M O N A S H

U N I V E R S I T Y

AUSTRALIA

COMMENTS ON TESTING ECONOMIC THEORIES AND THE USE OF MODEL SELECTION CRITERIA

Clive W.J. Granger, Maxwell L. King and Halbert White

Working Paper No. 17/92

December 1992

DEPARTMENT OF ECONOMETRICS

ISSN 1032-3813

ISBN 0 7326 0367 6

COMMENTS ON TESTING ECONOMIC THEORIES AND THE USE OF MODEL SELECTION CRITERIA

Clive Granger, Maxwell L. King and Halbert White

Working Paper No. 17/92

December 1992

DEPARTMENT OF ECONOMETRICS, FACULTY OF ECONOMICS COMMERCE & MANAGEMENT

MONASH UNIVERSITY, CLAYTON, VICTORIA 3168, AUSTRALIA.

COMMENTS ON TESTING ECONOMIC THEORIES AND THE USE OF MODEL SELECTION CRITERIA *

by

Clive Granger

University of California, San Diego, La Jolla, CA 92093, USA

Maxwell L. King

Monash University, Clayton, Victoria 3168, Australia

and

Halbert White

University of California, San Diego, La Jolla, CA 92093, USA

Abstract

This paper outlines several difficulties with testing economic theories, particularly that the theories may be vague, may relate to a decision interval different from the observation period and may need construction of a metric to convert a complicated testing situation to an easier one. We argue that it is better to use model selection procedures rather than formal hypothesis testing when asking the data to decide on model specification. This is because testing favors the null hypothesis, typically uses an arbitrary choice of significance level and researchers working with the same data could easily end up with different final models, which would make policy recommendations difficult.

* This research was supported by NSF grant SES 9023037 and by an ARC grant. Work on this paper commenced while the second author was a visitor at UCSD. He is grateful to the Department of Economics for its hospitality and support.

Address for correspondence: Professor Clive Granger,
Department of Economics 0508,
University of California, San Diego,
La Jolla, CA 92093-0508,
U.S.A.

1. Introduction

The basic paradigm for scientific research is the construction of an abstract theory, based on fundamental principles and sensible assumptions, from which can be derived propositions that should hold for the actual world, if the theory is correct. These propositions can be translated into specific hypotheses about properties of estimated models which can then be tested using actual data and statistical procedures. A model is here taken to mean an approximation to the generating mechanism of the variables occurring in the real world and which is also capable of containing the hypothesis of interest. In this paper we point out, using a fairly simple example, some of the difficulties that arise when trying to test a hypothesis (and thus a proposition or theory). It is then suggested that an alternative approach is at least worth discussing, in which a "best" model is selected, from a wide class of models, using a model selection criterion applied to actual data. This model is selected without any attention being paid to the hypothesis being investigated, except to insure that the data set used is sufficient for consideration of the hypothesis. The question of whether or not the hypothesis is correct thus becomes one of whether the model selected supports the hypothesis or not.

Our use of words such as theories, models, and hypothesis are standard in advanced statistical and econometric texts and formal definitions of "proposition" can be found in texts on the philosophy of science, such as Gärdenfors (1988). It is clear that the procedure suggested will not necessarily lead to a conclusion of either acceptance or rejection of the proposition - or hypothesis. This is again in accord with some aspects of modern philosophy and can be linked with the

idea of a belief function or a degree of belief B of the correctness of some particular hypothesis, again as discussed in Gärdenfors (1988). The purpose of these belief values is also to suggest why it is useful to analyze and "test" theories and hypotheses.

The following statements will be taken to be either self-evident (in the case of (a)) or, at least, to be reasonable working assumptions ((b), (c), (d)):

(a) Economics is a decision science. It is concerned with the decisions taken by economic agents, corporations, institutions and governments, and the effects of these decisions.

(b) Whether or not these decisions are optimal or optimizing, they are partially based on beliefs or individualistic "theories" about how the economy operates. To each theory, every economic agent has a "degree of belief" B, that the theory is correct. (Where "correct" can be taken to mean that the theory correctly specifies part of the generating mechanisms of the variables being considered, assuming such a mechanism exists). The values of the B's enter the decision process.

(c) The main, overt purpose of research in economics is to affect one's own degrees of belief or that of other researchers or of economic decision makers.

(d) Most economic agents will not change their B values if a theory is presented to them which has not been confronted with actual economic data. The use of test statistics is a helpful way of presenting evidence about the correctness of a theory or belief. They can be used to summarize this evidence in a possibly

uncontroversial way. Of course B need not change even if a theory has been confronted with data and been rejected.

It is convenient but not necessary to assume that B has the *properties* of a probability, so that $0 \leq B \leq 1$ and B is monotonically increasing as belief increases, but this does not imply that B *is* a probability. Unfortunately we are using the phrase "degree of belief" in almost the opposite sense to that used by Bayesians. For example, Judge et al. (1985, p.97) observed that "in a Bayesian framework probability is defined in terms of a degree of belief". (Also see Zellner, 1984, p.275.) Our use of the phrase corresponds more to the prior odds ratio of the hypothesis compared to a vague alternative. However B is used here merely as a pedagogical device and will not be treated formally in what follows.

Philosophers have discussed the dynamics of B-values (see Gärdenfors, 1988) as have Bayesians. Our attitude is that a statistical test is not a "final product" but rather an intermediate product, being an input to the decision process.

As a simple example, suppose that a government announces some general income tax cuts. There may be a theory that such cuts lead to an increase in the growth of GNP. A B-value for this theory may affect decisions about decreased savings rates by agents or increased investments by companies. Presenting evidence about the effects of the "supply side economics" tax cuts by the Reagan government in 1981 may affect B-values. In fact real US GNP growth was 3.1% in the 1970's and 2.8% in the 1980's which could suggest to some agents that B-values for this theory should be reduced. However other statistics on changes in

real, disposable income or on who benefits may affect B in other directions.

There are many aspects of B-values which need consideration but which we shall not discuss here. There can be a multidimensional aspect, with a theory having many forms, each of which has an associated B. Thus B is now a vector and its components may be interrelated. Similarly, if there exists a pair of alternative theories T_1, T_2 with B-values B_1, B_2 then presumably $0 \leq B_1 + B_2 < 1$ where the second inequality allows for the belief that neither theory is correct. There is also a potential problem with the dynamics, as if everyone has a high B-value it may affect behaviors such that the theory almost becomes true. Similarly, if B-values have apparently fallen for an influential group of economists or agents, the theory will hopefully be considered for revision.

2. An Example: Hall's Consumption Theory

To illustrate the difficulties inherent in testing theories in economics, it is useful to consider a deceptively simple theory - that suggested by Hall (1978) for consumption. Suppose that an individual obtains utility u(c) from an amount c of consumption. The results are based on a life cycle theory in which the person maximizes discounted utility

$$E_t \sum_{k=0}^{T-1} (1+\delta)^{-k} u(c_{t+k})$$

subject to the constraint

$$\sum_{k=0}^{T-1} (1+r)^{-k} \left[c_{t+k} - w_{t+k} \right] = A_t ,$$

where r is the (constant) interest rate, w_t is earnings at time t, δ is the discount rate, A_t is assets apart from human capital and E_t is the mathematical expectation conditional on all information available at time t. If $u'(c)$ is the marginal utility (i.e., du/dc) it follows from this construction and the permanent income hypothesis that

$$E_t\left[u'(c_{t+1})\right] = \lambda u'(c_t) \tag{1}$$

where $\lambda = (1+\delta)/(1+r)$. (2)

It follows that if $\lambda = 1$ and u(c) is a quadratic function then

$$c_{t+1} = c_t + e_{t+1} \tag{3}$$

where e_t is a martingale difference sequence, i.e., c_t is a random walk. However, if

$$u(c) = c^{(1+\theta)}$$

then

$$c_{t+1}^{\theta} = \lambda c_t^{\theta} + e_{t+1} . \tag{4}$$

So that now if $\lambda = 1$, c_t^{θ} is a random walk. Before looking at the data, this proposition may (or may not) sound convincing. So one could start with B = 0.5, say. [We assume that the theory was formulated by Hall without any specific data set in mind for which it might hold.]

The random walk implication of a form of the theory, i.e. (3), is generally the one tested in the literature, probably because Hall (1978) says that this simple relationship is a "close approximation to the

stochastic behavior of consumption under the life cycle-permanent income hypothesis". At first sight this would seem to be an easy hypothesis to substantiate, as time series techniques are available to test if the change in (real) consumption has the properties of a martingale difference sequence. For example, with a consumption series c_t one could fit the AR(p) model

$$c_t = \sum_{j=1}^{p} \alpha_j c_{t-j} + \varepsilon_t$$

and then test the null hypothesis:

$$H_0 : \alpha_1 = 1 ,$$

$$\alpha_j = 0 , \quad j = 2,\ldots,p ,$$

$$\rho_j = \mathrm{corr}(\varepsilon_t, \varepsilon_{t-j}) = 0 , \quad j = 1,\ldots,q ,$$

for some arbitrary large p and q. Thus the null hypothesis requires p+q particular parameter values to hold, which makes it rather complicated. In practice, values of p and q are chosen that are satisfactorily large, so that the test results can be considered convincing. A further complication is that the power of the test typically will decline as p and q increase. Acceptable values for p and q may vary across individuals. An alternative is to ask if the spectrum of Δc_t is flat, but in theory a spectrum is a continuous curve containing an uncountably infinite number of points, which is also difficult to test.

Using aggregate quarterly data for US real consumption (of services and non-durable goods) for the period 1947I - 1984III, Ermini (1988) compared three models for the change of consumption:

(i) a series with zero autocorrelations;

(ii) a moving average of order one, finding

$$\Delta c_t = \varepsilon_t + 0.239\ \varepsilon_{t-1}\ , \qquad (5)$$

where ε_t is as in (i);

(iii) and an ARMA(3,3) series suggested by considering all ARMA(p,q) models with $p + q \leq 6$ and maximizing likelihood.

He reports that a likelihood ratio test prefers the MA(1) or ARMA(3,3) models to the uncorrelated series, but cannot distinguish between the two temporally structured models. It would appear that the theory is rejected as the change in consumption is forecastable and so B may drop to 0.3, say. However, anyone familiar with time series analysis would recognize (5) from the result by Holbrook Working (1960) that if a flow series (such as consumption) is a random walk but is then temporally aggregated over a long period, the resulting series is ARIMA(0,1,1) with coefficient 0.25. It follows that (5), estimated on quarterly data, is consistent with the random walk theory but with the individual's decision period much less than a quarter. This is pointed out in Ermini (1988). As this looks promising, B could go up to 0.6. Does this mean that the theory is accepted by the data? In a sense, the theory is not rejected but neither are various other models. It is also pointed out in Ermini (1989) that if Δc_t is MA(1) with a negative coefficient, then after sufficient temporal aggregation, consumption becomes an IMA(1,1) process with MA coefficient 0.25. Thus many models are consistent with the data within the simple class considered and the "test" is not decisive. [This could be rephrased as saying that the theory is too vague.]

However, these tests just consider a property which is suggested by the theory of the single series c_t. The theory also proposes a much

more complicated property, that there exists no vector of series $\underline{x}_t$ such that the regression

$$\Delta c_t = \sum_{j=1}^{p} \underline{\beta}_j' \underline{x}_{t-j} + \varepsilon_t \qquad (6)$$

has any β component that is significantly different from zero. Such a hypothesis is virtually impossible to test - there are too many variables to consider for inclusion in $\underline{x}_t$ and too many parameters to check. At best, one can use a limited set of likely variables for $\underline{x}_t$, suggested by theory or by common sense, to be tested in small groups and with lag values (i.e. size of p) chosen to be modest or by a model selection criterion such as AIC or Schwarz's (1978) BIC [ignoring the important problem of interpretation of multiple tests]. If the data support the theory, with no significant explanatory variables found, then at most one can say that the theory has not been falsified; it cannot be claimed to be verified. Even with such an apparently simple theory one can only try to falsify the theory, with verification impossible, in agreement with a recent attitude in the philosophy of science; for a history and discussion see Redman (1991).

What does one conclude if a significant coefficient is found in (6) or if c_t has a temporal structure that is not consistent with a random walk after temporal aggregation? Then one may reject the strict random walk form of the model, but there are other versions which have not been tested. The utility function need not be quadratic and λ need not equal one. There is also the problem of cross-sectional aggregation. The theory is about the behavior of an individual but it is tested on aggregate consumption. Suppose that the j^{th} individual or family has consumption c_{jt} and also suppose that all individuals have the same

utility function $u(c) = c^{1+\theta}$, although this is extremely implausible. This is called the "constant elasticity of substitution form" of the utility function. From (4) it follows that the aggregate relationship is

$$\sum_{j=1}^{N} c_{j,t+1}^{\theta} = \lambda \sum_{j=1}^{N} c_{j,t}^{\theta} + \bar{e}_{t+1}$$

under the extra strong assumption that the λ value is the same for every individual. The value $c_{j,t}^{\theta}$ is not observed, in general, if $\theta \neq 1$. What is usually observed (or estimated from a sample) is aggregate consumption, $\sum_{j=1}^{N} c_{j,t}$, where N is the number of individuals or families, which has a value near 100 million in the United States. It is unclear how much correlation there is between $\left(\sum_{1}^{N} c_{j,t}\right)^{\phi}$ and $\left(\sum_{1}^{N} c_{j,t}^{\theta}\right)$ for any value of ϕ, particularly if the $c_{j,t}$ series are interrelated with each other. Thus, with cross-sectional aggregation and non-quadratic utility functions, aggregate data that is readily available to econometricians, cannot be used for testing the theory. It would be necessary for economic statisticians to find plentiful panel data so that the original form of the theory can be investigated. It also seems that the theory is not very precise, having an unspecified utility function, and so it is very difficult ever to falsify it. It is seen that an apparently simple theory, based on a rather unlikely set of basic axioms, is very difficult to evaluate. This is related to the "Duhem-Quine Thesis" discussed by Cross (1982).

In this example it is seen that the B value can fluctuate as new "information" about the correctness of the hypothesis is accumulated. This information may consist of results achieved by others or by

oneself. If the analysis that changes it is conducted personally a formal Bayesian procedure may be considered, with the initial B a prior odds ratio of some form, the data being written as a likelihood and the outcome being a new B written as a posterior odds ratio, as discussed by Zellner (1984). This assumes that the proposition being considered can be simply translated as a statistical hypothesis, such as $\underline{\beta} = 0$ where $\underline{\beta}$ represents coefficients on some finite set of variables. The example discussed here shows that such a translation is not always easy.

We feel that the problems encountered in "testing" Hall's consumption theory are not at all uncommon when testing economic theories, although these difficulties are not often discussed - but see Stigum (1990). A further example is the efficient market theory for speculative prices, which may be taken to say that returns (after adjustment for risk and transaction costs) are unforecastable using publicly available data. As this data set is potentially huge, it is obviously impossible to test all variables in it as possible explanatory variables for future adjusted return. What can be done is to accumulate tests using different variables and, possibly different data sets, i.e. various exchanges and periods, and to thus accumulate information about the correctness of the theory and so affect the degree of belief B.

An alternative approach is to try to construct a metric M which measures the deviation of the data from the theory and to base a test on M. For example, if one wants to test that a series x_t is a martingale difference, the Box-Pierce (1970) statistic (based on the sum of the squares of the first p estimated autocorrelations), or the maximum deviation of the estimated spectrum, at p frequencies, from the mean of this spectrum, would be possible choices for M. Similarly, if there are

k possible explanatory variables of x_{t+1}, one could choose p variables at random and use R^2 from the corresponding regression as M. In each case p has to be chosen to make the test both practical to implement but also sufficiently convincing that degrees of belief can be affected. In the second case if k is small compared to the sample size, all variables could be used and R^2 used to measure the goodness of fit, but if k is very large, a selection procedure is necessary to prevent over-fitting of the model associated with an optimistic R^2 value.

A final example of an important but difficult testing situation is to ask if a relationship is linear or non-linear (in mean). A null of linearity allows many models to be considered, with potentially very many parameters. The alternative of non-linearity requires consideration of a huge number of possible models and consequently an immense number of possible parameters. See Lee, White and Granger (1992) for recent work in this area.

It may be noted that sometimes a detailed economic theory leads to no testable implications. The question "what restrictions does economic theory (the assumption that rational agents maximize) place on asset prices?" leads to the answer "almost none" according to Rothschild (1990).

3. Problems with Pre-Testing

While hypothesis testing has a role to play in terms of testing economic theories, it is frequently used in the model building process to make choices between competing models based purely on the data. For specific examples, see the literature on general-to-specific modelling (Hendry, 1979, Gilbert, 1986, Pagan, 1987), cointegration (Engle and

Granger, 1991) and pretesting (Wallace, 1977, Judge and Bock, 1978 and Judge, 1984). In our view this is an incorrect use of hypothesis testing. Whenever a hypothesis test is used to ask the data to choose between two models, one model must be selected as a null hypothesis. In most instances, this is usually the more parsimonious model and typically a nested test is applied. Often it is difficult to distinguish between the two models because of data quality (multicollinearity, near-identification or the models being very similar such as in testing for integration). In such cases, the model chosen to be the null hypothesis is unfairly favored.

This point can be illustrated by reference to the pre-test literature which mainly concentrates on issues of estimator accuracy. Typical findings of empirical or simulation studies are that pre-testing strategies produce estimators with reasonable properties but the usual choice of significance level[1] such as 5% or 1% in the pre-test is far from optimal. For example, Fomby and Guilkey (1978) suggest that the Durbin-Watson test in the linear regression model should be applied at a significance level of about 50% rather than 5% if the aim is to re-estimate with AR(1) errors if the test rejects H_0. This suggestion is hardly surprising. Given a well-defined loss function of estimator accuracy, we no longer have a classical hypothesis testing problem in which the null hypothesis has its special role. Instead we have a model selection problem in which the relative importance of the null and

---

[1] A choice of significance level for a given hypothesis test is essentially a choice of power curve. If one has a higher degree of belief in the null hypothesis then one should be happy with a lower significance level and hence a lower power curve. Other than this preference ordering, there is typically no relationship between the degree of belief in either hypothesis and the choice of significance level.

alternative hypotheses are determined by the loss function.

When the model building process involves non-nested testing, the choice of null hypothesis is not obvious. Some advocate applying a non-nested test twice with each model having a turn as the null hypothesis. This does not always result in an unambiguous outcome. A further problem with non-nested tests is that they typically aim for a constant probability of committing a Type I error at all points in the null hypothesis parameter space. Because the models are non-nested, it is possible to have data generated from a null model which could not have possibly come from an alternative model. For example, in testing an AR(1) null hypothesis against an MA(1) alternative, observe that the first-order autocorrelation coefficient ρ_1 can take values in the range $-1 < \rho_1 < 1$ under H_0 but is restricted to $-0.5 \leq \rho_1 \leq 0.5$ for an MA(1) process. As King (1983) pointed out, a test which has constant size for all values of ρ_1 in the range $-1 < \rho_1 < 1$ is undesirable. A sensible test[2] would have size reducing to zero as $|\rho_1|$ increases past 0.5.

Almost always, model building involves a series of tests, often with little regard to controlling overall size. Two investigators working on the same data could easily end up with different models purely because they performed their tests in different orders or used different levels of significance.

The above arguments point to three deficiencies with formal hypothesis testing when used as a tool in model building. The first

---

2 For testing an AR(1) process against an MA(1) process, Burke, Godfrey and Tremayne (1990) and Franses (1991) have suggested procedures that satisfy this requirement.

concerns the manner in which the trade-off between Type I and Type II errors is resolved by controlling the probability of a Type I error to be a small value such as 5%. The second is the pre-occupation with the construction of tests whose probability of a Type I error is constant for all parameter values of the null hypothesis model. While this may be good practice for nested testing problems, it is questionable for non-nested problems. The most prominent non-nested test procedure is the Cox (1961, 1962) test, which can be viewed as the standard likelihood ratio statistic adjusted to have an asymptotic standard normal distribution under the null hypothesis. This results in a constant probability of a Type I error, asymptotically. It seems that this adjustment may be unnecessary and in fact harmful. The third deficiency is that formal tests involve pairwise comparisons of possible specifications.

4. Model Selection Criteria

It is our view that model building should be based on well-thought-out model selection procedures rather than a series of classical pairwise tests. The use of an information criterion based on minus the maximized log-likelihood function plus a penalty function for the number of parameters in the model is most worthy of consideration. This number is calculated for each model and the model with the smallest value is chosen. Examples include AIC and Schwarz's (1978) BIC. No one model is favored because it is chosen as a "null hypothesis". The order in which calculations are done does not affect the final results. Also, as Pötscher (1991) points out, minimizing such an information criterion amounts to testing each model against all other models by means of a standard likelihood ratio test and selecting that model which is accepted against all other models; the critical values are determined

by the penalty function. Observe that when nonnested models are being tested, the standard likelihood ratio statistic is used rather than Cox's adjusted likelihood ratio statistic. Judgment on which significance level to use is no longer needed although there is the issue of what penalty function is appropriate.

This approach has an advantage in dealing with another difficulty in testing an economic theory which is that the theory may only deal with a partial aspect of the data. For example, a theory may try to explain a single stylized fact, whilst ignoring other facts such as seasonal or trend components in the data. By selecting the best, or at least a good model, there should be few major features of the data that have not been modelled.

The situation considered is as follows:

(i) Suppose that there are a number of model <u>types</u>, M_1, M_2, ..., M_k, (for example, autoregressive, moving average with ARCH, bilinear) which are not necessarily nested. Each model in each type has a number of parameters, q, associated with it. Thus, the models in type M_j consist of $M_j(1)$, $M_j(2)$, ..., $M_j(Q)$. [In practice, there may be different types of parameters in each model, so that q is really a vector, but this complication is not considered.] If a particular theory is being considered, it may suggest one type of model even before looking at data. The models are chosen to relate to a theory that one is interested in testing. It will be assumed that the models are being constructed to test a theory rather than for forecasting or policy uses, for example.

(ii) There is available a variety of model selection criteria (henceforth criteria), S_1, S_2, ..., S_J. Each is assumed to be a function of the maximized log likelihood $L_j(q)$ of the model $M_j(q)$ and also of the number of parameters q. A specific form might be the information criterion

$$S_i(d) = -L_j(q) + q^d f_i(n) \qquad (7)$$

where d is some positive parameter and $f_i(n)$ is a specific function of n, the sample size. If several models are considered, the one with the smallest value of the criterion is preferred. As q increases, $L_j(q)$ is non-decreasing and the second term in (7) is the penalty for using more parameters. A criterion S_1 will be said to be "parsimonious" with respect to S_2 if it gives a higher penalty to the size of q. Thus, if the two criteria have the same d value, S_1 is more parsimonious than S_2 if $f_1(n) > f_2(n)$. Clearly this ranking may change as n changes. Well-known examples with d = 1 are AIC, for which f(n) = 2/n and Schwarz's BIC, for which f(n) = log(n)/n. Clearly for n > 8, BIC is the more parsimonious. The parameter d is introduced in (7) to widen the variety of criteria usually considered. If d > 1 there will be a tendency to choose more parsimonious models than if d = 1.

It is easy to see that if two models M_1, M_2 are such that $L_1 > L_2$ and $q_1 < q_2$ or $L_1 > L_2$ and $q_1 = q_2$ or $L_1 = L_2$ and $q_1 < q_2$ then all criteria of the form (7) will prefer M_1 to M_2. Many other forms of criteria than (7) can also be considered and a similar result will hold. Different choices for f_i will be appropriate depending upon whether models are nested or non-nested. Admissible choices for f_i are discussed by Sin and White (1992).

To implement the procedure, for a data set $\underline{X}_t$, t = 1,...,n, every model of type j is fitted up to parameter value Q and, for some particular criteria S_i, the best model chosen, $M_j(q_{io})$. Repeating this for each model type, the set of best models can be compared using S_i and the overall best model $M_{io}(q_o)$ chosen, with "o" denoting optimum.

When comparing models $M_1(q_1)$, $M_2(q_2)$, with the first preferred according to the criterion $S_i(d)$, then the difference in log likelihoods from (7) is

$$L_1(q_1) - L_2(q_2) > (q_1^d - q_2^d)f_i(n) . \qquad (8)$$

The LHS is the log of a likelihood ratio test statistic. Thus we are able to see the point made by Pötscher (1991) that minimizing (7) amounts to testing each model against all other models by means of a standard likelihood ratio test and selecting that model which is accepted against all others. The RHS of (8) shows how the critical values for these tests are determined by the penalty function.

One can ask how well the model selection criteria work asymptotically. Of the class of models considered, that is the union of all of the types of models, define the "best" model to be either

a) the true generating mechanism of the data (assuming this exists) corresponds to one of the models, or

b) it is the model, within the class considered, that is in a specific sense the closest to the generating mechanism, or, if two models are equally close, the more parsimonious model. The

distance measure used is analogous to the Kullback-Leibler criterion that is relevant for comparing distributions.

Nishi (1988) and Sin and White (1992) show that asymptotically, information criteria such as (7), with d = 1, consistently find the "best" model in the sense just defined provided

$$\lim_{n\to\infty} \frac{f(n)}{n} = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{f(n)}{\text{loglog } n} = +\infty .$$

It follows that AIC does not have good asymptotic properties but Schwarz's BIC does. (It is an open question whether this result continues to hold if d > 1.) Pötscher (1991) considers the asymptotic effects of using these types of model selection criteria on the estimation and parameter testing properties of the model chosen. If the criterion is such that the correct model is selected with probability approaching one, then there is no asymptotic effect of the model selection.

An alternative to information criteria as just discussed are "cross-validation" approaches to estimating the Kullbach-Leibler information or expected log-likelihood. These techniques give sample-based estimates of (7) that adjust for the biases contained in the sample estimate of $L_j(q)$. Because such techniques generally are asymptotically equivalent to criteria of the form (7), we shall not discuss them further here. However, the fact that cross-validation techniques can provide direct sample-based measures of bias in $L_j(q)$ makes them attractive as practical alternatives to (7).

An obvious question is how to decide which criterion to use. It is

clear that one cannot make a choice on a single data set as this would require the use of a super-criterion, but if this existed, it would be used directly as a model selection criterion rather than having to choose between criteria. The best criterion may be selected from a simulation study. If the data is generated from a model included in the set of models considered and with a finite q_0, a cost function can be constructed based on the distribution of the estimated q values from the criterion around the true q_0. A major purpose of a criterion is to limit the number of parameters used in a model for two reasons. The first because when estimating, parsimony is an advantage - better estimates can be expected for fewer parameters - and because the dangers of model over-fitting or data mining will hopefully be reduced. Ideally, if one has an objective in mind, such as getting the best forecasting model, a good criterion will predict from in-sample what is the best model for this objective.

A standard criterion is that suggested by Rissanen (1987) based on considerations of model complexity, leading to essentially the familiar BIC criterion. This criterion can be used with nonlinear and ARCH models, for example.

What, then, should be the respective roles of model selection and diagnostic testing? Should one first select a model and then perform diagnostic tests on the selected model, or should one perform diagnostic tests on all candidate models, and then select a model from those that pass the diagnostic tests, using an appropriate criterion? Because computation of the model selection criterion is usually much simpler than computation of the diagnostic test statistics, the first approach has the advantage of computational simplicity. Further, if the correct

model is in the candidate set, it will be selected with probability one asymptotically by a well-behaved criterion. Also, because diagnostic tests may often be interpreted as tests of particular restriction on a given model, the asymptotic size of such tests will be correct when one does model selection first.

There nevertheless appears to be some appeal to doing diagnostic testing first. The source of this appeal seems to us to stem from the insight that diagnostic testing may give into alternative models not formally included in the original candidate set. The candidate set may be expanded as a result of this insight. But there is no justification for then restricting the model selection to the subset that pass the diagnostic test. Certainly there is no justification on grounds of computational burden, as performing the diagnostic tests is already more burdensome. But in addition, unless the diagnostic tests have asymptotic size zero, a correctly specified model may be wrongly rejected by a diagnostic test and thereby excluded from further consideration when selection is limited to models that pass the diagnostic tests.

We therefore prefer (ideally) to do model selection first. In a perfect world of unlimited data and complete foresight about possible forms of misspecification, a consistent strategy is to consider a wider group of initial models, including the original ones plus those including the terms which the diagnostic tests would look for, such as missing variables, ARCH heteroskedasticity, lagged residuals and so forth. The criterion is then applied to this wider group of models and the overall best model determined.

A related question is whether it is useful to start with a large

group of model types. It is obviously more expensive to analyse many models but, in a perfect world, it makes it more likely that the good approximation to the true generating mechanism will be found. It is also possible that the criterion will have difficulty in deciding between a few models. This may suggest new combined models which further increases the number of models under consideration.

Unfortunately we do not live in a perfect world. We have limited data which leads to the following concern. If a large number of models are considered, there is a possible problem with "data mining", that is, a high probability of accidentally finding a model which happens to fit the particular data set very well. Clearly there is a trade-off between the accuracy of our model selection procedure and the number of models considered. As the pool of models increases, the chances of selecting the correct one declines. An important practical question is how should we position ourselves on this trade-off. The following three alternative strategies may help in this regard:

i) If only model classes that are not nested are considered, let the number of parameters in class M_j be limited to be no more than Q_j. Let $\bar{Q} = \Sigma Q_j$ denote the total number of parameters considered overall. As the number of model types considered increases, $\bar{Q}$ may become unacceptably large. One may decide to limit $\bar{Q}$ and have some rule which distributes the possible number of parameters between the models.

ii) A second alternative is to constrain the set of models under consideration to only those that are distinct possibilities and after selection, test for outside chances. This testing should perhaps be applied to a model that encompasses all models in the model

selection procedure. This would reduce problems caused by the incorrect model being selected. Note that such diagnostic tests will favor the encompassing model because of the choice of null hypothesis.

iii) A third alternative is to adopt some rule such that the parsimony parameter d in (7) is made an increasing function of $\bar{Q}$, so that as more models, and thus parameters, are considered, the penalty for having more parameters increases. Consideration is required of this possibility and what function $d(\bar{Q})$ is helpful.

Once the best model is found, there may still be a need to test it if only because we can never have perfect foresight about all possible models. We favor the use of a "portmanteau" test rather than several specific tests. It is worth bearing in mind that such a test might reject for all sorts of reasons. It may be best to interpret such a rejection only as indicating that the set of models being selected from needs augmenting.

Obviously, there are no easy answers. Considerable judgment is needed and there is much room for further research. So far we have assumed that the sample size n is fixed. However, in practice further data accumulate through time, so that a sequential model selection procedure is required. This is clearly another rich area for further research.

The criteria considered here are based completely on statistical properties of the data. Any particular researcher may want to add economic considerations to the criteria, such as an expected sign on a coefficient or a belief in homogeneity. This is certainly a real

possibility but asymptotically at best an improvement in efficiency will be gained; at worst the economic beliefs could be wrong, and the model selected will be deflected away from the "best" one, where "best" could be measured in terms of the model's ultimate purpose, such as providing relatively good forecasts.

If model selection is to be based on more than one criterion, this should be explicitly recognised. Selection should then proceed according to a coherent set of requirement criteria. This approach is discussed in the next section.

A problem that we have not faced is that models are built for a variety of purposes. Ideally, the best model according to a criterion should be best for all purposes, provided all appropriate variables have been considered in its construction. There would be no point in asking if a model is good for policy purposes if the policy variables were not included in the modelling process. The model may well not be designed to change a B-value, but an individual can use its results for that purpose. How this is done is up to the individual and may not be a formal process. However, if a Bayesian approach is used, with a prior odds ratio, and a likelihood leading to a posterior odds ratio, the evolution of the B-value can occur formally. Zellner (1978) points out the link between this procedure and a particular criterion, the AIC, but points out that the linkage is by no means exact even in a Gaussian linear regression context.

5. Model Selection by Testing for Requirements

A researcher may be able to provide a list of required properties for a model and an econometrician can then suggest tests of whether or

not any particular model meets these requirements. Such a set of tests can be considered as a model selection procedure, and this has been discussed by White (1990). One set of such requirements are those for a model to be "congruent with the evidence" according to Hendry and Richard (1982) and Hendry (1987). A model is said to be congruent if and only if:

a) it encompasses all rival models,

b) its error process is a "mean innovation process",

c) its "parameters of interest" are constant,

d) it is data admissible, and

e) its current conditioning variables are weakly exogenous for the parameters of interest.

Denote these requirements by C_0. Of course, not all researchers would agree that C_0 are the necessary requirements. White (1990) proposes various sets of requirements, with C_1 being (a) and (b) of C_0, C_2 replaces encompassing by "correct model specification", so that C_2 includes C_1. C_3 replaces correct specification by an information matrix equality and C_4 is the union of C_1 and C_3. Each requirement is associated with an m-test. White also provides conditions such that asymptotically the procedure chooses all models that satisfy the requirement. In a given application, one may find one model that satisfies the requirements, or many models or no model. If there are several models, then further preferred conditions can be added, such as parsimony. If no model is satisfactory this implies that a wider class of models should be investigated.

An obvious problem is that one researcher has to justify a

particular set of requirements as being reasonable to other researchers. Nevertheless, the test should be helpful for affecting degrees of beliefs. Experience is needed to see how this approach performs compared to other model selection methods.

The two methods of selection discussed in this and in the previous sections are different but clearly can be related. The approach in this section can be viewed as a complement to the model selection method of the previous section; one could use a selection criterion to select a model as in section 4, and the selected model can be subjected to the requirements described in this section. If it passes, it is accepted; if not, one might search over "near best" models according to the criteria until one is found that meets the requirements. The best way to conduct the search is unclear and whether or not some relaxation of the requirements is considered worthwhile to achieve a more parsimonious model is an individual decision. It is clear that further work is also required to make these ideas practical and capable of implementation.

6. Conclusions

We have pointed out several difficulties with testing economic theories, particularly that the theories may be vague, may relate to a decision interval that is different from the observation period and may need construction of a "metric" to convert a complicated testing situation to an easier one. The metric should also be designed to communicate empirical results that can change degrees of beliefs and consequently affect decisions.

A key component of econometric practice is the building of econometric models. Frequently researchers are forced to use the data to

make decisions about the particular form of a model. We argue that it is better to use well-thought-out model selection procedures rather than formal hypothesis testing in such situations. This is because formal testing favors the model chosen to be the null hypothesis, the choice of significance level is typically arbitrary and different researchers working with the same data could easily end up with different models purely because they performed their tests in different orders or used different levels of significance. In contrast, the use of an information criterion such as (7) means that no model is favored because it has been chosen as a "null hypothesis", judgment on the level of significance to be used is not required and the order of computation is irrelevant. There are, however, some unsolved problems such as the choice of penalty function in (7), how to guard against data-mining and how to ensure that an important model specification has not been overlooked. We also considered model selection based on testing for desirable properties of models. The two approaches can be combined to yield a comprehensive model selection strategy. Further research is needed to determine how these procedures might best be applied in practice.

References

Box, G.E.P. and D.A. Pierce, 1970, Distribution of residual auto-correlations in autoregressive-integrated-moving average time series models, Journal of the American Statistical Association 65, 1509-1526.

Burke, S.P., L.G. Godfrey and A.R. Tremayne, 1990, Testing AR(1) against MA(1) disturbances in the linear regression model: An alternative approach, Review of Economic Studies 57, 135-145.

Cross, R., 1982, The Duhem-Quine thesis, Lakatos and the appraisal of theories in macroeconomics, Economic Journal 92, 320-340.

Cox, D.R., 1961, Tests of separate families of hypothesis, in: Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, Vol.1 (University of California Press, Berkeley, CA) 105-123.

Cox, D.R., 1962, Further results on tests of separate families of hypotheses, Journal of the Royal Statistical Society B 24, 406-424.

Engle, R.F. and C.W.J. Granger, 1991, Long-run economic relationships: Readings in cointegration (Oxford University Press, Oxford).

Ermini, L., 1988, Temporal aggregation and Hall's model of consumption behavior, Applied Economics 20, 1317-1320.

Ermini, L., 1989, Some new evidence on the timing of consumption decisions and on their generating process, Review of Economics and Statistics 71, 643-650.

Fomby, T.B. and D.K. Guilkey, 1978, On choosing the optimal level of significance for the Durbin-Watson test and the Bayesian alternative, Journal of Econometrics 8, 203-213.

Franses, P.H., 1991, Model selection and seasonality in time series. Unpublished doctoral dissertation (Erasmus University, Rotterdam).

Gärdenfors, P., 1988, Knowledge in flux (MIT Press, Cambridge, MA).

Gilbert, C.L., 1986, Professor Hendry's econometric methodology, Oxford Bulletin of Economics and Statistics 48, 283-307.

Hall, R., 1978, Stochastic implications of the life-cycle permanent income hypothesis: Theory and evidence, Journal of Political Economy 86, 971-987.

Hendry, D.F., 1979, Predictive failure and econometric modelling in macroeconomics: The transactions demand for money, in P. Ormerod, ed., Modelling the Economy (Heinemann, London).

Hendry, D.F., 1987, Econometric methodology: A personal perspective, in T. Bewley, ed., Advances in Econometrics, Fifth World Congress, Vol. 2, 29-48 (Cambridge University Press, Cambridge).

Hendry, D.F. and J.-F. Richard, 1982, On the formulation of empirical models in dynamic econometrics, Journal of Econometrics 20, 3-33.

Judge, G.G., 1984, Pre-test and Stein-rule estimators: Some new results, Special Issue, Journal of Econometrics 25, 1-239 (North-Holland, Amsterdam).

Judge, G.G. and M.E. Bock, 1978, The statistical implications of pre-test and Stein-rule estimators in econometrics (North-Holland, Amsterdam).

Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lütkepohl and T.-C. Lee, 1985, The theory and practice of econometrics, 2nd ed. (Wiley, New York, NY).

King, M.L., 1983, Testing for autoregressive against moving average errors in the linear regression model, Journal of Econometrics 21, 35-51.

Lee, T.-H., H. White, and C. Granger, 1992, Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests, Journal of Econometrics, forthcoming.

Nishi, R., 1988, Maximum likelihood principle and model selection when the true model is unspecified, Journal of Multivariate Analysis 27, 392-403.

Pagan, A.R., 1987, Three econometric methodologies: A critical appraisal, Journal of Economic Surveys 1, 3-24.

Pötscher, B.M., 1991, Effects of model selection on inference, Econometric Theory 7, 163-185.

Redman, D.A., 1991, Economics and the philosophy of science (Oxford University Press, Oxford).

Rissanen, J., 1987, Stochastic complexity and the MDL principle, Econometric Reviews 6, 85-102.

Rothschild, M., 1990, Economic theory teaches us that economic theory teaches us nothing: The case of asset prices, Working Paper, Department of Economics, University of California, San Diego.

Schwarz, G., 1978, Estimating the dimension of a model, Annals of Statistics 6, 461-4364.

Sin, C.-Y. and H. White, 1992, Information criteria for selecting parametric models: A misspecification analysis, Working Paper, Department of Economics, University of California, San Diego.

Stigum, B., 1990, Towards a formal science of economics (MIT Press, Cambridge, MA).

Wallace, T.D., 1977, Pre-test estimation in regression: A survey, American Journal of Agricultural Economics 59, 431-443.

White, H., 1990, A consistent model selection procedure based on m-testing, in C.W.J. Granger, ed., Modelling economic series: Readings in econometric methodology (Oxford University Press, Oxford).

Working, H., 1960, Note on the correlation of first differences of averages in a random chain, Econometrica 28, 916-918.

Zellner, A., 1978, Jeffreys-Bayes posterior odds ratio and the Akaike information criterion for discriminating between models, Economics Letters 1, 337-341.

Zellner, A., 1984, Basic issues in econometrics (The University of Chicago Press, Chicago).