Computer Automation of General-to-Specific Model Selection Procedures

By HANS-MARTIN KROLZIG and DAVID F. HENDRY *
Department of Economics and *Nuffield College, Oxford*.

May 28, 2000

Abstract

Disputes about econometric methodology partly reflect a lack of evidence on alternative approaches. We reconsider econometric model selection from a computer-automation perspective, focusing on general-to-specific reductions, embodied in *PcGets*. Starting from a general congruent model, standard testing procedures eliminate statistically-insignificant variables, with diagnostic tests checking the validity of reductions, ensuring a congruent final selection. Since jointly selecting and diagnostic testing has eluded theoretical analysis, we study modelling strategies by simulation. The Monte Carlo experiments show that *PcGets* recovers the DGP specification from a general model with size and power close to commencing from the DGP itself.
JEL Classification: C51, C22.
Keywords: Econometric methodology; Model selection; Encompassing; Data mining; Monte Carlo experiments; Money demand; Consumption function.

1 Introduction

Despite the controversy surrounding econometric methodology, the 'LSE' approach (see Hendry, 1993, for an overview) has emerged as a leading approach to modelling. One of its main tenets is the concept of general-to-specific modelling: starting from a general dynamic statistical model, which captures the essential characteristics of the underlying data set, standard testing procedures are used to reduce its complexity by eliminating statistically-insignificant variables, checking the validity of the reductions at every stage to ensure the congruence of the selected model.

We discuss computer automation of such an econometric model-selection process, called *PcGets* (**ge***neral*-**t***o*-**s***pecific*). *PcGets* is an Ox Package (see Doornik, 1998, and Hendry and Krolzig, 1999) designed for general-to-specific modelling, currently focusing on reduction approaches for linear, dynamic, regression models. The development of *PcGets* has been stimulated by Hoover and Perez (1999), who sought to evaluate the performance of *Gets*. To implement a 'general-to-specific' approach in a computer algorithm, all decisions must be 'mechanized'. In doing so, Hoover and Perez made some important advances in practical modelling, and our approach builds on these by introducing further improvements. Given an initial general model, many reduction paths could be considered, and different selection strategies adopted for each path. Some of these searches may lead to different terminal specifications, between which a choice must be made. Consequently, the reduction process is inherently

*We grateful to Jurgen Doornik, Neil Ericsson, Jon Faust, Marianne Sensier, Rolf Tschernig, and seminar participants at the CEF99, Norges Bank, and the Bank of Argentina as well as an unknown referee for their comments. All the computations reported in this paper were carried out with the PcGets class in Ox. Financial support from the UK Economic and Social Research Council under grant L138251009 is gratefully acknowledged.

iterative. Should multiple congruent contenders eventuate after a reduction round, encompassing can be used to test between them, with only the surviving – usually non-nested – specifications retained. If multiple models still remain after this *'testimation'* process, a new general model is formed from their union, and the simplification process re-applied. Should that union repeat, a final selection is made using information criteria, otherwise a unique congruent and encompassing reduction has been located.

Our attempt to automate *Gets* throws further light on several methodological issues, and prompts some new ideas, which will be discussed in section 2. While the joint issue of variable selection and diagnostic testing using multiple criteria has eluded most attempts at theoretical analysis, computer automation of the model-selection process allows us to evaluate econometric model-selection strategies by simulation. Section 3 presents the results of some Monte Carlo experiments to investigate if the model-selection process works well or fails badly; their implications for the calibration of *PcGets* are also analyzed. The empirical illustrations presented in section 4 demonstrate the usefulness of *PcGets* for applied econometric research.

2 The econometrics of model selection

2.1 Concepts

The key issue for any model-selection procedure is the cost of search, since there are always bound to be mistakes in statistical inference: specifically, how bad does it get to search across many alternatives?

On the one hand, the conventional statistical analysis of repeated testing provides a pessimistic background: every test has a non-zero null rejection frequency (or size, if independent of nuisance parameters), and so type I errors accumulate. Setting a small size for every test can induce low power to detect the influences that really matter. The study by Lovell (1983) of trying to select a small relation (0 to 5 regressors) from a large database (up to 40 variables) suggested search had very high costs, leading to an adverse view of 'data mining'. Although Lovell did not consider a structured reduction approach among his methods, *Gets* has been criticized by Pagan (1987) on the grounds that the selection path may matter, and so the result is not 'path independent'. Indeed, Leamer (1983) claimed that 'the mapping is the message'. Moreover, 'pre-testing' is known to bias estimated coefficients, and may distort inference: see *inter alia*, Bock, Yancey and Judge (1973) and Judge and Bock (1978).

On the other hand, White (1990) showed that with sufficiently-rigorous testing, the selected model will converge to the data generating process (DGP). Thus, any 'overfitting' and mis-specification problems are primarily finite sample. Moreover, Mayo (1981) emphasized the importance of diagnostic test information being effectively independent of the sufficient statistics from which parameter estimates are derived. Also, Hendry (1995) argued that congruent models are the appropriate class within which to search, that encompassing resolves many instances of 'data mining', and that in econometrics, theory dependence has as many drawbacks as sample dependence, so modelling procedures are essential. Finally, Hoover and Perez (1999) reconsidered the Lovell (1983) experiments to evaluate the performance of *Gets*. Most important is their notion of commencing from the congruent general model by following a number of reduction search paths, terminated by either no further feasible reductions or significant diagnostic tests occurring. Hoover and Perez select among the surviving models the one which fits best. They show how much better a structured approach is than any method Lovell considered, suggesting that modelling *per se* need not be bad. Indeed, overall, the size of their selection procedure is close to that expected, and the power is reasonable. Moreover, re-running their experiments using our version (*PcGets*) delivered substantively better outcomes (see Hendry and Krolzig, 2000). Thus, the case against model selection is far from proved.

There is little research on how to design model-search algorithms in econometrics. To reduce search costs, any model-selection process must avoid getting stuck in a search path that initially inadvertently deletes variables that really matter in the DGP, thereby retaining other variables as proxies. Thus, it is imperative to explore multiple paths, although it is not known whether all possible paths should be searched. To meet this requirement, *PcGets* builds on the multi-path approach to *Gets* model selection in Hoover and Perez (1999). Equally, the search procedure must have a high probability of retaining variables that do matter in the DGP. To achieve that, *PcGets* uses encompassing tests between alternative reductions. Balancing these objectives of small size and high power still involves a trade-off, but one that is dependent on the algorithm: the upper bound is probably determined by the famous lemma in Neyman and Pearson (1928). Nevertheless, to tilt the size-power balance favourably, sub-sample information is exploited, building on the further development in Hoover and Perez of investigating split samples for significance (as against constancy). Since non-central 't'-values diverge with increasing sample size, whereas central 't's fluctuate around zero, the latter have a low probability of exceeding any given critical value in two sub-samples, even when those sample overlap. Thus, adventitiously-significant variables may be revealed by their insignificance in one or both of the sub-samples.

PcGets embodies some further developments. First, *PcGets* undertakes 'pre-search' simplification F-tests to exclude variables from the general unrestricted model (GUM), after which the GUM is re-formulated. Since variables found to be irrelevant on such tests are excluded from later analyses, this step uses a loose significance level (such as 10%). Next, many possible paths from that GUM are investigated: reduction paths considered include both multiple deletions as well as single, so t and/or F test statistics are used as simplification criteria. The third development concerns the encompassing step: all distinct contending valid reductions are collected, and encompassing is used to test between these (usually non-nested) specifications. Models which survive encompassing are retained; all encompassed equations are rejected. If multiple models survive this *'testimation'* process, their union forms a new general model, and selection path searches recommence. Such a process repeats till a unique contender emerges, or the previous union is reproduced, then stops. Fourthly, the diagnostic tests require careful choice to ensure they characterize the salient attributes of congruency, are correctly sized, and do not overly restrict reductions. A further improvement concerns model choice when mutually-encompassing distinct models survive the encompassing step. A minimum standard error rule, as used by Hoover and Perez (1999), will probably 'over-select' as it corresponds to retaining all variables with $|t| > 1$. Instead, we employ information criteria which penalize the likelihood function for the number of parameters. Finally, sub-sample information is used to accord a 'reliability' score to variables, which investigators may use to guide their model choice. In Monte Carlo experiments, a 'progressive research strategy' (PRS) can be formulated in which decisions on the final model choice are based on the outcomes of such reliability measure.

2.2 The multi-path reduction process of *PcGets*

The starting point for *Gets* model-selection is the general unrestricted model, so the key issues concern its specification and congruence. The larger the initial regressor set, the more likely adventitious effects will be retained; but the smaller the GUM, the more likely key variables will be omitted. Further, the less orthogonality between variables, the more 'confusion' the algorithm faces, leading to a proliferation of mutual-encompassing models, where final choices may only differ marginally (e.g., lag 2 versus 1).[1] Finally, the initial specification must be congruent, with no mis-specification tests failed at the outset.

[1] Some empirical examples for autoregressive-distributed lag (ADL) models and single-equation equilibrium-correction models (EqCM) are presented in section 4.

Empirically, the GUM would be revised if such tests rejected, and little is known about the consequences of doing so (although *PcGets* will enable such studies in the near future). In Monte Carlo experiments, the program automatically changes the significance levels of such tests.

The reduction path relies on a classical, sequential-testing approach. The number of paths is increased to try all single-variable deletions, as well as various block deletions from the GUM. Different critical values can be set for multiple and single selection tests, and for diagnostic tests. Denote by η the significance level for the mis-specification tests (diagnostics) and by α the significance level for the selection t-tests (we ignore F tests for the moment). The corresponding p-values of these are denoted $\widehat{\eta}$ and $\widehat{\alpha}$, respectively. During the specification search, the current specification is simplified only if no diagnostic test rejects its null. This corresponds to a likelihood-based model evaluation, where the likelihood function of model M is given by the density:

$$L_{\mathsf{M}}(\boldsymbol{\theta}_{\mathsf{M}}) = \left\{ \begin{array}{ll} f_{\mathsf{M}}(\mathbf{Y};\boldsymbol{\theta}_{\mathsf{M}}) & \text{if } \min\left(\hat{\boldsymbol{\eta}}_{\mathsf{M}}(\mathbf{Y};\tilde{\boldsymbol{\theta}}_{\mathsf{M}}) - \boldsymbol{\eta}\right) \left\{ \begin{array}{l} \geq \\ < \end{array} \right\} 0, \\ -\infty & \end{array} \right.$$

where $f_{\mathsf{M}}(\mathbf{Y};\boldsymbol{\theta}_{\mathsf{M}})$ is the probability density function (pdf) associated with model M at the parameter vector $\boldsymbol{\theta}_{\mathsf{M}}$, for the sample $\mathbf{Y}$. The vector of test statistics p-values, $\hat{\boldsymbol{\eta}}_{\mathsf{M}}(\mathbf{Y};\tilde{\boldsymbol{\theta}}_{\mathsf{M}})$, is evaluated at the maximum likelihood estimate $\widetilde{\boldsymbol{\theta}}_{\mathsf{M}}$ under model M, and mapped into its marginal rejection probabilities. So the pdf of model M is only accepted as the likelihood function if the sample information coheres with the underlying assumptions of the model itself.

In Monte Carlo experiments, *PcGets* sets the significance levels of the mis-specification tests endogenously: when a test of the DGP (or 'true model') reveals a significant diagnostic outcome (as must happen when tests have a non-zero size), the significance level is adjusted accordingly. In the event that the GUM fails a mis-specification test at the desired significance level $\overline{\eta}^{''}$, a more stringent critical value is used. If the GUM also fails at the reduced significance leve $\overline{\eta}^{'} < \overline{\eta}^{''}$, the test statistic is excluded from the test battery during the following search. Thus for the k^{th} test we have that:

$$\eta_k = \left\{ \begin{array}{c} \overline{\eta}^{''} \\ \overline{\eta}^{'} \\ 0 \end{array} \right\} \text{ if } \widehat{\eta}_{k,\mathrm{GUM}}(\mathbf{Y},\tilde{\boldsymbol{\theta}}_{\mathrm{GUM}}) \in \left\{ \begin{array}{cc} [\overline{\eta}^{''},1] & \text{'desired significance level'} \\ [\overline{\eta}^{'},\overline{\eta}^{''}) & \text{'reduced significance level'} \\ [0,\overline{\eta}^{'}) & \text{'test excluded'} \end{array} \right\}$$

where $0 < \overline{\eta}^{'} < \overline{\eta}^{''} < 1$. [2]

Each set of search paths is ended by an encompassing step (see e.g., Mizon and Richard, 1986, and Hendry and Richard, 1989). Used as the last step of model selection, encompassing seems to help control the 'size' resulting from many path searches. When a given path eliminates a variable x that matters, other variables proxy such an effect, leading to a 'spuriously large' – and mis-specified – model. However, some other paths are likely to retain x, and in the encompassing tests, the proxies will frequently be revealed as conditionally redundant, inducing a smaller final model, focused on the genuine causal factors.

[2] In contrast, Hoover and Perez (1999) drop such a test from the checking set (so an ever-increasing problem of that type may lurk undetected). Their procedure was justified on the grounds that if the GUM failed a specification test in a practical application, then an 'LSE' economist would expand the search universe to more variables, more lags, or transformations of the variables. In a Monte Carlo setting, however, it seems better to initially increase the nominal level for rejection, and if during any search path, that higher level is exceeded, then stop; we find that sometimes such GUM tests cease to be significant as reduction proceeds, and sometimes increase to reveal a flawed path.

Table 1 The *PcGets* algorithm.

Stage I

(1) **Estimation and testing of the GUM**

(a) If all variables are significant, the GUM is the *final* model, and the algorithm stops;
(b) if a diagnostic test fails for the GUM, its significance level is adjusted or the test is excluded from the test battery during simplifications of the GUM;
(c) otherwise, search paths start by removing an insignificant variable, or a set of insignificant variables.

(2) **Multiple reduction paths**: sequential simplification and testing of the GUM

(a) If any diagnostic tests fail, that path is terminated, and the algorithm returns to the last accepted model of the search path:
(i) if the last accepted model cannot be further reduced, it becomes the *terminal* model of the particular search path;
(ii) otherwise, the last removed variable is re-introduced, and the search path continues with a new reduction by removing the next least-insignificant variable of the last accepted model.
(b) If all tests are passed, but one or more variables are insignificant, the least significant variable is removed: if that specification has already been tested on a previous path, the current search path is terminated;
(c) if all diagnostic tests are passed, and all variables are significant, the model is the *terminal* model of that search path.

(3) **Encompassing**

(a) If none of the reductions is accepted, the GUM is the *final* model;
(b) if only one model survives the testimation process, it is the *final* model;
(c) otherwise, the *terminal* models are tested against their *union*:
(i) if all *terminal* models are rejected, their *union* is the *final* model;
(ii) if exactly one of the *terminal* models is not rejected, it is the *final* model;
(iii) otherwise, rejected models are removed, and the remaining *terminal* models tested against their *union*:
1. if all remaining *terminal* models are rejected, their *union* is the *final* model;
2. if exactly one remaining *terminal* model is not rejected, it is the *final* model;
3. otherwise, the *union* of the 'surviving' models becomes the GUM of Stage II.

Stage II

(1) **Estimation and testing of the GUM as in** *Stage I* (significance levels remain fixed)
(2) **Multiple reduction paths as in** *Stage I*
(3) **Encompassing and final model selection**

(a) If only one model survives the testimation process of *Stage II*, it is the *final* model;
(b) otherwise, the *terminal* models of stage II are tested against their *union*:
(i) if all *terminal* models are rejected, their *union* is the *final* model.
(ii) if exactly one *terminal* model is not rejected, it is the *final* model.
(iii) otherwise, the set of non-dominated terminal models are reported or information criteria are applied to select a unique *final* model.

The selection of the final model also improves upon Hoover and Perez (1999). Instead of selecting the best-fitting equation, *PcGets* focuses on encompassing testing between the candidate congruent selections. If a unique choice occurs, then the algorithm is terminated, otherwise, the union of the variables is formed as a new starting point for the reduction. Should that coincide with the previous union, then a model is selected by an information criterion (AIC, HQ, SC); otherwise the algorithm retries all the available paths again from that smaller union: if no simpler encompassing congruent model appears, final choice is by AIC, HQ or SC, etc.[3] Table 1 records details of the basic algorithm.

To control the overall size of the model-selection procedure, two extensions of the original algorithm were taken. First, the introduction before *Stage I* of block (F) tests of groups of variables, ordered by their t-values in the GUM (but potentially according to economic theory). This set includes the overall F-test of all regressors to check that there is something to model. Variables that are insignificant at this step, usually at a liberal critical value, are eliminated from the analysis, and a smaller GUM is formulated. Secondly, following Hoover and Perez (1999), *Stage III* was introduced as a check for potential over-selection in *Stage II* by a sub-sample split to eliminate problematic variables from the reduction search. This mimics the idea of recursive estimation, since a central t statistic 'wanders' around the origin, while a non-central t diverges. Thus, the i^{th} variable might be significant by chance for T_1 observations, yet not for $T_2 > T_1$ – whereas the opposite holds for the j^{th} – if there is not too much sample overlap. Consequently, a progressive research strategy (shown as PRS below) can gradually eliminate 'adventitiously-significant' variables. Hoover and Perez (1999) found that by adopting a progressive search procedure (as in *Stage III*), the number of spurious regressors can lowered (inducing a lower overall size), without losing much power. Details of the resulting algorithm are shown in Table 2.

2.3 Calibration of *PcGets*

The *'testimation'* process of *PcGets* depends on the choice of:

- pre-search F-test simplification;
- the significance levels κ of such tests;
- the simplification tests (t and/or F);
- the significance levels $\boldsymbol{\alpha}$ of the simplification tests;
- the n diagnostic checks in the test battery;
- the parameters of these diagnostic tests;
- the significance levels $\boldsymbol{\eta}$ of the n diagnostics;
- the significance levels $\boldsymbol{\gamma}$ of the encompassing tests;
- the sub-sample split;
- the significance levels $\boldsymbol{\delta}$ of the sub-sample tests.

The choice of mis-specification alternatives determines the number and form of the diagnostic tests. Their individual significance levels in turn determine the overall significance level of the test battery. Since significant diagnostic-test values terminate search paths, they act as constraints on moving away from the GUM. Thus, if a search is to progress towards an appropriate simplification, such tests must

[3] The information criteria are defined as follows:

$$\begin{aligned} AIC &= -2\log L/T + 2n/T, \\ SC &= -2\log L/T + n\log(T)/T, \\ HQ &= -2\log L/T + 2n\log(\log(T))/T, \end{aligned}$$

where L is the maximized likelihood, n is the number of parameters and T is the sample size: see Akaike (1985), Schwarz (1978), and Hannan and Quinn (1979).

Table 2 Additions to the basic *PcGets* algorithm.

Stage 0

(1) **Pre-simplification and testing of the GUM**

(a) If a diagnostic test fails for the GUM, the significance level of that test is adjusted, or the test is excluded from the test battery during simplifications of the GUM;
(b) if all variables are significant, the GUM is the *final* model, and the algorithm stops;
(c) otherwise, F-tests of sets of individually-insignificant variables are conducted:
(i) if one or more diagnostic tests fails, that F-test reduction is cancelled, and the algorithm returns to the previous step;
(ii) if all diagnostic tests are passed, the blocks of variables that are insignificant are removed and a simpler GUM specified;
(iii) if all diagnostic tests are passed, and all blocks of variables are insignificant, the null model is the *final* model.

Stage III

(1) **Post-selection sub-sample evaluation**

(a) Test the significance of every variable in the final model from *Stage II* in two overlapping sub-samples (e.g., the first and last r%):
(i) if a variable is significant overall and both sub-samples, accord it 100% reliable;
(ii) if a variable is significant overall and in one sub-sample, accord it 75% reliable;
(iii) if a variable is significant overall and in neither sub-sample, accord it 50% reliable;
(iv) if a variable is insignificant overall but in both sub-samples, accord it 50% reliable;
(v) if a variable is insignificant overall and in only one sub-sample, accord it 25% reliable;
(vi) if a variable is insignificant overall and in neither sub-sample, accord it 0% reliable.

be well 'focused' and have the correct size. The choice of critical values for pre-selection, selection and encompassing tests is also important for the success of *PcGets*: the tighter the size, the fewer the 'spurious inclusions of irrelevant', but the more the 'false exclusions of relevant' variables. In the final analysis, the calibration of *PcGets* depends on the characteristics valued by the user: if *PcGets* is employed as a first 'pre-selection' step in a user's research agenda, the optimal values of κ, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ may be higher than when the focus is on controlling the overall size of the selection process.

In section 3, we will use simulation techniques to investigate the calibration of *PcGets* for the operational characteristics of the diagnostic tests, the selection probabilities of DGP variables, and the deletion probabilities of non-DGP variables. However, little research has been undertaken to date to 'optimize' any of the choices, or to investigate the impact on model selection of their interactions.

2.4 Limits to *PcGets*

Davidson and Hendry (1981, p.257) mentioned four main problems in the general-to-specific methodology: (i) the chosen 'general' model can be inadequate, comprising a very special case of the DGP; (ii) data limitations may preclude specifying the desired relation; (iii) the non-existence of an optimal sequence for simplification leaves open the choice of reduction path; and (iv) potentially-large type-II error probabilities of the individual tests may be needed to avoid a high type-I error of the overall sequence. By adopting the 'multiple path' development of Hoover and Perez (1999), and implementing

a range of important improvements, *PcGets* overcomes some of problems associated with points (iii) and (iv). However, the **empirical** success of *PcGets* must depend crucially on the creativity of the researcher in specifying the general model and the feasibility of estimating it from the available data – aspects beyond the capabilities of the program, other than the diagnostic tests serving their usual role of revealing model mis-specification.

There is a central role for economic theory in the modelling process in 'prior specification', 'prior simplification', and suggesting admissible data transforms. The first of these relates to the inclusion of potentially-relevant variables, the second to the exclusion of irrelevant effects, and the third to the appropriate formulations in which the influences to be included are entered, such as log or ratio transforms etc., differences and cointegration vectors, and any likely linear transformations that might enhance orthogonality between regressors. The 'LSE approach' argued for a close link of theory and model, and explicitly opposed 'running regressions on every variable on the database' as in Lovell (1983) (see e.g., Hendry and Ericsson, 1991a). *PcGets* currently focuses on general-to-simple reductions for linear, dynamic, regression models, and economic theory often provides little evidence for specifying the lag lengths in empirical macro-models. Even when the theoretical model is dynamic, the lags are usually chosen either for analytical convenience (e.g., first-order differential equation systems), or to allow for certain desirable features (as in the choice of a linear second-order single-equation model to replicate cycles). Therefore, we adopt the approach of starting with an unrestricted rational-lag model with a maximal lag length set according to available evidence (e.g., as 4 or 5 for quarterly time series, to allow for seasonal dynamics). Prior analysis remains essential for appropriate parameterizations; functional forms; choice of variables; lag lengths; and indicator variables (including seasonals, special events, etc.). The present performance of *PcGets* on previously-studied empirical problems is impressive, even when the GUM is specified in highly inter-correlated, and probably non-stationary, levels. Orthogonalization helps notably in selecting a unique representation; as does validly reducing the initial GUM. Hopefully, *PcGets*' support in automating the reduction process will enable researchers to concentrate their efforts on designing the GUM: that could again significantly improve the empirical success of the algorithm.

2.5 Integrated variables

To date, *PcGets* conducts all inferences as I(0). Most selection tests will in fact be valid even when the data are I(1), given the results in, say, Sims, Stock and Watson (1990). Only t- or F-tests for an effect that corresponds to a unit root require non-standard critical values. The empirical examples on I(1) data provided below do not reveal problems, but in principle it would be useful to implement cointegration tests and appropriate transformations after stage 0, and prior to stage I reductions.

Similarly, Wooldridge (1999) shows that diagnostic tests on the GUM (and presumably simplifications thereof) remain valid even for integrated time series.

3 Some Monte Carlo results

3.1 Aim of the Monte Carlo

Although the sequential nature of *PcGets* and its combination of variable-selection and diagnostic testing has eluded most attempts at theoretical analysis, the properties of the *PcGets* model-selection process can be evaluated in Monte Carlo (MC) experiments. In the MC considered here, we aim to measure the 'size' and 'power' of the *PcGets* model-selection process, namely the probability of inclusion in the final model of variables that do not (do) enter the DGP.

First, the properties of the diagnostic tests under the potential influence of nuisance regressors are investigated. Based on these results, a decision can be made as to which diagnostics to include in the test battery. Then the 'size' and 'power' of *PcGets* is compared to the empirical and theoretical properties of a classical t-test. Finally we analyze how the 'success' and 'failure' of *PcGets* are affected by the choice of: (i) the significance levels $\boldsymbol{\eta}$ of the diagnostic tests; and (ii) the significance levels $\boldsymbol{\alpha}$ of the specification tests.

3.2 Design of the Monte Carlo

The Monte Carlo simulation study of Hoover and Perez (1999) considered the Lovell database, which embodies many dozens of relations between variables as in real economies, and is of the scale and complexity that can occur in macro-econometrics: the rerun of those experiments using *PcGets* is discussed in Hendry and Krolzig (2000). In this paper, we consider a simpler experiment, which however, allows an analytical assessment of the simulation findings. The Monte Carlo reported here uses only stages I and II in table 1: Hendry and Krolzig (2000) show the additional improvements that can result from adding stages 0 and III to the study in Hoover and Perez (1999).

The DGP is a Gaussian regression model, where the strongly-exogenous variables are Gaussian white-noise processes:

$$\begin{aligned} y_t &= \sum_{k=1}^{5} \beta_{k,0} x_{k,t} + \varepsilon_t, \qquad & \varepsilon_t &\sim \mathsf{IN}\left[0,1\right], \\ x_t &= v_t, & v_t &\sim \mathsf{IN}_{10}\left[\mathbf{0}, \mathbf{I}_{10}\right] \quad \text{for } t = 1,\ldots,T, \end{aligned} \tag{1}$$

where $\beta_{1,0} = 2/\sqrt{T}$, $\beta_{2,0} = 3/\sqrt{T}$, $\beta_{3,0} = 4/\sqrt{T}$, $\beta_{4,0} = 6/\sqrt{T}$, $\beta_{5,0} = 8/\sqrt{T}$.

The GUM is an $ADL(1,1)$ model which includes as non-DGP variables the lagged endogenous variable y_{t-1}, the strongly-exogenous variables $x_{6,t},\ldots,x_{10,t}$ and the first lags of all regressors:

$$y_t = \pi_{0,1} y_{t-1} + \sum_{k=1}^{10}\sum_{i=0}^{1} \pi_{k,i} x_{k,t-i} + \pi_{0,0} + u_t, \quad u_t \sim \mathsf{IN}\left[0,\sigma^2\right]. \tag{2}$$

The sample size T is 100 or 1000 and the number of replications M is 1000.

The orthogonality of the regressors allows an easier analysis. Recall that the t-test of the null $\beta_k = 0$ versus the alternative $\beta_k \neq 0$ is given by:

$$\mathsf{t}_k = \frac{\widehat{\beta}_k}{\widehat{\sigma}_{\beta_k}} = \frac{\widehat{\beta}_k}{\sqrt{\widehat{\sigma}_\varepsilon^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}_{kk}}} = \frac{\widehat{\beta}_k / \sqrt{\sigma_\varepsilon^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}_{kk}}}{\sqrt{\widehat{\sigma}_\varepsilon^2/\sigma_\varepsilon^2}}.$$

The population value of the t-statistic is:

$$\mathsf{t}_k^* = \frac{\beta_k}{\sigma_{\beta_k}} = \frac{\beta_k}{T^{-\frac{1}{2}} Q_{kk}^{-1/2} \sigma_\varepsilon},$$

where the moment matrix $Q = \lim_{T\to\infty}\left(T^{-1}\mathbf{X}'\mathbf{X}\right)$ is assumed to exist. Since the regressors are orthogonal, we have that $\widehat{\beta}_k = \widehat{\sigma}_{x_k y}/\widehat{\sigma}_k^2$ and $\widehat{\sigma}_{\beta_k}^2 = \widehat{\sigma}_\varepsilon^2/(T\widehat{\sigma}_k^2)$:

$$\mathsf{t}_k = \frac{\widehat{\beta}_k}{\widehat{\sigma}_{\beta_k}} = \sqrt{T}\widehat{\beta}_k \frac{\widehat{\sigma}_k}{\widehat{\sigma}_\varepsilon}.$$

Thus the non-zero population t-values are $2, 3, 4, 6, 8$. In (2), 17 of 22 regressors are nuisance.

3.3 Evaluation of the Monte Carlo

The evaluation of Monte Carlo experiments always involves measurement problems: see Hendry (1984). A serious problem here is that, with some positive probability, the GUM – and the 'truth' – will get rejected *ab initio* on diagnostic tests. Tests are constructed to have non-zero nominal size under their null, so sometimes the truth will be rejected: and the more often, the more tests that are used. Three possible strategies suggest themselves: one rejects that data sample, and randomly re-draws; one changes the rejection level of the 'offending' test; or one specifies a more general GUM which is congruent. We consider these alternatives in turn.

Hoover and Perez (1999) use a '2-significant test rejections' criterion to discard a sample and redraw, which probably slightly favours the performance of $Gets$. In our Monte Carlo with $PcGets$, the problem is 'solved' by endogenously adjusting the significance levels of tests that reject the GUM (e.g., 1% to 0.1%). Such a 'solution' is feasible in a Monte Carlo, but metaphysical in reality, as one could never know that a sample from an economy was 'unrepresentative', since time series are not repeatable. Thus, an investigator could never 'know' that the DGP was simpler empirically than the data suggest (although such a finding might gradually emerge in a PRS), and so would probably generalize the initial GUM. We do not adopt that solution here, partly because of the difficulties inherent in the constructive use of diagnostic-test rejections, and partly because it is moot whether the $PcGet$ algorithm 'fails by overfitting' on such aberrant samples, when in a non-replicable world, one would conclude that such features really were aspects of the DGP. Notice that fitting the 'true' equation, then testing it against such alternatives, would also lead to rejection in this setting, unless the investigator knew the truth, and knew that she knew it, so *no* tests were needed. While more research is needed on cases where the DGP would be rejected against the GUM, here we allow $PcGets$ to adjust significance levels endogenously.

Another major decision concerns the basis of comparison: the 'truth' seems to be a natural choice, and both Lovell (1983) and Hoover and Perez (1999) measure how often the search finds the DGP exactly – or nearly. Nevertheless, we believe that 'finding the DGP exactly' is not a good choice of comparator, because it implicitly entails a basis where the truth is known, and one is *certain* that it is the truth. Rather, to isolate the costs of selection *per se*, we seek to match probabilities with the same procedures applied to testing the DGP. In each replication, the correct DGP equation is fitted, and the same selection criteria applied: we then compare the retention rates for DGP variables from $PcGets$ with those that occur when no search is needed, namely when inference is conducted once for each DGP variable, and additional (non-DGP) variables are never retained.

3.4 Diagnostic tests

PcGets records the rejection frequencies of both specification and mis-specification tests for the DGP, the initial GUM, and the various simplifications thereof based on the selection rules. Figure 1 displays quantile–quantile (QQ) plots of the empirical distributions of seven potential mis-specification tests for the estimated correct specification, the general model, and the finally-selected model. Some strong deviations from the theoretical distributions (diagonal) are evident: the portmanteau statistic (see Box and Pierce, 1970) rejects serial independence of the errors too often in the correct specification, never in the general, and too rarely in the final model. The hetero-x test (see White, 1980) was faced with degrees of freedom problems for the GUM, but anyway does not look good for the true and final model either. Since this incorrect finite-sample size of the diagnostic tests induces an excessively-early termination of any search path, resulting in an increased overall size for variable selection, we decided to exclude the portmanteau and the hetero-x diagnostics from the test battery of statistics. Thus, the following results use the five remaining diagnostic tests in table 3.

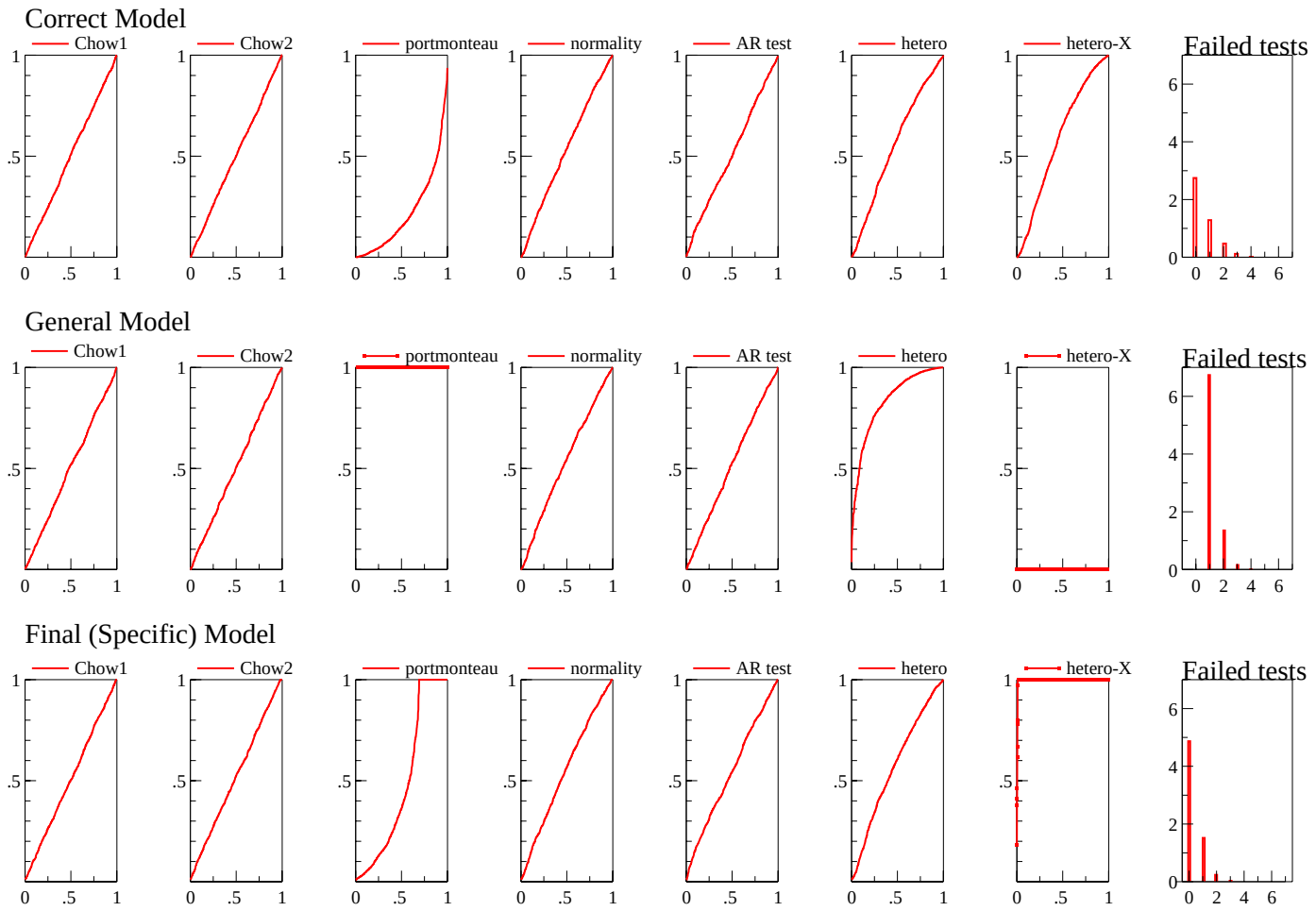

Figure 1 Selecting diagnostics: QQ Plots for $M = 1000$ and $T = 100$..

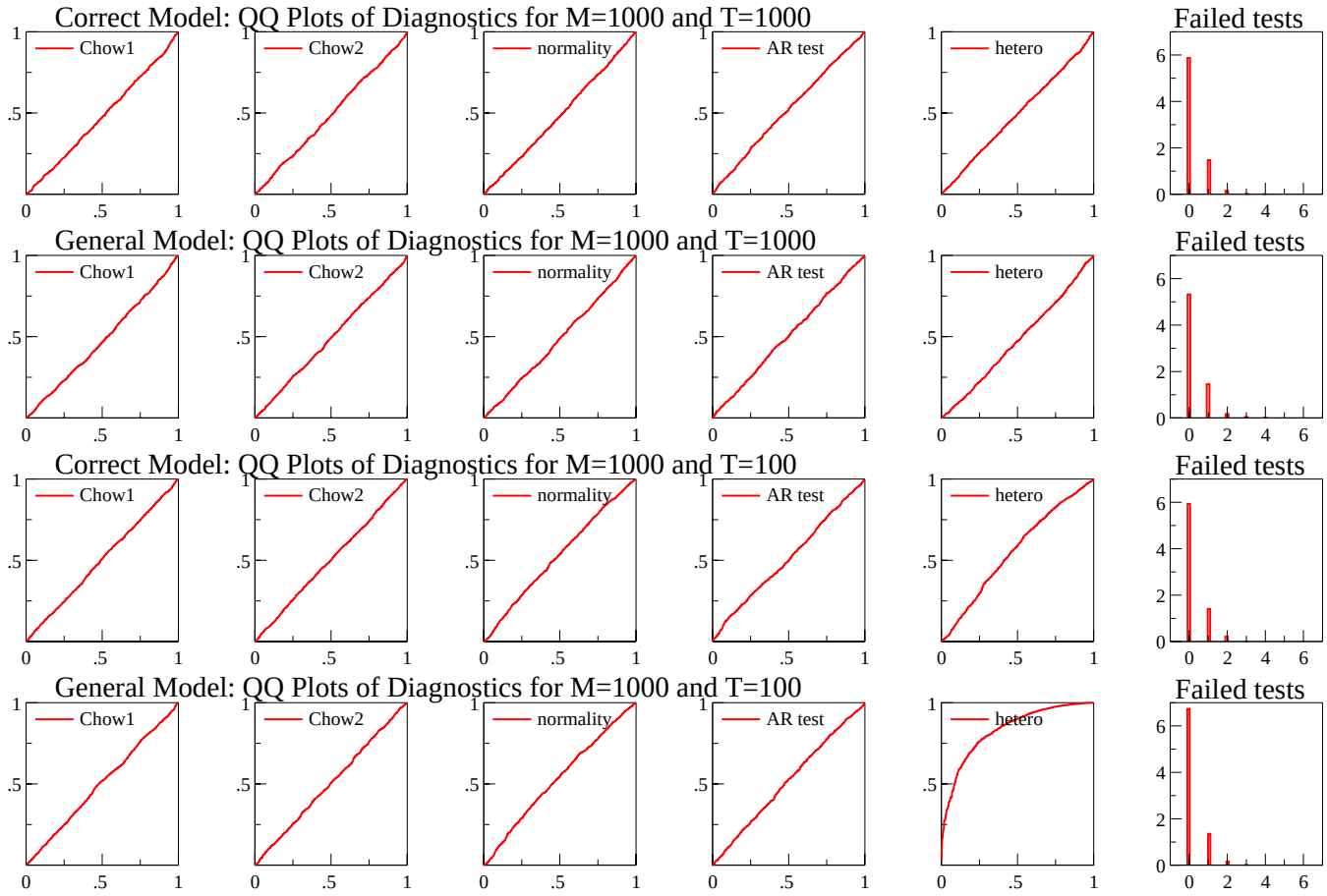

Figure 2 Diagnostics for small and large samples.

Table 3 Test battery.

Test	*Alternative*	*Statistic*	*Sources*
Chow($\tau_1 T$)	Predictive failure over a subset of $(1-\tau_1)T$ obs.	$F((1-\tau_1)T, \tau_1 T - k)$	Chow (1960, p.594-595),
Chow($\tau_2 T$)	Predictive failure over a subset of $(1-\tau_2)T$ obs.	$F((1-\tau_2)T, \tau_2 T - k)$	Hendry (1979)
portmanteau(r)	$r-$th order residual autocorrelation	$\chi^2(r)$	Box and Pierce (1970)
normality test	Skewness and excess kurtosis	$\chi^2(2)$	Jarque and Bera (1980), Doornik and Hansen (1994)
AR 1-p test	$p-$th order residual autocorrelation	$F(p, T-k-p)$	Godfrey (1978), Harvey (1981, p.173)
hetero test	Heteroscedasticity quadratic in regressors x_i^2	$F(q, T-k-q-1)$	White (1980), Nicholls and Pagan (1983),
hetero-x test	Heteroscedasticity quadratic in regressors $x_i x_j$	$F(q, T-k-q-1)$	Hendry and Doornik (1996)
F-test	General	$F(q, T-k-q)$	

There are T observations and k regressors in the model under the null. The value of q may differ across statistics, as may those of k and T across models. By default, *PcGets* sets $p = 4$, $r = 12$, $\tau_1 = [0.5T]/T$, and $\tau_2 = [0.9T]/T$.

Figure 2 demonstrates that for large samples ($T = 1000$), the empirical distributions of the test statistics are unaffected by the strongly-exogenous nuisance regressors. For small samples ($T = 100$), the properties of the mis-specification tests are still satisfactory, and except for the heteroscedasticity test, close to the distributions of the test statistics under the null of the true model.

3.5 Size and power of variable selection

Simplification can at best eliminate the nuisance regressors all or most of the time (size), yet retain the substance nearly as often as the DGP (power). The metric to judge the costs of reduction and of mis-specification testing was noted above. The probability is low of detecting an effect that has a scaled population t-value less than 2 in absolute value when the empirical selection criterion is larger. This suggests weighting the 'failure' of *PcGets* in relation to a variable's importance, statistically and economically. Then, 'missing' a variable with $|\mathtt{t}| < 2$ would count for less than missing an effect with $|\mathtt{t}| > 4$ (say). With such a baseline, low signal-noise variables will still rarely be selected, but that is attributable as a cost of inference, not a flaw of *Gets* type searches.

In the following, we measure the outcome of $PcGets$ by comparing its power and size with that of classical t-tests applied once to the correct DGP equation. The power function of a t-test of size α for including the k^{th} variable $x_{k,t}$ with coefficient $\beta_{k,0} \neq 0$ is given by:

$$\Pr\left(\text{'Include } x_{k,t}\text{'} \mid \beta_{k,0} \neq 0\right) = \Pr\left(|\mathtt{t}_k| \geq c_\alpha \mid \beta_{k,0} \neq 0\right),$$

when:

$$\Pr\left(|\mathtt{t}_k| \geq c_\alpha \mid \beta_{k,0} = 0\right) = \alpha.$$

The rejection probability is given by a non-central t distribution with ν degrees of freedom and non-centrality parameter ψ, which can be approximated by a normal distribution $\Phi(x)$ (see Abramowitz and Stegun, 1970), where:

$$x = \frac{\mathtt{t}\left(1 - \frac{1}{4\nu}\right) - \psi}{\left(1 + \frac{\mathtt{t}^2}{2\nu}\right)^{\frac{1}{2}}}.$$

The power of a t-test with size α and ν degrees of freedom is then given by the parametric variation of the population t-value in:

$$\Pr\left(\text{'Include } x_{k,t}\text{'} \mid \beta_{k,0} \neq 0\right) = \Pr\left(-\mathtt{t}_k \leq c_\alpha \mid \beta_{k,0} \neq 0\right) + \Pr\left(\mathtt{t}_k \geq c_\alpha \mid \beta_{k,0} \neq 0\right).$$

For $\alpha = 0.01$ and 0.05, and $\nu = 100$ and 1000, the power function is depicted in figure 3.

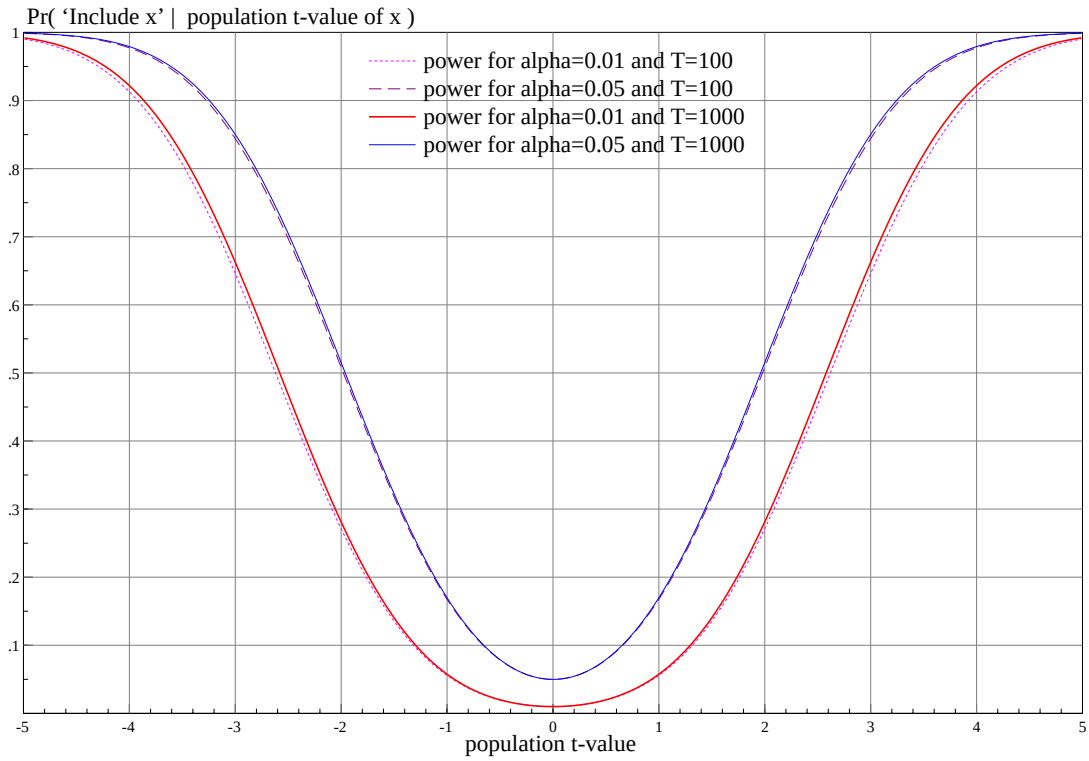

Figure 3 Power function of a t test for $\alpha = 0.01$ and 0.05, and $T = 100$ and 1000.

Table 4 shows that for large samples ($T = 1000$), the size (0.052 versus 0.05) and power (0.51 versus 0.515) are nearly at the **theoretical** levels one would expect from a t-test of the 'true' model. Hence the loss in size and power from the *PcGets* search is primarily a small-sample feature. But even for 100 observations and 22 regressors, the results from *PcGets* are promising. For $\alpha = \eta = 0.01$, the loss in power is less than 0.025 and the size is 0.019. The difference in the Monte Carlo experiments between *PcGets* and the **empirical** size and power of the t-test is even smaller. All of these experiments used AIC when more than one model survived both the model search and the encompassing process: using SC had little effect on the outcome, but improved the size slightly (see table 5). This match of actual and theoretical size suggests that the algorithm is not subject to 'spurious overfitting', since in a search over a large number of regressors, one must expect to select some adventious effects. Table 5 also shows the dramatic size improvements that can result from adding stage 0 (pre-search reduction) to the algorithm, at some cost in retaining relevant variables till $|\mathsf{t}| > 4$.

Smaller significance levels ($\alpha = 0.01$ versus 0.05) receive some support from the size-power trade-offs here when there are many irrelevant variables. Tighter significance levels for the t-tests reduce the empirical sizes accordingly, but lower the power substantially for variables with population t-values of 2 or 3, consistent with the large vertical differences between the power functions in figure 3 at such t-values. Otherwise, using 1% for all tests does well at the sample sizes current in macroeconomics, and dominates 5% dramatically on overall size, without much power loss for larger values of t. For example, on conventional t-tests alone, the probability of selecting no regressors when all 22 variables are irrelevant in (2) is:

$$\Pr\left(\left|\mathsf{t}_{\pi_{k,i}}\right| \leq c_\alpha \; \forall k, i \mid \pi_{k,i} = 0\right) = (1-\alpha)^{22},$$

which is 0.80 when $\alpha = 0.01$ but falls to 0.32 when $\alpha = 0.05$. Figure 4 clarifies the 'success' and 'failure' of *PcGets* for the 22 regressors of the GUM (the coefficients are in the order of equation 2).

Table 4 Power and Size I: Impact of diagnostic tests on *PcGets* t-tests.

	PcGets					t-test: simulated			t-test: theoretical		
α	0.05	0.05	0.05	0.05	0.01	0.05	0.05	0.01	0.05	0.05	0.01
η	0.05	0.01	0.00	0.01	0.01						
	AIC	AIC	AIC	AIC	AIC						
T, ν	100	100	100	1000	100	100	1000	100	100	1000	100
$\mathtt{t}=0$	0.0812	0.0686	0.0646	0.0521	0.0189				0.0500	0.0500	0.0100
	0.0022	*0.0021*	*0.0019*	*0.0017*	*0.0010*						
$\mathtt{t}=2$	0.5090	0.4930	0.4910	0.5100	0.2820	0.4730	0.5010	0.2580	0.5083	0.5152	0.2713
	0.0158	*0.0158*	*0.0158*	*0.0158*	*0.0142*						
$\mathtt{t}=3$	0.8070	0.8020	0.8010	0.8340	0.6210	0.8120	0.8360	0.6130	0.8440	0.8502	0.6459
	0.0125	*0.0125*	*0.0126*	*0.0118*	*0.0153*						
$\mathtt{t}=4$	0.9750	0.9720	0.9710	0.9880	0.9000	0.9760	0.9850	0.9020	0.9773	0.9791	0.9127
	0.0049	*0.0049*	*0.0053*	*0.0034*	*0.0095*						
$\mathtt{t}=6$	0.9990	0.9990	0.9990	1.0000	0.9990	1.0000	1.0000	0.9990	1.0000	1.0000	0.9996
	0.0010	*0.0010*	*0.0010*	*0.0000*	*0.0010*						
$\mathtt{t}=8$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0.0000	*0.0000*	*0.0000*	*0.0000*	*0.0000*						

The table reports the size (selection probability when population $\mathtt{t}=0$) and the power (non-deletion probability for population $\mathtt{t}>0$) of a standard t-test and *PcGets*. Standard errors shown in italics.

Table 5 Power and Size II: Information criteria and pre-search reduction.

	PcGets							
α	0.05	0.05	0.05	0.05	0.01	0.01	0.01	0.01
IC	AIC	HQ	SC	SC	AIC	HQ	SC	SC
Pre-search	—	—	—	yes	—	—	—	yes
$\mathtt{t}=0$	0.0686	0.0679	0.0677	0.0477	0.0189	0.0183	0.0185	0.0088
	0.0021	*0.0019*	*0.0019*	*0.0015*	*0.0010*	*0.0010*	*0.0010*	*0.0006*
$\mathtt{t}=2$	0.4930	0.4930	0.4930	0.4080	0.2820	0.2810	0.2820	0.1538
	0.0158	*0.0158*	*0.0158*	*0.0140*	*0.0142*	*0.0142*	*0.0142*	*0.0106*
$\mathtt{t}=3$	0.8020	0.8020	0.8020	0.7330	0.6210	0.6220	0.6200	0.4278
	0.0125	*0.0126*	*0.0126*	*0.0124*	*0.0153*	*0.0153*	*0.0154*	*0.0147*
$\mathtt{t}=4$	0.9720	0.9720	0.9720	0.9390	0.9000	0.9000	0.8980	0.7645
	0.0049	*0.0052*	*0.0053*	*0.0061*	*0.0095*	*0.0096*	*0.0096*	*0.0125*
$\mathtt{t}=6$	0.9990	0.9990	0.9990	0.9980	0.9990	0.9990	0.9990	0.9865
	0.0010	*0.0010*	*0.0010*	*0.0012*	*0.0010*	*0.0010*	*0.0010*	*0.0034*
$\mathtt{t}=8$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0.0000	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*

All MC experiments use $\eta = 0.01$ and $T = \nu = 100$. Standard errors shown in italics.

3.6 Test size analysis

The Monte Carlo reveals strong effects from the choice of the significance levels on the outcome of *Gets*. As table 4 shows, it is not only the significance level of the t-tests (α) that matters, but also those of the diagnostics (η): lowering the significance level of the diagnostic tests from 0.05 to 0.01 reduces the size by 0.0126 without affecting the power (e.g. loss of 0.0050 at $\mathtt{t}=3$ and $T=100$). This is a striking effect which merits closer examination.

It is important to distinguish between the individual significance levels (η), and the overall significance level of the test battery – which can be difficult to determine. Suppose we have a battery of n mis-specification tests each evaluated at the significance level η. Assuming independence of the tests, the overall rejection probability under the null is given by:

$$1-(1-\eta)^n.$$

For example if $n=5$ and $\eta=0.05$, then the probability of rejecting the DGP is 0.2262, which is

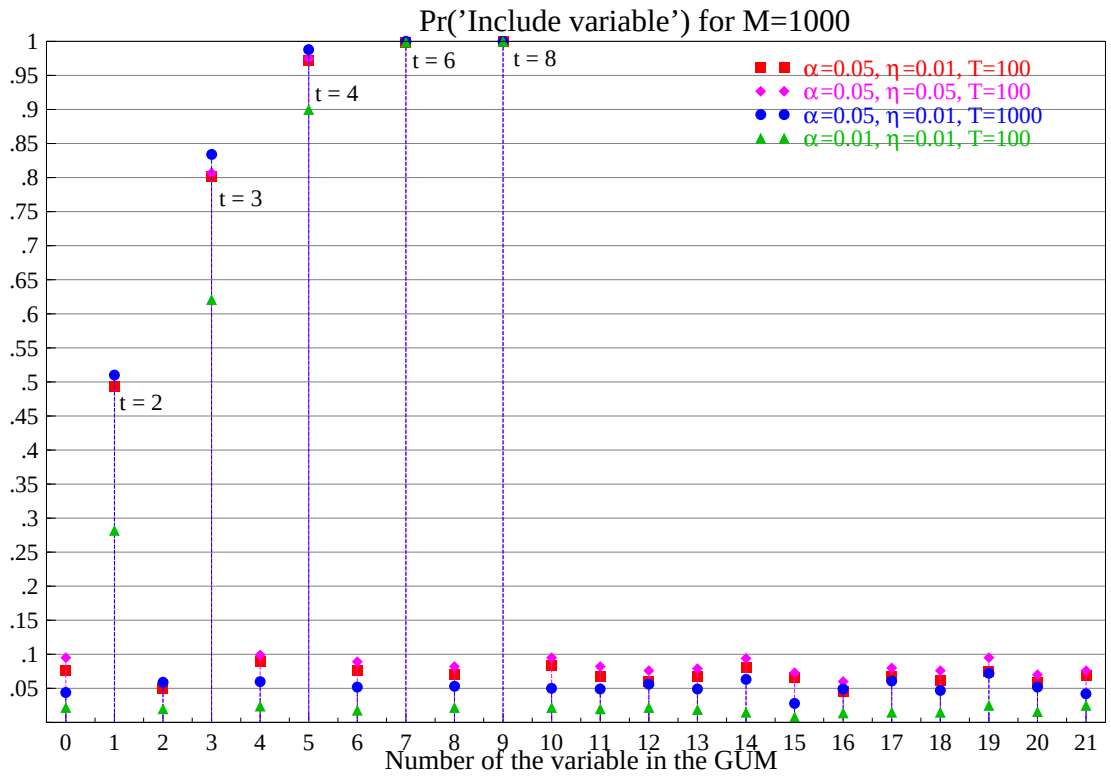

Figure 4 Probability of including each variable.

substantial. To ensure an overall rejection probability of 0.05 under the null, the individual significance level η has to satisfy $(1-\eta)^n = 0.95$. For, say, $n = 5$ mis-specification tests, then $\eta \simeq 0.01$ is necessary.

The combined variable-selection and specification-testing approach implies that the significance levels, η, of the diagnostics also affect the deletion probabilities for any non-DGP variable. The probability of selecting a non-DGP variable must be higher than the nominal size α (see figure 4), since for pure random sampling, excluding even a nuisance variable can alter the outcome of one or more diagnostic tests to a rejection. Thus, despite an insignificant t-statistic, *PcGets* would keep such a variable in the model.

The joint issue of variable selection and diagnostic testing in *PcGets* is hard to analyse, but consider the following one-step reduction from a GUM denoted M_1 to the DGP M_2:

$$\mathsf{M}_1 \text{ (GUM): } y_t = \mathbf{x}_t'\boldsymbol{\delta}_1 + \mathbf{z}_t'\boldsymbol{\delta}_2 + u_t,$$

$$\mathsf{M}_2 \text{ (DGP): } y_t = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t,$$

where ε_t is a homoscedastic innovation process. The reduction step consists of a specification test for $\boldsymbol{\delta}_2 = \mathbf{0}$ (t for scalar $\mathbf{z}_t$, or F for the vector case) and a set of n mis-specification tests. The individual significance levels are α and η, respectively: we assume that these tests are independent.

Let $\mathcal{R}_{\mathsf{M}}$ ($\overline{\mathcal{R}}_{\mathsf{M}}$) denote a (non-)rejection on the set of mis-specification tests for model M. For M_1:

$$\Pr(\overline{\mathcal{R}}_{\mathsf{M}_1}) = (1-\eta)^n = 1 - \Pr(\mathcal{R}_{\mathsf{M}_1}), \tag{3}$$

which is the initial probability that the GUM is not rejected by the mis-specification test battery, discussed above. Then, the overall size, α^* (probability of no reduction from M_1), is:

$$\alpha^* = \alpha + (1-\alpha)\Pr\left(\mathcal{R}_{\mathsf{M}_2} \mid \overline{\mathcal{R}}_{\mathsf{M}_1}\right) \simeq \alpha + (1-\alpha)\left(1-(1-\lambda(\eta))^n\right), \tag{4}$$

where the second term shows the inflation of the nominal size α due to diagnostic testing. It seems reasonable to let $\Pr\left(\mathcal{R}_{\mathsf{M}_2}|\overline{\mathcal{R}}_{\mathsf{M}_1}\right) = 1-(1-\lambda(\eta))^n$, as rejection is increasing in n and decreasing in η. Equation (4) provides a lower bound for the model-search procedure. For example, it does not take into account that DGP variables might be eliminated by selection tests, and there might be a sequence of t tests rather than a single F test for the same set of variables. The multi-step, multi-path reduction process of *PcGets* makes an analytical assessment intractable, but we get insights from the Monte Carlo findings. From table 4 (where $n = 5$), the empirical sizes of $PcGets$ selection at a nominal selection-test size of 5% are 0.065, 0.069, and 0.081 when conducted without diagnostic testing (i.e., $\eta = 0$), $\eta = 0.01$ and $\eta = 0.05$ respectively. Solving (4) for these effects gives estimates of $\lambda\left(0.01\right) = 0.0008$ and $\lambda\left(0.05\right) = 0.0035$. Although Monte Carlo is always problem dependent, such findings cohere with the established theory: the additional checking does not increase size greatly so long as we control the overall size of the test battery.

4 Empirical Illustrations

In this section, we apply *PcGets* to two well-known macro-econometric models: Davidson, Hendry, Srba and Yeo (1978) and Hendry and Ericsson (1991b).

4.1 DHSY

We reconsider the single-equation equilibrium-correction model of Davidson *et al.* (1978) (DHSY), who proposed the following model of consumers' expenditure in the UK:

$$\Delta_4 \widehat{c}_t = -0.09(c-y)_{t-4} + 0.48\Delta_4 y_t - 0.23\Delta_1\Delta_4 y_t - 0.12\Delta_4 p_t - 0.31\Delta_1\Delta_4 p_t + 0.006\Delta_4 D_t, \quad (5)$$

where y_t denotes real personal disposable income, c_t is real consumers' expenditures on non-durable goods and services, p_t is the implicit deflator of c_t (all in logs), and D is a dummy variable with the value unity in $68(i)$ and $73(i)$, -1 in the following quarter, and zero otherwise. We started with a GUM that generalizes (5) by allowing for 5 lags of all differenced variables, their EqCM, and a constant.

The results are summarized in table 6. All three final models are valid non-dominated reductions of the GUM, only differing slightly in their dynamic specification. Model 1 would be preferable if one applied the AIC criterion, model 2 in the case of HQ and SC. It is worth noting that introducing centered seasonal dummies into the general model (GUM 2) eliminates model 3 from the set of final models, but otherwise results in identical models. This is surprising, as there are significant seasonal effects in GUM 2. A comparison with the model chosen by DHSY shows the potential power of an automated general-to-specific model selection: *PcGets* detects ones that dominate DHSY's final selection. The DHSY results reflect their decision to delete the Δc_{t-3} term, which possibly leads to an economically more sensible model.

The value of structuring the problem before specifying the GUM becomes obvious in table 7 where we try up to 5 lags of each for c_t on y_t, $\Delta_4 p_t$, and a constant to see what *PcGets* finds. In table 8, seasonals are added.

It might be reassuring to see how the power of an 'artificial intelligence' model-specification system is limited by the researcher's specification of the GUM. This is very much in line with the ideas presaged in section 2.4.

Table 6 DHSY 1959 (2) - 1975 (4) .

	DHSY		GUM 1		GUM 2		Final model 1 *(GUM1, GUM2)*		Final model 2 *(GUM1, GUM2)*		Final model 3 *(GUM1)*	
	Coeff	t-value	Coeff	t-value	Coeff	t-value	Coeff	t-value	Coeff	t-value	Coeff	t-value
$\Delta_4 c_{t-1}$			0.0743	0.5011	0.0041	0.0273						
$\Delta_4 c_{t-2}$			-0.0472	-0.3502	-0.0275	-0.2069						
$\Delta_4 c_{t-3}$			0.2304	1.7445	0.2317	1.7767	0.1743	3.0144	0.2572	3.7712	0.1706	2.7910
$\Delta_4 c_{t-4}$			-0.0545	-0.3959	-0.0399	-0.2939						
$\Delta_4 c_{t-5}$			0.0821	0.6450	-0.0045	-0.0338						
$\Delta_4 y_t$	0.2508	7.0717	0.2332	5.2662	0.2594	5.6941	0.2486	7.5433	0.2614	8.1151	0.2362	6.8964
$\Delta_4 y_{t-1}$	0.2300	5.7883	0.1618	2.5850	0.1787	2.8231	0.1891	4.8657	0.1505	4.3485	0.1874	4.7449
$\Delta_4 y_{t-2}$			0.0267	0.4318	0.0353	0.5757						
$\Delta_4 y_{t-3}$			-0.0226	-0.3912	-0.0261	-0.4551						
$\Delta_4 y_{t-4}$			0.0069	0.0938	-0.0265	-0.3489						
$\Delta_4 y_{t-5}$			-0.0300	-0.4333	-0.0130	-0.1853						
$\Delta_4 p_t$	-0.4227	-4.3869	-0.4309	-3.3889	-0.4339	-3.4581	-0.4284	-4.7064	-0.2930	-6.5579	-0.3708	-3.9073
$\Delta_4 p_{t-1}$	0.3051	2.9585	0.2633	1.3715	0.2170	1.1377	0.2668	2.5528			0.2307	2.1328
$\Delta_4 p_{t-2}$			-0.0210	-0.1146	0.0322	0.1745						
$\Delta_4 p_{t-3}$			0.2005	1.0743	0.1677	0.9034			0.2490	3.9625		
$\Delta_4 p_{t-4}$			-0.2537	-1.3480	-0.2317	-1.2414						
$\Delta_4 p_{t-5}$			0.2110	1.4990	0.1215	0.8370	0.1150	2.3280			0.1250	2.5184
$(c-y)_{t-4}$	-0.0930	-7.7312	-0.0349	-1.1461	-0.1315	-2.3522	-0.0591	-3.9155	-0.0504	-3.0787		
Constant			0.0026	0.5100	-0.0069	-1.0139					0.0087	3.6101
$\Delta_4 D_t$	0.0065	2.8869	0.0083	2.7285	0.0074	2.4178	0.0063	3.0116	0.0068	3.2809	0.0060	2.8834
CSeason_1					0.0085	2.0815						
CSeason_2					0.0051	1.6312						
CSeason_3					0.0043	1.5988						
RSS		0.0023		0.0018		0.0016		0.0019		0.0020		0.0020
sigma		0.0062		0.0062		0.0061		0.0057		0.0058		0.0058
R^2		0.8528		0.8867		0.8970		0.8770		0.8720		0.8731
R^2_{adj}		0.7764		0.6220		0.5891		0.7723		0.7809		0.7688
LogLik		344.0657		352.8531		356.0237		350.0929		348.7563		349.0408
AIC		-10.0915		-9.9359		-9.9410		**-10.2117**		-10.2017		-10.1803
HQ		-10.0134		-9.6755		-9.6415		-10.1076		**-10.1105**		-10.0762
SC		-9.8941		-9.2778		-9.1842		-9.9485		**-9.9713**		-9.9171
Chow(1967:4)	1.0142	0.4887	0.9630	0.5573	0.7620	0.7388	0.8585	0.6641	0.8551	0.6687	0.9570	0.5527
Chow(1973:4)	0.7281	0.6945	0.7187	0.7017	0.4718	0.8966	0.6256	0.7847	0.7414	0.6824	0.7863	0.6417
normality test	0.1120	0.9455	0.6559	0.7204	0.0374	0.9815	0.5170	0.7722	0.8055	0.6685	1.4087	0.4944
AR 1-5 test	0.2239	0.9507	0.3498	0.8795	0.4442	0.8148	0.1000	0.9917	0.3536	0.8778	0.4999	0.7750
hetero test	0.6478	0.7906	0.3855	0.9772	0.1980	0.9891	0.6231	0.8469	0.6810	0.7801	0.4172	0.9611

(1) DHSY corresponds to equation (8.45)** in Davidson *et al.* (1978).

(2) $\Delta_4 D_t$ is the fourth difference of a dummy variable which is +1,-1 in 1968(i), 1968(ii) and 1973(i),1973(ii) reflecting budget effects in 1968 and the introduction of VAT in 1973 (see footnote 5 in Davidson *et al.*, 1978)

(3) CSeason_1,2,3 are centralised seasonal dummies.

Table 7 DHSY without SEASONALS, 1959 (2) - 1976 (2) .

	General model		Final model 1		Final model 2		Final model 3		Final model 4	
	Coeff	t-value	Coeff	t-value	Coeff	t-value	Coeff	t-value	Coeff	t-value
$\Delta_4 c_{t-1}$	-0.0900	-0.6248								
$\Delta_4 c_{t-2}$	-0.0377	-0.2660								
$\Delta_4 c_{t-3}$	0.2406	1.7255	0.2554	3.1972	0.1471	2.4153	0.1871	2.7581	0.1524	2.3475
$\Delta_4 c_{t-4}$	-0.0079	-0.0551								
$\Delta_4 c_{t-5}$	0.0395	0.2943								
$\Delta_4 y_t$	0.2489	5.3523	0.2630	7.5598	0.2762	7.8876	0.2859	8.2106	0.2738	7.8156
$\Delta_4 y_{t-1}$	0.2030	3.1752	0.1896	4.8375	0.1879	4.6804	0.1585	4.1983	0.1346	3.7473
$\Delta_4 y_{t-2}$	0.0540	0.8294								
$\Delta_4 y_{t-3}$	-0.0064	-0.1074								
$\Delta_4 y_{t-4}$	-0.0934	-1.3259	-0.0804	-2.0297						
$\Delta_4 y_{t-5}$	0.0186	0.2645								
$\Delta_4 p_t$	-0.3919	-3.0655	-0.3999	-5.0674	-0.3919	-4.8537	-0.2742	-5.3097	-0.1693	-5.8668
$\Delta_4 p_{t-1}$	0.2399	1.1928	0.3032	3.6100	0.2868	3.3490				
$\Delta_4 p_{t-2}$	0.0186	0.0958					0.1829	3.0574	0.1829	3.0574
$\Delta_4 p_{t-3}$	0.0601	0.3192								
$\Delta_4 p_{t-4}$	-0.2125	-1.0761								
$\Delta_4 p_{t-5}$	0.2154	1.4561							0.1427	3.3997
$(c-y)_{t-4}$	-0.0514	-1.6258	-0.0696	-4.6275	-0.0699	-4.5376	-0.0643	-3.8717	-0.0643	-3.8717
Constant	0.0033	0.6162							0.0099	3.9229
RSS		0.0022		0.0023		0.0025		0.0026		0.0025
sigma		0.0066		0.0061		0.0063		0.0064		0.0063
R^2		0.8804		0.8705		0.8619		0.8583		0.8626
Radj^2		0.6380		0.7822		0.7869		0.7837		0.7876
LogLik		357.8109		355.0706		352.8512		351.9716		353.0330
AIC		-9.8206		-10.0890		-10.0537		-10.0282		-10.0589
HQ		-9.5765		-9.9991		-9.9766		-9.9511		-9.9819
SC		-9.2054		-9.8624		-9.8594		-9.8339		-9.8647
Chow(1967:4)	1.3920	0.2434	1.1658	0.3414	1.1699	0.3354	1.0505	0.4494	1.3654	0.1978
Chow(1973:4)	0.9134	0.5305	0.6604	0.7551	0.7546	0.6706	0.8571	0.5776	0.8929	0.5459
normality test	0.2589	0.8786	0.3234	0.8507	0.0989	0.9517	0.4123	0.8137	0.3682	0.8319
AR 1-5 test	0.2791	0.9222	0.2078	0.9579	0.2303	0.9478	0.2538	0.9362	0.4904	0.7821
hetero test	0.2646	0.9992	0.6009	0.8505	0.7030	0.7409	0.7415	0.7047	0.4176	0.9318

Table 8 DHSY with SEASONALS, 1959 (2) - 1976 (2) .

	General model		Final model 1		Final model 2		Final model 3	
	Coeff	t-value	Coeff	t-value	Coeff	t-value	Coeff	t-value
$\Delta_4 c_{t-1}$	-0.1582	-1.1301						
$\Delta_4 c_{t-2}$	-0.0260	-0.1915						
$\Delta_4 c_{t-3}$	0.2234	1.6707	0.2256	2.8686	0.2539	2.9652	0.1315	2.0653
$\Delta_4 c_{t-4}$	0.0117	0.0851						
$\Delta_4 c_{t-5}$	-0.0609	-0.4550						
$\Delta_4 y_t$	0.2726	5.9369	0.2587	7.6527	0.2694	7.9734	0.2888	8.7700
$\Delta_4 y_{t-1}$	0.2221	3.5704	0.1839	4.8264	0.1555	4.3204	0.1331	3.7959
$\Delta_4 y_{t-2}$	0.0588	0.9378						
$\Delta_4 y_{t-3}$	-0.0047	-0.0831						
$\Delta_4 y_{t-4}$	-0.1178	-1.7262	-0.0836	-2.1744	-0.0801	-2.0466		
$\Delta_4 y_{t-5}$	0.0426	0.6174						
$\Delta_4 p_t$	-0.4303	-3.4959	-0.3794	-4.9201	-0.2646	-5.3552	-0.2035	-7.0134
$\Delta_4 p_{t-1}$	0.2027	1.0457	0.2661	3.1994				
$\Delta_4 p_{t-2}$	0.0557	0.2932			0.1622	2.7605		
$\Delta_4 p_{t-3}$	0.0612	0.3375						
$\Delta_4 p_{t-4}$	-0.1809	-0.9521						
$\Delta_4 p_{t-5}$	0.1200	0.8230					0.1244	2.9540
$(c-y)_{t-4}$	-0.1711	-3.2051	-0.0827	-5.2521	-0.0784	-4.5545	-0.0765	-4.4559
Constant	-0.0092	-1.3280						
CSeason_1	0.0107	2.7376	0.0039	2.2111	0.0039	2.1400	0.0039	2.1593
CSeason_2	0.0065	2.0817						
CSeason_3	0.0046	1.6746						
RSS		0.0019		0.0022		0.0022		0.0023
sigma		0.0063		0.0060		0.0061		0.0061
R^2		0.8972		0.8801		0.8755		0.8710
$Radj^2$		0.6111		0.7781		0.7740		0.7826
LogLik		363.0433		357.7306		356.4398		355.2075
AIC		-9.8853		-10.1371		-10.0997		-10.0930
HQ		-9.6027		-10.0344		-9.9969		-10.0031
SC		-9.1730		-9.8781		-9.8407		-9.8663
Chow(1967:4)	1.0230	0.5082	1.0827	0.4201	1.1253	0.3797	1.1338	0.3697
Chow(1973:4)	0.5628	0.8330	0.5121	0.8736	0.5742	0.8272	0.8614	0.5739
normality test	0.0787	0.9614	0.4538	0.7970	0.7825	0.6762	0.6811	0.7114
AR 1-5 test	1.2453	0.3054	0.1856	0.9669	0.2376	0.9442	0.4448	0.8153
hetero test	0.1703	0.9999	0.6109	0.8504	0.7793	0.6924	0.8515	0.6062

4.2 UK Money Demand

We now reconsider the Hendry and Ericsson (1991b) model (HE) of narrow money demand in the UK:

$$\Delta\widehat{(m-p)}_t = -0.093\,(m-p-x)_{t-1} - 0.17\Delta\,(m-p-x)_{t-1} - 0.69\Delta p_t - 0.63R_t + 0.023 \quad (6)$$

where the lower-case data are in logs: m is M1, x is real total final expenditure in 1985 prices, p is its deflator, and R is the opportunity cost of holding money (3-month local-authority interest rate minus the retail sight-deposit rate). The results using *PcGets* are given in table 10.

Two GUMs are considered, both nesting (6) without imposing the homogeneity restriction on $\Delta(m-p)$ and Δx. As the GUMs A and B are linked by linear transformations of the same set of regressors leading to an identical fit, A and B are observationally equivalent. Although transformations *per se* do not entail any associated reduction, the different structuring of the information set affects the reduction process. This highlights the role of variable orthogonalization.

Just one model survived the selection process in each case. Final model A corresponds to the HE model without imposing the homogeneity restriction and, hence, leaves a further valid reduction: the Δx_{t-1} coefficient is dropped by *PcGets* after acceptance by the corresponding t-test (also supported by HQ and SC). The final model resulting from GUM B is also essentially the HE model as:

$$\begin{aligned} -0.8\Delta^2 p_t - 0.7\Delta p_{t-1} &= -0.8\Delta p_t + 0.8\Delta p_{t-1} - 0.7\Delta p_{t-1} \\ &= -0.8\Delta p_t + 0.1\Delta p_{t-1}, \end{aligned}$$

and the last term is irrelevant. But only an 'expert system' would notice the link between the regressors to cancel the redundant term. However, the initial formulation of regressors clearly matters, supporting EqCM forms, and confirming that orthogonalization helps. Alternatively, if the cointegrating vector is used:

$$(m-p-x)_{t-1} - 7\Delta p_{t-1} - 0.7R_{t-1},$$

then both $\Delta^2 p$ and ΔR enter (see Hendry and Doornik, 1994), so that is also a correct feature detected. *PcGets* seems to be doing remarkably well as an expert on the empirical problems, as well as mimicking the good size and power properties which Hoover and Perez (1999) claim.

To illustrate the benefits from structuring the problem, we consider a simple unrestricted autoregressive-distributed lag model for UK M1, regressing m on p, x, and R with a lag length of 4. As shown in table 9, three reduced models survive the model-selection process, and differ regarding their dynamic specification, but are close regarding their long-run effects. Model 1 uniformly dominates 2 and 3. When rewritten in an EqCM form, the selected outcome is again similar to HE:

$$\begin{aligned} \Delta m_t &= -0.33m_{t-1} + 0.21m_{t-4} + 0.33p_t - 0.20p_{t-3} + 0.13x_t - 0.58R_t - 0.34R_{t-2} \\ &\simeq -0.11(m-p-x)_{t-1} - 0.21\Delta_3 m_{t-1} + 0.20\Delta_3 p_t + 0.11\Delta x_t - 0.92R_t + 0.34\Delta_2 R_t \quad (7) \end{aligned}$$

If pre-search F-tests are used (at 10%), the final model is the same as (7) other than omitting R_{t-2}. It must be stressed that these cases benefit from 'fore-knowledge' (e.g., of dummies, lag length etc.), some of which took the initial investigators time to find.

Table 9 UKM1 Money Demand, 1964 (1) - 1989 (2).

	HE		Unrestricted HE		General A		Final model A		General B		Final model B	
	Coeff	t-value	Coeff	t-value	Coeff	t-value	Coeff	t-value	Coeff	t-value	Coeff	t-value
$(m-p-x)_{t-1}$	-0.0928	-10.8734	-0.0938	-10.6160	-0.1584	-6.0936	-0.0934	-10.5108	-0.1584	-6.0936	-0.1035	-9.1048
Δp_t	-0.6870	-5.4783	-0.6952	-5.4693	-1.0499	-4.8670	-0.7005	-5.4831				
Δp_{t-1}									-1.0499	-4.8670	-0.7021	-4.8215
r_t	-0.6296	-10.4641	-0.6411	-9.8391	-1.1121	-6.1012	-0.6468	-9.8893				
r_{t-1}									-1.1121	-6.1012	-0.7223	-9.2106
$\Delta(m-p)_{t-1}$	-0.1746	-3.0102	-0.1926	-2.7637	-0.2827	-2.7449	-0.1858	-2.6569	-0.2827	-2.7449	-0.2520	-2.7609
$\Delta(m-p)_{t-2}$					-0.0407	-0.3696			-0.0407	-0.3696		
$\Delta(m-p)_{t-3}$					-0.2906	-2.6800			-0.2906	-2.6800		
$\Delta(m-p)_{t-4}$					-0.1446	-1.3519			-0.1446	-1.3519		
Δx_t					-0.0623	-0.5509			-0.0623	-0.5509		
Δx_{t-1}	0.1746	3.0102	0.1384	1.4392	0.0718	0.5870			0.0718	0.5870		
Δx_{t-2}					0.0083	0.0720			0.0083	0.0720		
Δx_{t-3}					-0.2274	-1.8802			-0.2274	-1.8802		
Δx_{t-4}					-0.0925	-0.7815			-0.0925	-0.7815		
$\Delta^2 p_t$					0.3222	1.0738			-0.7276	-3.5087	-0.8010	-4.3777
$\Delta^2 p_{t-1}$					0.3483	1.2135			0.3483	1.2135		
$\Delta^2 p_{t-2}$					0.6813	2.4708			0.6813	2.4708		
$\Delta^2 p_{t-3}$					0.2944	1.1908			0.2944	1.1908		
$\Delta^2 p_{t-4}$					-0.0430	-0.2134			-0.0430	-0.2134		
Δr_t					0.6884	3.1647			-0.4236	-3.6397	-0.4842	-4.5733
Δr_{t-1}					0.3293	1.7151			0.3293	1.7151		
Δr_{t-2}					0.2038	1.2647			0.2038	1.2647		
Δr_{t-3}					0.1631	1.1558			0.1631	1.1558		
Δr_{t-4}					0.0872	0.7119			0.0872	0.7119		
Constant	0.0234	5.8186	0.0244	5.3756	0.0434	5.1072	0.0262	5.9862	0.0434	5.1072	0.0276	6.1696
RSS		0.0164		0.0163		0.0130		0.0167		0.0130		0.0160
sigma		0.0131		0.0132		0.0130		0.0133		0.0130		0.0131
R^2		0.7616		0.7622		0.8103		0.7569		0.8103		0.7664
$Radj^2$		0.7235		0.7165		0.6240		0.7191		0.6240		0.7128
LogLik		435.8552		435.9734		447.2861		434.8837		447.2861		436.8711
AIC		-8.6171		-8.5995		-8.4857		-8.5977		-8.4857		-8.5974
HQ		-8.5644		-8.5362		-8.2432		-8.5450		-8.2432		-8.5236
SC		-8.4868		-8.4432		-7.8865		-8.4674		-7.8865		-8.4151
Chow(1967:4)	0.5871	0.9658	0.5712	0.9717	0.6029	0.9404	0.5464	0.9805	0.6029	0.9404	0.4777	0.9937
Chow(1973:4)	0.6367	0.8267	0.6204	0.8406	0.6135	0.8439	0.5489	0.8958	0.6135	0.8439	0.5379	0.9031
normality test	1.9766	0.3722	2.4432	0.2948	6.2503	0.0439	5.2260	0.0733	6.2503	0.0439	6.9513	0.0309
AR 1-5 test	1.6810	0.1473	1.7672	0.1278	0.4112	0.8395	1.4427	0.2167	0.4112	0.8395	1.4285	0.2219
hetero test	1.7820	0.0916	1.1874	0.3112	0.5902	0.9479	1.7668	0.0948	0.5553	0.9649	1.0354	0.4256

Table 10 UKM1 Money Demand, 1964 (1) - 1989 (2).

	General model		Final model 1		Final model 2		Final model 3	
	Coeff	t-value	Coeff	t-value	Coeff	t-value	Coeff	t-value
m_{t-1}	0.6265	5.8463	0.6661	9.3474	0.6963	9.9698	0.6309	6.4812
m_{t-2}	0.1744	1.4059					0.2708	2.9985
m_{t-3}	-0.2084	-1.6438						
m_{t-4}	0.2815	2.7652	0.2083	3.5952	0.1847	3.2485		
p_t	0.1466	0.6854	0.3322	5.9563				
p_{t-1}	0.3099	0.8998			0.4484	5.7072	0.3868	5.4743
p_{t-2}	-0.0557	-0.1613						
p_{t-3}	-0.4272	-1.2779	-0.2049	-4.0702	-0.3278	-4.4808	-0.2880	-4.1340
p_{t-4}	0.1470	0.7631						
x_t	-0.0140	-0.1297	0.1290	6.6776				
x_{t-1}	0.2946	2.2751			0.1222	6.3752	0.1008	6.6021
x_{t-2}	-0.1351	-1.0353						
x_{t-3}	-0.1585	-1.2075						
x_{t-4}	0.1693	1.5595						
r_t	-0.4164	-3.6849	-0.5812	-8.6121	-0.5219	-7.9668	-0.4217	-4.4029
r_{t-1}	-0.3253	-1.9202					-0.2880	-2.3268
r_{t-2}	-0.0726	-0.4207	-0.3380	-3.0200	-0.3400	-2.9697		
r_{t-3}	-0.0346	-0.2030						
r_{t-4}	-0.0282	-0.2363						
Constant	-0.3145	-0.7918						
RSS		0.0135		0.0153		0.0157		0.0159
sigma		0.0128		0.0127		0.0128		0.0129
R^2		0.9998		0.9998		0.9998		0.9998
Radj^2		0.8038		0.9312		0.9311		0.9311
LogLik		455.5085		449.0368		447.8882		447.2097
AIC		-8.5394		-8.6674		-8.6449		-8.6316
HQ		-8.3310		-8.5944		-8.5719		-8.5586
SC		-8.0247		-8.4872		-8.4647		-8.4514
Chow(1967:4)	0.4367	0.9958	0.5440	0.9814	0.5418	0.9820	0.5243	0.9864
Chow(1973:4)	0.5427	0.9064	0.4743	0.9470	0.4374	0.9628	0.4716	0.9483
normality test	6.1768	0.0456	6.1584	0.0460	5.8507	0.0536	6.0494	0.0486
AR 1-5 test	0.9351	0.4631	1.2622	0.2872	1.3347	0.2569	1.4750	0.2058
hetero test	0.8275	0.7223	1.3887	0.1781	1.4050	0.1703	1.4021	0.1717

5 Conclusions and new directions

The aim of the paper was to evaluate computerized model-selection strategies to see if they worked well, indifferently, or failed badly. The results come much closer to the first: the diagnostic-test operational characteristics are fine; selection-test probabilities match those relevant to the DGP; and deletion-test probabilities show 1% retention at a nominal 1% when no sub-sample testing is used.

We also found that, although estimates are 'biased' on average, conditional on retaining a variable, the parameter estimates were close to unbiased. This is essential for economic policy – if a variable is included, *PcGets* delivers the right response; otherwise, when it is excluded, one is simply unaware that such an effect exists.

On two empirical modelling problems, given the GUM that earlier investigators used, *PcGets* selects either closely similar, or somewhat improved specifications. Thus, we deem the implementation successful, and deduce that the underlying methodology is appropriate for model selection in econometrics.

Nevertheless, this is a first attempt: consequently, we believe it is feasible to circumvent the baseline nominal selection probabilities. First, since diagnostic tests must be insignificant at every stage to proceed, *PcGets* avoids spurious inclusion of a variable simply because wrong standard errors are computed (e.g., from residual autocorrelation). Thus, it could attain the same lower bound as in a pure white-noise setting, since every selection must remain both congruent and encompassing. Secondly, following multiple paths reduces the overall size, relative to (say) stepwise regression, despite the hugely increased number of selection (and diagnostic) tests conducted. Such an outcome highlights that an alternative statistical theory of model selection is needed than the conventional 'Bonferroni'-type of analysis, and Hendry and Krolzig (1999) present the basics thereof. Intuitively, the iterative loops around sequences of path searches could be viewed as 'sieves' of ever-decreasing meshes filtering out the relevant from the irrelevant variables: as an analogy, first large rocks are removed, then stones, pebbles, so finally only the gold dust remains. Thirdly, post-selection tests may further reduce the probability of including non-DGP variables below the nominal size of selection t-tests, at possible costs in the power of retaining relevant variables, and possibly the diagnostics becoming significant. Although the costs of missing 'real' influences rise, the power-size trade off in Hoover and Perez (1999) is quite flat around an 80-20 split.

So far, we have not discussed the role of structural breaks, particularly in regressors, both in-sample, and after selection. In general, breaks in regressors in-sample should not alter the selection probabilities: there is still an $\alpha\%$ chance of false inclusion, but different variables will be selected. However, breaks after selection, as the sample grows, should help to eliminate adventitious influences. *PcGets* tests for constancy as one diagnostic, and conducts sub-sample evaluation of reliability, but does not make 'decisions' based on the latter information. In a PRS, however, such accrual of information is essential.

Various other extensions of the algorithm developed in this paper are being exploried. One possible extension of *PcGets* concerns the pre-selection of variables: for example, one might fix economically-essential variables, then apply *PcGets* to the remaining orthogonalized variables. Forced search paths also merit inclusion (e.g., all even-numbered variables in a Monte Carlo; economically-relevant selected sets in empirical modelling). Suppose two investigators commenced with distinct subsets of the GUM, would they converge to the same reduction path if, after separate reduction exercises, they conducted encompassing tests, and recommenced from the union of their models? Such a search could be a 'forced' path in the algorithm, and may well be a good one to follow, but remains a topic for future research. Implementing cointegration reductions and other linear transforms could also help here.

Further work comparing *Gets* on the same data to simple-to-general approaches, and other strategies – such as just using information criteria to select – are merited. More detailed Monte Carlo studies

are required to investigate the impacts of breaks (interacting with the effects of sub-sample selection), collinearity, integration and cointegration. But the door is open – and we anticipate some fascinating developments will follow for model selection.

References

Abramowitz, M., and Stegun, N. C. (1970). *Handbook of Mathematical Functions*. New York: Dover Publications Inc.

Akaike, H. (1985). Prediction and entropy. In Atkinson, A. C., and Fienberg, S. E. (eds.), *A Celebration of Statistics*, pp. 1–24. New York: Springer-Verlag.

Bock, M. E., Yancey, T. A., and Judge, G. C. (1973). Statistical consequences of preliminary test estimators in regression. *Journal of the American Statistical Association*, **68**, 109–116.

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

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, **28**, 591–605.

Davidson, J. E. H., and Hendry, D. F. (1981). Interpreting econometric evidence: Consumers' expenditure in the UK. *European Economic Review*, **16**, 177–192. Reprinted in Hendry, D. F. (1993), *Econometrics: Alchemy or Science?* Oxford: Blackwell Publishers.

Davidson, J. E. H., Hendry, D. F., Srba, F., and Yeo, J. S. (1978). Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal*, **88**, 661–692. Reprinted in Hendry, D. F. (1993), *op. cit.*.

Doornik, J. A. (1998). *Object-Oriented Matrix Programming using Ox 2.0*. London: Timberlake Consultants Press.

Doornik, J. A., and Hansen, H. (1994). A practical test for univariate and multivariate normality. Discussion paper, Nuffield College.

Godfrey, L. G. (1978). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica*, **46**, 1303–1313.

Hannan, E. J., and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, **B, 41**, 190–195.

Harvey, A. C. (1981). *The Econometric Analysis of Time Series*. Deddington: Philip Allan.

Hendry, D. F. (1979). Predictive failure and econometric modelling in macro-economics: The transactions demand for money. In Ormerod, P. (ed.), *Economic Modelling*, pp. 217–242. London: Heinemann. Reprinted in Hendry, D. F. (1993), *op. cit.*.

Hendry, D. F. (1984). Monte Carlo experimentation in econometrics. In Griliches, Z., and Intriligator, M. D. (eds.), *Handbook of Econometrics*, Vol. 2–3, C.H. 16. Amsterdam: North-Holland.

Hendry, D. F. (1993). *Econometrics: Alchemy or Science?* Oxford: Blackwell Publishers.

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

Hendry, D. F., and Doornik, J. A. (1994). Modelling linear dynamic econometric systems. *Scottish Journal of Political Economy*, **41**, 1–33.

Hendry, D. F., and Doornik, J. A. (1996). *Empirical Econometric Modelling using PcGive 9 for Windows*. London: Timberlake Consultants Press.

Hendry, D. F., and Ericsson, N. R. (1991a). An econometric analysis of UK money demand in 'Monetary Trends in the United States and the United Kingdom by Milton Friedman and Anna J. Schwartz'. *American Economic Review*, **81**, 8–38.

Hendry, D. F., and Ericsson, N. R. (1991b). Modeling the demand for narrow money in the United Kingdom and the United States. *European Economic Review*, **35**, 833–886.

Hendry, D. F., and Krolzig, H.-M. (1999). General-to-specific model specification using PcGets for Ox. Discussion paper, Economics Department, Oxford University.

Hendry, D. F., and Krolzig, H.-M. (2000). Improving on *'data mining reconsidered'* by K.D. Hoover and S.J. Perez. *Econometrics Journal*, **2**, 41–58.

Hendry, D. F., and Richard, J.-F. (1989). Recent developments in the theory of encompassing. In Cornet, B., and Tulkens, H. (eds.), *Contributions to Operations Research and Economics. The XXth Anniversary of CORE*, pp. 393–440. Cambridge, MA: MIT Press.

Hoover, K. D., and Perez, S. J. (1999). Data mining reconsidered: Encompassing and the general-to-specific approach to specification search. *Econometrics Journal*, **2**, 1–25.

Jarque, C. M., and Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, **6**, 255–259.

Judge, G. G., and Bock, M. E. (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam: North Holland Publishing Company.

Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review*, **73**, 31–43. Reprinted in Granger, C. W. J. (ed.) (1990), *Modelling Economic Series*. Oxford: Clarendon Press.

Lovell, M. C. (1983). Data mining. *Review of Economics and Statistics*, **65**, 1–12.

Mayo, D. (1981). Testing statistical testing. In Pitt, J. C. (ed.), *Philosophy in Economics*, pp. 175–230: D. Reidel Publishing Co. Reprinted as pp. 45–73 in Caldwell B. J. (1993), *The Philosophy and Methodology of Economics*, Vol. 2, Aldershot: Edward Elgar.

Mizon, G. E., and Richard, J.-F. (1986). The encompassing principle and its application to non-nested hypothesis tests. *Econometrica*, **54**, 657–678.

Neyman, J., and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, **20A, 175–240**, 263–294.

Nicholls, D. F., and Pagan, A. R. (1983). Heteroscedasticity in models with lagged dependent variables. *Econometrica*, **51**, 1233–1242.

Pagan, A. R. (1987). Three econometric methodologies: A critical appraisal. *Journal of Economic Surveys*, **1**, 3–24. Reprinted in Granger, C. W. J. (ed.) (1990), *op. cit.*.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Sims, C. A., Stock, J. H., and Watson, M. W. (1990). Inference in linear time series models with some unit roots. *Econometrica*, **58**, 113–144.

White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**, 817–838.

White, H. (1990). A consistent model selection. In Granger, C. W. J. (ed.), *op. cit.*.

Wooldridge, J. M. (1999). Asymptotic properties of some specification tests in linear models with integrated processes. In Engle, R. F., and White, H. (eds.), *Cointegration, Causality and Forecasting*, pp. 366–384. Oxford: Oxford University Press.