Annals of Economic and Social Measurement, 1/3, 1972

CRITERIA FOR EVALUATION OF ECONOMETRIC MODELS*

BY PHOEBUS J. DHRYMES, E. PHILIP HOWREY, SAUL H. HYMANS, JAN KMENTA, EDWARD E. LEAMER, RICHARD E. QUANDT, JAMES B. RAMSEY, HAROLD T. SHAPIRO AND VICTOR ZARNOWITZ

This multi-authored article develops a framework for systematically evaluating large scale econometric models. Reasonably self-contained aspects of model evaluation include parametric evaluation prior to the "release" of the model (model selection, parameter estimation, and pseudo-forecasts and structural stability tests) and evaluation after "release" of the model. Many operational procedures for parametric evaluation are noted; alternative, ad hoc procedures are necessary in some cases, given the present state of the art. Non-parametric "validation" procedures are then outlined. These include single-variable measures, tracking measures, error decomposition, and cyclical and dynamic properties. A statistical appendix sketches some of the theoretical results used in the paper.

I. INTRODUCTION

For purposes of this paper an econometric model is considered to be an analytical representation of one or more statements about economic behavior, which representation relies upon statistical implementation for the purposes of hypothesis testing, parameter estimation, or use in prediction or simulation circumstances. A model in this sense may be anything from a single linear equation to a complicated set of simultaneous, non-linear equations. The term "model evaluation" is here used to encompass a broad set of tests to which a model can and should be subjected at many different stages during the process of construction and subsequent use.

During the past decade econometric models have come in for increasingly widespread use by government (for policy analysis and forecasting), by industry (largely as a forecasting tool), and by universities (for instructional use and a wide variety of research purposes). Despite the growing importance of such models in various decision-making situations, the process of systematic model evaluation has—with some noteworthy exceptions—lagged seriously behind the process of multi-model proliferation. Within the past few years, however, a handful of significant attempts have been made—with respect to large scale econometric models—to conduct serious cross-model comparisons. Building on a series of pioneering efforts by Carl Christ [10], Irma Adelman [1], Henri Theil [50], and others, the studies of Zarnowitz, Boschan and Moore [57], and Evans, Haitovsky and Treyz [21] are examples of current research work in this area. Particular model builders, of course, have also subjected their own models to careful "audits" both on sample and post-sample data. At the level of subsector and single equation

* This paper is a joint effort of the authors listed and was undertaken as a project of the Seminar on Criteria for the Evaluation of Econometric Models (S. H. Hymans and H. T. Shapiro, co-chairmen) of the Conference on Econometrics and Mathematical Economics, sponsored by the National Bureau of Economic Research and the National Science Foundation. An earlier version of this paper was presented at the Brookings Model Conference, Washington, D.C. February 11–12, 1972. The authors are grateful to Professors C. Christ, R. A. Gordon, L. R. Klein and L. D. Taylor for their continuing help during many stages of the writing of this paper.

models recent work by Bischoff [7], Hymans [37], [38], and Jorgenson, Hunter, and Nadiri [39] may be cited as examples of cross-model evaluations. What stands out most clearly from all these evaluation exercises is that, aside from the simplest single-equation cases we suffer the lack of a clear and accepted analytical basis for the selection of proper criteria for model evaluation. This is true with respect to the criteria by which a single model should be evaluated and holds a-fortiori in the case of cross-model evaluations. This state of affairs has been the motivation for several recent papers, [21] [36], and is the *raison-d'etre* for the NBER-NSF sponsored seminar which has led to this paper.

In the next section of this paper, we shall outline a framework which decomposes the evaluation set into fairly natural subsets, and thus permits the orderly discussion of reasonably self-contained aspects of model evaluation. These are discussed in turn in succeeding sections of the paper.

It has been our aim to suggest operational procedures for evaluation whenever possible, and to compare alternative procedures whenever our knowledge permits. To this end, a number of statistical derivations and proofs have been relegated to an appendix in order that the flow of discussion in the body of the paper may be more easily digested. While we have succeeded in arriving at some useful "recipes" for particular evaluation circumstances, there are still gaping holes in our knowledge. For some evaluation problems we simply have nothing to suggest for a "best practice" procedure, and we have had to be content with a brief and general enumeration of the alternative, often ad hoc, procedures which are in current use or under current study. Most of what we have to say is in direct reference to time series econometric models, but much of what follows applies to cross-section models with perhaps minor rephrasing.

II. Aspects of Model Evaluation

What we (as builders, users or judges of models) choose to do in the process of evaluating an econometric model is heavily dependent on what we have chosen to axiomatize. At an early stage in the life of a model we may regard its functional form as "up for grabs," as something yet to be determined. At a later stage, after the model has already been "certified" with respect to functional form, we may choose to test hypotheses about parameter values within the confines of the functional form already settled upon or axiomated.[1] Alternatively, we may take the approach which one of the authors has called "Sherlock Holmes inference," a process of data analysis in which Sherlock the econometrician weaves together all the bits of evidence into a plausible story. In this view, it is taken as axiomatic that the process being modeled is far too complicated and the data available far too weak to be able to specify and implement a structurally and behaviorally sound representation. Such notions as parametric hypothesis testing, best linear unbiased estimators, and the like are then wholly irrelevant, if not dangerously misleading. Nearly all that remains is a series of evaluative measurements specified in the light of the particular uses to which it is desired to put the model. At best,

[1] This is the basic set-up in the classical statistical procedures based on the work of Fisher, Neyman, Pearson and others.

the model can tentatively be certified as a reasonable tool for specific uses until it errs seriously and is found to have a fatal uncorrectable flaw, or until it is replaced by a better "untrue" model.[2] Sherlock Holmes' inference leads naturally to evaluation procedures heavily geared to the specific potential uses of the model, that is, to the calculation of performance statistics with generally unknown probability characteristics (and a strong presumption of stochastic dependence which even eliminates the possibility of conducting distribution-free statistical tests). Procedures of this kind have also had to be employed in the evaluation of models originally constructed under a strong stochastic axiomatization. This has been necessitated, for example, by the fact that we have not yet succeeded in identifying a uniquely proper way to evaluate a matrix of dynamically generated time series forecasts of all the endogenous variables in a macroeconometric model. Nor do we fully understand the stochastic properties of such a matrix,[3] a necessary first step in the generation of any statistically valid inference procedure.

To break this formidable evaluation process down into a series of manageable problems, we propose first a binary split into categories which we shall refer to as *parametric* and *non-parametric* evaluation. An evaluation procedure is said to be parametric if it relies on a formal statistical test based on the stochastic specification assumed to apply to the econometric model. Non-parametric evaluation is concerned with specialized and descriptive procedures such as those mentioned in the previous paragraph. Such procedures are not derived from the stochastic assumptions of the model, and they rarely depend on formal tests of significance. It is our view that non-parametric evaluation can be important and valid under many different axiomatizations, and we shall discuss this matter more fully in section V below. Our discussion of parametric evaluation will proceed according to the following outline:

Parametric Evaluation

1. Prior to "release" of the model
   (a) Model selection
   (b) Hypothesis tests and parameter estimation
   (c) Pseudo-forecasts and structural stability tests
2. Subsequent to "release" of the model
   (a) Availability of a small post-sample data set: predictive testing, pooling of sample and post-sample data.
   (b) Availability of a large post-sample data set.

III. Parametric Evaluation: Prior to Model Release

In this section we discuss a number of aspects of evaluation which are considered as taking place during the process of model construction and continuing through to the first time the model builder actually "puts his money" on the results generated by the model.

[2] This is not the first time that economists have heard such arguments.

[3] Except possibly for some very simple cases.

(*a*) *Model Selection*

The term "model selection" here refers to the problem of choosing between alternative functional representations of an economic relation. The classical statistical procedures which most economics graduate students are required to internalize depend very heavily on a specification axiom. These procedures yield likelihood ratio tests, minimum variance estimators and predictors, and other such munificent benefits all under the assumption that $Y = X\beta + \varepsilon$ and its familiar accompanying probability statements accurately reflect the true state of affairs. As practicing economists we are well aware that a logically prior problem exists. Economic theory gives preciously few clues as to the functional forms appropriate to the specification of economic relationships, and the presence of random error terms in stochastically specified equations adds an additional element of functional ambiguity. In certain cases, known in the literature as situations of "nested hypotheses," classical statistical techniques provide sound discriminating procedures limited in power "only" by the quantity and richness of the sample evidence. Classical techniques are woefully silent in the case of non-nested hypotheses, or disparate families of hypotheses, but research is being done in this area and there is also the possibility of a useful Bayesian approach to such problems.

Techniques for the handling of pairs of nested hypotheses in a linear econometric model are by now second nature in the profession. They are well-documented in our standard textbooks and there is little to be gained by any review here. Let us turn directly to the less understood problem of selecting among alternative model specifications which cannot be represented in the framework of nested hypotheses.

Ramsey has made an interesting beginning in the analysis of non-nested linear models [46]. Suppose we consider two alternative specifications of a linear-in-the-parameters model to explain the dependent variable Y:

$$H_0: E[Y\ X] = X\beta$$

$$H_A: E[Y|Z] = Z\gamma,$$

where $Z = g(X)$, and the function g represents a non-stochastic, non-linear transformation. H_0 is the maintained hypothesis, while H_A is the alternative hypothesis. If H_A is true, then the regression calculated under H_0 has used an incorrect functional form for the regressors. Letting u denote the vector of residuals from the least squares regression of Y on X, it is easily shown [46; pp. 353–354] that

$$E[u|X, H_0] = 0,$$

and

$$E[u|X, H_A] = MZ\gamma$$

where $M = [I - X(X'X)^{-1}X']$. Using $Z = g(X)$, the second relation can be written as

$$E[u|X, H_A] = Mg(X)\gamma = h(X)\gamma,$$

where $h(X) = Mg(X)$.

Ramsey reasons[4] that

(i) $h(X)$ can be approximated as a multivariate power series in the X variables,

(ii) The predicted values of Y from the regression of Y on X, say $\hat{Y}$, are linear functions of X, and therefore,

(iii) It should be possible to approximate $h(X)$ by a power series in $\hat{Y}$. It is therefore approximately true that

$$E[u|X, H_A] \cong \sum_{j=2}^{J} \hat{Y}^j \alpha_j,$$

where

(i) the number J represents a Jth degree power series approximation to $h(X)$,

(ii) the index j begins with $j = 2$ since the least squares residuals are uncorrelated with $\hat{Y}$, and

(iii) $\hat{Y}^j$ refers to the jth power of $\hat{Y}$, element by element.

Under H_0, all the α_j should be zero; under H_A at least some of the α_j should be non-zero. Ramsey's idea, then, is to regress the residuals on powers of $\hat{Y}$ and test the hypothesis that the vector $\alpha = (\alpha_2, \alpha_3, \ldots, \alpha_J)'$ is null. Rejecting the null hypothesis on α is equivalent to rejecting H_0 in favor of some hypothesis of the form H_A.[5] In point of fact, Ramsey carries out the above test, not on the least squares residuals, but on Theil's BLUS residuals [51; chapter 5]. The idea is the same, but the BLUS residuals yield more convenient stochastic properties which permit the test on the vector α to be carried out by the usual multiple regression F-test, provided one begins with the assumption of (conditional) normality of the vector Y.[6]

An alternative approach to the problem, one not limited to the linear model framework and not requiring any condition analogous to the $Z = g(X)$ requirement in the Ramsey approach, may be formulated as follows. Let two alternative specifications of an economic relation be represented by the hypotheses H_f and H_g. According to H_f the random variable Y has probability density function (p.d.f.) $f(y; \alpha)$, with the parameter α specified to be an element of the space Ω_α. According to H_g, Y has p.d.f. $g(y; \beta)$ with $\beta \in \Omega_\beta$, and furthermore

$$\Omega_\alpha \cap \Omega_\beta \neq \Omega_\alpha,$$
$$\Omega_\alpha \cap \Omega_\beta \neq \Omega_\beta.$$

In such a case the usual (variants of) likelihood ratio tests are not available and the asymptotic chi-square test on $-2 \ln \lambda$ (where λ is the likelihood ratio) cannot

[4] The reader is referred to the Ramsey paper [46] for a more rigorous discussion.

[5] Note that the test depends only on the alternative hypothesis that the X variables should have been transformed via *some* $g(X)$ before running the regression. The function g is not used specifically in carrying out the test. The test is therefore quite general, but probably sacrifices power relative to a test which might have been constructed for a specific alternative such as $Z_i = \ln X_i$.

[6] In [46] Ramsey reports the results of several applications of his test procedure. An entirely similar procedure can be used to obtain tests for heteroskedasticity, omitted variables, and simultaneity, as Ramsey indicates, but such tests do not necessarily pinpoint the cause of rejection of the maintained hypothesis.

be performed. Problems of this type have been studied by D. R. Cox [14] [15] who has suggested various procedures—within the framework of classical statistics—for testing H_f against H_g.

One possibility is to transform the problem into a more familiar framework by introducing a new parameter γ. The probability density function of the random variable can then be written as

$$h(y;\alpha,\beta) = k[f(y;\alpha)]^{\gamma}[g(y;\beta)]^{1-\gamma},$$

where the factor of proportionality required for h to be a p.d.f. is given by

$$\frac{1}{k} = \int_{-\infty}^{\infty} [f(y,\alpha)]^{\gamma}[g(y,\beta)]^{1-\gamma}\,dy.$$

Employing $h(y;\alpha,\beta)$ one can, at least in principle, obtain maximum likelihood estimators for α, β and γ. Because of the presence of the factor k, the maximization of the likelihood function may pose considerable numerical problems. It appears possible to use the asymptotic theory of likelihood ratio tests for testing hypotheses about γ. Clearly, confirmation that γ is (close to) zero or unity supports one hypothesis and tends to discredit the other; intermediate values of γ are ambiguous and awkward in economics since the two hypotheses may be incompatible. Perhaps such an outcome suggests the interpretation that both hypotheses are suspect.[7]

Cox's main procedure is based on the (generalized) likelihood ratio

$$e^{l_{fg}} = \frac{\sup\limits_{\alpha\in\Omega_\alpha} L_f^*(\alpha)}{\sup\limits_{\beta\in\Omega_\beta} L_g^*(\beta)}$$

where $L_f^*(\alpha)$ and $L_g^*(\beta)$ are the sample likelihoods under H_f and H_g respectively. Since it is not true in the present case that $\Omega_\alpha \subset \Omega_\beta$, it is not true in general that $l_{fg} \leq 0$; hence standard procedures cannot be applied. Let $\hat{\alpha}$ and $\hat{\beta}$ be the maximum likelihood estimators under H_f and H_g respectively. The natural logarithm of the generalized likelihood ratio is

$$\begin{aligned} l_{fg} &= \ln L_f^*(\hat{\alpha}) - \ln L_g^*(\hat{\beta}) \\ &= L_f(\hat{\alpha}) - L_g(\hat{\beta}) \\ &= \{L_f(\alpha) - L_g(\beta_\alpha)\} + \{L_f(\hat{\alpha}) - L_f(\alpha)\} - \{L_g(\hat{\beta}) - L_g(\beta_\alpha)\} \end{aligned}$$

where

$$\beta_\alpha = \operatorname{plim} \hat{\beta},$$

the probability limit taken on the assertion that H_f is true. That a large value for l_{fg} constitutes evidence against H_g may be seen as follows. Under H_f and the usual regularity conditions,

$$\operatorname{plim}\,[L_f(\hat{\alpha}) - L_f(\alpha)] = \operatorname{plim}\,[L_g(\hat{\beta}) - L_g(\beta_\alpha)] = 0,$$

[7] Recent work by Atkinson [5], elaborates the results given by Cox. Moreover, it shows that in instances where multiple hypotheses (exceeding two) are employed, or when the exponential combination of the distributions involves two parameters, γ_1, γ_2 (instead of γ, $1-\gamma$) it may not be possible to identify the "mixing" parameters.

while

$$\text{plim}\,[L_f(\alpha) - L_g(\beta_\alpha)] > 0,^{8}$$

and therefore a "large" l_{fg} renders evidence against H_g.

The test statistic considered by Cox is a variant of l_{fg}, namely

$$\begin{aligned} S_f &= l_{fg} - E_{\hat{\alpha}}\{L_f(\hat{\alpha}) - L_g(\hat{\beta})\} \\ &= \{L_f(\hat{\alpha}) - L_g(\hat{\beta})\} - E_{\hat{\alpha}}\{L_f(\hat{\alpha}) - L_g(\hat{\beta})\}, \end{aligned}$$

where $E_{\hat{\alpha}}$ denotes the expectation operator conditional on the hypothesis H_f.

It is shown by Cox that S_f is asymptotically normally distributed and its variance is obtained. Clearly the test is not symmetric and the roles of H_f and H_g can be interchanged. The results of the test on S_f may indicate consistency with H_f, departure from H_f in the direction of H_g or departure away from H_g. If the test is performed on both S_f and S_g (obtained by interchanging the roles of H_f and H_g), there are nine possible outcomes and care must be taken to employ the correct qualitative interpretation. In appendix section A.1 we give an example of an application of this procedure. Unfortunately, the test cannot be performed routinely since, as we show in the appendix, the form of the test statistic depends crucially on the nature of the hypotheses to be tested and can easily involve nuisance parameters. Further, carrying out the test requires computations of substantial analytical difficulty.

Finally, we turn to a Bayesian approach to the problem of model selection. While the classical approach of Cox uses the generalized likelihood ratio

$$e^{l_{fg}} = \frac{\sup\limits_{\alpha\in\Omega_\alpha} L_f^*(\alpha)}{\sup\limits_{\beta\in\Omega_\beta} L_g^*(\beta)}$$

as a measure of whether the data generally favor hypothesis f relative to hypothesis g, the Bayesian approach considers, instead, a weighted likelihood ratio of the form

$$R = \int_\alpha L_f^*(y;\alpha)W(\alpha, f)\,d\alpha \bigg/ \int_\beta L_g^*(y;\beta)W(\beta, g)\,d\beta,$$

where $W(\alpha, f)$ and $W(\beta, g)$ are "appropriately" defined weights relating to the parameters (α, β) and hypotheses (H_f, H_g) under consideration. It is perhaps simplest to illustrate the meaning of such weights in the likelihood function in the following way.

Let $\tilde{\omega}_f$ and $\tilde{\omega}_g (= 1 - \tilde{\omega}_f)$ represent the model builder's "prior probabilities" attaching to (initial degrees of belief in) H_f and H_g respectively. Let $p_f(\alpha)$ be the prior density on α, given that H_f is true; similarly let $p_g(\beta)$ be the prior density on β given that H_g is true. Let $w_f(\alpha)$ be the "cost" of rejecting H_f when true and $w_g(\beta)$ the "cost" of rejecting H_g when it (H_g) is true. The (expected) cost of rejecting H_f, on the basis of information y, when in fact H_f is true, is

$$\tilde{\omega}_f \int_\alpha L_f^*(y;\alpha)p_f(\alpha)w_f(\alpha)\,d\alpha.$$

[8] Recall that the probability limits are taken conditional on H_f.

Similarly the (expected) cost of rejecting H_g when it is, in fact, true is

$$\tilde{\omega}_g \int_\beta L_g^*(y;\beta)p_g(\beta)w_g(\beta)\,d\beta.$$

In this context the weight $W(\alpha, f)$ is given by $\tilde{\omega}_f p_f(\alpha)w_f(\alpha)$, and similarly for $W(\beta, g)$. The usual rule derived from minimizing expected loss is:

$$\text{Accept } H_f \text{ if } \tilde{\omega}_f \int_\alpha L_f^*(y;\alpha)p_f(\alpha)w_f(\alpha)\,d\alpha \geq \tilde{\omega}_g \int_\beta L_g^*(y;\beta)p_g(\beta)w_g(\beta)\,d\beta,$$

otherwise reject.

Now if $w_f(\alpha) = w_g(\beta) = c$, a constant independent of α and β, then the rule reduces to:

Accept H_f (on the basis of information y) if:

$$\frac{\tilde{\omega}_f \int_\alpha L_f^*(y;\alpha)p_f(\alpha)\,d\alpha}{\tilde{\omega}_g \int_\beta L_g^*(y;\beta)p_g(\beta)\,d\beta} \geq 1.$$

The left-hand quantity, of course, is the usual definition of posterior odds.

Current activity in this area of Bayesian research, e.g., Geisel [24], Zellner [58], Leamer [41], Dickey [19], is aimed at exploring the implications of alternative weighting functions (prior densities). There are several important substantive implications of the Bayesian literature on this topic, including (a) Minor differences in R^2's among the competing models allow considerable discriminatory power depending on the degrees-of-freedom, (b) An appropriate criterion statistic for choice among models is (roughly) an average of the sample R^2 and an "*a priori*" R^2 computed using *a priori* likely values of the parameters. (That is, it does not matter if an R^2 is high if it implies absurd values of the parameters.)

Economic model builders rarely view themselves in the role of decision maker. Generally, the model builder concentrates on the estimation of many parameters and the pure testing of relatively few hypotheses.[9] But here, in the crucial area of model selection, is a circumstance clearly defined as a decision problem, whether to select H_f or H_g as the axiom on which to proceed in subsequent analysis.[10] And this clearly represents an area for which Bayesian analysis

[9] In current practice, most of the pure statistical tests carried out by model builders involve either the omitted variables specification analysis of Theil [50], or the test for structural change discussed by Chow [9], or various tests for the presence of autocorrelation. These major exceptions aside, it seems clear that far more time and attention is given to estimation than to the statistical testing of hypotheses.

[10] We recognize a logical problem here; having chosen H_f on the basis of the data available, subsequent estimates of parameters, tests of hypotheses etc. are to be understood as conditional on the "truth" of H_f. But given that the choice of H_f is itself the outcome of a statistical test the probabilistic properties of the subsequent estimators, the levels of significance, *are not the stated (nominal) ones.* The latter would hold only if H_f were in fact true, and would be valid in the present case conditionally on H_f. Indeed, empirical research ought to differentiate sharply between the test and "discovery" of hypotheses. Thus, if after a long "data mining" process one decides that a given model fits the data well, this exercise ought not to be understood as a test of the hypothesis that the world is described by such a model; at least not at the stated level of significance. It may, however, and indeed ought to be thought of as the discovery or the formulation of a hypothesis to be subsequently tested on an independent body of data. An early reference to this problem is T. A. Bancroft [6].

is tailor-made. After all, we do approach model selection with strong prior attachments even now. Only we tend—as a group—to apply these attachments in rather ad hoc, if not haphazard, and surely not reproducible ways. There may be a great deal to be gained by formalizing these procedures along Bayesian lines.

(b) Estimation and Testing

At this point we assume that some model selection procedure has gotten the researcher to the point at which it is appropriate to seek optimal parameter estimates (or to test hypotheses) under the usual specification axiom regarding appropriateness of the form of the model being analyzed. The existing econometric literature is more explicit in this area and in recent years econometricians have begun to pay increasing attention to the estimation of parameters which are subject to constraints [33] [52] and to various problems involving non-linear estimation [18] [26]. There would seem to be little purpose in our reviewing this literature which is quite familiar to most of those who engage in the construction (and testing) of econometric models. Rather, we have chosen to call attention to two strands of thought which exist in the literature of mathematical statistics, which seem to us to be potentially useful in economic problems, and which are on the whole not at all well-known to econometric model builders. We refer to two different situations involving restrictions on parameter values. The first—to which we now turn—is a case of intermediate hypotheses involving successively more severe restrictions on the admissable parameter space.[11] Here the problem has not yet been satisfactorily solved and we mention it briefly to draw attention to a research area which could yield a substantial payoff for econometric model building.

Suppose it is desired to test

$$H_0 : \theta \varepsilon \omega$$

against

$$H_1 : \theta \varepsilon (\Omega - \omega)$$

where θ is a vector of parameters, Ω is the admissable parameter space, and $\omega \subset \Omega$. It may be meaningful to conduct a sequence of tests on the intermediate hypotheses $\omega_1, \omega_2, \ldots, \omega_n$, where

$$\Omega = \omega_0 \supset \omega_1 \supset \omega_2 \supset, \ldots, \supset \omega_n = \omega,$$

in order to be able to pinpoint the reason, say, for the failure of hypothesis H_0 above.[12]

Suppose, in other words, that we employ the following procedure: Test ω_1 against $\omega_0 - \omega_1$. If ω_1 is not rejected, text ω_2 against $\omega_1 - \omega_2$. If ω_2 is not

[11] Economists are familiar with a special case of this problem involving a *single* subset hypothesis, and Chow [9] has provided a useful method for dealing with a two-sample problem within the subset hypothesis framework.

[12] Thus, a Chow test [9] may lead to the inference of structural change either because the coefficient vector, β, in the model $Y = X\beta + \varepsilon$ differs between the two sample periods under investigation, *or* because the variance of ε has changed (or both). It would therefore be desirable to be able to handle an intermediate hypothesis regarding the stability of the variance of ε.

rejected, test ω_3 against $\omega_2 - \omega_3$, and so on. If no rejections occur, then $H_0(\theta \varepsilon \omega = \omega_n)$, is accepted. If, however, some subhypothesis is rejected, say we reject $\theta \varepsilon \omega_k$ and thus accept $\theta \varepsilon(\omega_{k-1} - \omega_k), 0 < k < n$, we know that $\theta \notin \bigcup_{i=k}^{n} \omega_i$ and $\theta \varepsilon(\omega_{k-1} \cap \bar{\omega}_k)$, $\bar{\omega}_k$ being the complement of ω_k (in Ω). Since the sequence of intermediate hypotheses represents successively more severe restrictions upon the parameter space, the test tells us at what point the severity of the restriction becomes incompatible with the sample and, consequently, we know "why" H_0 is rejected.

Problems of this type have been discussed extensively by, among others, Darroch and Silvey [16], Hogg [31], Larson and Bancroft [40], and Seber [47]. To this point no easy solutions have yet been identified, a principal stumbling block involving the problem of statistical dependence of the successive hypothesis tests.

A more satisfactory result can be displayed in the case, to which we now turn, involving a Lagrange multiplier approach to the testing of a set of restrictions on the parameters being estimated. In general terms, the problem can be stated as follows. Let Y be a random variable (or vector) with p.d.f. $f(y; \theta)$ depending on a k-dimensional vector of parameters denoted by θ. It is asserted that certain restrictions hold, say $h(\theta) = 0$, where $h(\theta)$ is an r-dimensional vector valued function with $r < k$. The parameters can, in general, be estimated by first imposing the restrictions on the vector θ or, alternatively, by maximizing the expression

$$\mathscr{L}(\theta, \lambda) = L(y; \theta) + \lambda' h(\theta)$$

with respect to θ and λ, where $L(y; \theta)$ is the log likelihood corresponding to a sample on Y and λ is an r-dimensional vector of Lagrange multipliers.

The latter approach can be shown to yield a test of the validity of the restrictions, while the former does not. One could, of course, estimate unrestricted parameters and then derive statistics appropriate to testing the restrictions. If the restrictions are thereby rejected, then the unrestricted parameter estimates are the appropriate ones. On the other hand, if the hypothesis $h(\theta) = 0$ is accepted one would want to have the estimates obtained from a procedure which observes the restrictions—presumably on grounds of efficiency. The Lagrangian procedure yields both restricted parameters *and* the estimated Lagrange multipliers. In this case the test on the validity of the restrictions may be carried out on the Lagrange multipliers. If the restrictions are, in fact, valid the Lagrange multipliers should be zero since the restrictions imposed on the procedure are not binding—the data already incorporate such restrictions. Thus, a test on the estimated multipliers should lead to acceptance of the hypothesis that they are "insignificantly different from zero."

On the other hand, if the restrictions are invalid then the restrictions imposed by the procedure are, in fact, binding and a test based on the estimates of the Lagrange multipliers should yield the conclusion that they are "significantly different from zero." Thus, insignificance of Lagrange multipliers leads to *acceptance* of the restricted model, while significance leads to *rejection* of the restricted model and thus *acceptance* of the unrestricted model. If the unrestricted model is accepted, however, the restricted estimates are no longer appropriate—on grounds of possible inconsistency due to misspecification.

Such problems have been investigated by Aitchison and Silvey [2], [3], [48], who have shown that under the usual regularity conditions underlying maximum likelihood estimation, the appropriate test statistic for the hypothesis

$$H_0 : \lambda = 0$$

is

$$A = -\frac{1}{T}\hat{\lambda}' D^{-1}\hat{\lambda} = \frac{1}{T}\left(\frac{\partial L(y;\theta)}{\partial\theta}\right)' V^{-1}\left(\frac{\partial L(y;\theta)}{\partial\theta}\right),$$

where T is the sample size,

$$D^{-1} = -(RV^{-1}R')$$

$$R' = \left[\frac{\partial h(\theta)}{\partial\theta}\right]$$

and V is the so-called "information matrix,"

$$V = -\frac{1}{T}E\left[\frac{\partial^2 L(y;\theta)}{\partial\theta\partial\theta'}\right].$$

In the test statistic A all unknown parameters have been replaced by their *restricted* maximum likelihood estimates. If the statistic is "small" we accept the restricted model; if "large" we reject. Notice that if the restricted model were, in fact, valid then we would expect the restricted estimates to be "close" to the unrestricted ones. But the unrestricted estimates imply $\partial L/\partial\theta = 0$; thus, if both are close then for the restricted estimates we would have $\partial L/\partial\theta \cong 0$. Such considerations make this test intuitively quite attractive. Aitchison and Silvey have shown that the statistic A is, asymptotically, distributed as Chi-square with r degrees-of-freedom under the hypothesis $\lambda = 0$.

It is instructive to specialize the Aitchison–Silvey test to the linear model framework and compare it with the more familiar F-test based on the unrestricted estimates. Suppose

$$Y = X\beta + \varepsilon,$$

where Y is $(T \times 1)$; X is $(T \times K)$, nonstochastic, and of rank K; β is $(K \times 1)$; and ε is a $(T \times 1)$ multivariate normal vector with mean zero and covariance matrix $\sigma^2 I$. The log-likelihood function is

$$L = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\varepsilon'\varepsilon,$$

and, for subsequent reference, we note that $\partial^2 L/\partial\beta\partial\beta' = -(1/\sigma^2)(S)$, where $S = (X'X)$. The restrictions on β are given by

$$R\beta = r,$$

where r is a $J \times 1$ vector of known constants; R is a $(J \times K)$ matrix of known constants with the rank of R equal to $J < K$. We then form the Lagrangean function,

$$\mathcal{L} = L + \lambda'(R\beta - r).$$

Maximizing $\mathscr{L}$ with respect to β, σ^2, and λ yields the estimators (see [25, pp. 256–258]:

$$\hat{\beta} = b + S^{-1}R'(RS^{-1}R')^{-1}(r - Rb) \tag{1}$$

$$\hat{\lambda} = \frac{1}{\hat{\sigma}^2}(RS^{-1}R')^{-1}(r - Rb) \tag{2}$$

and

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{T}, \tag{3}$$

where b is the unrestricted Least Squares estimator, $b = S^{-1}X'Y$; and $\hat{\varepsilon}$ is the (restricted estimator) residual vector, $\hat{\varepsilon} = (Y - X\hat{\beta})$.

The Aitchison–Silvey test-statistic, A, is

$$A = -\frac{1}{T}\hat{\lambda}'D^{-1}\hat{\lambda}. \tag{4}$$

In this case D^{-1} is given by $-T\sigma^2(RS^{-1}R')$, since R' is itself the derivative of the constraint function with respect to the parameter vector β, and the information matrix is given by

$$\begin{aligned} V &= -\frac{1}{T}E\left[\frac{\partial^2 L}{\partial\beta\partial\beta'}\right] = -\frac{1}{T}E\left(-\frac{1}{\sigma^2}S\right) \\ &= \frac{1}{T\sigma^2}E(S) \\ &= \frac{S}{T\sigma^2}, \end{aligned}$$

since S is a non-stochastic matrix. Thus,

$$\begin{aligned} D^{-1} &= -(RV^{-1}R') \\ &= -\left[R\left(\frac{S}{T\sigma^2}\right)^{-1}R'\right] \\ &= -T\sigma^2(RS^{-1}R'). \end{aligned}$$

Substituting the latter into A in equation (4) yields:

$$A = \sigma^2\hat{\lambda}'(RS^{-1}R')\hat{\lambda}. \tag{5}$$

This statistic is asymptotically distributed as (central) chi-square with J degrees-of-freedom under the hypothesis $\lambda = 0$, as shown in [48]. With σ^2 unknown, $\hat{\sigma}^2$ can be substituted, yielding the observable test-statistic

$$\hat{A} = \hat{\sigma}^2\hat{\lambda}'(RS^{-1}R')\hat{\lambda} \tag{6}$$

which converges in distribution to the asymptotic distribution of A (since $\hat{\sigma}^2$ is consistent for σ^2) and is therefore also asymptotically chi-square with J degrees-of-freedom, if $\lambda = 0$.

The common test of the hypothesis $R\beta = r$ is based on an F-distributed statistic the derivation of which may be motivated as follows. The specification of the model implies that the unrestricted Least Squares estimator, b, is distributed multivariate $\mathcal{N}(\beta, \sigma^2 S^{-1})$, so that

$$R(b - \beta) = (Rb - R\beta) \sim \mathcal{N}(0, \sigma^2 RS^{-1}R').$$

But if $R\beta = r$, it follows that

$$(Rb - r) \sim \mathcal{N}(0, \sigma^2 RS^{-1}R'),$$

and therefore the statistic

$$\begin{aligned} C &= (Rb - r)'[\sigma^2 RS^{-1}R']^{-1}(Rb - r) \\ &= \frac{1}{\sigma^2}(Rb - r)'(RS^{-1}R')^{-1}(Rb - r) \end{aligned} \tag{7}$$

is distributed as (central) chi-square with J degrees-of-freedom. The statistic C contains the nuisance parameter σ^2, but

$$\frac{e'e}{\sigma^2} = \frac{(Y - Xb)'(Y - Xb)}{\sigma^2} = \frac{T\mathscr{S}^2}{\sigma^2}$$

is independent of the estimator b and is distributed as (central) chi-square with $(T - K)$ degrees-of-freedom. Thus,

$$\mathscr{F} = \frac{C/J}{T\mathscr{S}^2/\sigma^2(T - K)} = \frac{(Rb - r)(RS^{-1}R')^{-1}(Rb - r)}{\mathscr{S}^2}\frac{(T - K)}{TJ} \tag{8}$$

is distributed as (central) F with J and $(T - K)$ degrees-of-freedom, if $R\beta = r$.

To compare the latter with the Aitchison–Silvey test, substitute the expression for $\hat{\lambda}$ from (2) into the expression for A given in (5) to yield

$$A = \frac{\sigma^2}{\hat{\sigma}^2}\frac{1}{\hat{\sigma}^2}(r - Rb)'(RS^{-1}R')^{-1}(r - Rb).$$

Suppose now that σ^2 is known and does not have to be estimated, then A becomes

$$\begin{aligned} A &= \frac{1}{\sigma^2}(r - Rb)'(RS^{-1}R')^{-1}(r - Rb) \\ &= \frac{1}{\sigma^2}(Rb - r)'(RS^{-1}R')^{-1}(Rb - r), \end{aligned} \tag{9}$$

which is precisely the statistic C given in (7). Thus, if σ^2 were known, the Aitchison–Silvey test would coincide with the usual test on the unrestricted estimators, for the latter would then be based on the statistic C, there being no need to employ $T\mathscr{S}^2/\sigma^2$ to get rid of any nuisance parameter. From this we obtain the conclusions that, within the linear model framework as specified

(i) the two tests are (mathematically) equivalent if σ^2 is known,

and

(ii) the Aitchison–Silvey test is a valid small sample test under the normality assumption on ε, provided σ^2 is known.

If σ^2 is unknown, we then have the choice between the small sample F-test and the asymptotic chi-square test. Two additional results can be proven for the case of unknown σ^2:

(iii) the two tests are asymptotically equivalent in the sense that $J\mathscr{F}$ and $\hat{A}$ have the same asymptotic distribution (see appendix section A.2),

(iv) If ε is normally distributed, then the usual F-test is the appropriate test because the other is only asymptotically valid, while the F-test is valid for any sample size, enjoys the properties of a "Neyman-Structure" test [42; chapter 4] and so on. Furthermore, although

$$\frac{T\hat{\sigma}^2}{\sigma^2}$$

is distributed as chi-square with $(T - K + J)$ degrees-of-freedom, it is not independent of the estimator b, and thus *cannot* be used to convert A into an F-statistic with more denominator degrees-of-freedom (hence higher power) than $\mathscr{F}$ (see appendix section A.2).

Finally, and probably most important from an econometric model point of view, it appears that in the *absence* of a normality assumption on ε the Aitchison–Silvey test based on $\hat{A}$ is preferable to the test based on $\mathscr{F}$ for the following considerations. If ε is not normally distributed, the statistic C given in equation (7) will be distributed as chi-square with J degrees-of-freedom asymptotically, since it is mathematically equivalent to A.[13] Further the asymptotic distribution of C will be unaffected if the σ^2 in (7) is replaced by *any* consistent estimator. In effect, the standard statistic $\mathscr{F}$ results from replacing σ^2 by $\mathscr{S}^2$, a consistent estimator derived from b, while the Aitchison-Silvey statistic $\hat{A}$ results from replacing σ^2 by $\hat{\sigma}^2$, a consistent estimator derived from $\hat{\beta}$ which contains the restrictions $R\beta = r$. If the restrictions are valid then $\hat{\sigma}^2$ should be preferable to $\mathscr{S}^2$ (on grounds of efficiency), in the same way that any full information estimator is to be preferred to its corresponding limited information estimator. Although it does not matter asymptotically, for any finite sample size the estimator $\mathscr{S}^2$ can be considered to be based on a sample of size $(T - K)$ while $\hat{\sigma}^2$ can be considered to be based on a sample of size $(T - K + J) > (T - K)$.[14]

[13] This could be proven directly without appealing to the equivalence of C and A. If ε is not normally distributed, we can consider a quasi-maximum likelihood estimation problem, as though ε were normally distributed, or simply minimize the residual sum of squares subject to $R\beta = r$ and still obtain the same results including asymptotic normality.

[14] If the test is to be based on asymptotic principles, there is no purpose to running the test on $\mathscr{F}$ in any case. One should use either

$$\hat{C} = \frac{1}{\mathscr{S}^2}(Rb - r)'(RS^{-1}R')(Rb - r)$$

or

$$\hat{A} = \frac{1}{\hat{\sigma}^2}(Rb - r)'(RS^{-1}R')^{-1}(Rb - r),$$

each of which is asymptotically χ_J^2 if $R\beta = r$. We are arguing that $\hat{A}$ is preferable because $\hat{\sigma}^2$ is a "better" fixed sample estimator of σ^2 by virtue of its using more information about the structural model.

The reader will have noted that the discussion in this section has been predicated on a single-equation approach with non-stochastic regressors. In the case of stochastic regressors little difficulty is introduced if the regressors are fully independent of the error term ε. The small-sample F-test based on equation (8) would become a conditional F-test (conditional on the observed X's). In the Aitchison–Silvey test, the information matrix would be given by

$$V = \frac{ES}{T\sigma^2} = \frac{E(T^{-1}S)}{\sigma^2}.$$

This results in

$$\hat{A} = \frac{1}{T}\hat{\sigma}^2\hat{\lambda}'[R\{E(T^{-1}S)\}^{-1}R']\hat{\lambda},$$

which can be consistently estimated by

$$\begin{aligned}\hat{\hat{A}} &= \frac{1}{T}\hat{\sigma}^2\hat{\lambda}'[R(T^{-1}S)^{-1}R']\hat{\lambda} \\ &= \hat{\sigma}^2\hat{\lambda}'[RS^{-1}R']\hat{\lambda},\end{aligned} \tag{10}$$

precisely as in equation (6). The Aitchison–Silvey Test is thus completely unaffected by the presence of random regressors if they are independent of ε.[15] If the regressors include a lagged dependent variable (and we maintain the assumption of independent error terms) it becomes necessary to rely on a central limit theorem for dependent random variables to establish the asymptotic distribution of the Aitchison–Silvey statistic. Theil [51; p. 487] refers to one such central limit theorem which would apparently justify use of the Aitchison–Silvey test in the case of a lagged dependent variable.

Finally, suppose we are dealing with a simultaneous-equations model. If β is a vector of reduced-form parameters, then all of the foregoing applies. We are more apt, however, to be concerned about restrictions applying to behavioral (structural) parameters of the model. In that case, suppose the regressors in the equation for Y contain predicted values of some endogenous variables obtained from a directly estimated reduced form, so that b and $\hat{\beta}$ become, respectively, unrestricted and restricted $2SLS$ estimators of the structural parameters β. If the structural error terms are serially independent and the predetermined variables are either non-stochastic or fully independent of the structural error terms, then the Aitchison–Silvey test can be performed on the $2SLS$ estimators with unchanged asymptotic justification, precisely as discussed in the immediately preceding paragraph.[16]

[15] $\sqrt{T}(\hat{\lambda}/T)$ would still be asymptotically normally distributed, or—equivalently—$\sqrt{T}(b - \beta)$ would be asymptotically normally distributed with zero mean and covariance matrix $\sigma^2 \operatorname{Plim}(T^{-1}X'X)^{-1}$, which would again result in the statistic C in (7) being asymptotically χ_J^2 if $R\beta = r$.

[16] The Aitchison–Silvey test-statistic would still be consistently estimated by the $\hat{\hat{A}}$ of equation (10), which would still yield the statistic

$$\frac{1}{\hat{\sigma}^2}(Rb - r)'(RS^{-1}R')^{-1}(Rb - r)$$

upon substitution for $\hat{\lambda}$, though b is now the unrestricted $2SLS$ estimator. It is shown in [14; pp. 190–191] that under the conditions stated above,

$$\sqrt{T}(b - \beta) \text{ is asymptotically } \mathcal{N}\left(0, \sigma^2 \operatorname{plim}\left(\frac{X'X}{T}\right)^{-1}\right),$$

where X contains "predicted" endogenous variables. This is all that is needed to establish that the above statistic is asymptotically χ_J^2 (if $R\beta = r$), with $\hat{\sigma}^2$ being the variance estimator based upon $\hat{\beta}$ (the restricted $2SLS$ estimator of β).

The presence of lagged endogenous variables would again lead to the need for a central limit theorem for dependent variables.

(c) *Pseudo-Forecasts and Structural Stability Tests*

We assume now that an econometric model has been estimated and is ready for a "forecasting" evaluation prior to actual use as an operating model. A number of evaluation methods are available and several will be discussed in section V below. Here we should like to concentrate on the use of a data set which could have been pooled with the sample used to estimate the model, but was instead "saved" for a post-construction test of the model. We are well aware that under strong specification axioms it makes more sense to use all the available data in estimation, than to save some of it for later testing. This view is argued persuasively by Christ [11; pp. 546–548]. But in a realistic situation in which model selection procedures, hypothesis tests of various kinds, and a number of other "experiments" all amount to considerable data-mining, it would seem wise to have saved some data on which to evaluate the resulting model.[17]

Suppose, then, that the model-builder has available a set of *m* observations on each of the independent and dependent variables of the model. These data are assumed to lie outside the sample used to estimate the model, and it is further assumed that the *m* observations are too few in number to permit re-estimation of the model.[18] The model is to be used along with the *m* observations on the independent variables to generate *m* forecasts of the dependent variable(s) which can then be compared with the *m known* values of the dependent variable(s). For the case of a single equation and $m = 1$, a normality assumption on the error term (plus serial independence of the error term) permits the familiar *t*-test which can be considered equivalently either as a predictive test of the model or as a test of structural stability. For the single equation case with $m > 1$, it is possible to calculate a root mean squared error of forecast (the square root of the average of the squared forecasting errors) and it is tempting to think that such a statistic should be approximately the same as the standard error of estimate of the fitted equation if the structure has not changed. That this is not so, is alluded to in a recent paper by Jorgenson, Hunter and Nadiri [39].

Suppose the relation $Y = X\beta + \varepsilon$, with the same assumptions as previously given (including normality), is estimated by Least-Squares. The residual vector, say e, is given by

$$e = M\varepsilon,$$

where

$$M = I - XS^{-1}X',$$

and $e'e/(T - K)$ has expectation σ^2. The standard error of estimate is, of course, the square root of $e'e/(T - K)$. Now suppose that X_0 is the $(m \times K)$ matrix of

[17] Obviously if the model builder "knows" the data set which has been saved, he may find it impossible to prevent it from influencing his specification of the model. To that extent, a test on saved data is biased in favor of the model being tested. Subsequent testing on data which could not have been known at the time of model construction is clearly more desirable.

[18] In section IV we discuss the case in which there are enough new data to re-estimate the model on the new data set.

observations to be used in the predictive test of the model. If the structure of the model is correct, then

$$Y_0 = X_0\beta + \varepsilon_0$$

and the vector of forecast errors, say e_0, is given by

$$e_0 = Y_0 - X_0 b,$$

where $b = S^{-1}X'Y$. It is well known that under the stated assumptions e_0 is distributed as multivariate Normal with mean zero and covariance matrix $\sigma^2(I_m + X_0 S^{-1} X_0')$, where I_m is an $(m \times m)$ identity matrix. Denoting the matrix $(I_m + X_0 S^{-1} X_0')$ by Q, it follows that

$$\frac{e_0'(I_m + X_0 S^{-1} X_0')^{-1} e_0}{\sigma^2} = \frac{e_0' Q^{-1} e_0}{\sigma^2}$$

is distributed as (central) chi-square with m degrees-of-freedom. Thus

$$E[e_0' Q^{-1} e_0/m] = \sigma^2.$$

The mean squared error of forecast, however, is given by $e_0' e_0/m$, not $e_0' Q^{-1} e_0/m$, and the difference between these two measures is

$$e_0' e_0/m - (e_0' Q^{-1} e_0/m = e_0'(I_m - Q^{-1}) e_0/m.$$

It can be shown (see appendix section A.3) that $(I_m - Q^{-1})$ is a positive definite matrix. Thus $e_0'(I_m - Q^{-1})e_0/m$ is always positive which implies that

$$E(e_0' e_0/m) > E(e_0' Q^{-1} e_0/m) = \sigma^2.$$

The root mean squared error of forecast, which is the square root of $e_0' e_0/m$, should thus be *expected* to exceed the standard error of estimate of the fitted equation. Intuitively, this result is due to the fact that the variance of the forecast error arises not only from the residual variance, σ^2, but also from the discrepancy between b and β. The proper predictive test involves the ratio

$$(11) \qquad \frac{e_0' Q^{-1} e_0/m}{e'e/(T-K)} = \frac{e_0'(I_m + X_0 S^{-1} X_0')^{-1} e_0/m}{e'e/(T-K)}$$

which "corrects" for the component of the prediction error due to imprecision in the estimation of β, and is distributed as (central) F with m and $(T - K)$ degrees-of-freedom, if the structure is unchanged [40].

It is interesting that this predictive testing procedure can be generalized to the situation in which the reduced form of a linear simultaneous equations model is used to forecast m new observations on each of G endogenous variables. We make the following assumptions:

(i) the predetermined variables are non-stochastic,

(ii) the reduced form error terms are Normally distributed, serially independent, but contemporaneously dependent with contemporaneous covariance matrix denoted by $\boldsymbol{\Sigma}$.

(iii) the reduced form parameters are estimated by ordinary least squares.

The covariance matrix $\boldsymbol{\Sigma}$ is estimated by $\hat{\boldsymbol{\Sigma}}$ with typical element $e_i'e_j/(T - K)$ where e_i is the vector of residuals from the reduced form equation corresponding to the ith endogenous variable, e_j is the residual vector corresponding to the reduced form equation of the jth endogenous variable, and K is the number of predetermined variables (the same, of course, in all G reduced form equations). Now define e_0^G as an $(mG \times 1)$ vector of forecast errors, where the first m elements correspond to the first endogenous variable, the second m elements correspond to the second endogenous variable, and so on. We show in appendix section A.3 that the statistic

$$(e_0^G)'[\hat{\boldsymbol{\Sigma}}^{-1} \otimes (I_m + X_0 S^{-1} X_0')^{-1}](e_0^G)\frac{(T - K - G + 1)}{mG(T - K)}, \tag{12}$$

where $\otimes$ represents the Kronecker product, is distributed as (central) F with mG and $(T - K - G + 1)$ degrees-of-freedom if the structure is unchanged. It is obvious that for $G = 1$ the expression in (12) collapses to the single equation statistic given in (11).[19]

The assumption of non-stochastic predetermined variables can be relaxed in two ways. If the predetermined variables are stochastic but fully independent of the reduced form error terms, then the test-statistic given in (12) is appropriate for an F-test conditional on *both* X and X_0. More interesting is the case of predetermined variables which include lagged endogenous variables. Suppose we make a series of m one-period forecasts, that is, always using *actual* values for the lagged endogenous variables. It is then possible to consider the forecasts to be conditional on the observed matrix X_0, even though X_0 contains lagged endogenous variables. In this case, *if T is large* (the size of m does *not* matter)

$$(e_0^G)'[\hat{\boldsymbol{\Sigma}}^{-1} \otimes (I_m + X_0 S^{-1} X_0')^{-1}](e_0^G) \tag{13}$$

can be considered to have an *approximate* chi-square distribution with mG degrees-of-freedom if the structure is unchanged (see Appendix section A.3).[20] Unfortunately, we do not at this time know of any analogous statistical test for a sequence of dynamic forecasts in which the model generates its own lagged endogenous variables. We conclude this section by observing that if the model passes its predictive test evaluation, the m saved observations should then presumably (but see footnote 10) be incorporated into the data set to reestimate the model on all $(T + m)$ observations. If the model fails, then, of course, it's "back to the drawing board."

[19] Except for a recursive model, it makes little sense to assume that $\boldsymbol{\Sigma}$ is diagonal, for each reduced form error term is, in general, a linear combination of all the structural error terms. On the other hand, if we consider the set of G equations to be "seemingly unrelated regressions," $\boldsymbol{\Sigma}$ might be diagonal in which case (12) can be simplified to

$$\sum_{i=1}^{G}\left\{\frac{e_{0,i}'(I_m + X_0 S^{-1} X_0')^{-1} e_{0,i}/m}{e_i'e_i/T - K}\right\}\frac{(T - K - G + 1)}{G(T - K)},$$

where $e_{0,i}$ is the set of m forecast errors corresponding to the ith dependent variable. In this case, the test-statistic is proportional to the sum of the single-equation test-statistics as given in (11).

[20] The statistic in (12) yields a small sample test and would be proportional to a χ^2 if $\boldsymbol{\Sigma}$ were known. The F distribution arises because $\hat{\boldsymbol{\Sigma}}$ has been used as a Wishart-distributed estimator of $\boldsymbol{\Sigma}$. In equation (13), which is only approximately valid for large samples, no such correction is appropriate. If $\boldsymbol{\Sigma}$ itself were in (13) the statistic would still be only an approximate χ^2, and since $\hat{\boldsymbol{\Sigma}}$ is a consistent estimator of $\boldsymbol{\Sigma}$, the same should hold for the statistic containing $\hat{\boldsymbol{\Sigma}}$.

IV. Parametric Evaluation: Subsequent to Model Release

In this section we present a brief set of comments related to the evaluation of econometric models which are already at an operating stage. This section is quite brief for two primary reasons. First, the procedures discussed in this section depend on a sufficiently strong axiomatization to permit statistical testing in the familiar classical sense; there is not a great deal of scope for discussion here because our current knowledge is not terribly extensive. Secondly, much of what there is to say can be said by referring the reader back to discussions already presented in the previous section.

(a) *Availability of a Small Data Set*

Here we have reference to the continual flow of new data which, in the case of time series models, accrues a point at a time. Existing models can be checked against small sets of new data very frequently. Indeed, most of the operating macro forecasting models are subjected to a "residual analysis" check at least once per calendar quarter as new national income account data are issued by the government. These and other models, however, could in principle be put through a regularly scheduled predictive testing procedure along the lines discussed in section III, part (*c*). The only differences lie in the fact that the test procedure would be conducted on a data set which, obviously, could not have been incorporated into the original sample. Such predictive testing is especially valuable because it involves data successively further separated from the sample data used in the initial specification of the model.

A clearly useful procedure would be to incorporate each new data set into the model's estimation sample each time a predictive test is passed.[21] Most model-builders stop far short of such a procedure and re-estimate, indeed re-build, their models on a much looser schedule. It is not quite so obvious whether failure to pass a given predictive test, based on a small data set, should be grounds for immediate rejection of a model, for a number of reasons. Newly released data are frequently subject to substantial subsequent revision; it may be the new data which have failed the test, not the model. Small data sets can be heavily dominated by unique events which are outside the model's specified structure. Such circumstances have to be recognized as a *limitation* of the model, not as an indication that those processes which *are* represented within the model have been proven to be inadequately specified.

(*b*) *Availability of a Large Data Set*

Some econometric models are constructed in order to test hypotheses, not to be in continual use as forecasting or policy-analysis models. In such cases, they may well lie dormant over periods of time long enough for substantial new bodies of data to emerge. In the case of cross-section models, large sets of new data continually appear or can be obtained. In these circumstances it is possible to use the new data set, by itself, to re-estimate the model. This, of course, puts the model-builder (or someone else, for that matter) into the position of being able to conduct

[21] Ray Fair, for example, is one of the few model operators who actually re-estimates his model each quarter. See [22].

a rather powerful test of structural change. Economists are quite familiar with the use of the analysis of variance test discussed by Gregory Chow [9] for this situation. Here, especially, it would be useful if the series-tests on successively more restrictive nested hypotheses[22] were to become operational.

The predictive tests as discussed above are not, of course, limited in application to small data sets and are therefore alternatives to the Chow test. The latter, however, is a more powerful test when the new data set is large enough to be used by itself to re-estimate the mode!. Indeed, the Chow test is the classical likelihood ratio test for this situation.[23]

V. Non-Pa.·.ametric Evaluation

In view of the nature of the preceding discussion, it is useful to remind the reader once again that no pejorative intent is to be inferred from our use of the term non-parametric evaluation, or its connection with the process of Sherlock Holmes inference which we identified earlier. Indeed, we firmly believe that the need for somewhat descriptive kinds of evaluation procedures points as much to the richness of the areas of application of econometric models as it does to any inability of economists to put forth a strong axiomatization for their models. The spirit of our discussion here may be stated as follows. In the current state of our knowledge and analytical needs, to concentrate our attention solely on proving or disproving the "truth" of an econometric model is to choose an activity virtually guaranteed to suppress the major benefits which can flow from the proper use of econometric models. Having constructed the best models of which we are capable,[24] we ought to concern ourselves directly with whether or not particular models can be considered to be reliable tools for particular uses, regardless of the strict faithfulness of their specification.

In this context, "validation" becomes a problem-dependent or decision-dependent process, differing from case to case as the proposed use of the model under consideration changes. Thus a particular model may be validated for one purpose and not for another. In each case the process of validation is designed to answer the question: Is this model fulfilling the stated purpose? We can then speak of the evaluation of these models as the process of attempting to validate them for a series of purposes.[25] Thus the motivation of model-builders or users becomes directly relevant to the evaluation of the models themselves. The "success" of a model can then be measured by the extent to which it enables its user to decrease the frequency and consequences of wrong decisions. As Zarnowitz [55] has pointed

[22] See section III, part (*b*).

[23] The Chow test is a fixed sample F-test based on the same strict axiomatization as the predictive test discussed in section III, part (*c*). We have not here concerned ourselves with generalizations in the direction of lagged dependent variables, reduced-forms vs. structural models, and so on. Presumably this could be done along the lines of our previous discussions, with substantial benefits accruing to the process of econometric model evaluation.

[24] And while continuing the search for ever closer approximations to economic reality.

[25] Howrey et. al. [36] have pointed out that the method of estimation itself may also be partially a function of the use to which the model is to be put. The evaluation of any model should, of course, include an evaluation of the estimating procedures used. We do not comment on this aspect of the evaluation process here. For an interesting discussion of this issue, see Howrey [36].

out, however, the full application of even this more limited goal still poses very high informational requirements, namely: (i) the errors must be identifiable, (ii) the preferences of the decision maker and the constraints under which he operates must be available, (iii) the cost of providing the model must be ascertained. Large macroeconometric models, for example, are frequently used for both forecasting and policy analysis. In the role of a forecasting instrument, a model's usefulness is directly related to the accuracy of its *ex ante* forecasts. In the case of the policy analysis role, the main criterion is how well the model performs with respect to conditional forecasts based on particular configurations of policy options. In this case, especially, the user of the model typically possesses some—at least qualitative—knowledge about the policy maker's preferences concerning growth rates, inflation, unemployment, and so on. Such knowledge provides a natural set of criteria by which to judge the model's adequacy as a tool of policy analysis.[26]

But even here it is dangerous to polarize the evaluation too strongly onto specific use-oriented criteria. Our tests or evaluation procedures should—initially at least—center on the ability of the model to generate "historical" simulations which conform to the actual data. These simulations might be either deterministic or stochastic, and either static (one period) or dynamic (multi-period) in nature. A minimal requirement would involve a broad consistency of the data generated by a deterministic single-period simulation with the data from the actual historical record (both within and outside the sample period).[27]

However, even if a model "passed" a more demanding test of its ability to "track" the historical record (e.g., a deterministic multi-period historical simulation), economists would normally also want to investigate whether or not the model responded to various types of stimuli in the fashion anticipated or suggested by economic theory or independent empirical observation. Quite aside from the individual hypotheses underlying particular equations in the system, economists have certain (not entirely independent) "*reduced form*" *hypotheses* to which they would demand "acceptable" models to conform. That is, as a profession we seem to have developed some more or less vague ideas about the magnitudes of various impact, dynamic and steady-state multipliers as well as some prior notions about other dynamic characteristics that the model "should" exhibit. Despite Haavelmo's early warning [27], however, we have, at least until the recent work of Howrey [34], failed to realize just how difficult such tests are to design and carry out. This set of issues was finally confronted again at a recent NBER conference concerned with whether or not an existing set of models adequately reproduced the cyclical swings observed in our economic system.[28] It is difficult to catalogue what seems to be a

[26] Thus, a model which accurately predicts the employment effects of alternative tax policies may be considered "successful" even if its prediction of the composition of GNP is poor by the standards for other uses of a model.

[27] Especially, perhaps, in the simulation of historical episodes which involve policy changes or initial conditions relevant to current interests and decisions. It should be emphasized, however, that consistency of the data generated by a deterministic multi-period simulation with historical data is in general too strong a requirement. Howrey and Kalejian [35] have shown that under certain circumstances the dynamic deterministic simulation path of a correctly specified non-linear model may differ substantially from the historical time path.

[28] Conference on Research in Income and Wealth, Harvard University, November 14–15, 1969. For a summary introduction to these issues as they arose at this conference see Hickman [30].

minimal set of demands of this sort as needs and requirements vary according to the preferences and prejudices of the researcher and the actual needs of the user. In any case, the constraints imposed by these demands are, given the current state of knowledge, not overly stringent. Even if we consider the case of the government expenditure multiplier—where a relatively large amount of evidence has accumulated, "acceptable" estimates of its magnitude (both impact and steady state) vary widely among different "accepted" models of the U.S. economy.

We should also briefly consider whether in all types of experiments the simulated data should be generated by stochastic or non-stochastic simulation procedures. Certainly stochastic simulation, if we have the necessary extra information (in practice we often ignore the problem of obtaining good estimates of the variance-covariance matrix of the disturbance process), will yield a more informative characterization of the model being used and thus increase the quality of the evaluation procedure. Further, if the model is non-linear, and most macroeconometric models are these days, then the reduced form of the model *cannot* be inferred from the results of a non-stochastic solution [35]. That is, the application of non-stochastic simulation procedures yields results that should be expected to differ from those implied by the properties of the actual reduced form of the model. Although some preliminary experiments with the Wharton model suggested that the differences were not large, the results of the more extensive multi-model study by Haitovsky and Wallace [29] suggest a strong contrary conclusion regarding the ability of non-stochastic simulations to represent the reduced form properties of existing non-linear models.

The evaluation of the predictive ability of a model is essentially a goodness-of-fit problem. Because the statistical techniques available for this purpose normally require a strong axiomatization of the structure, econometric model builders have often found themselves restricted to simple graphical techniques (the fit "looks good") or simple summary measures (root mean square error, Theil's U-Statistic..., etc.),[29] of the performance of certain key variables. In a recent paper, Haitovsky and Treyz [28] have proposed an interesting descriptive decomposition of the forecast error for an endogenous variable in a large econometric model. The decomposition identifies error components involving: (a) the structural equation explaining the variable in question, (b) the rest of the estimated structural system, (c) incorrect values of lagged endogenous variables (in the case of dynamic simulations), (d) incorrect guesses about exogenous variables (in the case of an *ex ante* forecast), and (e) failure to make serial correlation "adjustments" for observed errors. Some attention has also been given to the development of a statistic analogous to the single-equation R^2, to be used to test the hypothesis that $\beta = 0$, where β is the coefficient vector of the system of equations under consideration. An interesting and complete discussion of this issue can be found in Dhrymes [17; Ch. 5]. Dhrymes defines such a statistic, but finds that it is dependent on the unknown covariance parameters of the joint distribution of the error terms of the system. Dhrymes [17] also derives an alternate test procedure

[29] Howrey et. al. [36] have recently suggested some difficulty with the root mean square error statistic (where small sample properties are unknown), particularly when used to compare structural versus autoregressive models, or sample versus post sample performance of a given model. See also our section III, part (*b*), and the discussion of Theil's U-Statistic in Jorgenson et. al. [39].

regarding the goodness-of-fit of the reduced form model (the fraction of the generalized variance of the jointly dependent variables explained by the reduced form), but this procedure involves the restriction that the number of variables in the model (endogenous plus predetermined) be less than the total number of observations—a restriction not generally fulfilled by large econometric models. The trace *correlation* statistic suggested by Hooper (based on the estimates of the canonical correlations) is closely related to the statistic discussed by Dhrymes, but its distribution seems quite intractable—although Hooper has given an approximate expression for the asymptotic variance of the statistic [32]. Perhaps this is an area of research that holds some promise.

Many interesting applications with large econometric models involve what is known as a "multiple response problem." That is we are interested in more than one characterization of the outcome of the experiment. This raises the question of whether to treat the outcome as one of many experiments each with a single response, or to combine all the responses (endogenous variables of interest) into a single response. This latter procedure, of course, involves the explicit formulation of the utility function of the user—a difficult situation.[30]

Other techniques which are in common use in the evaluation of a model's predictive performance are regression analysis and spectral analysis. In the former case we simply regress actual values on the predicted values of a series and test whether the resulting equations have zero intercepts and slopes not significantly different from unity (see Cohen and Cyert [12] and Hymans [37]). This general technique has also been used extensively by Theil [50], but as usual he has extended it and forced it to yield additional information. By regressing predicted values on actual values and actual values lagged one period, Theil is also able to investigate whether or not predicted changes tend to be biased toward recent actual changes. Theil's inequality coefficient and its decomposition into elements of bias, variance and covariance is very closely related to this type of analysis (although it refers to a regression of actual *changes* on predicted *changes*) and offers a great deal more information including some information on the tendency of the model to make turning point errors. Mincer and Zarnowitz [43] have provided some further development of Theil's procedure and have also suggested an additional measure of forecast error: the relative mean squared error. The latter is particularly interesting by virtue of its attempt to compare the costs and benefits of forecasts derived from alternative models of the economic process.

Spectral (cross-spectral) analysis is a statistical technique that can be used to obtain a frequency decomposition of the variance (covariance) of a univariate (bivariate) stochastic process. There are several ways in which spectrum analytic techniques might be used in the evaluation of econometric models. Naylor et al. [44] suggest that the spectra estimated from simulation data be compared with the spectra estimated directly from actual data. Howrey [34] has pointed out that for linear models the implied spectrum can be derived directly from the model and the stochastic simulation of the model is therefore not needed to make this comparison. Another application of spectral techniques is to test estimates of the

[30] For an interesting attempt to solve the multiple response problem see Fromm and Taubman [23] and Theil [49], [50].

structural or reduced-form disturbances for serial correlation, an important step in the Box–Jenkins modeling procedure [8].[31]

Cross-spectral analysis can also be used to investigate the relationship between predicted and actual values. That is, the Theil procedures can be extended to the frequency domain using cross-spectral analysis. This permits statistical testing of some more general hypotheses about the relationship of actual and predicted values.

An important advantage of spectral analysis is that it is a nonparametric approach to data analysis. Thus it is a particularly useful device in situations involving a weak axiomatization of the relationships under investigation. In addition, spectral methods do not depend on the statistical independence of the generated data points; they require only that the process generating the data be stationary to the second order. The significance tests that are available, however, depend on the assumption of Normality of the underlying process or on a sample size that is large enough that a form of the central limit theorem can be invoked. What little empirical experience has been accumulated in connection with the use of spectral analysis to investigate econometric models suggests that the technique can be used quite effectively to investigate certain dynamic properties of econometric models.

By way of tieing up the strands of this necessarily broad discussion, we should like to sketch, in outline form, the range of descriptive measures which have been found to yield useful insights into the performance and realiability characteristics of large scale econometric models. While some of these measures can be subjected to classical statistical tests, many are—at this stage of our knowledge—merely descriptive and geared to specialized model uses. A large number of these procedures can be traced to the writings of Zarnowitz and his co-workers [53], [54], [56], [57], Evans, Haitovsky and Treyz [21], Box and Jenkins [8], and Theil [50].

An Outline of Non-Parametric Measures

A. *Single-Variable Measures*
   1. Mean forecast error (changes and levels)
   2. Mean absolute forecast error (changes and levels)
   3. Mean squared error (changes and levels)
   4. Any of the above relative to
      (a) the level or variability of the variable being predicted
      (b) a measure of "acceptable" forecast error for alternative forecasting needs and horizons

B. *Tracking Measures*
   1. Number of turning points missed
   2. Number of turning points falsely predicted

[31] If one is primarily interested in forecasting (as opposed to explaining the behavior of the economic system) the conceptual simplicity of the Box-Jenkins procedure (essentially a battery of sophisticated smoothing techniques) has some appeal. This is particularly so if there is only one variable of interest as these procedures do not treat the output variables as being "tied" together in a system of interdependent relationships. Thus, forecasts of output, employment and prices, for example, need have no particular relatlonship to each other. Further, since the procedures are void of economic theory, they cannot, of course, be used to test hypotheses. Currently research is being done on developing procedures for building more information and constraints (exogenous and policy variables) into these models [8] [20] [45]. These investigations, if successful, may prove fruitful to econometricians.

3. Number of under- or overpredictions
4. Rank correlation of predicted and actual changes (within a subset of "important" actual movements)
5. Various tests of randomness
   (a) of directional predictions
   (b) of predicted turning points

C. *Error Decompositions*
1. Bias and variance of forecast error
2. Errors in start-up position vs. errors in the predicted changes
3. Identification of model subsectors transmitting errors to other sectors

D. *Comparative Errors*
1. Comparison with various "naive" forecasts[32]
2. Comparison with "judgmental," "consensus," or other non-econometric forecasts
3. Comparison with other econometric forecasts

E. *Cyclical and Dynamic Properties*
1. Impact and dynamic multipliers
2. Frequency response characteristics

The measures just outlined have been found to be suitable for a wide variety of purposes, and—surely—a user's confidence in any particular model would grow in proportion to the number of positive results yielded by such of these measures as seem relevant to the use in question. Several recent studies, [29], [39], and especially the Cooper–Jorgenson study [13], have made a valuable contribution by standardizing both the period of fit and the technique of estimation across alternative models prior to conducting inter-model comparisons. While model builders have in some measure tended to resent such activity on the part of "outsiders,"[33] the controversy certainly shows signs of producing improved procedures on all sides.

Models will be used for decision making, and their evaluation, therefore, ought to be tied to optimization of these decisions. The question we have to ask ourselves, then, is what series of tests and/or procedures will be sufficient to achieve a particular level of confidence in the use of a model for a certain specified purpose? What model builders have done, to date, is to catalogue the properties of their models, concentrating on those aspects of the system which seemed useful to them. There are two difficulties. First, model *users* may or may not find these properties to be relevant to their decision making. Second, we have not yet standardized the "list" of properties studied. A remedy for the latter situation would be most helpful to all users, is certainly feasible, and ought to receive high priority. The former issue is much more formidable and requires a greater degree of cooperation and candid communication than has to date taken place between model builders and the growing population of model users.

[32] The procedures of Box and Jenkins [8] may be particularly powerful in helping to identify the autoregressive procedures which would best serve as "Naive" alternatives to a structural model.

[33] See Howrey, Klein and McCarthy [36] who present arguments regarding the controls needed in such standardization attempts.

Appendix

This appendix serves to *sketch* some of the less familiar theoretical results which are the basis for statements made in the body of the paper.

A.1 An Illustration of the Cox Procedure for Non-Nested Hypotheses

Hypothesis H_f:

$$L_f(y;\alpha) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(y - X\alpha)'(y - X\alpha),$$

where

$$y = (y_1, y_2, \ldots, y_T)', \quad X = (e, x), \quad x = (x_1, x_2, \ldots, x_T)',$$
$$\alpha = (\alpha_1, \alpha_2), \quad e = (1, 1, \ldots, 1)'.$$

Hypothesis H_g:

$$L_g(y:\beta) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(y - X^*\beta)'(y - X^*\beta)$$

where

$$X^* = (e, x^*), \quad x^* = (x_1^*, x_2^*, \ldots, x_T^*)', \quad \beta = (\beta_0, \beta_1)'.$$

Define

$$\hat{\sigma}_f^2 = \frac{1}{T}y'[I - N]y, \quad N = X(X'X)^{-1}X'$$

$$\hat{\sigma}_g^2 = \frac{1}{T}y'[I - N^*]y, \quad N^* = X^*(X^{*\prime}X^*)^{-1}X^{*\prime}.$$

Then

$$l_{fg} = -\frac{T}{2}[\ln\hat{\sigma}_f^2 - \ln\hat{\sigma}_g^2]$$

$$\beta_\alpha = \operatorname{plim}(X^{*\prime}X^*)^{-1}X^{*\prime}[X\alpha + u] = \operatorname{plim}\left(\frac{X^{*\prime}X^*}{T}\right)^{-1}\left(\frac{X^{*\prime}X}{T}\right)\alpha$$

on the assumption that H_f is true and that accordingly

$$y = X\alpha + u,$$

$u = (u_1, u_2, \ldots, u_T)'$, $\{u_t : t = 1, 2\ldots\}$ being a sequence of identically and independently distributed random variables with mean zero and variance σ^2. In the preceding it is assumed that the x's are either a sequence of fixed constants or if they are random variables they are distributed independently of u.

We observe that, under H_f,

$$\frac{1}{T}L_f(\hat{\alpha}) = \tfrac{1}{2}[\ln(2\pi) + 1] - \tfrac{1}{2}\ln\hat{\sigma}_f^2$$

$$\frac{1}{T}L_f(\alpha) = -\tfrac{1}{2}\ln(2\pi) - \tfrac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\frac{u'u}{T}.$$

Because $\hat{\sigma}_f^2$ is a consistent estimator of σ^2 and so is $u'u/T$, we conclude that

$$\operatorname{plim}\frac{1}{T}[L_f(\hat{\alpha}) - L_f(\alpha)] = 0.$$

Further, since $\operatorname{plim}[L_g(\hat{\beta})|H_f] = L_g(\beta_\alpha)$,

$$\operatorname{plim}\frac{1}{T}[L_g(\hat{\beta}) - L_g(\beta_\alpha)] = 0.$$

Moreover,

$$\frac{1}{T}L_g(\beta_\alpha) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\frac{(y - X^*\beta_\alpha)'(y - X^*\beta_\alpha)}{T}.$$

But we see that

$$\operatorname{plim}\frac{1}{T}(y - X^*\beta_\alpha)'(y - X^*\beta_\alpha) = \sigma^2 + \operatorname{plim}\left[\frac{1}{T}\alpha' X'(I - N^*)X\alpha\right].$$

Thus

$$\operatorname{plim}\frac{1}{T}[L_f(\alpha) - L_g(\beta_\alpha)] = \frac{1}{2\sigma^2}\operatorname{plim}\left[\frac{1}{T}\alpha' X'(I - N^*)X\alpha\right] \geq 0.$$

In general we would expect a strict inequality except for special x-sequences.

Turning now to the test statistic $(1/T)S_f$ (as defined in the text, supra), we obtain

$$\frac{1}{T}S_f = -\tfrac{1}{2}[\ln\hat{\sigma}_f^2 - \ln\hat{\sigma}_g^2] + \tfrac{1}{2}E_{\hat{\alpha}}[\ln\hat{\sigma}_f^2 - \ln\hat{\sigma}_g^2].$$

Under H_f, $(T\hat{\sigma}_f^2/\sigma^2)$ is (central) chi-square with $(T - 2)$ degrees of freedom, and $(T\hat{\sigma}_g^2/\sigma^2)$ is non-central chi-square with $(T - 2)$ degrees of freedom. Thus, in principle, this expectation may be carried out. In general, it will involve the unknown parameter α and for purposes of the test we would have to insert the maximum likelihood estimate (MLE) $\hat{\alpha}$, in its stead. Further, such tests require specification of the distribution of the data under consideration and the derivation of the MLE under the two alternatives.

A.2 The Aitchison–Silvey Test

1. *$J\mathcal{F}$ and $\hat{A}$ have the same asymptotic distribution*

$$J\mathcal{F} = \frac{(Rb - r)'(RS^{-1}R')^{-1}(Rb - r)}{\mathcal{S}^2}\frac{(T - K)}{T},$$

while

$$\hat{A} = \frac{1}{\hat{\sigma}^2}(Rb - r)'(RS^{-1}R')^{-1}(Rb - r).$$

Hence

$$(J\mathcal{F} - \hat{A}) = \hat{A}\left[\frac{\hat{\sigma}^2}{\mathcal{S}^2}\frac{(T - K)}{T} - 1\right]$$
$$= \hat{A}\left[\frac{\hat{\sigma}^2}{\mathcal{S}^2}\left(1 - \frac{K}{T}\right) - 1\right].$$

Since $\hat{A}$ has an asymptotic distribution,

$$\hat{A}\left[\frac{\hat{\sigma}^2}{\mathcal{S}^2}\left(1 - \frac{K}{T}\right) - 1\right] = (J\mathcal{F} - \hat{A})$$

will have a zero probability limit if

$$\operatorname{plim}\left[\frac{\hat{\sigma}^2}{\mathcal{S}^2}\left(1 - \frac{K}{T}\right) - 1\right] = 0.$$

But

$$\operatorname{plim}\left[\frac{\hat{\sigma}^2}{\mathcal{S}^2}\left(1 - \frac{K}{T}\right) - 1\right] = \left[\operatorname{plim}\frac{\hat{\sigma}^2}{\mathcal{S}^2}\lim\left(1 - \frac{K}{T}\right) - 1\right]$$
$$= (1 - 1) = 0,$$

since $\hat{\sigma}^2$ and $\mathcal{S}^2$ are both consistent estimators of σ^2. Hence, plim $(J\mathcal{F} - \hat{A}) = 0$, and since $\hat{A}$ has an asymptotic distribution this condition implies that $J\mathcal{F}$ has the same asymptotic distribution as $\hat{A}$. Q.E.D.

1. *$T\hat{\sigma}^2$ is not independent of b*

$$T\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}.$$
$$\hat{\varepsilon} = Y - X\hat{\beta}$$
$$= (M + N)\varepsilon,$$

where

(i) $M = I - XS^{-1}X'$, idempotent of rank $(T - K)$,
(ii) $N = XS^{-1}R'(RS^{-1}R')^{-1}RS^{-1}X'$, idempotent of rank J, and therefore
(iii) $M + N$ is idempotent of rank $(T - K + J)$.
It follows that

$$T\hat{\sigma}^2 = \varepsilon'(M + N)\varepsilon.$$
$$b = S^{-1}X'Y = \beta + S^{-1}X'\varepsilon,$$

hence, $\qquad b - \beta = S^{-1}X'\varepsilon.$

Thus $(b - \beta)$ is a linear form in the Normally distributed vector ε, and $T\hat{\sigma}^2$ is an idempotent quadratic form in ε. Independence of the linear and quadratic forms requires $S^{-1}X'(M + N) = 0$. But

$$\begin{aligned} S^{-1}X'(M + N) &= S^{-1}(X'M + X'N) \\ &= S^{-1}[X'(I - XS^{-1}S') + X'(XS^{-1}R'(RS^{-1}R')^{-1}RS^{-1}X')] \\ &= S^{-1}[0 + R'(RS^{-1}R')^{-1}RS^{-1}X'] \\ &= S^{-1}R'(RS^{-1}R')^{-1}RS^{-1}X' \neq 0. \end{aligned}$$

Hence b and $T\hat{\sigma}^2$ are not independent. Q.E.D.

A.3 Predictive Testing

1. *$(I_m - Q^{-1})$ is a positive definite matrix*

$$Q = I_m + X_0S^{-1}X_0'.$$

Clearly, I_m is positive definite. Q is positive definite if $X_0S^{-1}X_0'$ is positive definite. Let z be any nonzero m-dimensional vector, then $z'X_0S^{-1}X_0'z = (z'X_0)S^{-1}(z'X_0)' > 0$, by virtue of S^{-1} being positive definite.

Since Q is positive definite, so is its inverse, thus I_m and Q^{-1} are positive definite and we can apply the theorem given in [17; pp. 581–583] which implies that $(I_m - Q^{-1})$ will be positive definite if and only if the roots of Q^{-1} are smaller than unity.

But the roots of Q^{-1} are the inverses of the roots of Q. Denote a root of Q by $(1 + \alpha)$, so that

$$\begin{aligned} 0 &= [Q - (1 + \alpha)I_m]z \\ &= [I_m + X_0S^{-1}X_0' - (1 + \alpha)I_m]z \\ &= [X_0S^{-1}X_0' - \alpha I_m]z. \end{aligned}$$

Thus α is a root of $X_0S^{-1}X_0'$ and must be positive since $X_0S^{-1}X_0'$ is positive definite. But $\alpha > 0$ implies $(1 + \alpha) > 1$, which implies $(1 + \alpha)^{-1} < 1$.

Q.E.D.

2. *The Distribution of*

$$(e_0^G)'[\hat{\Sigma}^{-1} \otimes (I_m + X_0S^{-1}X_0')^{-1}](e_0^G)\frac{(T - K - G + 1)}{mG(T - K)}$$

The vector of forecast errors, say $e_{0,g}$, corresponding to the gth endogenous variable is given by

$$e_{0,g} = X_0(b_g - \beta_g) - \varepsilon_{0,g},$$

where

(i) β_g is the vector of reduced form coefficients corresponding to the gth endogenous variable.
(ii) b_g is the Least Squares estimator of β_g.
(iii) $\varepsilon_{0,g}$ is the gth reduced form disturbance vector in the forecast period.

Then

$$\begin{pmatrix} e_{0,1} \\ e_{0,2} \\ \vdots \\ e_{0,G} \end{pmatrix} = \begin{pmatrix} X_0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & X_0 & \cdot & \cdot & \cdot & 0 \\ \vdots & & & & & \\ 0 & \cdot & \cdot & \cdot & \cdot & X_0 \end{pmatrix} \begin{pmatrix} b_1 - \beta_1 \\ b_2 - \beta_2 \\ \vdots \\ b_G - \beta_G \end{pmatrix} + \begin{pmatrix} \varepsilon_{0,1} \\ \varepsilon_{0,2} \\ \vdots \\ \varepsilon_{0,G} \end{pmatrix}$$

or

$$(e_0^G) = Z_0(b - \beta) + (\varepsilon_0^G).$$

Conditional on X and X_0, e_0^G is clearly normally distributed with mean zero and the following covariance matrix.

$$E[(e_0^G)(e_0^G)'|X, X_0] = Z_0[\text{cov}\,(b - \beta)]Z_0' + \text{cov}\,(\varepsilon_0^G),$$

where

(i) cov $(b - \beta)$ is the covariance matrix of $(b - \beta)$ conditional on X and X_0;

$$\text{cov}\,(b - \beta) = \Sigma \otimes S^{-1},$$

(ii) cov (ε_0^G) is the covariance matrix of (ε_0^G);

$$\text{cov}\,(e_0^G) = \Sigma \otimes I_m,$$

and

(iii) Σ is the contemporaneous covariance matrix of ε.
Combining terms above yields

$$\begin{aligned} \text{cov}\,(e_0^G) &= E[(e_0^G)(e_0^G)'|X, X_0] \\ &= Z_0(\Sigma \otimes S^{-1})Z_0' + \Sigma \otimes I_m \\ &= \Sigma \otimes X_0 S^{-1} X_0' + \Sigma \otimes I_m \\ &= \Sigma \otimes (I_m + X_0 S^{-1} X_0'). \end{aligned}$$

Thus, $(e_0^G) \sim \mathcal{N}[0, \Sigma \otimes (I_m + X_0 S^{-1} X_0')]$,
which implies that

$$(e_0^G)'[\Sigma^{-1} \otimes (I_m + X_0 S^{-1} X_0')^{-1}](e_0^G)$$

is distributed as χ^2_{mG}.

Now $\hat{\Sigma}$, as defined in the body of the paper, is based only on the residuals in the period of fit which, it can be shown, are independent of e_0^G.

It follows [4; pp. 105–107, 181–183] that

(i) $(T - K)\hat{\Sigma}$ is a Wishart distributed matrix, independent of (e_0^G), with $(T - K)$ degrees-of-freedom, and

(ii) $$(e_0^G)'[\hat{\Sigma}^{-1} \otimes (I_m + X_0 S^{-1} X_0')^{-1}](e_0^G)\frac{(T - K - G + 1)}{mG(T - K)}$$

is distributed as $F_{mG, (T-K-G+1)}$.

3. The Approximate Distribution of $(e_0^G)'[\hat{\Sigma}^{-1} \otimes (I_m + X_0 S^{-1} X_0')^{-1}](e_0^G)$ for Large T, for Static Forecasting with Lagged Endogenous Variables.

Again $(e_0^G) = Z_0(b - \beta) + (\varepsilon_0^G)$, but Z_0 contains lagged endogenous variables. Since b can be written as $b = \beta + (Z'Z)^{-1}Z'\varepsilon^G$, where ε^G and Z are the fit period analogues of (ε_0^G) and Z_0 respectively, it follows that

$$\begin{aligned}(e_0^G - \varepsilon_0^G) &= Z_0(Z'Z)^{-1}Z'\varepsilon \\ &= Z_0(T^{-1}Z'Z)^{-1}T^{-1}Z'\varepsilon.\end{aligned}$$

assuming that the observed moment matrix of the predetermined variables in the fit period converges in probability to their population moments, i.e., plim $(T^{-1}Z'Z)$ exists and is non-zero, then

$$\begin{aligned}&\text{plim}\,[\sqrt{T}(e_0^G - \varepsilon_0^G)|Z_0] \\ &\qquad = Z_0\,\text{plim}\,(T^{-1}Z'Z)^{-1}\,\text{plim}\,\sqrt{T}(T^{-1}Z'\varepsilon).\end{aligned}$$

But plim $\sqrt{T}(T^{-1}Z'\varepsilon)$ is asymptotically distributed as Normal with mean zero and covariance matrix $\Sigma \otimes M_X^*$, where M_X^* is the matrix of population moments of the predetermined variables. (See [51; p. 487].) Further, by the definition of Z,

$$\text{plim}\,(T^{-1}Z'Z)^{-1} = I_{GK} \otimes (M_X^*)^{-1},$$

where I_{GK} is a $(GK \times GK)$ identity matrix. Thus, $\sqrt{T}(e_0^G - \varepsilon_0^G)$, conditional on Z_0, is asymptotically distributed as $\mathcal{N}(0, H)$, where

$$\begin{aligned}H &= Z_0[I_{GK} \otimes (M_X^*)^{-1}](\Sigma \otimes M_X^*)[I_{GK} \otimes (M_X^*)^{-1}]Z_0' \\ &= \Sigma \otimes X_0(M_X^*)^{-1}X_0'.\end{aligned}$$

For large T, it should therefore be approximately true that

$$(e_0^G - \varepsilon_0^G) \text{ is approximately } \mathcal{N}(0, T^{-1}H).[1]$$

Since e_0^G and ε_0^G are independent (e_0^G depending only on ε's prior to the fit period),

$$e_0^G \text{ is approximately } \mathcal{N}[0, (T^{-1}H) + (\Sigma \otimes I_m)],$$

for large T. But

$$\begin{aligned}(T^{-1}H) + (\Sigma \otimes I_m) &= (T^{-1}\Sigma \otimes X_0(M_X^*)^{-1}X_0') + (\Sigma \otimes I_m) \\ &= \Sigma \otimes [I_m + T^{-1}X_0(M_X^*)^{-1}X_0'] \\ &= \Sigma \otimes [I_m + X_0(TM_X^*)^{-1}X_0'].\end{aligned}$$

Hence, for large T,

$$(e_0^G)'[\Sigma^{-1} \otimes (I_m + X_0(TM_X^*)^{-1}X_0')^{-1}](e_0^G)$$

is approximately χ^2_{mG}.

Since

$$\text{plim}\,\hat{\Sigma}^{-1} = \Sigma^{-1}$$

[1] $(e_0^G - \varepsilon_0^G)$ has of course a degenerate limiting distribution. We are arguing here that as T increases $(e_0^G - \varepsilon_0^G)$ "degenerates" through the normal limiting distribution of $\sqrt{T}(e_0^G - \varepsilon_0^G)$.

and $\text{plim}\,(T^{-1}S) = \text{plim}\,(T^{-1}X'X) = M_X^*$, the above statistic can be consistently estimated by

$$(e_0^G)'[\hat{\Sigma}^{-1} \otimes (I_m + X_0 S^{-1} X_0')^{-1}](e_0^G)$$

which, for large T, is also approximately χ^2_{mG}.

REFERENCES

[1] Adelman, I. and F. Adelman, "The Dynamic Properties of the Klein–Goldberger Model," *Econometrica*, Vol. 27 (1959).

[2] Aitchison, J. "Large Sample Restricted Parametric Tests," *JRSS* (series B), Vol. 24, 1962.

[3] Aitchison, J. and S. D. Silvey. "Maximum Likelihood Estimation of Parameters Subject to Restraints," *Annals of Mathematical Statistics*, Vol. 29, 1958.

[4] Anderson, T. W. *An Introduction to Multivariate Statistical Analysis*, J. Wiley and Sons, 1958.

[5] Atkinson, A. C. "A Method for Discriminating Between Models," *JRSS* (Series B) Vol. 32, 1970.

[6] Bancroft, T. A. "On Biases in Estimation Due to the Use of Preliminary Tests of Significance," *Annals of Mathematical Statistics*, Vol. 15, 1944.

[7] Bischoff, Charles W. "Business Investment in the 1970's: A Comparison of Models," *Brookings Papers on Economic Activity*, 1:1971.

[8] Box, G. and G. Jenkins, *Time Series Analysis*, Holden-Day, 1970.

[9] Chow, Gregory. "Tests of the Equality Between Two Sets of Coefficients in Two Linear Regressions," *Econometrica*, Vol. 28 (1960).

[10] Christ, Carl F. "A Test of an Econometric Model for the U.S., 1921–1947," in Universities—National Bureau Committee for Economic Research *Conference on Business Cycles*, New York, National Bureau of Economic Desearch, 1951, pp. 35–107.

[11] Christ, Carl F. *Econometric Models and Methods*, J. Wiley and Sons, 1966.

[12] Cohen, Kalman J. and R. M. Cyert. "Computer Models in Dynamic Economics," *QJE*, February 1961.

[13] Cooper, Ronald L. and D. W. Jorgenson. "The Predictive Performance of Quarterly Econometric Models of the United States," in *Econometric Models of Cyclical Behavior*, B. Hickman, Editor, Conference on Research in Income and Wealth, Vol. 36, National Bureau of Economic Research, 1972.

[14] Cox, D. R. "Further Results on Tests of Separate Families of Hypotheses," *JRSS* (Series B), Vol. 24 (1962).

[15] Cox, D. R. "Tests of Separate Families of Hypotheses," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, 1961.

[16] Darroch, N. N. and S. D. Silvey. "On Testing More than One Hypothesis," *Annals of Mathematical Statistics*, Vol. 34 (1963).

[17] Dhyrmes, Phoebus J. *Econometrics*, Harper and Row, 1970.

[18] Dhrymes, Phoebus J. "On the Treatment of Certain Recurrent Nonlinearities in Regression Analysis," *SEJ*, October 1966.

[19] Dickey, J. M. "Bayesian Alternatives to the F Test," Research Report No. 50, SUNY at Buffalo.

[20] Dunn, D. M., W. H. Williams, and A. Spivey, "Analysis and Prediction of Telephone Demand in Local Geographic Areas," *The Bell Journal of Economics and Management Science*, Vol. 2, No. 2, Autumn 1971, p. 561.

[21] Evans, Michael K., Y. Haitkovsky, and G. Treyz. "An Analysis of the Forecasting Properties of U.S. Econometric Models," in *Econometric Models of Cyclical Behavior*, B. Hickman, Editor, Conference on Research in Income and Wealth, Vol. 36, National Bureau of Economic Research, 1972.

[22] Fair, Ray C. *A Short-Run Forecasting Model of the U.S. Economy*, Heath and Co., Lexington, Mass., 1971.

[23] Fromm, Gary and P. Taubman. *Policy Simulations with an Econometric Model*. Washington, The Brookings Institution, 1968.

[24] Geisel, M. S. Comparing and Choosing Among Parametric Statistical Models: A Bayesian Analysis with Microeconomic Applications, Unpublished Ph.D. Dissertation, University of Chicago, 1969.

[25] Goldberger, Arthur S. *Econometric Theory*, J. Wiley and Sons, 1964.

[26] Goldfeld, Stephen M., R. E. Quandt, and H. F. Trotter. "Maximization by Quadratic Hill Climbing," *Econometrica*, July 1966.

[27] Haavelmo, T. "The Inadequacy of Testing Dynamic Theory by Comparing Theoretical Solutions and Observed Cycles," *Econometrica*, October 1940.

[28] Haitovsky, Yoel, and G. Treyz. "The Decomposition of Econometric Forecast Error," Mimeo.

[29] Haitovsky, Yoel, and N. Wallace. "A Study of Discretionary and Nondiscretionary Fiscal and Monetary Policies in the Context of Stochastic Macro-Econometric Models," in V. Zarnowitz, ed., *The Business Cycle Today*, N.B.E.R., 1972.

[30] Hickman, Bert G. "Introduction and Summary," in *Econometric Models of Cyclical Behavior*, B. Hickman, editor, Conference on Research in Income and Wealth, Vol. 36, National Bureau of Economic Research, 1972.

[31] Hogg, R. B. "On the Resolution of Statistical Hypotheses," *Journal of American Statistical Association*, Vol. 56 (1961).

[32] Hooper, J. W. "Partial Trace Correlations," *Econometrica*, Vol. 30 (1962).

[33] Houthakker, Hendrik S. and L. D. Taylor. *Consumer Demand in the United States*, Second edition, Harvard University Press, 1970.

[34] Howrey, E. Philip, "Dynamic Properties of a Condensed Version of the Wharton Model," *Econometric Models of Cyclical Behavior*, B. Hickman, editor, Conference on Research in Income and Wealth, Vol. 36, National Bureau of Economic Research, 1972.

[35] Howrey, E. Philip and H. H. Kalejian. "Computer Simulation Versus Analytical Solutions," in T. H. Naylor, ed., *The Design of Computer Simulation Experiments*, Duke University Press, 1969.

[36] Howrey, E. Philip, L. R. Klein, and M. D. McCarthy. "Notes on Testing the Predictive Performance of Econometric Models," Discussion Paper No. 173, Wharton School, Department of Economics, University of Pennsylvania, Philadelphia, 1970.

[37] Hymans, Saul H. "Consumption: New Data and Old Puzzles," *Brookings Papers on Economic Activity*, 1:1970.

[38] Hymans, Saul H. "Prices and Price Behavior in Three U.S. Econometric Models," Paper prepared for the Conference on the Econometrics of Price Determination, October 30–31, 1970. Washington, D.C.

[39] Jorgenson, Dale, W., J. Hunter, and M. Nadiri. "The Predictive Performance of Econometric Models of Quarterly Investment Behavior," *Econometrica*, March 1970.

[40] Larson, H. J. and T. A. Bancroft. "Sequential Model Building for Prediction in Regression Analysis," *Annals of Mathematical Statistics*, Vol. 34 (1963).

[41] Leamer, Edward E. "Model Selection Searches: A Bayesian View," Discussion Paper 151, Harvard Institute of Economic Research, December 1970.

[42] Lehmann, E. L. *Testing Statistical Hypotheses*, J. Wiley and Sons, 1959.

[43] Mincer, Jacob and V. Zarnowitz. "The Evaluation of Economic Forecasts," in J. Mincer, ed., *Economic Forecasts and Expectations: Analyses of Forecasting Behavior and Performance*, N.B.E.R., 1969.

[44] Naylor, Thomas H., K. Wertz and T. H. Wonnacott. "Spectral Analysis of Data Generated by Simulation Experiments with Econometric Models," *Econometrica*, April 1969.

[45] Pierce, D. A. "Fitting Dynamic Time Series Models: Some Considerations and Examples," Federal Reserve Bank of Cleveland, Mimeo.

[46] Ramsey, J. B. "Tests for Specification Errors in Classical Least-Squares Regression Analysis," *Journal of the Royal Statistical Society*, Series B, 1969, pt. 2, pp. 350–371.

[47] Seber, G. A. F. "Linear Hypotheses and Induced Tests," *Biometrika*, Vol. 51 (1964).

[48] Silvey, S. D. "The Lagrangian Multiplier Test," *Annals of Mathematical Statistics*, Vol. 30, 1959.

[49] Theil, Henri. *Applied Economic Forecasting*, Rand-McNally, 1966.

[50] Theil, Henri. *Economic Forecasts and Policy*, North-Holland Publishing Co., 1961.

[51] Theil, Henri. *Principles of Econometrics*, John Wiley and Sons, 1971.

[52] Theil, Henri and A. S. Goldberger, "On Pure and Mixed Statistical Estimation in Economics," *IER*, Vol. 2 (1961).

[53] Zarnowitz, Victor. "Forecasting Economic Conditions: The Record and the Prospect," in V. Zarnowitz, editor, *The Business Cycle Today*, National Bureau of Economic Research, 1972.

[54] Zarnowitz, Victor. *An Appraisal of Short-Term Economic Forecasts*, N.B.E.R., 1967.

[55] Zarnowitz, Victor. "New Plans and Results of Research in Economic Forecasting," *Fifty-First Annual Report*, National Bureau of Economic Research, 1971, pp. 53–70.

[56] Zarnowitz, Victor. "Prediction and Forecasting: Economic," *International Encyclopedia of the Social Sciences*, The Macmillan Co. and the Free Press, 1968.

[57] Zarnowitz, Victor, C. Boschan, and G. H. Moore, with the assistance of Josephine Su. "Business Cycle Analysis of Econometric Model Simulations," in *Econometric Models of Cyclical Behavior*, B. Hickman, Editor, Conference on Research in Income and Wealth, Vol. 36, National Bureau of Economic Research, 1972.

[58] Zellner, Arnold. *An Introduction to Bayesian Inference in Econometrics*, John Wiley and Sons, New York, 1971.