau, 1975, Transcendental logarithmic utility functions, American Economic Review 65, 367–383.
Cox, D.R., 1961, Tests of separate families of hypotheses, in: Proceedings of the fourth Berkeley symposium of mathematical statistics and probability, Vol. 1 (University of California Press, Berkeley, CA).
Cox, D.R., 1962, Further results on tests of separate families of hypotheses, Journal of the Royal Statistical Society, Series B, 24, 406–424.
Cragg, J.G., 1983, More efficient estimation in the presence of heteroscedasticity of unknown form, Econometrica 51, 751–763.
Crain, W.M., W.F. Shughart II and R.D. Tollison, 1984, The convergence of satisficing to marginalism: An empirical test, Journal of Economic Behavior and Organization 5, 375–385.
Dastoor, N.K., 1983, Some aspects of testing non-nested hypotheses, Journal of Econometrics 21, 213–228.
Dastoor, N.K., 1985, A classical approach to Cox's test for non-nested hypotheses, Journal of Econometrics 27, 363–370.
Dastoor, N.K., 1990, A note on model discrimination after testing, Canadian Journal of Economics 23, no. 1, 236–244.
Davidson, R. and J.G. MacKinnon, 1981, Several tests for model specification in the presence of alternative hypotheses, Econometrica 49, 781–793.
Day, R.H., 1967, Profits, learning and the convergence of satisficing to marginalism, Quarterly Journal of Economics 81, 302–311.
Deaton, A. and J. Muellbauer, 1980, Economics and consumer behavior (Cambridge University Press, New York).
Diewert, W.E., 1969, Functional form in the theory of production and consumer demand, Ph.D. dissertation (University of California, Berkeley, CA).
Diewert, W.E., 1971, An application of the Shephard duality theorem: A generalized Leontief production function, Journal of Political Economy 79, 461–507.
Diewert, W.E., 1973, Functional forms for profit and transformation functions, Journal of Economic Theory 6, 284–316.
Diewert, W.E., 1974, Applications of duality theory, in: M.D. Intriligator and D.A. Kendrick, eds., Frontiers of quantitative economics, Vol. II (North-Holland, Amsterdam).

Diewert, W.E., 1986, The measurement of the economic benefits of infrastructure services (Springer-Verlag, Heidelberg).
Dufour, J.-M., 1989, Nonlinear hypotheses, inequality restrictions, and non-nested hypotheses: Exact simultaneous tests in linear regressions, Econometrica 57, no. 2, 335–355.
Duleep, H.O., 1986a, Poverty and inequality in mortality, Ph.D. dissertation (MIT, Cambridge, MA).
Duleep, H.O., 1986b, Measuring income's effect on adult mortality, Journal of Human Resources 21, 238–250.
Duleep, H.O., 1986c, Incorporating longitudinal aspects into mortality research using social security administrative record data, Journal of Economic and Social Measurement 14, no. 2, 121–133.
Edwards, J.B. and G.H. Orcutt, 1969, The reliability of statistical indicators of forecasting ability, Working paper, Social Systems Research Institute (University of Wisconsin, Madison, WI).
Engle, R.F., 1982, A general approach to Lagrange multiplier model diagnostics, Journal of Economics 20, no. 1, 83–104.
Ezekiel, M. and K.A. Fox, 1959, Methods of correlation and regression analysis, linear and curvilinear, 3rd ed. (Wiley, New York).
M.S. Feldstein, 1974, Errors in variables: A consistent estimator with smaller MSE in finite samples, Journal of the American Statistical Association 69, 990–996.
Fisher, G.R. and M. McAleer, 1981, Alternative procedures and associated tests of significance of nonnested hypotheses, Journal of Econometrics, 103–119.
Frost, P.A., 1979, Proxy variables and specification bias, Review of Economics and Statistics 61, 323–325.
Gallant, A.R., 1981, On the bias in flexible functional forms and an essentially unbiased form, Journal of Econometrics b15, 211–245.
Goldberger, A.S., 1964, Econometric theory (Wiley, Inc., New York).
Griliches, Z., 1985, Data and econometricians – The uneasy alliance, American Economic Review 75, no. 2, 196–200.
Harvey, A., 1976, Estimating regression models with multiplicative heteroskedasticity, Econometrica 44, 461–465.
Hausman, J.A. and D.A. Wise, 1985, Social experimentation (University of Chicago Press, Chicago, IL).
Heckman, J.J., Heterogeneity and state dependence, in: S. Rosen, ed., Studies in labor markets (University of Chicago Press, Chicago, IL).
Heckman, J.J., 1984, The $\chi^2$ goodness of fit statistic for models with parameters estimated from microdata, Econometrica 52, 1543–1547.
Heckman, J.J., Foreword, in: The second paycheck: A socioeconomic analysis of earnings (Academic Press, Orlando, FL) pp. xi–xii.
Heckman, J.J. and J.R. Walker, 1988, Using goodness of fit and other criteria to choose among competing duration models: A case study of Hutterite data, C.C. Clogg, ed., Sociological Methodology, 1987 (American Sociological Association, Washington, D.C.)
Hendry, D.F., 1979, Predictive failure and econometric modelling in macroeconomics: The transactions demand for money, in: P. Ormerod, ed., Economic modelling: Current issues and problems in macroeconomic modelling in the UK and the US (Heinemann, London).
Hendry, D.F., 1983, Comment, Econometric Reviews 2, 111–114.
Hoerl, A.E. and R.W. Kennard, 1970a, Ridge regression: Biased estimation for non-orthogonal problems, Technometrics 12, 55–67.
Hoerl, A.E. and R.W. Kennard, 1970b, Ridge regression: Application to nonorthogonal problems, Technometrics 12, 69–82.
Horowitz, J.L., 1985, Testing probabilistic discrete choice models of travel demand by comparing predicted and observed aggregation choice shares, Transportation Research B 19, 17–38.
Johnston, J., 1984, Econometric methods, 3rd ed. (McGraw-Hill, New York).
Jorgenson, D.W., 1984, Econometric methods for modeling producer behavior, Discussion paper 1086 (Harvard Institute for Economic Research, Harvard University, Cambridge, MA).
Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lutkepohl and T.-C. Lee, The theory and practice of econometrics, 2nd ed. (Wiley, Inc., New York).

Kmenta, J., 1986, Elements of econometrics, 2nd ed. (Macmillan Publishing Company, New York).
Kramer, W. and H. Sonnberger, 1986, The linear regression model under test (Physica-Verlag, Heidelberg).
Kullback, S., 1959, Information theory and statistics (Wiley, Inc., New York).
Leamer, E.E., 1978, Specification searches: Ad hoc inference with nonexperimental data (Wiley, Inc., New York).
Leamer, E.E., 1983, Model choice and specification analysis, in: Z. Griliches and M. Intriligator, eds., Handbook of econometrics Vol. 1 (North-Holland, Amsterdam).
Leamer, E.E., 1985, Sensitivity analyses would help, American Economic Review 75, no. 3, 308–313.
Lovell, M.C., 1983, Data mining, Review of Economics and Statistics 65, no. 1, 1–12.
MacKinnon, J.G., 1983, Model specification tests against non-nested alternatives, Econometric Reviews 2, 85–110.
Mallows, C.L., 1964, Choosing variables in a linear regression: A graphical aid. Presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, KS, May.
Mallows, C.L., 1973, Some comments on $C_p$, Technometrics 15, 661–676.
Massy, W.F., D.B. Montgomery and D.G. Morrison, 1970, Stochastic models of buying behavior (MIT Press, Cambridge, MA).
McAleer, M., 1987, Specification tests for separate models: A survey, in: M.L. King and D.E.A. Giles, eds., Specification analysis in the linear model (Routledge & Kegan Paul, London).
McAleer, M. and M.H. Pesaran, 1986, Statistical inference in non-nested econometric models, Applied Mathematics and Computation 20, 271–311.
McCallum, B.T., 1972, Relative asymptotic bias from errors of omission and measurement, Econometrica 40, 757–758.
McCloskey, D.N., 1985, The loss function has been mislaid: The rhetoric of significance tests, American Economic Review 75, no. 2, 201–205.
McFadden, D., 1974, Conditional logit analysis of qualitative choice behavior, in: P. Zarembka, ed., Frontiers of Econometrics (Academic Press, New York).
Mizon, G. and J.F. Richard, 1986, The encompassing principle and its application to non-nested hypotheses, Econometrica 54, 657–678.
Nakamura, A. and M. Nakamura, 1973, Estimating the variances of sample correlations. Proceedings of the Social Statistics Section of the Annual Meeting of the American Statistical Association, 365–369.
Nakamura, A. and M. Nakamura, 1978, On the impact of the tests for serial correlation upon the test of significance for the regression coefficient, Journal of Econometrics 7, 199–210.
Nakamura, A. and M. Nakamura, 1981, On the relationship among several specification error tests presented by Durbin, Wu and Hausmann, Econometrica 49, no. 6, 1583–1588.
Nakamura, A. and M. Nakamura, 1983, Part-time and full-time work behavior of married women: A model with a doubly truncated dependent variable, Canadian Journal of Economics 16, no. 2, 229–257.
Nakamura, A. and M. Nakamura, 1985a, Dynamic models of the labor force behavior of married women which can be estimated using limited amounts of past information, Journal of Econometrics 27, 273–298.
Nakamura, A. and M. Nakamura, 1985b, The second paycheck: A socioeconomic analysis of earnings (Academic Press, Orlando, FL).
Nakamura, A. and M. Nakamura, 1985c, On the performance of tests by Wu and by Hausman for detecting the ordinary least squares bias problem, Journal of Econometrics 29, 213–227.
Nakamura, A. and M. Nakamura, 1985d, Rational expectations and the firm's dividend behavior, Review of Economics and Statistics 67, no. 4, 606–615.
Nakamura, A. and M. Nakamura, 1986, Review of social experimentation by J.G. Hausman and D.A. Wise, Journal of the American Statistical Association 81, 566–567.
Nakamura, A. and M. Nakamura, 1988, The profit behavior of U.S. and Japanese firms, Working paper (University of Alberta, Edmonton, Alta.).
Nakamura, A., M. Nakamura and G.H. Orcutt, 1976, Testing for relationships between time series, Journal of the American Statistical Association 71, no. 353, 214–222.

Nelson, R.R. and S.G. Winter, An evolutionary theory of economic change (Harvard University Press, Cambridge, MA).
Orcutt, G.H., 1948, A study of the autoregressive nature of the time series used for Tinbergen's model of the economic system of the United States, 1919–1932, Journal of the Royal Statistical Society, Series B 10, no. 1, 1–45.
Orcutt, G.H., 1952, Toward partial redirection of econometrics, Review of Economics and Statistics 34, 195–200.
Orcutt, G.H., 1968, Research strategy in modeling economic systems, in: The future of statistics (Academic Press, New York), 71–95.
Orcutt, G.H. and D. Cochrane, 1949, A sampling study of the merits of autoregressive and reduced form transformations in regression analysis, Journal of the American Statistical Association 44, 356–372.
Orcutt, G.H., M. Greenberger, J. Korbel and A.M. Rivlin, 1961, Microanalysis of socioeconomic systems: A simulation study (Harper and Brothers, New York).
Orcutt, G.H. and S.F. James, 1948, Testing the significance of correlation between time series, Biometrika 35 (Parts 3 and 4), 397–413.
Orcutt, G.H., A. Nakamura and M. Nakamura, 1980, Poverty research on family determination of labor income, in: V.T. Covello, ed., Poverty and public policy: An evolution of social science research (Schenkman, Cambridge, MA) 53–77.
Orcutt, G.H. and A.G. Orcutt, 1968, Incentive and disincentive experimentation for income maintenance policy purposes, American Economic Review 58, no. 4, 754–772.
Orcutt, H.G. and H.S. Winokur Jr., 1969, First-order autoregression: Inference, estimation, and prediction, Econometrica 37, 1–14.
Ohtani, K., 1985, A note on the use of a proxy variable in testing hypotheses, Eonomics Letters 17, 107–110.
Pesaran, M.H., 1974, On the general problem of model selection, Review of Economic Studies 4, 153–171.
Poirier, D.J., 1974, The econometrics of structural change (North-Holland, Amsterdam).
Pollak, R.A. and T.J. Wales, The likelihood dominance criterion: A new approach to model selection, Journal of Econometrics, forthcoming.
Ruud, P.A., 1984, Tests of specification in econometrics, Econometric Reviews 3, no. 2, 211–242.
Sawa, T., 1978, Information criteria for discriminating among alternative regression models, Econometrica 46, no. 6, 1273–1291.
Schultz, T.P., 1978, The influence of fertility on labor supply of married women: Simultaneous equation estimates, in: R. Ehrenberg, ed., Research in Labor Economics, Vol. 2 (JAL Press, Greenwich, CT), 273–351.
Theil, H., 1961, Economic forecasts and policy, 2nd ed. (North-Holland, Amsterdam).
Theil, H., 1971, Principles of econometrics (Wiley, Inc., New York).
Thursby, J.G., 1979, Alternative specification error tests: A comparative study, Journal of the American Statistical Association 74, 222–225.
Tiao, G.C. and A. Zellner, 1964, Bayes' theorem and the use of prior information in regression analysis, Biometrica 51, 219–230.
Tsurumi, H., 1976, A Bayesian test of the product life cycle hypothesis applied to Japanese crude steel production, Journal of Econometrics 4, 371–392.
Tsurumi, H. and Y. Tsurumi, 1983, U.S.-Japanese automobile trade: A Bayesian test of a product life cycle, Journal of Econometrics 23, 193–210.
Vuong, Q.H., 1989, Likelihood ratio tests for model selection and non-nested hypotheses, Econometrica 57, 307–333.
White, H., 1980a, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, Econometrica 48, 817–838.
White, H., 1980b, Using least squares to approximate unknown regression functions, International Economic Review 21, no. 1, 149–170.
White, H., 1982, Maximum likelihood estimation of misspecified models, Econometrica 50, no. 1, 1–25.
Wickens, M.R., 1972, A note on the use of proxy variables, Econometrica 40, 759–761.
Zellner, A., 1971, An introduction to Bayesian inference in econometrics (Wiley, Inc., New York).

Zellner, A., 1978, Jeffreys–Bayes posterior odds ratio and the Akaike information criterion for discriminating between models, Economic Letters 1, 337–342.
Zellner, A., 1979, Statistical analysis of econometric models, Journal of the American Statistical Association 74, 628–651.
Zellner, A. (ed.), 1984, Basic issues in econometrics (University of Chicago Press, Chicago, IL).
Zellner, A., 1985, Bayesian econometrics, Econometrica 53, no. 2, 253–269.
Zellner, A. and P.E. Rossi, 1984, Bayesian analysis of dichotomous quantal response models, Journal of Econometrics 27, 365–393.