University of
California, Davis

DATA MINING RECONSIDERED: ENCOMPASSING AND THE GENERAL-TO-SPECIFIC APPROACH TO SPECIFICATION SEARCH

Kevin D. Hoover & Stephen J. Perez

Working Paper #97-27

Department
of Economics

Data Mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search

Kevin D. Hoover
University of California, Davis

Stephen D. Perez
Virginia Commonwealth University

Working Paper Series No. 97-27
October, 1997

We thank Clinton Greene, James Hartley, David Hendry, Edward Leamer, Michael Lovell, Thomas Mayer, Steven Sheffrin and the participants in the University of California, Davis, Macroeconomics Workshop for helpful comments on an earlier draft.

Abstract

The effectiveness of one aspect of the London School of Economics (LSE) approach to econometrics is assessed in a simulation study. The paper uses a data set and nine models analogous to those in Lovell's (1983) study of data mining. A simplified general-to-specific algorithm is tested in a simulation framework. While the study documents some of the pitfalls of the general-to-specific approach, it is, on the whole, supportive of the effectiveness of the LSE methodology as applied to stationary data with relatively simple dynamics. The general-to-specific methodology clearly dominates the alternative search methodologies investigated by Lovell.

Data Mining Reconsidered:
Encompassing and the General-to-Specific Approach to Specification Search

Occasionally, a bold moral inversion may startle us out of our dogmatic slumbers: a Proudhon proclaims, "Property is Theft;" a David Hendry (e.g., 1995, ch. 15, section 1) proclaims, *data mining is good econometrics*. Although Proudhon's slogan commands our attention, in the end we reject it. Can the same be said for Hendry's support of data mining? We propose to evaluate Hendry's data mining procedures - not philosophically, theoretically or methodologically, but practically: in a simulation study in which we know what the underlying process is that generated the data, do the methods advocated by Hendry and other practitioners of the LSE [London School of Economics] econometric methodology in fact recover the true specification?[1]

To practice data mining is to sin against the norms of econometrics: of that there can be little doubt. That few have attempted to justify professional abhorrence to data mining signifies nothing: few have felt any pressing need to justify our abhorrence of theft either. What is for practical purposes beyond doubt needs no special justification; and we learn that data mining is bad econometric practice, just as we learn that theft is bad social practice, at our mothers' knees as it were. Econometric norms, like social norms, are internalized in an environment in which explicit prohibitions, implicit example, and many subtle pressures to conformity mold our *morés*. Models of "good" econometric practice, stray remarks in textbooks or lectures, stern warnings from supervisors and referees, all teach us that data mining is abhorrent. All agree that theft is wrong, yet people steal . . .

---

[1]The adjective "LSE" refers to an tradition of time-series econometrics that began in the 1960s at the London School of Economics; see Mizon (1995) for a brief history. The practitioners of LSE econometrics are now widely dispersed among academic institutions throughout Britain and the world.

and they mine data. So, from time to time moralists, political philosophers and legal scholars find it necessary to raise the prohibition against theft out of its position as a background presupposition of social life and to scrutinize its ethical basis and to discriminate among its varieties and to categorize various practices as falling inside or outside the strictures that proscribe it. Similarly, the practice of data mining has itself been scrutinized only infrequently (e.g., Leamer 1978, 1983; Mayer 1980, 1993; Lovell 1983; Hoover 1995).

Lovell (1983) makes one of the few attempts that we know to evaluate specification search in a simulation framework. Unfortunately, none of the search algorithms that he investigates comes close to approximating LSE methodology. Still, Lovell's simulation framework provides a neutral test-bed on which we evaluate LSE methods, one in which there is no question of our having "cooked the books." Within this framework, we pose a straightforward question: *does the LSE approach work?*

I. Encompassing and the Problem of Data Mining

Specification Search

The relevant LSE methodology is the *general-to-specific modeling approach*.[2] It relies on an intuitively appealing idea. A sufficiently complicated model can, in principle, describe the economic world.[3] Any more parsimonious model is an improvement on such a complicated model if it conveys *all* of the same information in a simpler, more compact

---

[2] The LSE approach is described sympathetically in Gilbert (1986), Hendry (1987; 1995, esp. chs. 9-15), Pagan (1987), Phillips (1988), Ericsson, Campos and Tran (1990), and Mizon (1995). For more sceptical accounts, see Faust and Whiteman (1995) and Hansen (1996).

form. Such a parsimonious model would necessarily be superior to all other models that are restrictions of the completely general model except, perhaps, to a class of models nested within the parsimonious model itself. The art of model specification in the LSE framework is to seek out models that are valid parsimonious restrictions of the completely general model and that are not redundant in the sense of having an even more parsimonious model nested within them that also are valid restrictions of the completely general model.

The name "general-to-specific" itself implies the contrasting methodology. The LSE school stigmatizes much of common econometric practice as *specific-to-general*. Here one starts with a simple model, perhaps derived from a simplified (or highly restricted theory). If one finds econometric problems (e.g., serial correlation in the estimated errors) then one complicates the model in a manner intended to solve the problem at hand (e.g., one postulates that the error is AR(1) of a particular form, so that estimation using a Cochrane-Orcutt procedure makes sense).

The general-to-specific modeling approach is related to the theory of *encompassing*.[4] Roughly speaking, one model encompasses another if it conveys all of the information conveyed by another model and more besides. It is easy to understand the fundamental idea by considering two non-nested models of the same dependent variable. Which is better? Consider a more general model that uses the non-redundant union of the regressors of the two models. If model I is a valid restriction of the more general model

---

[3] This is a truism. Practically, however, it involves a leap of faith; for models that are one-to-one, or even distantly approach one-to-one, with the world are not tractable.

[4] For general discussions of encompassing, see, for example, Mizon (1984), Mizon and Richard (1986), Hendry and Richard (1987), Hendry (1988; 1995, ch. 14).

(e.g., based on an F-test), and model II is not, then model I encompasses model II. If model II is a valid restriction and model I is not, then model II encompasses model I. In either case, we know everything about the joint model from one of the restricted models; we therefore know everything about the other restricted model from the one. There is, of course, no necessity that either model will be a valid restriction of the joint model: each could convey information that the other failed to convey. In population, a necessary, but not sufficient, condition for one model to encompass another is that it have a lower standard error of regression.[5]

A hierarchy of encompassing models arises naturally in a general-to-specific modeling exercise. A model is tentatively admissible on the LSE view if it is congruent with the data in the sense of being: (i) consistent with the measuring system (e.g., not permitting negative fitted values in cases in which the data are intrinsically positive); (ii) coherent with the data in that its errors are innovations that are white noise as well as a martingale difference sequence relative to the data considered; and (iii) stable (cf. Phillips (1988, pp. 352-353); Mizon (1995, pp. 115-122); White (1990, pp. 370-374)). Further conditions (e.g., consistency with economic theory, weak exogeneity of the regressors with respect to parameters of interest, orthogonality of decision variables) may also be required for economic interpretability or to support policy interventions or other particular purposes, but they are not our focus here. If a researcher begins with a tentatively admissible general model and pursues a chain of simplifications, at each step maintaining admissibility and checking whether the simplified model is a valid restriction of the more

---

[5] Economists, of course, do not work with populations but samples, often relatively small ones. Issues about the choice of the size of the tests and related matters are as always of great practical importance.

general model, then the simplified model will be a more parsimonious representation of all the models higher on that particular chain of simplification and will encompass all of the models lower along the same chain.

Criticisms

The first charge against the general-to-specific approach as an example of invidious data mining points out that the encompassing relationships that arise so naturally apply only to a specific path of simplifications. There is no automatic encompassing relationship between the final models of different researchers who have wandered down different paths in the forest of models nested in the general model. One answer to this is that any two models can be tested for encompassing either through the application of non-nested hypothesis tests or through the approach described above of nesting them within a joint model. Thus, the question of which, if either, encompasses the other can always be resolved. Nevertheless, critics may object - with some justification - that such playoffs are rare and do not consider the entire range of possible termini of general-to-specific specification searches. We believe that this is an important criticism and we will return to it presently.

A second objection notes that variables may be correlated either because there is a genuine relation between them or because - in short samples - they are adventitiously correlated. Thus, a methodology that emphasizes choice among a wide array of variables based on their correlations is bound to select variables that just happen to be related in the particular data set to the dependent variables, even though there is no economic basis for the relationship. This is the objection of Hess, Jones and Porter (1994) that the general-

to-specific specification search of Baba, Hendry and Starr (1992) selects an "overfitting" model.

By far the most common reaction of critical commentators and referees to the general-to-specific approach questions the meaning of the test statistics associated with the final model. The implicit argument runs something like this: Conventional test statistics are based on independent draws. The sequence of tests (F- or t-tests) on the same data used to guide the simplification of the general model, as well as the myriad of specification tests used repeatedly to check tentative admissibility, are necessarily not independent. The test statistics for any specification that has survived such a process are necessarily going to be "significant". They are "Darwinian" in the sense that only the fittest survive. Since we know in advance that they pass the tests, the critical values for the tests could not possibly be correct. The critical values for such Darwinian test statistics must in fact be much higher; but just how much higher no one can say.

Responses

Hoover (1995) defends data mining in general against the more simple-minded versions of these criticisms. Common criticisms of data mining fail to distinguish adequately between measures of sampling distribution and measures of epistemic warrant. By "measures of epistemic warrant", we mean measures of our justification for believing a particular specification to be the truth or measures of the nearness of a particular specification to the truth. The two are completely different, and in some contexts that difference is clear. For example, in Monte Carlo experiments no question arises about the truth of the specification. It is true by construction, and what we seek to discover is its

sampling distribution. Similarly, when we bootstrap the standard errors of the coefficients in a regression equation, the issue is not whether this specification is the truth; but, given this specification, what does it imply for the sampling distribution.[6] Every specification is statistically correct in the sense that, if one wants to know the sampling properties of that distribution *per se*, one gets them from that specification. But not every specification is economically correct in the sense that it recapitulates the underlying process that generated the data. Measures of sampling distribution do not bear directly on economic correctness; they do not measure the distance between a particular specification and the true specification.

The goal of specification search or data mining is to find the true specification. Various methods search over possible specifications and judge one "best" according to some criterion. Suppose that such a search were successful, what would the sample statistics indicate? The sample statistics are distributed up to the their own sampling error (for they too are random variables) in whatever manner that one would derive from repeated realizations of the error terms. Interpreted this way, these distributions are independent of the search method or of how much search has been engaged in, for the appropriate counterfactual experiment is to hold the specification fixed and to re-estimate over many realizations of the error not to hold the errors fixed and re-estimate over many specifications.

To clarify the issues consider a data mining exercise that searches for the process

---

[6] On bootstrapping, see *inter alia* Jeong and Maddala (1993).

that generates **y**, where **y** = [y_t], an N x 1 vector of observations, t = 1, 2, . . . N. Let **X** = {$\mathbf{X}_j$}, j = 1, 2, . . . M, be the universe of variables over which a search might be conducted. Let $\mathbf{X}^P = \times\mathbf{X}$, the power set of **X**. If **y** were generated from a linear process, then the actual set of variables that generated it is an element of $\mathbf{X}^P$. Call this set of true determinants $\mathbf{X}_T \in \mathbf{X}^P$, and let the true data-generating process be:

$$\mathbf{y}^k = \mathbf{X}_T^k \beta_T + \omega^k , \tag{1}$$

where $\omega^k = [\omega_t^k]$, the vector of error terms, and k indicates the different realizations of both errors and the variables in **X**. Now, let $\mathbf{X}_i \in \mathbf{X}^P$ be any set of variables; these define a model:

$$\mathbf{y}^k = \mathbf{X}_i^k \beta_i + \varepsilon^k , \tag{2}$$

where $\varepsilon^k = [\varepsilon_t^k]$ includes ω^k, as well as every other factor by which equation (2) deviates from the true underlying process in equation (1). Typically in economics only one realization of these variables is observed; k is degenerate and takes only a single value. In other fields, for example in randomized experiments in agricultural and elsewhere, k truly ranges over multiple realizations, each realization, it is assumed, coming from the same underlying distribution. While in general, regressors might be random (the possibility indicated by the superscript k on $\mathbf{X}_i$), many analytical conclusions require the assumption that $\mathbf{X}_i$ remains fixed in repeated samples of the error term. What this amounts to $\mathbf{X}_i^k = \mathbf{X}_i^h \quad \forall k,h$, while $\varepsilon^k \neq \varepsilon^h, \quad \forall k \neq h$, except on a set of measure zero.

We can estimate the model in equation (2) for a given i and any particular realization of the errors (a given k). From such estimations we can obtain various sample statistics. For concreteness, consider the estimated standard errors that correspond to $\hat{\beta}_i$ the estimated coefficients of equation (2) for specification i.[7] What we would like to have are the population standard errors of the elements of $\hat{\beta}_i$ about $\bar{\hat{\beta}}_i$. Conceptually, they are the dispersion of the sampling distribution of the estimated coefficients while $\mathbf{X}_i$ remains fixed in repeated samples of the error term. Ideally, sample distributions would be calculated over a range of k's, and as k approached infinity, the sample distributions would converge to the population distributions. In practice there is a single realization of ε^k. While conceptually this requires a further assumption that the errors at different times are drawn from the same distribution (the ergodic property), the correct counterfactual question remains: what would the distribution be if it were possible to obtain multiple realizations with fixed regressors? Sample statistics derive their distribution from repeated sample of the residual within a constant specification.

Those critics of data mining who would penalize t-statistics or standard errors according to the amount of search, misconceive the issue when they treat sample statistics, not according to their natural interpretation (*viz.*, what is the distribution as k becomes large?), but as answering a different question: what is the distribution as $\mathbf{X}_i$ ranges over the elements of $\mathbf{X}^P$? Sample statistics obviously do not answer that question, but it is hoped that the penalties imposed will bring them closer to doing so. It is not, however,

---

[7] To keep the discussion simple, we will often speak of the standard errors and t-statistics of regression coefficients as exemplars of the sampling distribution. Our points are *mutatis mutandis* also relevant to other statistics.

even easy to specify a common set of statistics over the different $\mathbf{X}_i \in \mathbf{X}^P$. For example, let $\hat{\beta}_g^s$ be the coefficient on the variable $X_g \in \mathbf{X}$ for a particular specification $\mathbf{X}_s \in \mathbf{X}^P$. Since X_g will not be a member of every subset of $\mathbf{X}$ (i.e., of every element of $\mathbf{X}^P$), neither $\hat{\beta}_g^s$ nor its standard error or t-statistic will be defined in many cases.[8]

Criticisms of Darwinian tests statistics suppose that it makes sense to track the evolution of the coefficients on particular variables (e.g., the evolution of $\hat{\beta}_g^s$ the coefficient on $X_g \in \mathbf{X}_S$ as $\mathbf{X}_S$ ranges over the elements of $\mathbf{X}^P$). But that compares apples and oranges: the sampling distributions of any $\hat{\beta}_g^i$ and $\hat{\beta}_g^j$, the coefficients on X_g in the regressions in which $\mathbf{X}_i$ and $\mathbf{X}_j$ are regressors, are unaffected by whether or not the other regression or a regression using any other set of regressors including X_g has been run; the sampling distributions refer to repeated realizations of the errors, not to respecifications.

The point can be brought home with a parable. There are twin brothers. One has a theory that implies a particular specification like equation (2). He estimates it and reports the test statistics. The other has no theory, but searches over many specifications and finally lights upon the "best" specification, which happens to be the same one his twin estimates. The critics of Darwinian test statistics would say that the first twin reports the correct test statistics, while the second reports the "fittest," which should be discounted to reflect the degree of search. Yet, because they are the same specification, their test statistics answer the same counterfactual: what would the distributions be after repeated

---

[8] To avoid this problem, Leamer's (1983) extreme-bounds analysis eschews the sample statistics and considers only those elements of $\mathbf{X}^P$ that include a particular set of variables of interest. He then computes the maximum and the minimum estimates of the coefficients as the "extreme bounds" and treats these as measures of specification uncertainty for the particular variables of interest.

realizations of the errors? The critics treat the test statistics as measures of epistemic warrant, but there is no feature of those statistics that bears any special relationship to epistemic warrant; they reflect sampling distributions.

Lovell (1983, pp. 2-4) is an example of a critic of Darwinian test statistics, who argues that critical values must be adjusted to reflect the degree of search. Lovell considers a regression like equation (2) in which the elements of $\mathbf{X}$ are mutually orthogonal. He considers sets of regressors that include exactly two members (i.e., a fairly narrow subset of $\mathbf{X}^P$. The dependent variable $\mathbf{y}$ is actually purely random and unrelated to any of the variables in $\mathbf{X}$ (i.e., $\mathbf{X}_T = \varnothing$). Using 5-percent critical values, he demonstrates that one or more significant *t*-statistics occur more than five percent of the time. He proposes a formula to correct the critical values to account for the amount of search. Here Lovell slides between the interpretation of the *t*-test as a measure of sampling distribution and a measure of epistemic warrant: the critical values or the size of the test refers to the probability of a particular *t*-statistic on repeated draws from the same distribution (that is the significance of the textbook assumption that the regressors are fixed in repeated samples), while Lovell's experiment is to alter the distribution with every new $\mathbf{X}_i$. Lovell's numbers are correct, but he poses the wrong question.

Lovell is correct that the *t*-statistics are treacherous guides to the correct specification: high *t*-statistics do not tell us that we have a correct specification; and this is true whether they are high after a little search or a lot of search. Nevertheless, if we could be assured that we had the correct specification, they would be the correct guides to the accuracy of the coefficient estimates and should not be faced with "corrected" critical values as Lovell proposes. The *t*-statistic is a joint result of the correlation between the

dependent variable and the independent variables, the sample variation of the independent variable to which it refers, and the number of observations. An insignificant *t*-statistic may mean that a variable has no economic effect or that there has not been enough sample variation to make the effect observable; a significant *t*-statistic may mean that a variable has an economic effect or it may mean that the sample is too small to control for adventitious correlation. At its best, the *t*-statistic measures the ability of the regression to extract a signal from the background noise, but it does not guarantee that the signal will not be distorted to the point of uninterpretability.

An illustration: Suppose that variable *D* is in fact generated by a linear combination of mutually orthogonal variables *A*, *B* and *C* and a residual error. Suppose that the coefficient on *C* relative to its variance is small compared to the coefficients on *A* and *B* compared to their variances, and that the variance of C is relatively small compared to the variance of the error term. Now suppose that a regression is run of *D* on *A*, *B* and *C* and that the *t*-statistic on *C* is significant at a little over the 5 percent level. When regressors *A* and *B* are dropped, *C* becomes insignificant, not because it is not a determinant of *D* in fact, but because so weak a determinant either requires lots of data or good controls on other sources of variation (which is what *A* and *B* provide) in order to stand out. The *t*-statistic indicates how well the influence of *C* is measured, not how economically important it is. Equally, high *t*-statistics are no indication of correct specification. A regression of *B* on *A*, *C* and *D* might easily show a *t*-statistic on *D* higher than that on *B* in the economically correct regression of *D* on *A*, *B* and *C*.

Useful economic interpretation of regression results is limited by the fact that test statistics are measures of signal extraction, not measures of economic influence. The

tradition of regarding them as measures of economic influence begins with what Leamer (1978, p. 4) has called the axiom of correct specification, implicit in econometrics textbooks (e.g., in proofs of the Gauss-Markov theorem). The "axiom of correct specification" is the assumption that the model in question is the true model: that is that $\mathbf{X}_i = \mathbf{X}_T$. In one sense, it is always fulfilled; for if the questions that interest us are questions about the relationship between sample and population distributions of a particular specification (such as equation (2)), repeated draws from the same error distribution provide the correct answers. Test statistics would be relevant (though not decisive) in the role of measuring economic influence if we had the assurance that the specification were correct.

The idea of Darwinian test statistics arises, as it does for Lovell, because test statistics that are well-defined only under correct specification are compared across competing (and, therefore, necessarily not all correct) specifications. Similarly, the idea that an econometrician should follow a one-shot procedure in which a single, theoretically-derived specification is estimated and either accepted or rejected on the basis of the test statistics and not sequentially modified in light of the test results is again a concession to the axiom of correct specification. The equivocation between sampling and specification shows up here again in the guise of an equivocation on the adjective "correct." A stochastic property of the statistic is confused with a measure of epistemic warrant. The hope of the one-shot strategy is that theory can deliver the correct specification, so the only questions left concern measurement of the sampling properties. But this is a leap of faith without any justification. The economic correctness of a specification is a question of it matching the economic behavior of the world. On the one hand, as empiricists we

want our theories to be justified by facts, but where is such justification to come from if they are not to be modified in the face of empirical results. On the other hand, typical economic theories are not restrictive enough to choose a single specification as the only one consistent with economic principles. Thus, to derive tight specifications auxiliary assumptions are made. Different auxiliary assumptions produce different specifications. The problem of choosing among these specifications is just the problem of data mining - only acknowledged *sotto voce*.

The general-to-specific approach is more straight-forward. It accepts that choice among specifications is unavoidable, that an economic interpretation requires correct specification, and that correct specification is not likely to be given *a priori*. It conflates the problem of sampling and the problem of specification search - but in a benign way. Each specification is taken on probation. The questions asked are: *Are the sampling properties of this specification those that the true specification would have?* (For example, the true specification, by virtue of recapitulating the underlying data-generating process, should show errors that are white noise innovations.) *And, is this specification nested within a higher dimensional specification with the same sampling properties?* (That is, is the specification a parsimonious representation of a general specification that has the desirable statistical properties?) The idea of Darwinian test statistics is irrelevant because we are not tracking an "evolving" specification from context to context, but rather asking of different specifications how they perform when estimated *as if* in the context of a one-shot test.

White (1990, pp. 379-380) reports a remarkable theorem the upshot of which is this: for a fixed set of specifications and a battery of specification tests, as the sample size

grows toward infinity and increasingly smaller test sizes are employed, the test battery will - with a probability approaching unity - select the correct specification from the set. In such cases, White's theorem implies that type I and type II error both fall asymptotically to zero. White's theorem says that, given enough data, only the true specification will survive a stringent enough set of tests. This turns the criticism of Darwinian test statistics on its head. The critics fear that the survivor of sequential tests survives accidentally and therefore that the critical values of such tests ought to be adjusted to reflect the likelihood of an accident. White's theorem suggests that the true specification survives precisely because the true specification is necessarily, in the long run, the fittest specification.

Approaches that focus on sampling, say on the standard errors of the coefficients on a vector of variables miss the point. White's theorem suggests that we envisage the problem differently. An analogy is the fitting together of a jig-saw puzzle. Even if a piece duplicates another in part or all of its shape, as more pieces are put into place, the requirement that the surface picture as well as the geometry of the pieces cohere implies that each piece has a unique position. Inferences about the puzzle as a whole, or the piece in relation to the puzzle, can be made soundly only conditional on getting the pieces into their proper positions.

Although the logic of the general-to-specific approach is seen to be sound once one recognizes that its claims are conditional (i.e., "as if"), White's theorem is not enough to assure us that LSE methods generate good results in practice. To investigate its practical properties we use Lovell's simulation framework.

II. The "Mine": Lovell's Framework for the Evaluation of Data Mining

To investigate data mining in a realistic context Lovell (1983) begins with twenty annual macroeconomic variables covering various measures of real activity, government fiscal flows, monetary aggregates, financial market yields, labor market conditions and a time trend. These variables form the "data mine," the universe for the specification searches that Lovell conducts. The advantage of such a data set is that it presents the sort of naturally occurring correlations (true and adventitious, between different variables and between the same variables through time) that practicing macroeconomists in fact face.

The test-bed for alternative methods of specification search are nine econometric models. The dependent variable for each specification is a "consumption" variable artificially generated from a subset of between zero and three of the variables from the set of twenty variables plus a random error term. The random error term may be either independently normally distributed or autoregressive of order one. Except for one specification in which the dependent variable is purely random, the coefficients of Lovell's models were initially generated by regressing actual consumption on the various subsets of dependent variables or as linear combinations of models so generated. These subsets emphasize either monetary variables or fiscal variables. These coefficients are then used together with a random number generator to generate simulated dependent variables.[9]

For each of the nine specifications, Lovell created 50 separate artificial dependent "consumption" variables corresponding to 50 independent draws for the random error

---

[9] This was an attempt to add a bit of realism to the exercise by echoing the debate in the 1960's between Friedman and Meiselman and the Keynesians over the relative importance of monetary and fiscal factors in the economy. While this is no longer a cutting-edge debate in macroeconomics, that in no way diminishes the usefulness of Lovell's approach as a method of evaluating specification search techniques.

terms. For each of these replications he then compared the ability to recover the true specification of three algorithms searching over the set of twenty variables. The three algorithms were stepwise regression, maximum $\overline{R}^2$, and choosing the subset of variables for which the minimum *t*-statistic of the subset is maximized relative to the minimums of other subsets.

Lovell presents detailed analyses of the relative success of the different algorithms. A quick summary is provided in Table 1, which reproduces Lovell's (1983, p. 10) Table 7.[10] It is fair to say that the results were not in general favorable to the success of data mining. The best of the three algorithms, step-wise regression, chooses the correct variables only 70 percent of the time; with a nominal test size of 5 percent, it is, nevertheless, subject to a 30 percent rate of type I error.

To evaluate the general-to-specific approach, we modify Lovell's framework in three respects. First, we update his data to 1995. Using annual observations, as Lovell does, we repeated his simulations and found closely similar results on the new data set. Second, we substituted quarterly for annual data for each series to render the data similar to the most commonly used macroeconomic time-series. Again, we repeated Lovell's simulations on quarterly data and found results broadly similar to his. Finally, it has become more widely appreciated since Lovell's paper that numerous econometric problems arise from failing to account for the fact that many macroeconomic series are nonstationary.[11] To avoid the issues associated with nonstationarity and cointegration, we

---

[10] Although the right-hand most column provides a hint of things to come, the reader should ignore it for the time being.

[11] For surveys of nonstationary econometrics, see Stock and Watson (1988), Dolado, Jenkinson and Rivero (1990); Campbell and Perron (1991); Banerjee (1995).

differenced each series as many times as necessary to render it stationary as indicated by Dickey-Fuller tests. The universe of data for the evaluation of the general-to-specific approach is reported in Table 2. Notice that there are now only 18 primary variables reported: the time trend (one of Lovell's variables) is no longer relevant because the data are constructed to be stationary; furthermore, because of limitations in the sources of data, we omit Lovell's variable "potential level of GNP in $1958".[12] Corresponding to each of the variables 1-18 are their lagged values numbered 19-36. In addition, variables 37-40 are the first to fourth lags of the "consumption" variable.[13] Table 3 is the correlation matrix for variables 1-18 plus actual personal consumption expenditure.

Table 4 presents nine models constructed in the same manner as Lovell's but using the new stationary, quarterly data-set.[14] Model 1 is purely random. Model 3 takes the log of simulated consumption as the dependent variable and is an AR(2) time-series model. Model 4 relates consumption to the M1 monetary aggregate; model 5 to government purchases; and model 6 to both M1 and government purchases. The dynamic models 2, 7, 8, and 9 are the same as the static models 1, 4, 5, and 6 except that an AR(1) error term replaces the identically, independently normally distributed error term. The principal question of this paper is, how well does the general-to-specific approach do at recovering these nine models in the universe of variables described in Table 2?

---

[12] We also replaced Lovell's variables "index, 5 coincident indicators" with "index, 4 coincident indicators" and "expected investment expenditure" with "gross private investment."

[13] As lags of the artificially generated dependent variables, these variables differ from model to model in the simulations below. Actual personal consumption expenditure is used in calibrating the models in Table 4.

[14] All simulations are conducted using Gauss386i Ver. 3.01 and its normal random number generator.

II. The "Mining Machine": An Algorithm for a General-to-Specific Specification Search

The practitioners of the general-to-specific approach usually think of econometrics as an art, the discipline of which comes not from adhering to recipes but from testing and running horse-races among alternative specifications. Nevertheless, in order to test the general-to-specific approach in Lovell's framework we are forced to first render it into a mechanical algorithm. The algorithm that we propose is, we believe, a close approximation to a subset of what practitioners of the approach actually do.[15] A number of their concerns, such as appropriate measurement systems and exogeneity status of the variables, are moot because of the way in which we have constructed our nine test models. Also, because we have controlled the construction of the test models in specific ways, considerations of compatibility with economic theory can be left to one side.

The search algorithm can be described as follows.

The Search Algorithm

A. The data run 1960.3 to 1995.1. *Candidate variables* include current and one lag of independent variables and four lags of the dependent variable. A *replication* is creation of a set of simulated consumption values using one of the nine models in Table 4 and one draw from the random number generator. *Nominal size* governs the conventional critical values used in all of the tests employed in the search: it is either 1, 5, or 10 percent.

B. A *general specification* is estimated on a replication using the observations from 1960.3 to 1995.1 on the full set of candidate variables, while retaining the observations from 1990.2 to 19995.1 for out-of-sample testing. The following battery of tests is run on the general specification:

a. normality of residuals

---

[15] See, in addition to the general discussions as indicated in fn. 1 above, Hendry and Richard (1982), White (1990) and Hendry (1995, ch. 15).

b. autocorrelation of residuals up to second order[16]
c. autocorrelated conditional heteroscedasticity (ARCH) up to second order
d. in-sample stability test (Chow test of first half of the sample against the second half)
e. out-of-sample stability (Chow test of specification estimated against re-estimation using 20 data points retained for the test.

If the general specification fails any one of the tests at the nominal size, then this test is not used in subsequent steps of the specification search for the current replication only. If the general specification fails more than one test, the current replication is eliminated and the search begins again with a general specification of a new replication.

C. The variables of the general specification are ranked in ascending order according to their *t*-statistics. For each replication, ten search paths are examined. Each path begins with the elimination of one of the variables in the subset with the ten lowest (insignificant) *t*-statistics. The first search begins by eliminating the variable with the lowest *t*-statistic and re-estimating the regression. This re-estimated regression becomes the *current specification*. The search continues until it reaches a *terminal specification.*

D. Each current specification is subjected to the battery of tests described in step B with the addition of:

f. An *F*-test of the hypothesis that the current specification is a valid restriction of the general specification.

If the current specification passes all of the tests, the variable with the next lowest *t*-stat is eliminated. The resultant current specification is then subjected to the battery of tests. If the current specification fails any one of these tests, the last variable eliminated is replaced and the current specification is re-estimated eliminating the variable with the next lowest insignificant *t*-statistic. The process of variable elimination ends when a current specification passes the battery of tests and either has all variables significant or cannot eliminate any remaining insignificant variable without failing one of the tests.

F. The resultant specification is then estimated over the full sample.
I. If all variables are significant the current specification is the terminal specification.

---

[16] In using AR(2) and ARCH(2) tests we trade on our knowledge that for every model except Model 3, which has a two period lag, the longest true lag is only one period. As the number of search variables increases with the number of lags, tractability requires some limitation on our models. Given that fact, the limitation of the test statistics to order 2 is probably harmless.

II. If any variables are insignificant, they are removed as a block and the battery of tests is performed.

a. If the new model passes and all variables are significant the new model is the terminal model and go to G.

b. If the new model does not pass, replace the block and go to G.

c. If the new model passes and some variables are insignificant, return to II.

G. After a terminal specification has been reached, it is recorded and the next search path is tried until all ten have been searched.

H. Once all ten search paths have ended in a terminal specification, the *final specification* for the replication is the terminal specification with the lowest standard error of regression.

The general-to-specific search algorithm here is a good approximation to what actual practitioners do with the exception, perhaps, of the explicit requirement to try several different search paths. We added this feature because preliminary experimentation showed that without it the algorithm very frequently got stuck far from any sensible specification. While in this respect our attempt to mechanize LSE econometric methodology may have in fact suggested an improvement to the standard LSE practice, we do not regard this modification as invidious to that practice or as a particularly radical departure. Typically, LSE practitioners regard econometrics as an art informed by both econometric and subject-specific knowledge. We have no way of mechanizing individual econometric craftsmanship. We regard the use of multiple search paths as standing in the place of two normal LSE practices that we are simply unable to model in a simulation study: First, LSE practitioners insist on consistency with economic theory to eliminate some absurd specifications. Since we control the data-generating processes completely, there is no relevant theory to provide an independent check. Second, LSE practitioners typically require that final specifications encompass rival specifications that may or may

not have been generated through a general-to-specific search. While the ultimate goal is, of course, to find the truth, the local, practical problem is to adjudicate between specifications that economists seriously entertain as possibly true. We have no set of serious rival specifications to examine. But if we did, they would no doubt reside at the end of different search paths; so we come close to capturing the relevant practice in considering multiple search paths.[17]

IV. Does the General-to-Specific Approach Pick the True Specification?

To assess the general-to-specific approach we follow Lovell in conducting a specification search for 50 replications of each of the nine specifications listed in Table 4. Specifications could be evaluated as either picking out the correct specification or not. We believe, however, that acknowledging degrees of success provide a richer understanding of the successes and failures of data mining. We present the results in five categories. Each category compares the final specification with the correct specification. The sensibility of the encompassing approach informs the categories. It is a necessary condition that the standard error of regression for an encompassing specification is lower (in population) than every specification that it encompasses. Thus, in population, the true specification must have the lowest standard error of regression. We use this criterion in our search algorithm; but, unfortunately, it need not be satisfied in small samples. We

---

[17]There is nothing sacred about ten paths; it is an entirely pragmatic choice. The proof of the pudding is in the eating. The simulation data themselves suggest that we would not do substantially better if we considered every possible path: there turn out to be few rejections in which the true model dominates the final model. One reason for not trying every path is that to do so would emphasize the mechanical nature of what is in practice not a mechanical procedure.

therefore ask: Does the algorithm find the correct model? If not, does it fail because the small sample properties of the data indicate that a rival specification is statistically superior or because the algorithm simply misses? The latter is a serious failure; the former, especially if the true specification, is nested within the chosen specification, is a near success. We focus on the question of whether or not the true model is nested within the chosen model, because ideally the algorithm would always select the true regressors (i.e., have high power), but nevertheless is subject to type I error (i.e., it sometimes selects spurious additional regressors). The five categories are:

Category 1: *The correct specification is chosen.* (The algorithm is an unqualified success.)

Category 2: *The correct specification is nested in the chosen specification and the chosen specification has the lower standard error of regression.* (The algorithm has done its job perfectly, but it is an (adventitious) fact about the data that additional regressors significantly improve the fit of the regression. The chosen specification appears to encompass the correct specification and there is no purely statistical method of reversing that relationship on the available data set.)

Category 3: *The correct specification is nested in the chosen specification and the correct specification has the lower standard error of regression.* (The algorithm fails badly. Not only does the correct specification in fact encompass the chosen specification, but it could be found if the algorithm had not stopped prematurely on the search path.)

Category 4: *An incorrect specification is chosen, the correct specification is not nested in the chosen specification, and the chosen specification has a lower standard error of regression than the correct specification.* (The algorithm fails to pick the correct specification, but does so for good statistical reasons: given the sample the chosen specification appears to encompass the true specification. It is like category 2 except that, rather than simply including spurious variables, it (also) omits correct variables.)

Category 5: *An incorrect specification is chosen, the correct specification is not nested in the chosen specification, and the correct specification has a lower standard error of regression than the chosen specification.* (This is, like category 3, a serious failure of the algorithm - even worse, because the chosen

specification does not even define a class of specifications of which the true specification is one.)

These categories are still too coarse to provide full information about the success of the algorithm. Even category 5 need not always represent a total specification failure. It is possible that a specification may not nest the correct specification but may overlap with it substantially - including some, but not all, of the correct variables, as well as some incorrect variables. We will therefore track for each replication how many times each correct variable was included in the final specifications, as well as the number of additional significant and insignificant variables included.

A Benchmark Case: Nominal Size 5 Percent

Table 5 presents the results of specification searches for 50 replications on nine specifications for nominal size of 5 percent (i.e., the critical values based on this size are used in the test battery described in step D of the search algorithm described in Section III).[18] A 5 percent size, as the most commonly used by empirical researchers, will serve as our benchmark case throughout this investigation. According to Table 5, the general-to-specific search algorithm chooses exactly the correct specification (category 1) only a small fraction of the time: on average over nine models almost 9 in 50 replications. Its

---

[18] Models 2, 7, 8, 9 involve an AR(1) error term of the form $u_t^* = \rho u_{t-1}^* + u_t$. Each of these models can be expressed as a dynamic form subject to common-factor restrictions. Thus if $y_t = \mathbf{X}_t\beta + u_t^*$, this is equivalent to (a) $y_t = \rho y_{t-1} + \mathbf{X}_t\beta - \mathbf{X}_{t-1}(\rho\beta) + u_t$, so that an estimated regression conforms to (a) if it takes the form (b) $y_t = \pi_1 y_{t-1} + \mathbf{X}_t\Pi_2 - \mathbf{X}_{t-1}\Pi_3 + u_t$, subject to the common-factor restriction $\pi_1\Pi_2 = -\Pi_3$. We present the alternative expressions of the models as Models 2', 7', 8' and 9'. Although many LSE econometricians regard the testing of common-factor restrictions an important element in specification search, we count a search successful if it recovers the model in the form (b) without regard to the common-factor restriction. See Hoover (1988) and Hendry (1995. ch. 7, section 7), for discussions of common-factor restrictions.

success rate varies with the model, the best results (17 of 50) found for Models 1, 5, and 8. Models 3 and 4 have nearly as good a success rate; while Models 2 and 9 strike out and Models 6 and 7 place only 1 of 50 in category 1.

Still, the general-to-specific algorithm is by no means a total failure. Most of the specifications are classed in category 2: the final specification is overparameterized relative to the true model, but that is the best one could hope to achieve on purely statistical grounds, because the chosen final specification in fact statistically encompasses the true specification. On average 30 of 50 searches end in category 2 and nearly 39 of 50 in categories 1 and 2 combined. If category 2 is a relative success, the price is overparameterization: an average of just over two extra variables spuriously and significantly retained in the specification. (In addition, one extra insignificant variable is retained on average.) In one sense, this is bad news for the search algorithm as it suggests that searches will quite commonly include variables that do not correspond to the true data-generating process. But, we can look at it another way. Each falsely included (significant) variable represents a case of type I error. The search is conducted over 40 variables and 50 replications. The table represents the empirical rate of type I error (size) for the algorithm: on average 5.7 percent, only slightly above the 5 percent nominal size used in the test battery.

These averages mask considerable variation across models. At one extreme, every search over Models 1 and 2 ends in category 1 or 2; at the other extreme only 5 of 50 searches over Model 9 end in category 2 and none in category 1. In between 47 and 49 searches end in categories 1 and 2 for Models 3, 4, 5, 7, and 8. Model 6, like model 9, is hard to find or even to nest in the chosen model. So, how do these models fail?

Weak Signals, Strong Noise

Searches for both Models 6 and 9 most frequently end in category 4: the true specification is not nested within the final specification, but the final specification (statistically) encompasses the true specification. This suggests, not a failure of the algorithm, but unavoidable properties of the data. Table 5 indicates that Models 6 and 9 correctly choose most of the true variables most of the time, but that they appear to have special difficulty in capturing government purchases of goods and services (Variable 3) or its first lagged value (Variable 21). We conjecture that the difficulty in this case is that these variables have relatively low variability compared to the dependent variables and the other true independent variables in Models 6 and 9. They therefore represent a common and unavoidable econometric problem of variables with a low signal-to-noise ratio.[19] It is always problematic how to discriminate between cases in which such variables are economically unimportant and cases in which they are merely hard to measure.

Consider Model 6 in more detail. The signal-to-noise ratio in the true model can be defined as $|\text{coefficient}_j \times \text{standard deviation of independent variable}_j| /$ standard deviation of the random error. In Model 6, the signal-to-noise ratio for Variable 3 is 0.04, while for Variable 11 (the M1 monetary aggregate) it is 1.16. By adjusting the coefficient value for Variable 3, the signal-to-noise ratio can be increased. We formulate two additional models (6A and 6B) in which the coefficient on Variable 3 is raised (in absolute

---

[19] The reader will notice that in Models 5 and 8, these variables appear to present no special difficulties. There is, however, no paradox. The relevant factors are not only the absolute variability of the dependent variable, but also the size of the coefficient that multiplies it; and these must be judged relative to the other independent variables in the regression, as well as to the dependent variable (and therefore, finally, to the error term). The fact that these variables are easily picked up in cases in which there are no competing variables merely underlines the fact that it is the *relative* magnitudes that matter.

value) from -0.02 to -0.32 and to -0.67, yielding signal-to-noise ratios of 0.58 (half of that for Variable 11) and 1.16 (the same as that for Variable 11). Table 6 presents the results of 50 replications of the search at a nominal size of 5 percent for Models 6, 6A, and 6B. With even half the signal-to-noise ratio of Variable 11, the final specification for Model 6A ends up 43 of 50 replications in categories 1 or 2, and Variable 3 is correctly selected 44 of 50 times. With an equal signal to noise ratio, the final specification for Model 6B ends up 50 of 50 replications in categories 1 and 2, and Variable 3 is selected correctly 50 of 50 times.

Estimated standard errors fall as sample size increases. We might then imagine that the general-to-specific search algorithm would itself pick up weaker signals better as sample size increases. To test this proposition, we expanded the data set to five times its original length (i.e., to 695 observations). To preserve time series properties, we resampled from the original data using a procedure analogous to that employed in block bootstrapping (see, e.g., Li and Maddala (1995)). Table 7 presents the results of 50 replications over nine models with a nominal size of 5 percent. Comparing Table 7 to Table 5, we see that the power is higher for every model except Model 1, for which it is an irrelevant measure as there are no true regressors, and Model 6. In fact for six of the nine models it achieves the maximum 100 percent. But power was relatively high to begin with, and the average increase is small (90.1 percent compared to 89.4 percent). The results suggest that more data alone is not a solution to the problem of low signal-to-noise ratios: the problematic Variables 3 and 21 in Models 6 and 9 are not chosen more frequently on the longer data set.

Predictably as the standard errors fall with the longer sample size, the number of type I errors for a fixed nominal size rises. The number of falsely included significant regressors has nearly doubled (an average of 3.84 compared to 2.23), which is reflected in the true size almost doubling (10.3 percent compared to 5.7 percent). As a practical matter, this reinforces the view that smaller nominal sizes should be used as sample size increases, which is consistent with White's theorem (see Section III above).

Size and Power

How do the properties of the general-to-specific search algorithm change as the nominal size used in the test battery changes? Tables 8 and 9 present analogous results to those in Table 5 (nominal size 5 percent) for nominal sizes of 10 percent and 1 percent.

Some general patterns are clear in comparing the three tables. As nominal size falls the number of final specifications in category 1 rises sharply from an average of under 2 of 50 at a nominal size of 10 percent to an average of over 16 of 50 at a nominal size of 1 percent. At the same time, the relationship between nominal size and category 2 is direct not inverse, so that the total in categories 1 and 2 together is lower (average almost 31 of 50) for a nominal size of 1 percent than for a nominal size of 5 percent (almost 39 of 50) or 10 percent (just over 38 of 50). Similarly, a smaller nominal size sharply reduces the average number of both falsely significant variables and retained insignificant variables. All these features are indications of the tradeoff between size and power. The average true size corresponding to a 10 percent nominal size is 10.1 percent - almost identical - and is associated with an average power of 89.9 percent. The true size corresponding to a nominal size of 5 percent is nearly the same, 5.7 percent; but the reduction in size also implies a slight loss of power (down to 89.4 percent): the smaller size implies fewer cases

of incorrectly chosen variables, but more cases of omitted correct variables. The true size corresponding to a nominal size of 1 percent is almost double at 1.9 percent; and there is a further loss of power to 86.5 percent. The tradeoff between size and power seems to be pretty flat, although as nominal size becomes small the size distortion becomes relatively large. This may argue for a smaller conventional size in practical specification searches than the 5 percent nominal size commonly used.

V. How Does the General-to-Specific Approach Compare to Alternative Search Methods?

Table 1 reports Lovell's summary of his results for three search algorithms with the addition of a column corresponding to the general-to-specific search algorithm with a nominal size of 5 percent. Lovell's summary is restricted to four static models (1, 4, 5, and 6). The last column is computed for the corresponding four models from Table 5. These are very rough comparisons, as Lovell's results and those of Table 5 are based on different data sets as described in Section II above.

The general-to-specific algorithm dominates Lovell's Max-R^2 and Max-min $|t|$ algorithms. Stepwise regression provides closer competition, but, here too, the general-to-specific algorithm appears better: its true size (on these four models) is actually smaller than the nominal size and 1/7 the true size of stepwise regression; it correctly identifies the true coefficients 13 percent more frequently (79 versus 70 percent of the time).[20]

---

[20] Lovell reports type II error in the last row of the table: the frequency with which the null of zero coefficients are accepted when it is in fact false. Since the true hypothesis for Model 1 (which is just a random error) is in fact a zero coefficient, there can be no type II error for that model. If a selection method chooses the wrong variable but finds that variable to be significant, no type II error will be recorded (the null of a zero coefficient is rejected albeit for the wrong reason). This explains why the max-min $|t|$ commits no type II error even though it *never* choose the correct variables. We might more usefully count the number of times each method fails to select the correct variables (this is the complement

VI. What Do Test Statistics Mean After Extensive Search?

The most common doubt expressed about the final specifications reported from general-to-specific specification searches is over the interpretation of test statistics. How are we to interpret the *t*-statistics of a regression that involves massive (and not easily quantified) amounts of pre-test selection and (it is pejoratively but wrongly argued) arbitrarily directed search? Should we not, following Lovell for example, discount the test statistics in proportion to the degree of search? In one sense, the answer to this is just to point out (as in Section I above) the equivocation of sampling properties with measures of epistemic warrant. The commonly computed test statistics of regressions correctly measure sampling properties independently of their provenance. The connection between these properties and the verisimilitude or lack of verisimilitude of a regression requires further argument. Many anti-data-mining arguments are pure assertion, or trade on the previously mentioned equivocation. To recognize that, however, does not establish the effectiveness of any particular specification search algorithm.

It would be desirable to be assured that an algorithm converged on the true data-generating process; for in that case, the sampling properties of the final specification would be the sample properties of the true specification. The results of the previous section, however, indicate a number of pitfalls that might vitiate the success of the general-to-specific algorithm. It is only relatively infrequently that it converges on the exactly

---

of the power measures reported in Tables 5-10 of the current paper). On this basis, max-min $|t|$ has a 100 percent error rate, while the general-to-specific algorithm has a 14 percent error rate. It is not possible from the information in Lovell's (1983) tables to compute the error rate for the other algorithms as one would need to know both which variables were selected and what their *t*-statistics were.

correct specification. Commonly, a relatively large number of extra significant regressors are included in the final specification, and extra insignificant regressors are often apparently needed to obtain desirable properties for the estimated residuals. In the face of these common departures from a precise match between the chosen final specifications and the true specification, the question posed in this section is, to what degree does the final specification reflect the sampling properties of the true specification?

To investigate this question we conduct specification searches on 1000 replications of Model 9. Model 9 was chosen because it is the most difficult of Lovell's nine models for the search algorithm to uncover. It is both a dynamic model and one that suffers from low signal-to-noise ratios for some of its variables. Table 10 presents the results of this exercise for the universe of variables in Table 2 for searches with a nominal size of 5 percent.

Although every variable in the universe of search is chosen in some replications and therefore have nonzero mean values, incorrect inclusion is relatively rare. This shows up in the fact that every correctly excluded coefficient except that on Variable 38 has a median value of zero. Variable 38 is the second lag of the dependent variable (artificial consumption expenditures). It is chosen (incorrectly) in over 88 percent of the replications, while its brother the (correct) first lag (Variable 37) is chosen in nearly 100 percent of the replications. We will return to this phenomenon presently.

Concentrating now on the properly included variables, we measure the accuracy of the estimates as the absolute values of the mean and median coefficient biases as a percentage of the true value. Variable 11 appears to be fairly accurately measured with mean bias of 2.4 percent and median bias of 3.1 percent. The biases of Variables 29 and

37 are substantially higher but still moderate. In contrast, the two variables with low signal-to-noise ratios (Variables 3 and 21) have very large mean biases of 107 percent and 75 percent and median biases of 100 percent.

To evaluate the interpretation of *t*-statistics, we kept track of the estimated *t*-statistics for each final specification. We measured the type I error for the properly excluded variables as the number of times that the *t*-statistic was outside the 95 percent confidence interval (i.e. the number of times a variable was improperly included with $|t\text{-statistic}| > 1.96$) and the type II error for the properly included variables as the number of times the *t*-statistic was inside the 95 percent confidence interval. From these data we can compute the empirical size and power of the *t*-test against the null hypothesis that the coefficient on a variable is zero (exclusion of a variable from the search is treated as equivalent to a coefficient value of zero).

The empirical sizes of the properly excluded variables range from 3 percent (Variable 10) to 87 percent (Variable 38), with an average size of about 8.5%. Variable 38 is the second lagged value of the dependent variable. This variable, as we noted previously, is the only variable that is incorrectly chosen more often than not. It is highly correlated with the first lag of the dependent variable (correlation coefficient 0.75).[21] This multicollinearity is the likely source of the large empirical size. While we should regard this example as a warning of one of the pitfalls of dynamic specification search, it may say more about the inadequacy of our algorithm in mimicking the recommended practice of

---

[21] The correlation is measured using actual personal consumption expenditure rather than the simulated dependent variable, which varies from replication to replication. The correlation should be close in any case.

the LSE approach. The LSE methodology stresses the importance of *orthogonal* regressors and the need to find reparameterizations to insure orthogonality. If we exclude the three properly excluded lags of the dependent variable (Variables 38, 39, and 40), then the range of empirical sizes for the remaining properly excluded variables is 3 percent to just under 9.5 percent, and the average empirical size is 5.4 percent, very close to the nominal size of 5 percent used in the search algorithm.

Since we know the true specification of Model 9, it is possible to compute the power against the null that the coefficient on any properly included variable is zero for any single replication. In order to account for the fact that the dependent variable (and its lagged value) varies with each replication, we compute the power from 1000 replications and estimates of the true model. This is indicated in Table 10 as the "True Power." We compare the estimated empirical power of the search algorithm against this true power. While the empirical power varies tremendously with the variable (100 percent for Variable 11 but just over 9 percent for Variable 21), there is a close conformity between the empirical power and the true power. The largest discrepancy occurs with Variable 29 (the first lag of Variable 11, the M1 monetary aggregate), which has an empirical power of 86 percent against a true power of 100 percent. Once again this may be the result of the high correlation between the current and lagged values of the variable (correlation coefficient = 0.682).

In summary, the size and power of final specifications from the general-to-specific search algorithm provide very good approximations to the size and power of the true specifications. We have also conducted, but do not report here, two further sets of 1000

replications for nominal sizes of 10 percent and 1 percent. The results are similar in character to those in Table 10.

VII. The Problem of Overfitting: An Extension to the LSE Methodology

Our investigations confirm the worry of some critics who believe that the general-to-specific search results in overparameterized models. Final specifications, more often than not, retain incorrectly significant variables and insignificant variables that appear to be needed to induce sensible properties in the error terms. Given that we have shown that the empirical size and power of t-tests are not much distorted by the search procedure, this is perhaps of less concern than it first appears. Furthermore, the problem appears to be substantially mitigated through the use of smaller nominal sizes in the search procedure. We have shown that the cost of using smaller nominal sizes in terms of power is relatively small. Thus, as well as evaluating the LSE approach, we make a constructive suggestion that practitioners should prefer smaller nominal test sizes.

Type I error in the search process occurs because the data possess adventitious properties in small samples. By their very nature these properties should not remain stable across subsamples. This suggests a possible method of reducing the number of incorrectly retained significant variables (i.e., reducing the empirical size of the algorithm), which, to the best of our knowledge, is not generally practiced by LSE econometricians, but which is consistent with the general philosophy of the LSE methodology. We consider splitting the sample into two (possibly overlapping) subsamples - one running from the beginning of the sample to a point some fraction of the way to the end, the second running from the end of the sample some fraction of the way backwards to the beginning. If, for example, the fraction is 1/2, the subsamples are the first half and the second half of the sample and

they do not overlap. If the fraction is 2/3, the subsamples are the first 2/3 and the last 2/3 of the sample and the two subsamples have the middle third in common. We run a modified version of the search algorithm on each subsample. The final model is then the intersection of the two subsample models; that is, only variables that are chosen in both subsamples appear in the final model, on the grounds that the others are there by accidents of the data.

The algorithm of Section II above is modified by omitting step B.d, the in-sample Chow test for coefficient stability and reducing the number of data points retained for out-of-sample stability testing in step B.e from 20 to 10. Both modifications are pragmatic responses to the loss of degrees of freedom from the use of shorter subsamples.

For 50 replications of the nine models with subsamples of 1/2 the data set, the average number of falsely included significant variables is only 0.45 compared to 2.23 for the full data set in Table 5. This is a fall in the empirical size to 1.2 percent from 5.7 percent. This improvement in size, however, comes at the cost of great loss of power: 69.7 percent compared to 89.4 percent for the full sample.

Figure 1 plots the tradeoff between size and power for subsamples with different degrees of overlap based on 50 replications of the nine models. The tradeoff is nonlinear: the highest power occurs naturally with the full undivided sample; the loss of power is relatively small up to the point at which the subsamples are 3/4 of the full sample and then falls rapidly to the point where the subsamples split the full sample without overlap.[22] The

---

[22] Two features of the tradeoff locus should be noted. First, the little pit associated with the subsamples of 6/7 and 7/8 of the sample compared to the adjacent 5/6 and 8/9 is probably the result of sampling variability from the relatively small number of replications. Second, the size associated with the full sample (the 1/1 split) at 6.8 percent is somewhat greater than the full sample size of 5.7 percent reported

tradeoff locus can be regarded as a possibility frontier, and an investigator's loss function would rank the various possibilities (higher indifference curves would lie to the northwest). Obviously, any point along the locus (except the concave pit (see fn. 21)) are conceivable optima. Still, for a large class of loss functions the kink at 3/4 would prove to be the optimum. At that point the average size is 1.8 percent (less than a third of the size reported in Table 5), and the average power is 83.6 percent (a loss of only 2.8 percentage points or about 3 percent compared to the power reported in Table 5). With a well chosen subsample split, the modified algorithm produces a large improvement in size (reduction in overparameterization) for a small loss of power.

VIII. Data Mining in Retrospect . . . and Prospect.

The results of our investigation of the general-to-specific search algorithm should be reasonably heartening to practitioners of the LSE approach. We have demonstrated that the general-to-specific approach performs better than the alternative data-mining methods stigmatized in Lovell's (1983) article. Beyond that, we have shown that the empirical size and power of specifications produced from general-to-specific searches, with one caveat, conform well to the theoretical size and power one would expect if one knew the true specification *a priori*. Test statistics based on such searched specifications therefore bear the conventional interpretation one would ascribe to one-shot tests. Of course, estimated standard errors are measures of sampling characteristics, not of epistemic virtue; this remains true with a searched specification. To state again a point

---

in Table 2. This is the result of the changes in the algorithm made for these simulations, which have the effect of making it a less stringent filter on each subsample separately.

made at the beginning of the paper, a *t*-statistic may be insignificant either because a variable is economically unimportant or because it has a low signal-to-noise ratio or small sample. The searched specification may, nevertheless possess epistemic virtues not open to the one-shot test: since the correct specification necessarily encompasses all incorrect specifications, the fact that the searched specification is naturally nested within a very general specification that in turns nests a wide class of alternative specifications strengthens the searched specification as a contender for the place of model-most-congruent-to-the-truth. The evidence of strength is not found in the *t*-statistics, but in the fact of the Darwinian survival of the searched specification against alternatives and in its natural relationship to the general specification.

The one caveat is that our evidence shows that size certainly and, to a lesser extent, power are distorted for lags of (especially the dependent) variables of the true specification. This appears to have to do with failures of orthogonality. At a minimum, it reminds the practitioner why the LSE approach stresses the importance of orthogonality and special care with respect to dynamic specification.

While generally supportive of the LSE approach, this study was able to confirm the risk often asserted by critics that practical general-to-specific searches could turn into arbitrary wanderings in the maze of specification possibilities that might terminate arbitrarily far from the correct specification. While the LSE approach in fact incorporates a number of elements (ignored in our mechanical rendering of the search procedure) that protect against false termini, we found that the simple expedient of trying a number of initial starting points in the search gave very good results. We recommend this to practitioners.

Finally, we would like to pursue two further extensions of the current study. First, we have restricted the models to stationary data. In the past decade, it has become increasingly important in macroeconometrics to deal with non-stationary data. Practitioners of the LSE approach were early contributors to this development, stressing the importance of error-correction modeling long before cointegration had been named or its intimate relationship to error-correction models understood. It is therefore natural that we should attempt to evaluate the success of the general-to-specific approach in non-stationary contexts.

Finally, an important alternative view of specification is provided by Leamer (1978, 1983). Leamer regards specification search as inevitable and makes a particular proposal, "extreme-bounds analysis," to guide practitioners on the epistemic virtues of estimated regressions. It would be useful to conduct a detailed comparison of the two approaches.[23]

---

[23] There are already several articles critical of Leamer's approach from an LSE perspective; see, for example, McAleer, Pagan and Volker (1985) (and Leamer's (1985) reply), Mizon and Hendry (1990), and Pagan (1987).

References

Baba, Yoshihisa, David F. Hendry, and Ross M. Starr (1992) "The Demand for M1 in the U.S.A.," *Review of Economic Studies*, 59, January, 25-61.

Banerjee, Anindya. (1995) "Dynamic Specification and Testing for Unit Roots and Cointegration," in Kevin D. Hoover (ed.) *Macroeconometrics: Developments, Tensions and Prospects*. Boston: Kluwer.

Campbell, John Y. and Philip Perron (1991) "Pitfalls and Opportunities: What Macroeconomists Should Know About Unit Roots," in Olivier J. Blanchard and Stanley Fischer (eds.) *NBER Macroeconomics Annual 1991*. Cambridge, MA: MIT Press.

Dolado, J., T.J. Jenkinson, and S.-S. Rivero (1990) "Cointegration and Unit Roots," *Journal of Economic Surveys* 4, 249-273.

Ericsson, Neil R., Julia Campos and H-A Tran (1990) "PC-GIVE and David Hendry's Econometric Methodology," *Revista de Econometrica* 10, 7-117.

Faust, Jon and Charles H. Whiteman. (1995) "Commentary [on Grayham E. Mizon "Progressive Modeling of Macroeconomic Times Series: The LSE Methodology]," in Kevin D. Hoover (ed.) *Macroeconometrics: Developments, Tensions and Prospects*. Boston: Kluwer, pp. 171-180.

Gilbert, Christopher L. (1986) "Professor Hendry's Econometric Methodology," *Oxford Bulletin of Economics and Statistics*, 48, August, 283-307. Reprinted in Granger (1990).

Granger, C.W.J.(ed.)(1990) *Modelling Economic Series: Readings in Econometric Methodology*. Oxford: Clarendon Press.

Hansen, Bruce E. (1996) "Methodology: Alchemy or Science?" *Economic Journal* 106(438), September, 1398-1431.

Hendry, David F. (1987) "Econometric Methodology: A Personal Viewpoint," in Truman Bewley (ed.) *Advances in Econometrics*, Vol. 2. Cambridge: Cambridge University Press.

Hendry, David F. (1988) "Encompassing," *National Institute Economic Review*, (August), 88-92.

Hendry, David F. (1995) *Dynamic Econometrics*. Oxford: Oxford University Press.

Hendry, David F. and Jean-Francois Richard (1982) "On the Formulation of Dynamic Models in Empirical Econometrics," *Journal of Econometrics* 20(Annals), 3-33. Reprinted in Granger (1990).

Hendry, David F. and Jean-François Richard. (1987) "Recent Developments in the Theory of Encompassing," in Bernard Cornet and Henry Tulkens (eds.) *Contributions to Operations Research and Economics: The Twentieth Anniversary of Core*. MIT Press: Cambridge, MA.

Hess, Gregory D., Christopher S. Jones, and Richard D. Porter (1994) "The Predictive Failure of the Baba, Hendry and Starr Model of the Demand for M1 in the United States," Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series, working paper 94-34.

Hoover, Kevin D. (1988) "On the Pitfalls of Untested Common-Factor Restrictions: The Case of the Inverted Fisher Hypothesis," *Oxford Bulletin of Economics and Statistics* 50(2), 135-139.

Hoover, Kevin D. (1995) "In Defense of Data Mining: Some Preliminary Thoughts," in Kevin D. Hoover and Steven M. Sheffrin (eds.) *Monetarism and the Methodology of Economics: Essays in Honour of Thomas Mayer*. Aldershot: Edward Elgar, 1995.

Jeong, J. and G.S. Maddala (1993) "A Perspective on Applications of Bootstrap Methods in Econometrics," in G.S. Maddala, C.R. Rao and H.D. Vinod (eds.) *Handbook of Statistics*, vol. 11, *Econometrics*. Amsterdam: North Holland, pp. 573-610.

Leamer, Edward. (1978) *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. John Wiley: Boston.

Leamer, Edward. (1983) "Let's Take the Con Out of Econometrics," *American Economic Review*, Vol. 73, No. 1 (March), 31-43. Reprinted in Granger (1990).

Leamer, Edward (1985) "Sensitivity Analysis Would Help," *American Economic Review* 75(3), June, 308-13. Reprinted in Granger (1990).

Li, Hongyi and G.S. Maddala. (1995) "Bootstrapping Cointegrating Regressions," unpublished typescript.

Lovell, Michael C. (1983) "Data Mining," *The Review of Economics and Statistics*, Vol. 65, No. 1 (February), 1-12.

Mayer, Thomas. (1980) "Economics as a Hard Science: Realistic Goal or Wishful Thinking?," *Economic Inquiry*, Vol. 18, No. 2 (April), 165-178.

Mayer, Thomas. (1993) *Truth versus Precision in Economics*. Edward Elgar: Aldershot.

McAleer, Michael, Adrian R. Pagan, and Paul A. Volker. (1983) "What Will Take the Con Out of Econometrics," *American Economic Review*, Vol. 75, No. 3 (June), 293-307. Reprinted in Granger (1990).

Mizon, Grayham E. (1984) "The Encompassing Approach in Econometrics," in D.F. Hendry and K.F. Wallis (eds.) *Econometrics and Quantitative Economics.* Oxford: Basil Blackwell, pp. 135-172.

Mizon, Grayham E. (1995) "Progressive Modelling of Macroeconomic Time Series: The LSE Methodology," in Kevin D. Hoover (ed.) *Macroeconometrics: Developments, Tensions and Prospects.* Boston: Kluwer.

Mizon, Grayham E. and David F. Hendry. (1990) "Procrustean Econometrics: Or Stretching and Squeezing Data," in Granger (1990), pp. 121-136.

Mizon, Grayham E. and Jean-François Richard (1986) "The Encompassing Principle and Its Application to Testing Non-nested Hypotheses," *Econometrica* 54, 657-678.

Pagan, Adrian (1987) "Three Econometric Methodologies: A Critical Appraisal," *Journal of Economic Surveys*, Vol. 1, No. 1, 3-24. Reprinted in Granger (1990).

Phillips, Peter C. B. (1988) "Reflections on Econometric Methodology," *Economic Record* 64, 334-359.

Stock, James H. and Mark W. Watson (1988) "Variable Trends in Economic Times Series," *Journal of Economic Perspectives* 2, 147-174.

White, Halbert. (1990) "A Consistent Model Selection Procedure Based on *m*-testing," in Granger (1990), pp. 369-383.

TABLE 1 [LOVELL'S TABLE 7]. DATA MINING PERFORMANCE SUMMARIZED
(Models 1, 4, 5, and 6; $\hat{\alpha} = 0.05$)

	Stepwise	Max $\bar{R}^2$	Max-min $\lvert t \rvert$	*General-to-Specific*
Correct Variable Selected	70%	52%	0%	*79%*
Correct or Related Variable Selected	82	75	36	*84*[1]
Type I Error (true significance level)	30	53	81	*4*
Type II Error	15	8	0	*1*

Note: This table summarizes the performance of the three alternative selection criteria for the 4 models satisfying the classical regression assumptions. The first row shows the frequency with which the correct explanatory variables were selected at the $\hat{\alpha} = 0.05$ significance level. The second row shows the frequency with which a member of the correct monetary or fiscal policy set was selected. The third reports the incidence of Type I errors (rejecting the null hypothesis when its was true). The fourth row reports the incidence of Type II errors (accepting the null hypothesis when false).

[Authors' note: This table reproduces the information in Lovell's (1983) Table 7 exactly and adds an (italicized) additional column summarizing the analogous results for the corresponding models as reported in Table 5 below.]

[1] Related variables are here Variables 3, 4, and 5 (fiscal variables; see Table 2) and their lagged values, Variables 21, 22, and 23.

TABLE 2. CANDIDATE VARIABLES FOR SPECIFICATION SEARCH

Variable	Variable Number					Times Differenced for Stationarity[1]	CITIBASE Identifier[2]
	Current	Lag					
		1	2	3	4		
4 Coincident indicators	1	19				1	DCOINC
GNP Price Deflator	2	20				2	GD
Government Purchases of Goods and Services	3	21				2	GGEQ
Federal Purchases of Goods and Services	4	22				1	GGFEQ
Federal Government Receipts	5	23				2	GGFR
GNP	6	24				1	GNPQ
Disposable Personal Income	7	25				1	GYDQ
Gross Private Domestic Investment	8	26				1	GPIQ
Total Member Bank Reserves	9	27				2	FMRRA
Monetary Base (Federal Reserve Bank of St. Louis))	10	28				2	FMBASE
M1	11	29				1	FM1DQ
M2	12	30				1	FM2DQ
Dow Jones Stock Price	13	31				1	FSDJ
Moody's AAA Corporate Bond Yield	14	32				1	FYAAAC
Labor Force (16 years+, civilian)	15	33				1	LHC
Unemployment Rate	16	34				1	LHUR
Unfilled Orders (Manufacturing, All Industries)	17	35				1	MU
New Orders (Manufacturing, All Industries)	18	36				2	MO
Personal Consumption Expenditure[3]	N/A	37	38	39	40	1	GCQ

Note: Data run 1959.1 - 1995.1. All data from CITIBASE: Citibank economic database (Floppy disk version); July 1995 release. All data converted to quarterly by averaging or summing as appropriate. All dollar denominated data in billions of constant 1987 dollars. Series FMRRA, FMBASE, GGFR, FSDJ, MU, and MO are deflated using the GNP price deflator (Series GD).

[1]Indicates the number of times the series had to be differencesd before a Dickey-Fuller test could not reject the null hypothesis of nonstationarity at a 5 percent significance level.

[2]Indicates the identifier code for this series in the CITIBASE economic database.

[3]For calibrating models in Table 4 actual personal consumption expenditure data is used as the dependent variables; for specification searches, actual data is replaced by artificial data generating according to models in Table 4. Variable numbers refer to these artificial data, which vary from context to context.

TABLE 3. CORRELATION MATRIX FOR SEARCH VARIABLES

Variable Name and Number	Variable Number																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Dep.*
1. 4 Coincident indicators	1.00	0.21	0.04	-0.07	0.21	0.83	0.57	0.76	-0.02	-0.02	0.24	0.20	-0.04	0.23	0.17	-0.85	0.21	0.23	0.60
2. GNP Price Deflator	0.21	1.00	-0.09	-0.08	0.28	0.16	0.07	0.19	0.24	0.49	-0.04	-0.06	-0.06	0.11	0.04	-0.13	0.24	0.12	-0.02
3. Government Purchases of Goods and Services	0.04	-0.09	1.00	0.54	0.03	0.13	0.07	0.03	0.07	-0.02	-0.04	-0.08	-0.06	-0.04	0.03	-0.01	-0.08	-0.29	-0.02
4. Federal Purchases of Goods and Services	-0.07	-0.08	0.54	1.00	0.01	0.03	-0.09	-0.18	0.14	0.07	0.00	0.07	-0.06	-0.05	-0.04	-0.02	0.04	-0.15	-0.02
5. Federal Government Receipts	0.21	0.28	0.03	0.01	1.00	0.20	0.06	0.13	0.40	0.25	0.16	0.11	-0.12	0.07	-0.03	-0.09	0.03	0.25	0.15
6. GNP	0.83	0.16	0.13	0.03	0.20	1.00	0.49	0.83	-0.03	-0.06	0.27	0.20	0.03	0.11	0.11	-0.73	0.16	0.22	0.65
7. Disposable Personal Income	0.57	0.07	0.07	-0.09	0.06	0.49	1.00	0.40	0.24	0.10	0.17	0.17	-0.03	0.07	0.09	-0.31	0.05	0.15	0.40
8. Gross Private Domestic Investment	0.76	0.19	0.03	-0.18	0.13	0.83	0.40	1.00	-0.16	-0.06	0.17	0.08	-0.02	0.20	0.07	-0.66	0.10	0.10	0.30
9. Total Member Bank Reserves	-0.02	0.24	0.07	0.14	0.40	-0.03	0.24	-0.16	1.00	0.54	0.25	0.21	-0.08	-0.16	-0.17	0.08	-0.10	0.21	0.07
10. Monetary Base (Federal Reserve Bank of St. Louis))	-0.02	0.49	-0.02	0.07	0.25	-0.06	0.10	-0.06	0.54	1.00	0.20	0.14	0.01	-0.06	0.01	0.07	0.09	0.01	-0.03
11. M1	0.24	-0.04	-0.04	0.00	0.16	0.27	0.17	0.17	0.25	0.20	1.00	0.60	0.27	-0.33	-0.04	-0.23	-0.39	0.28	0.47
12. M2	0.20	-0.06	-0.08	0.07	0.11	0.20	0.17	0.08	0.21	0.14	0.60	1.00	0.04	-0.33	-0.07	-0.22	-0.21	0.19	0.41
13. Dow Jones Stock Price	-0.04	-0.06	-0.06	-0.06	-0.12	0.03	-0.03	-0.02	-0.08	0.01	0.27	0.04	1.00	-0.26	0.13	0.02	0.06	0.06	0.18
14. Moody's AAA Corporate Bond Yield	0.23	0.11	-0.04	-0.05	0.07	0.11	0.07	0.20	-0.16	-0.06	-0.33	-0.33	-0.26	1.00	0.11	-0.22	0.27	0.12	-0.05
15. Labor Force (16 years+, civilian)	0.17	0.04	0.03	-0.04	-0.03	0.11	0.09	0.07	-0.17	0.01	-0.04	-0.07	0.13	0.11	1.00	0.02	0.14	0.01	0.13
16. Unemployment Rate	-0.85	-0.13	-0.01	-0.02	-0.09	-0.73	-0.31	-0.66	0.08	0.07	-0.23	-0.22	0.02	-0.22	0.02	1.00	-0.23	-0.12	-0.50
17. Unfilled Orders (Manufacturing, All Industries	0.21	0.24	-0.08	0.04	0.03	0.16	0.05	0.10	-0.10	0.09	-0.39	-0.21	0.06	0.27	0.14	-0.23	1.00	-0.04	-0.01
18. New Orders (Manufacturing, All Industries)	0.23	0.12	-0.29	-0.15	0.25	0.22	0.15	0.10	0.21	0.01	0.28	0.19	0.06	0.12	0.01	-0.12	-0.04	1.00	0.39
Dep.* Personal Consumption Expenditure	0.60	-0.02	-0.02	-0.02	0.15	0.65	0.40	0.30	0.07	-0.03	0.47	0.41	0.18	-0.05	0.13	-0.50	-0.01	0.39	1.00

Note: Corrleations are calculated for the variables in Table 2 for the period 1959.3-1995.1. Variables are differenced as indicated in Table 2.

*Dep. indicates that personal consumption expenditure is the dependent variable used in calibrating the models in Table 4. It is not a search variable. The dependent variables and its lags used in the simulations below are constructed accorded to those models.

TABLE 4.

MODELS USED TO GENERATE ALTERNATIVE ARTIFICIAL "CONSUMPTION" DEPENDENT VARIABLES

Random Errors
$u_t \sim N(0,1)$
$u_t^* = 0.75u_{t-1}^* + u_t\sqrt{7}/4$

Models	
Model 1:	$y1_t = 130.0\, u_t$
Model 2:	$y2_t = 130.0\, u_t^*$
Model 2':	$y2_t = 0.75\, y2_{t-1} + 85.99\, u_t$
Model 3:	$ln(y3)_t = 0.395\, ln(y3)_{t-1} + 0.3995\, ln(y3)_{t-2} + 0.00172\, u_t$
Model 4:	$y4_t = 1.33\, x11_t + 9.73\, u_t$
Model 5:	$y5_t = -0.046\, x3_t + 0.11\, u_t$
Model 6:	$y6_t = 0.67\, x11_t - 0.023\, x3_t + 4.92\, u_t$
Model 6A:	$y6_t = 0.67\, x11_t - 0.32\, x3_t + 4.92\, u_t$
Model 6B:	$y6_t = 0.67\, x11_t - 0.65\, x3_t + 4.92\, u_t$
Model 7:	$y7_t = 1.33\, x11_t + 9.73\, u_t^*$
Model 7':	$y7_t = 0.75\, y7_{t-1} + 1.33\, x11_t - 0.9975\, x29_t + 6.73\, u_t$
Model 8:	$y8_t = -0.046\, x3_t + 0.11\, u_t^*$
Model 8':	$y8_t = 0.75\, y8_{t-1} - 0.046\, x3_t + .00345\, x21_t + 0.073\, u_t$
Model 9:	$y9_t = 0.67\, x11_t - 0.023\, x3_t + 4.92\, u_t^*$
Model 9':	$y9_t = 0.75\, y9_{t-1} - 0.023\, x3_t + 0.01725 x21_t + 0.67\, x11_t - 0.5025\, x29_t + 3.25\, u_t$

Note: The variables $y\#_t$ are the artificial variables created by each model. The variables $x\#_t$ correspond to the variables with the same number in Table 2. The coefficients for models 3, 4, and 5 come from the regression of personal consumption expenditures (Dep. in Table 2) on independent variables as indicated by the models. Model 6 is the average of models 4 and 5. Models 7, 8, and 9 have same coefficients as models 4, 5, and 6 with autoregressive errors.

Models 2', 7', 8', and 9' are exactly equivalent expressions for Models 2, 7, 8, 9 in which lags of the variables are used to eliminate the autoregressive parameter in the error process.

TABLE 5. SPECIFICATION SEARCHES AT 5 PERCENT NOMINAL SIZE[1]

	True Model[2]									
	1	2	3	4	5	6	7	8	9	Means
Final Specification is:[3]										
Category 1	17	0	14	13	17	1	1	17	0	8.89
Category 2	33	50	33	36	32	7	43	31	5	30.00
Category 3	0	0	1	1	1	0	0	1	0	0.44
Category 4	0	0	1	0	0	29	6	0	44	8.89
Category 5	0	0	1	0	0	13	0	1	1	1.78
True Variable Number[4]	Null Set	37	37/38	11	3	3/ 11	11/29/37	3/21/37	3/11/21/ 29/37	
Times Included in 50 Replications		50	50/48	50	50	8/50	50/44/50	50/49/50	9/50/6/ 44/50	
Average Rate of Inclusion per Replication of:										
True Variables		1.00	1.96	1.00	1.00	1.16	2.88	2.98	3.18	
Insignificant Variables	1.18	0.72	0.68	0.48	0.18	0.94	2.36	0.92	1.74	1.02
Falsely Significant Variables	1.66	4.08	1.76	1.62	1.12	1.58	3.28	1.28	3.00	2.15
Type I Error (True Size) (percent)	4.2	10.5	4.6	4.2	2.9	4.2	8.9	3.5	8.6	5.7
Power (percent)	N/A	100.0	98	100.0	100.0	58.0	96.0	99.3	63.6	89.4

[1]Search algorithm described in text (Section III). Test batteries use critical values corresponding to two-tailed tests with the nominal size in title. The universe of variables searched over is given in Table 2. All regressions include a constant, which is ignored in evaluation of the successes or failures or searches. Sample runs 1960.3-1995.1 or 139 observations. The table reports the results of 50 replications.

[2]The artificial consumption variable is generated according to the specifications in Table 4.

[3] Categories of specification search results are described in the text (Section IV).

[4] Variable numbers correspond to those given in Table 2.

TABLE 6. SPECIFICATION SEARCHES ON MODEL 6 AND TWO VARIATIONS AT 5 PERCENT NOMINAL SIZE[1]

	True Model[2]		
	6	6A	6B
Final Specification is:[3]			
Category 1	1	6	10
Category 2	7	37	40
Category 3	0	0	0
Category 4	29	5	0
Category 5	13	2	0
True Variable Number[4]	3/11	3/11	3/11
Times Included in 50 Replications	8/50	44/49	50/50
Average Rate of Inclusion per Replication of:			
True Variables	1.16	1.86	2.00
Insignificant Variables	0.94	0.46	0.62
Falsely Significant Variables	1.58	2.18	1.54
Type I Error (True Size) (percent)	4.2	5.7	4.1
Power (percent)	58.0	93.0	100.0

[1]Search algorithm described in text (Section III). Test batteries use critical values corresponding to two-tailed tests with the nominal size in title. The universe of variables searched over is given in Table 2. All regressions include a constant, which is ignored in evaluation of the successes or failures or searches. Sample runs 1960.3-1995.1 or 139 observations. The table reports the results of 50 replications.

[2]The artificial consumption variable is generated according to the specifications in Table 4.

[3] Categories of specification search results are described in the text (Section IV).

[4] Variable numbers correspond to those given in Table 2.

TABLE 7. SPECIFICATION SEARCHES AT 5 PERCENT NOMINAL SIZE WITH LONG SAMPLE SIZE[1]

	True Model[2]									
	1	2	3	4	5	6	7	8	9	Means
Final Specification is:[3]										
Category 1	11	0	6	9	9	0	0	4	0	4.33
Category 2	38	50	44	41	41	5	50	46	2	35.22
Category 3	1	0	0	0	0	0	0	0	0	0.11
Category 4	0	0	0	0	0	36	0	0	48	9.33
Category 5	0	0	0	0	0	9	0	0	0	1.00
True Variable Number[4]	Null Set	37	37/38	11	3	3/ 11	11/29/ 37	3/21/ 37	3/11/21/ 29/37	
Times Included in 50 Replications		50	50/50	50	50	5/50	50/50/ 50	50/50/ 50	5/50/9/ 50/50	
Average Rate of Inclusion per Replication of:										
True Variables		1	2	1	1	1.1	3	3	3.28	
Insignificant Variables	0.92	0.04	0.24	0.06	0.56	0.16	0.26	1.38	0.26	0.43
Falsely Significant Variables	2.24	5.22	2.4	3.06	2.32	2.18	7.38	2.74	7.04	3.84
Type I Error (True Size) (percent)	5.6	13.4	4.7	7.8	5.9	5.7	19.9	7.4	20.1	10.3
Power (percent)	N/A	100.0	100.0	100.0	100.0	55.0	100.0	100.0	65.6	90.1

[1] Search algorithm described in text (Section III). Test batteries use critical values corresponding to two-tailed tests with the nominal size in title. The universe of variables searched over is given in Table 2. All regressions include a constant, which is ignored in evaluation of the successes or failures or searches. Sample is five times the length of data used in other replications (695 observations). The longer sample is created from the shorter samples as described in Tabless 2 and 4 using a procedure described in the text analogous to block bootstrapping. The table reports the results of 50 replications.

[2] The artificial consumption variable is generated according to the specifications in Table 4.

[3] Categories of specification search results are described in the text (Section IV).

[4] Variable numbers correspond to those given in Table 2.

TABLE 8. SPECIFICATION SEARCHES AT 10 PERCENT NOMINAL SIZE[1]

	True Model[2]									
	1	2	3	4	5	6	7	8	9	Means
Final Specification is:[3]										
Category 1	6	0	4	1	6	0	1	5	0	1.89
Category 2	44	50	42	47	43	10	46	44	2	36.44
Category 3	0	0	1	2	1	0	0	0	0	0.44
Category 4	0	0	2	0	0	38	3	1	48	10.22
Category 5	0	0	1	0	0	2	0	0	0	0.33
True Variable Number[4]	Null Set	37	37/38	11	3	3/ 11	11/29/ 37	3/21/ 37	3/11/ 21/29/37	
Times Included in 50 Replications		50	49/48	50	50	10/49	50/47/ 50	50/49/ 50	12/50/ 7/45/50	
Average Rate of Inclusion per Replication of:										
True Variables		1.00	1.94	1.00	1.00	1.18	2.94	2.98	3.28	
Insignificant Variables	0.9	2.84	1.18	0.52	1.44	1.10	0.76	0.44	1.56	1.19
Falsely Significant Variables	2.66	5.44	3.32	3.18	3.2	3.84	4.82	3.46	4.46	3.82
Type I Error (True Size) (percent)	6.7	13.9	8.7	8.2	8.2	10.1	13.0	9.4	12.7	10.1
Power (percent)	N/A	100.0	97.0	100.0	100.0	59.0	98.0	99.3	65.6	89.9

[1]Search algorithm described in text (Section III). Test batteries use critical values corresponding to two-tailed tests with the nominal size in title. The universe of variables searched over is given in Table 2. All regressions include a constant, which is ignored in evaluation of the successes or failures or searches.

[2]The artificial consumption variable is generated according to the specifications in Table 4. The table reports the results of 50 replications.

[3] Categories of specification search results are described in the text (Section IV).

[4] Variable numbers correspond to those given in Table 2.

TABLE 9. SPECIFICATION SEARCHES AT 1 PERCENT NOMINAL SIZE[1]

	True Model[2]									
	1	2	3	4	5	6	7	8	9	Means
Final Specification is:[3]										
Category 1	41	2	25	44	40	0	7	31	0	16.56
Category 2	9	48	14	6	10	0	37	16	1	15.67
Category 3	0	0	2	0	0	0	0	1	0	0.33
Category 4	0	0	0	0	0	30	5	0	45	8.89
Category 5	0	0	9	0	0	20	1	2	4	4.00
True Variable Number[4]	Null Set	37	37/38	11	3	3/ 11	11/29/37	3/21/37	3/11/21/ 29/37	
Times Included in 50 Replications		50	46/45	50	50	0/50	50/44/50	50/48/48	1/50/2/ 41/50	
Average Rate of Inclusion per Replication of:										
True Variables		1.00	1.82	1.00	1.00	1.00	2.88	2.92	2.88	
Insignificant Variables	0.12	0.18	0.22	0.12	0.78	0.08	1.24	0.32	1.5	0.51
Falsely Significant Variables	0.2	1.88	0.46	0.28	0.2	0.30	1.2	0.38	1.46	0.71
Type I Error (True Size) (percent)	0.5	4.8	0.9	0.7	0.5	0.8	3.2	1.0	4.2	1.9
Power (percent)	N/A	100.0	91.0	100.0	100.0	50.0	96.0	97.3	57.6	86.5

[1]Search algorithm described in text (Section III). Test batteries use critical values corresponding to two-tailed tests with the nominal size in title. The universe of variables searched over is given in Table 2. All regressions include a constant, which is ignored in evaluation of the successes or failures or searches.

[2]The artificial consumption variable is generated according to the specifications in Table 4. The table presents the results of 50 replications.

[3] Categories of specification search results are described in the text (Section IV).

[4] Variable numbers correspond to those given in Table 2.

TABLE 10. MONTE CARLO STATISTICS FOR SPECIFICATION SEARCH ON MODEL 9 (1000 REPLICATIONS)

	Variables												
	Correctly Included					Correctly Excluded							
	3	11	21	29	37	1	2	4	5	6	7	8	9
True Value[1]	-0.023	0.670	0.017	-0.500	0.750	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Estimated Coefficients													
Mean	0.002	0.686	0.004	-0.294	0.574	-0.204	0.174	-0.007	0.002	0.002	0.000	-0.001	0.000
Median	0.000	0.691	0.000	-0.322	0.578	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Max	0.329	0.960	0.166	0.000	0.859	3.578	9.763	0.291	0.080	0.066	0.060	0.086	0.004
Min	-0.308	0.307	-0.137	-0.611	0.000	-4.900	-4.491	-0.338	-0.074	-0.087	-0.062	-0.085	-0.004
Standard Deviation	0.044	0.091	0.027	0.140	0.106	0.858	1.079	0.048	0.012	0.013	0.010	0.014	0.001
Simulated Standard Deviation[2]	0.05	0.06	0.05	0.07	0.06								
Mean Bias[3] (percent)	106.7	2.4	75.0	41.3	23.5								
Median Bias[4] (percent)	100.0	3.1	100.0	35.6	22.9								
Empirical Size[5] (percent)						8.7	5.3	4.1	4.5	6.1	4.6	6.5	3.7
True Power[6] (percent)	10.0	100.0	8.0	100.0	100.0								
Empirical Power[7] (percent)	9.4	100.0	9.1	86.0	99.7								
Chosen But Insignificant (percent)						4.0	3.0	3.2	4.4	4.3	3.7	3.1	4.8

[1] Coefficients from Model 9', Table 4.
[2] Actual standard deviation of coefficients from 1000 replications of Model 9 (i.e., without search).
[3] | (mean estimated values - true value) | /true value expressed as percentage.
[4] | (median estimated values - true value) | /true value expressed as percentage.
[5] Proportion of *t*-statistics outside ±1.96 - the nominal 5 percent critical value.
[6] 1 - proportion of *t*-statistics inside ±1.96 (i.e., the nominal 5 percent critical value) for 1000 replications of Model 9' (i.e., without search).
[7] 1 - proportion of *t*-statistics inside ±1.96 (i.e., the nominal 5 percent critical value).

Table continues next page.

Continues TABLE 10. MONTE CARLO STATISTICS FOR SPECIFICATION SEARCH ON MODEL 9 (1000 REPLICATIONS)

	Variables Correctly Excluded												
	10	12	13	14	15	16	17	18	19	20	22	23	24
True Value[1]	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Estimated Coefficients													
Mean	-0.023	0.001	0.000	-0.033	0.000	-0.164	0.000	0.000	0.210	0.315	0.006	-0.001	0.001
Median	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Max	0.753	0.132	0.011	2.747	0.005	6.681	0.000	0.000	5.885	8.000	0.375	0.061	0.065
Min	-1.184	-0.088	-0.013	-3.039	-0.003	-8.343	0.000	-0.001	-2.918	-4.536	-0.417	-0.078	-0.059
Standard Deviation	0.153	0.020	0.002	0.525	0.001	1.204	0.000	0.000	0.815	1.147	0.053	0.011	0.011
Simulated Standard Deviation[2]													
Mean Bias[3] (percent)													
Median Bias[4] (percent)													
Empirical Size[5] (percent)	3.1	7.6	3.3	4.8	4.7	6.8	3.7	4.7	9.3	6.5	4.5	3.7	4.6
True Power[6] (percent)													
Empirical Power[7] (percent)													
Chosen But Insignificant (percent)	4.3	4.4	2.8	3.0	4.3	3.7	3.9	3.8	4.3	4.3	4.3	4.1	3.4

[1]Coefficients from Model 9', Table 4.
[2]Actual standard deviation of coefficients from 1000 replications of Model 9 (i.e., without search).
[3] | (mean estimated values - true value) | /true value expressed as percentage.
[4] | (median estimated values - true value) | /true value expressed as percentage.
[5]Proportion of *t*-statistics outside ±1.96 - the nominal 5 percent critical value.
[6] 1 - proportion of *t*-statistics inside ±1.96 (i.e., the nominal 5 percent critical value) for 1000 replications of Model 9' (i.e., without search).
[7] 1 - proportion of *t*-statistics inside ±1.96 (i.e., the nominal 5 percent critical value).

Table continues next page

Continues TABLE 10. MONTE CARLO STATISTICS FOR SPECIFICATION SEARCH ON MODEL 9 (1000 REPLICATIONS)

	Variables Correctly Excluded													
	25	26	27	28	30	31	32	33	34	35	36	38	39	40
True Value[1]	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Estimated Coefficients														
Mean	-0.003	0.001	0.000	-0.052	-0.004	0.000	0.096	0.000	0.001	0.000	0.000	-0.220	0.057	-0.013
Median	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.231	0.000	0.000
Max	0.056	0.074	0.003	0.564	0.118	0.013	4.316	0.005	7.375	0.000	0.000	0.000	0.363	0.145
Min	-0.073	-0.087	-0.003	-1.399	-0.130	-0.013	-4.184	-0.005	-6.074	0.000	0.000	-0.507	-0.180	-0.235
Standard Deviation	0.013	0.012	0.001	0.191	0.021	0.002	0.719	0.001	1.118	0.000	0.000	0.105	0.086	0.046
Simulated Standard Deviation[2]														
Mean Bias[3] (percent)														
Median Bias[4] (percent)														
Empirical Size[5] (percent)	8.2	5.2	6.8	4.9	8.5	4.5	5.4	3.5	6.2	4.0	3.5	87.3	29.5	8.4
True Power[6] (percent)														
Empirical Power[7] (percent)														
Chosen But Insignificant (percent)	4.1	3.3	3.2	5.0	4.1	3.7	3.0	2.8	3.2	4.6	3.4	1.1	3.8	3.8

[1]Coefficients from Model 9', Table 4.
[2]Actual standard deviation of coefficients from 1000 replications of Model 9 (i.e., without search).
[3] | (mean estimated values - true value) | /true value expressed as percentage.
[4] | (median estimated values - true value) | /true value expressed as percentage.
[5]Proportion of *t*-statistics outside ±1.96 - the nominal 5 percent critical value.
[6] 1 - proportion of *t*-statistics inside ±1.96 (i.e., the nominal 5 percent critical value) for 1000 replications of Model 9' (i.e., without search).
[7] 1 - proportion of *t*-statistics inside ±1.96 (i.e., the nominal 5 percent critical value).

Figure 1. Average Size-Power Tradeoff for Split Sample Searchs