**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

# 4 Autometrics

**Jurgen A. Doornik**

## 4.1 Introduction

David Hendry has developed and advocated the use of general-to-specific procedures for model selection over many years, see the collection of his earlier papers in 'Econometrics: Alchemy or Science?', Hendry (2000). As a result of this, the 'battery of diagnostic tests' has become a salient feature of econometric software. Nonetheless, many practitioners found the application of the 'LSE' or 'Hendry' methodology quite difficult. On top of this, 'data- mining' had become a pejorative term in econometrics, partially as a result of the experiments by Lovell (1983). [1]

[1] Despite the fact that Lovell only considered three simple search methods: forward selection (a form of stepwise regression), maximum $R^{-2}$, and max-min $t$, on a very small annual data set (at least from a current perspective).

A sea change has followed the work by Hoover and Perez (1999). They revisited the experiments of Lovell (1983), and, helped by a 1000-fold increase in computational power, implemented an automated version of the approach advocated by David Hendry. To their surpise, the artificial practitioner adopting the general-to-specific (GETS) procedure did very well.

As could be expected, David Hendry was greatly interested in these results, and, together with Hans-Martin Krolzig, immediately set about replicating the results. Hendry and Krolzig proposed some improvements to the algorithm of Hoover and Perez (Hendry and Krolzig, 1999; Hendry and Krolzig, 2005), and considered the theoretical aspects of model selection in several papers, see Hendry and Krolzig (2005) in particular. In addition, they created PcGets, a user-friendly computer program aimed at the empirical modeller, see Hendry and Krolzig (2001).

In this chapter, I introduce Autometrics, a third implementation of GETS model selection. Because Autometrics is based on the same principles, it

end p.88

has much in common with the approaches of Hoover—Perez and Hendry— Krolzig. In particular, the general properties should remain valid (Hendry and Krolzig, 2005). There are also some differences, as discussed below, which result in somewhat improved operational characteristics.

PcGets has many control parameters that need calibration, documented in Hendry and Krolzig (2003), resulting in a 'liberal' and 'conservative' strategy. [2]

[2] Although it was found that many of these were off-setting.

Autometrics does not have quite as many, but still needs extensive testing to set default choices for parameters, and to assess its practical operation. Reporting on these experiments is the second objective of this chapter.

## 4.2 Main Aspects of the Automated GETS Algorithms

There are five main ingredients in the algorithm proposed by Hoover and Perez (1999):

1. **General unrestricted model**. The GUM [3]

   [3] This terminology was introduced by Hendry—Krolzig.

   is the starting point for the model reduction procedure, and provides the initial information set. The GUM must be relevant, that is, provide sufficient information on the process that is modelled, and statistically well]behaved. The latter property is checked with a set of diagnostic tests. The objective is to create a 'congruent' initial model, see Hendry and Nielsen (2007, Ch. 20).
2. **Multiple path search**. Each insignificant variable in the GUM defines a reduction path. The first path is entered by deleting the most insignificant variable (that is, the one with the lowest absolute $t$-value). This path continues by deleting the most insignificant variable, each time re-estimating the model. The path terminates when all variables are significant. But note that other criteria will be used as well. With $k$ insignificant variables, there are $k$ paths. Hoover—Perez decided to follow 10 paths at most. [4]

   [4] Their motivation was to mimic the empirical modeller adopting GETS, rather than trying to improve on this using computer automation.

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

3. **Encompassing test**. A second criteria for a model reduction to fail is when the current model fails to encompass the GUM. The model is always nested in the GUM, and this test is implemented as a simple *F*-test on the removed variables. In that case, the variable is kept in the model (despite being insignificant), and the next variable in line is considered. I shall refer to this specific encompassing test as: 'backtesting with respect to the GUM'. The current reduction is not allowed to fail the backtesting criteria. Since such backtesting is at a user-defined level, this stage is intended to limit the loss of information (relative to the GUM) that is tolerated in the reduction.

end p.89

4. **Diagnostic testing**. In addition, every estimated model is subjected to a battery of diagnostic tests. When any test fails, the current reduction is rejected, and the next in line considered. Hoover—Perez use tests for normality, residual correlation, residual ARCH, as well as an in-sample Chow test. There is also an out-of-sample Chow test, which requires that some data is held back (10% here).
5. **Tiebreaker**. It is possible that every path reduces to the same terminal model. In this case there is a unique final model, at the cost of much redundant computation. In general, though, there can be multiple terminal models, all of which are valid reductions of the GUM. The user may have economic or aesthetic justifications to prefer one terminal over another. A fully automated procedure must decide on a final model; Hoover and Perez (1999) adopt the best fitting terminal model, while Hoover and Perez (2004) use the minimum Schwarz Criterion.

Hendry and Krolzig (1999) and Hendry and Krolzig (2001) extend the algorithm in three principal directions:

1. **Presearch**. Five types of presearches are added to reduce the computational effort as well as the empirical size: two on lags, one on variables (that is, the same variable at different lag lengths) and two more on groups of insignificant coefficients.
2. **Multiple-path search**. Hendry—Krolzig follow all paths, rather than the first 10. They also create additional search paths by working on blocks of regressors, selecting all whose individual insignificance exceeds a certain level. Hendry—Krolzig choose 6 levels for the 'liberal' (5%) strategy, adding 6 search paths, and 5 for the conservative (1%) strategy.
3. **Iteration**. Hoover—Perez perform one round of multiple-path searches. Hendry—Krolzig propose an iterative procedure as follows: form the union of the terminal candidate models after the first round, delete any model that is an invalid reduction of this union (an encompassing test), if necessary form the union again and run the encompassing tests again, until no further terminals are deleted. This union of surviving terminals makes a new 'GUM' [5]

   [5] Perhaps it would be better to call it a SUM: specific unrestricted model.

   which is the starting point for a completely new run of the whole algorithm. This process is iterated until convergence (the new GUM is the same as the previous GUM), after which the tiebreaker chooses a final model.

Hendry and Krolzig modify three additional aspects:

1. **Tiebreaker**. The minimum Schwarz Criterion as used as the default tiebreaker, with AIC and Hannan-Quinn as an option.
2. **Out-of-sample testing**. There is a trade-off between holding data back for out-of-sample testing, and using a larger sample for model estimation.

end p.90

Lynch and Vital-Ahuja (1998) study this in more detail, and conclude that the latter is preferable. Hendry and Krolzig (2004) revisit this within the context of model selection, reaching a similar conclusion: holding data back reduces the extent to which wrong variables are selected, but at the same time makes it harder to find those that matter. In the methods that we consider this trade-off can be controlled directly through the significance levels that are chosen. Then, split sample methods provide a less transparent mechanism, which tend to cost more than they gain (also see Hoover and Perez, 2004 footnote 12).

3. **Invalid GUM**. Hoover—Perez remove replications for which the GUM fails two or more diagnostic tests from their simulation experiments; if only one test fails, it is removed from the criterion set. Hendry—Krolzig retain all such cases, adjusting the level of diagnostic testing instead. If the GUM fails one or more test at p-value $p_d$ (with a default of $p_d = 0.01$), then each failed test is made less strict by raising its $p_d$ to a point where the test passes. The new level is then maintained throughout.

This approach allows model selection to work with an invalid GUM. But it should be noted that it invalidates a fundamental assumption of the method.

The Autometrics algorithm follows on from the Hendry—Krolzig algorithm, aiming to improve on the implementations by

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

Hoover—Perez and Hendry— Krolzig (as envisaged by Hoover and Perez, 2004, p. 790):

1. **Presearch**. The presearch is an ad hoc addition to the algorithm, which has been getting more complex over time. This is particularly true for variable- based presearch. The automated GETS algorithms are so involved because they are designed to handle complex correlations between (sets of) variables. The candidate sets for the presearches are constructed in simplistic ways, and once a variable is removed, it cannot reappear. This works well when the DGP consists of only a few highly significant variables, because the power is hardly affected and the size improves. Or when the DGP consists of orthogonal variables, when it is much easier to construct the relevant sets. However, in general we don't know the state of nature. This argument applies more weakly to lag-length reduction, because most modellers tend to have a preference for shorter over longer lags.
   The objective of Autometrics is to create an algorithm that can function without (but may still be improved somewhat by a presearch).
2. **Search paths**. The multiple-path search is an unstructured way of searching the model space (that is, all possible sets of the variables in the GUM): many paths may turn out to be the same, while other paths are left unsearched. Hoover —Perez motivate this approach by wanting to de-emphasize the mechanical nature of the search. However, this is one aspect at which computers are much better than humans. Autometrics considers the

end p.91

whole search space from the outset, but may discard parts in a systematic way. This tree-search method is considered in section 4.3.1.

3. **Scope**. The algorithm can have wider use than regression models only, therefore Autometrics is implemented entirely within the likelihood framework (of which regression is a special case). [6]

   [6] This is for future research: different model types have their own relevant tests, and may have different categories of parameters. For example, in an ARFIMA model with regressors the ARFIMA parameters may need to be treated separately.

4. **Efficiency**. The aim of Autometrics is to improve computational efficiency, for example by avoiding repeated estimation of the same model, delayed diagnostic testing, and remembering terminals between iterations.

The next section considers these aspects in more detail. Readers who are not interested in these details, or in the calibration experiments (section 4.4, 4.5), can jump straight to the comparison with Hoover—Perez and PcGets (section 4.6) or a practical illustration (section 4.7).

## 4.3 The Autometrics Algorithm

### 4.3.1 Tree Search

The starting point for Autometrics is the whole space of models generated by the variables in the initial model. At every node in this tree is a unique model which can be estimated. Then, the subnodes on the next level can be ordered according to increasing significance of the variables in the model.

Figure 4.1 provides an example. With four variables ABCD in the GUM, there are 4! possible unique models (the ordering of variables within a model is irrelevant), represented by solid dots in the figure. The GUM is at the root, the next model is obtained by removing A, then removing AB, etc. [7]

[7] Figure 4.1 is a tree that lies on its side. We move through this representation of the tree from left to right and top to bottom starting at the root (skipping through the nodes in parentheses): ABCD, BCD, CD, D, (CD), C, (BCD), BD, B, (BCD), BC, (ABCD), ACD, ….

However, models are invariant to the ordering of the variables they consist of. The redundant models are represented by an open circle; the first open circle in the rightmost column would be D. Only the unique $2^4$ models are labelled in Figure 4.1.

The root node in Figure 4.1 is the GUM, model ABCD. With A the most insignificant variable in the GUM, the next model would be without A, which is model BCD. This is followed by CD, when B is the most insignificant in BCD. If, on the other hand, D would be most insignificant in model BCD, then D would be the first candidate for removal. In that case we would need to change the labelling in the graph of the tree. Within the regression context, the significance of each variable in a model is based on the individual *t*-value.

end p.92

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

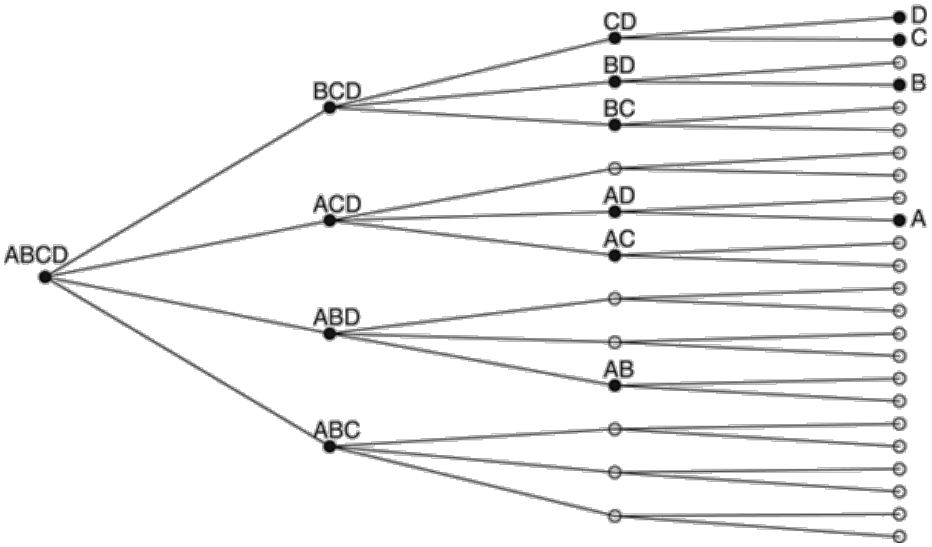


**Fig. 4.2. Multiple-path search: Representation for a GUM with variables ABCD.**

Once the complete branch that started with deleting A has finished, we can continue with the next major branch that starts from the GUM by deleting variable B. Note that variable A is always part of the models in this branch: ACD, AD, A, AC.

The resulting tree of Figure 4.1 is a unique representation of the model space, and, if we move through it from left to right and top to bottom all possible models will be estimated. [8]

[8] Of course, other orderings of the tree are feasible. When all regressors are orthogonal, this ordering is most likely to yield the final model first.

The Hoover—Perez and Hendry—Krolzig multiple-path search is depicted in Figure 4.2, assuming that the ordering always stays the same. The first path in the Autometrics search corresponds to the first of the paths considered by Hoover—Perez and Hendry—Krolzig, but after that the methods diverge.

### 4.3.2 Efficient Tree Search

For $k$ insignificant variables, there are $2^k$ models, so visiting each node is not feasible in practice. Autometrics implements several strategies to skip nodes and move efficiently through the tree. Figure 4.3 shows the order in which the nodes are visited. [9]

[9] Strictly speaking, the tree contains the empty model after node 4, which is omitted from the graph.

Remember that, at each node, the subnodes are reordered with the most insignificant variable first. This is not visible in the graph, unless we allow the letters to refer to different variables (so, as depicted here, in the model CD, C is the most insignificant variable).

end p.93

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

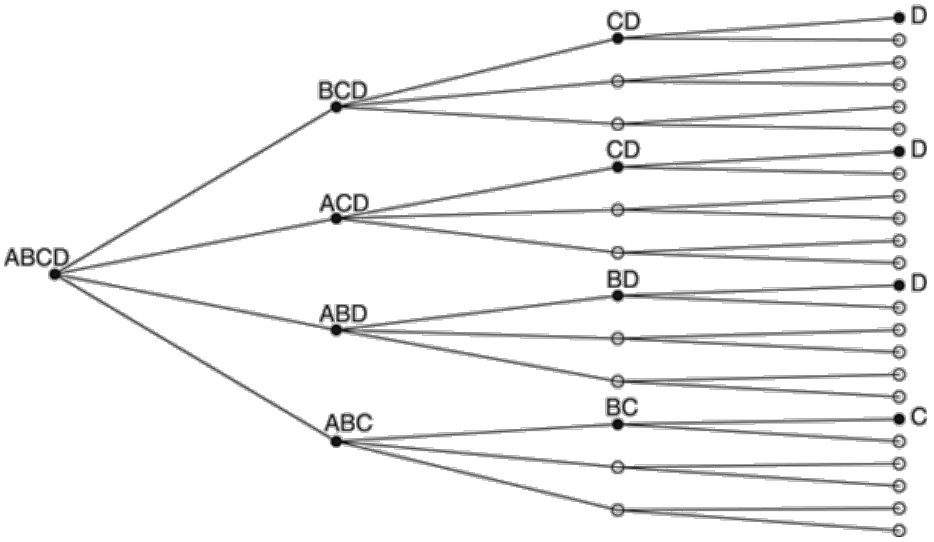


**Fig. 4.3. Search tree: All unique models starting from a GUM with variables ABCD, numbered according to the order in which the search algorithm considers them.**

Also note that the relevant information is the reduction path from the GUM: for model 12, AC, this is 1 → 9 → 12. The search algorithm can be designed in such a way that we always have the path back to the GUM in memory, $k$ models at most, rather than the full tree.

Autometrics uses the following principles to advance the tree search:

**Pruning**. By default, every reduction involves removing one variable. The first and most obvious principle is that, if a deletion fails or the reduced model fails (on backtesting or diagnostic testing), the node is invalid. Then all subsequent subbranches of the tree can be pruned (ignored).

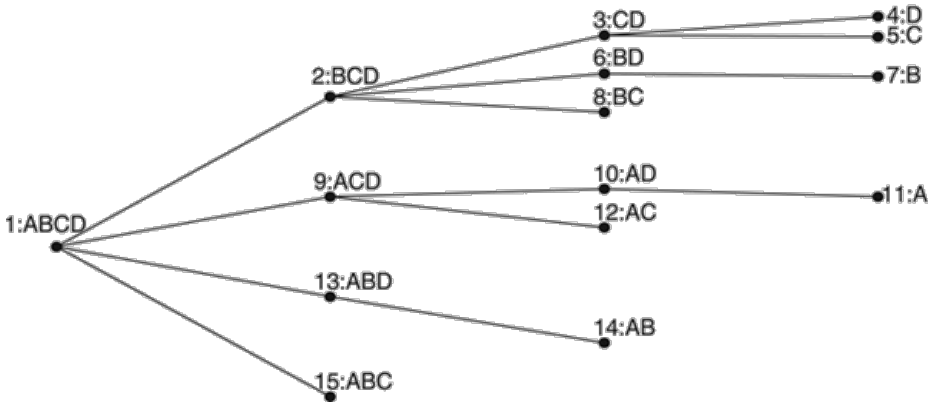


**Fig. 4.4. Chop & bunch factor: size (top three), power (middle three), and time (bottom three; in seconds for $M$ = 1, 000 replications) for chopping and bunching at $\alpha$pa: individual results and averages for each $p_a$ (squares for $p_a$ = 0.001, triangles for $p_a$ = 0.01, plusses for $p_a$ = 0.05, circles for $p_a$ = 0.1).**

end p.94

Starting from model BCD of Figure 4.3: if the variable B cannot be removed, there is no need to consider models CD, D, or C. As another example, if model CD fails on the backtest with respect to the GUM, subnodes of CD need not be considered, and we can fall through to the next node at the same level, BD.

This pruning is governed by the main Autometrics p-value $p_a$, which determines the significance level below which a variable cannot be deleted from the model. Ideally, $p_a$ defines the empirical behaviour of the procedure: the fraction of irrelevant variables retained (the 'size') is close to $p_a$.

**Bunching**. Instead of always removing a single variable at a time, it can be useful to consider groups. In other words, a

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

bunch of branches. Bunching works as follows: from the current node, along the generated search path, variables are grouped together provided their individual insignificance merits this. Deletion is then tried as a block. If successful, we can jump straight ahead, thus replacing several separate steps with a single blocked step. If deletion fails, the algorithm backtracks until a bunch is found that can be deleted—if necessary, to a bunch of size one.

As an illustration, assume that in the second node of Figure 4.3, model BCD, the variables BC are both insignificant enough to be bunching candidates. Then we start with an *F*-test on BC jointly. If successful, this gives us the next node D directly. Otherwise we backtrack to deleting B only (ie model CD).

The significance test of the bunch is done at $p_a$, while the p-value $p_b$ determines the amount of bunching, as described in section 4.4.3. If $p_b$ is too high, there will be excessive backtracking, which is costly. Also, terms that may matter can then hide between insignificant ones. Setting it too low effectively switches bunching off, at some computational cost. By default $p_b = \max\left\{\frac{1}{2}p_a^{1/2}, p_a^{3/4}\right\}$, see section 4.4.2.

**Chopping**. Chopping refers to the permanent removal of a 'highly' insignificant variable from the branches of the model. When a bunch is insignificant enough, the whole bunch could get the chop. Chopping saves on computation, but could possibly mean that some combinations of variables are missed.

Continuing the example for node BCD. If B is insignificant enough to be considered for chopping, then, after visiting CD, D, and C, we don't consider any nested model with B in it. In other words, after C we fall through to node ACD. If, on the other hand, BC can be chopped from BCD, then we only visit model D before falling through to node ACD.

The p-value $p_c$ determines whether chopping is applied. By default $p_c = p_b$, see section 4.4.2.

end p.95

**Model contrasts**. A terminal candidate model ('terminal') is a model that cannot be reduced any further on the adopted criteria. When a terminal has been found, there is no need to find the same model again. Because the tree is uniquely ordered, we can determine the minimal bunch along the current trajectory which must be deleted to give a different model. We can use this to our advantage to move more efficiently through the tree.

Suppose, for example, that model D was found as a terminal candidate, and the search has taken us to model 9:ACD. Whether A is significant or not, it is always kept in the models that originate from ACD; C and D are the 'free' variables. Model ACD nests D, and it is of no interest when D is already a terminal. Instead, at least CD has to be removed to find a different terminal, so this bunch can be tested immediately through an *F*-test at $p_a$.

Although the principles underlying pruning, chopping, and bunching are quite simple, implementing them in software involves some fairly complex administrative code.

The combination of bunching and chopping can be seen as a form of embedded presearch. However, here it is an integral part of the structure of the search procedure.

There is one fundamental aspect of the tree search which is very different from the multiple-path search. In the latter, all terminal candidate models have no insignificant variables left, unless further reduction failed because of diagnostic testing or backtesting with respect to the GUM. In the tree search, branches maintain insignificant variables by design. For example, any branch starting from model 13:ABD has AB in, which may or may not be significant. Therefore, we denote such a terminal candidate with insignificant variables as a 'semiterminal'. Only further refinement will turn a semiterminal into a proper terminal, which cannot be reduced any further. This is one of the issues considered in the next section.

### 4.3.3 Further Details of the Autometrics Algorithm

Several aspects are not yet specified, or handled differently in Autometrics:

**Backtracking on diagnostic failure**. Unlike the algorithms of Hoover— Perez and Hendry—Krolzig, diagnostic checking is not performed when the reduction is in progress. Instead, the diagnostic tests are only evaluated when a terminal is reached. Upon failure, we backtrack until a valid model is found. This is feasible because the path back to the GUM is available. There are two benefits to this approach. First, diagnostic testing is relatively costly (roughly five times more than model estimation). Second,

end p.96

diagnostic failure could be a temporary blockage on a reduction path: it is possible (and indeed happens in practice) that only some intermediate models fail.

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

**Diagnostic failure of the GUM**. The execution of diagnostic testing follows Hendry—Krolzig: if at the GUM one or more tests fail at $p_d$ (with a default of $p_d = 0.01$), the significance level of those tests is made less strict. Autometrics differs in that it tries to restore diagnostic validity along the way. When there are multiple terminal candidates and some pass at the original levels, then only the successful terminals are kept, and the p- values reset to the original level.

**Root branches**. The GUM is the root model, and the root branches are the models with just one variable deleted (BCD, ACD, ABD, ABC in Figures 4.1 and 4.3; these are also the starting point for the multiple-path searches of Hoover—Perez and Hendry—Krolzig). By default Autometrics starts applying the bunching and chopping only from a root branch, not directly from the root. This means that each variable is used as a 'pivot' for entering the tree search. As a consequence, each variable will have an opportunity to enter the set of final models.

**Model contrasts**. There are two ways in which model contrasts are used in Autometrics. *Union contrast* determines the contrasting bunch with respect to the union of the current set of terminal candidates. Union contrast is used while the current GUM (the new union of the set of terminals) still changes between iterations: there is no need yet to find all possible models—our primary interest is to find variables that should be in the GUM. *Terminal contrast* determines the smallest bunch that would yield a model that is different from any of the current terminals. This mode is used at the end when the current GUM is fixed.

**Branches with nested terminals**. A branch has a nested terminal if any restriction on that branch (possibly invalid) could yield that terminal candidate. Autometrics search is split in two parts. Paths with nested terminals are skipped in the first round. The second round then follows all root paths with nested terminals using union contrast (or terminal contrast at the end) to jump ahead in the branches.

**Insignificant variables**. All but one of the root branches start with keeping one or more insignificant variable in the model. Some of these may still be insignificant in the terminal candidate models that were found, that is they are semiterminals, as discussed in section 4.3.2. To turn these into terminal candidates, the tree search reduction procedure is applied to it (but without using root branches). Another potential source of semiterminals is the backtracking from diagnostic failure. Such semiterminals are refined in the same way, until they are proper terminals.

end p.97

Any irreduceable terminal candidate model can still have insignificant variables from failure in backtesting against the GUM or diagnostic failure.

**Presearch**. Only a presearch on lag length is switched on by default. The method used is the so-called extended lag reduction, see section 4.4.5.

### 4.3.4 Iterative Search

The following steps describe the iterative procedure as used in Autometrics. The steps that involve the tree search are labelled in bold.

0.0 Estimate the initial GUM.
This initializes the search procedure.

0.1 (optional) Add dummy variables for outliers.
If any are detected, the expanded model becomes the initial GUM.

0.2 (optional) Lag-length pre-search.

0.3 Test all regressors at a loose significance level.
If passed, accept the empty model as the final model, provided diagnostic testing is satisfied, then stop.

1.0 Set $i = 0$.
The starting point for the current iteration is GUM 0 (this may be the same as the initial GUM), which has $k$ free regressors.

1.1 (convergence) If all regressors in the GUM are significant then stop. This is at a slightly more stringent p-value to allow for 'squeezing', §4.5.2.

1.2 Update the diagnostic p-values.
Ideally, the user ensures that the initial GUM passes the diagnostic tests. However, when this is not the case, the p-value for each failed test statistic is increased. Subsequently, the p-values are adjusted downwards again if possible.

1.3 **Run reduction over the root branches.**
Terminal candidate models ('terminals') are collected as the search progresses. Any subtree that has a previously found terminal nested in it is skipped to speed up the search. This will result in one or more terminal.

1.4 **Run reduction to search for nested terminals**.

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

Revisit the subtrees that were skipped before. At each point it is possible to compute the minimal contrast with a known terminal to jump ahead to a possible new (non-nested) terminal. If the union of terminals after the previous step 1.3 is smaller than the union from the previous iteration, then use union contrast, otherwise use terminal contrast.

end p.98

1.5 Remove terminals that fail diagnostics.
If the p-values $p_d$ for diagnostic testing had to be adjusted downwards, and there are some terminals that pass at the original p-value, then keep only those terminal models which pass and reset $p_d$ to the original value.

1.6 Form the union of the terminal models.
The union is called the *current GUM* or *GUM* $i + 1$.

1.7 Remove terminals that fail backtesting.
When using the default Autometrics settings, this step is skipped, because backtesting with respect to GUM 0 has already been done as an integral part of the tree search: there are no terminals that fail. Optionally, the PcGets default of backtesting with respect to the current GUM can be adopted instead. In that case, there may be terminals that fail the encompassing test against the new GUM.

1.8 Remove terminals with insignificant variables.
When using the default Autometrics settings, this step is skipped, because a terminal remains a terminal candidate for subsequent iterations.
However, this step is relevant when the PcGets default is used in 1.7: backtesting is with respect to the current GUM, which changes between iterations. So a terminal candidate with insignificant variables may not be a terminal next time.

1.9 Increment $i$ and continue to step 1.1.
GUM $i$ (determined at 1.6, but now $i$ has been incremented), is the new base for the search. Note that steps 1.7 and 1.8 (when they are not skipped) remove terminals but do not modify the GUM. This is different from the PcGets implementation.

In addition to the tree search, notable differences with PcGets are: reduced presearch, an attempt to restore original diagnostic p-values during the iteration, use of GUM 0 for backtesting, a different way (more gentle) of constructing the new GUM for the next iteration, and survival of terminals between iterations.

## 4.4 Calibration

In addition to the choice of p-value, $p_a$, at which the Autometrics reduction is executed, there are some design parameters of the algorithm which can be changed. We run calibration experiments to find a balanced default setting for these, as well as design strategies that focus on different goals.

end p.99

**Table 4.1. Design aspects that are determined through calibration**

| | |
|---|---|
| *Bunch and chop factor* | |
| 1 | bunching and chopping is done at $p_a$, the deletion p-value; |
| $\alpha$ | bunching and chopping is done at $\alpha p_a$, $\alpha$ = 2, 4, 6, 9. |
| *Tree search method* | |
| 1 | prune only—delete one variable at a time; |
| 2 | prune & chop—delete one variable at a time and allow for chopping the deleted variable; |
| 3 | prune & bunch—delete one or more variable at a time, the criterion for addition becomes more stringent as the bunch grows; |
| 4 | prune & bunch & chop—as above, allowing for chopping of the bunch; |
| 5 | prune & bunch & chop aggressively—as above, but keeping the p-value for extending the bunch fixed. |
| *Backtesting* | |
| 0 | none; |
| 1 | initial GUM backtest; |
| 2 | GUM0 backtest (same as initial GUM when there is no presearch); |
| 3 | current GUM backtest. |
| *Presearch* | |
| 0 | none; |
| 1 | lag reduction; |
| 2 | lag reduction followed by variable reduction. |

In the evaluation, we define the 'size' of the experiments as the proportion of irrelevant variables that survives the reduction process. If we start with $n_1$ irrelevant regressors, and are left with $\hat{n}_1$, the size is $\hat{n}_1/n_1$. [10]

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

[10] Hoover and Perez (1999) define size as the proportion of *significant* irrelevant variables: $\hat{n}_1^* / n_1$, which is smaller or equal to $\hat{n}_1/n_1$.

Ideally, this is roughly equal to $p_a$. 'Power' [11]

[11] Power and size are used somewhat differently here than in statistical testing. More recently we have started to experiment with the terms *potency* and gauge to avoid confusion.

reflects the success of the reduction: the proportion of relevant variables that is retained. This can be written as $\hat{n}_0/n_0$, where $n_0$ is the number of regressors in the DGP, and $\hat{n}_0$ the (subset) of those that are in the final model. In our experiments the GUM always nests the DGP.

In most cases there is a trade-off between size and power: a gain in power is achieved at a cost in size (ie both increase). Then it can be difficult to make a choice. Only rarely is it possible to decrease size and increase power simultaneously.

Table 4.1 summarizes the parameters that are to be determined, while Table 4.2 indicates which aspects are kept fixed. Throughout, the calibration experiments are run at several reduction p-values $p_a$.

### 4.4.1 Experimental Design

The experimental design closely resembles some of the experiments reported in Hendry and Krolzig (2005). The first three, labelled *static GUM* in Table 4.3,

end p.100

**Table 4.2. Design aspects that are kept fixed during calibration**

| | |
|---|---|
| Outlier detection | switched off, |
| Pretest on all variables jointly | at* $p_p = f\left(p_a^{0.8}, 5\right) \approx 5p_a^{0.8}$; |
| Squeezing | at $p_s = f(p_a, 0.2) \approx p_a/5$, |
| Diagnostic testing | at $p_d = 0.01$, |
| Tiebreaker | Schwarz criterion (SC), |
| Iteration | until new GUM does not reduce any further, |
| Maximum terminals | 20, |
| Diagnostic test set | Normality, AR(2), ARCH(2), Chow(50%), Hetero, see Hendry and Doornik (2007). |

$$f(p, \boldsymbol{\alpha}) = \boldsymbol{\alpha}\, p \text{if} 0.94 \leq 1 + p(\boldsymbol{\alpha} - 1) \leq 1.06;\ f(p, \boldsymbol{\alpha}) = \frac{\alpha p}{1 + p(\boldsymbol{\alpha} - 1)}$$

otherwise.
correspond to Hendry and Krolzig's $S_2$, $S_3$, and $S_4$: the DGP has 8 regres- sors (all $t$-values are 2 in experiment 1, $t = 3$ in experiment 2, and $t$=4 in experiment 3), while the GUM contains the DGP variables and an additional 22 irrelevant regressors; all regressors are generated as i.i.d standard normals.

The next three experiments (*dynamic GUM* in Table 4.3) correspond to those labelled JEDC in Hendry and Krolzig (2005), but with a larger GUM. The DGP has 5 variables at lag 0, with $t$-values 8, 6, 4, 3, 2 respectively; the GUM also includes the (irrelevant) first lag of these 5, plus lags zero and one of 7 irrelevant variables, as well as $y_{t-1}$. In this case, the regressors are created with autocorrelation, as a way to introduce correlation between regressors.

**Table 4.3. Experimental design**
DGP

$$y_t = \sum_{i=1}^{k_0} \beta_j x_{i,t} u_t, \quad u_t \sim \text{IN}[0, 1],$$

$$x_{i,t} = \rho x_{i,t-1} + v_{i,t},\ x_{i,0} = 0,\ i = 1,\ \ldots,\ k_0 + k_1 \quad v_{i,t} \sim \text{IN}\left[0, (1-\rho)^2\right].$$

GUM

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

$$y_t = \gamma_0 + \delta_0 y_{t-1} + \sum_{i=1}^{k_0+k_1} \left(\gamma_j x_{i,t} + \delta_j x_{i,t-1}\right) + \epsilon_t, \epsilon_t \sim \mathrm{IN}\left[0, \sigma_\epsilon^2\right].$$

Experiments* 1, 2, 3 (*static GUM*)
*DGP* : $k_0 = 8,\ t_{\beta_1} = \ldots = t_{\beta_8} = \{2, 3, 4\},\ \rho = 0,$
*GUM*: $\delta_i = 0$, $k_1 = 26$ : 26 irrelevant regressors.
Experiments* 4, 5, 6 (*dynamic GUM*)
*DGP* : $k_0 = 5,\ t_{\beta_1} = \ldots = t_{\beta_5} = \{8, 6, 4, 3, 2\},\ \rho = \{0, 0.4, 0.8\},$
*GUM* : $k_1 = 7$ : 20 irrelevant regressors (12 at lag 1).
Experiments* 7, 8 (*small and large dynamic GUM*)
*DGP* : $k_0 = 5,\ t_{\beta_1} = \ldots = t_{\beta_5} = \{2, 3, 4, 6, 8\},\ \rho = 0,$
*GUM*: $k_1 = \{1, 19\}$ : 8 or 44 irrelevant regressors respectively.
* All regressions include a constant.

end p.101

The final two experiments are a smaller and larger version of experiments 4–6, but with i.i.d. regressors.

$M$ = 1, 000 Monte Carlo replications are used throughout for sample sizes $T$= 60, 100, 250. The reduction p-values $p_a$ are chosen as 0.001, 0.01, 0.05, 0.1, each requiring a separate run of the experiment. So the 8 experiments are run 12 times (3 sample size at 4 nominal p-values). The graphs show the result separated by sample size: 32 outcomes along each vertical line (8 experiments at 4 p-values). At each sample size, the experiments use exactly the same artificial data. The outcomes at $p_a$ = 0.1 are shown as a circle, and their averages as a solid line; those at $p_a$ = 0.05 are shown as a plus symbol, and their average as a dashed line; etc.

All simulations were implemented in Ox version 5, see Doornik (2007). Where timings are reported, it is relevant to know that the programs were run in single-threaded mode (−rp1 command line switch) on a PC with a Core 2 Duo Q6600 running at 2.4 Ghz.

### 4.4.2 Bunch and Chop Factor

First we consider how the bunch and chop factor impacts on the algorithm. For this we keep the remainder of Table 4.1 fixed: the tree search method is 4 (prune & bunch & chop, these were introduced in section 4.3.2; additional detail is given in the next section), and backtesting is against GUM 0. Presearch is switched off until section 4.4.5. The p-value at which we bunch and chop is $ap_a$ , with $a$ displayed on the horizontal axis in Figure 4.4. So, for example, when $p_a$ = 0.05, the bunching and chopping is at p-values 0.05, 0.1, ..., 0.45.

Figure 4.4 shows that the choice of $a$ has almost no impact on power (the middle three graphs, for sample sizes $T$ = 60, 100, 250) for $p_a$ = 0.1, but somewhat more as $p_a$ falls. The size plots (top three) follow a similar pattern. There is a clearer pattern in terms of the time it takes to run a reduction: for $p_a$ = 0.1, 0.05 it slopes upwards, even at $T$ = 250. But for the smallest p-value there is almost no impact.

The default value of $a$ is chosen from these experiments as close to 2 for larger p-values, and much larger as the p-values get very small:

$$p_b = p_c = \max\left\{\frac{1}{2}p_a^{1/2},\ p_a^{3/4}\right\}. \tag{4.1}$$

The following table shows which values of $a$ are implied by (4.1):

| $p_a$ | 0.25 | 0.1 | 0.05 | 0.01 | 0.001 | 0.0001 |
|---|---|---|---|---|---|---|
| $p_b$ (= $p_c$ ) | 0.354 | 0.178 | 0.112 | 0.050 | 0.016 | 0.005 |
| implicit $a$(= $p_b$ /$p_a$ ) | 1.414 | 1.778 | 2.236 | 5.000 | 15.811 | 50.000 |

end p.102

### 4.4.3 Tree Search Method

Although the bunching and chopping factor was chosen to balance time against the power/size performance, its overall impact is small. We now turn to the question of whether these operations matter at all.

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

We consider the 5 settings for the tree search method listed in Table 4.1. The first (*prune* = 1) is pruning only, which coincides with setting pb = $p_c = 1$. In this case, a regressor is considered for deletion if its p-value is above $p_a$.

The remainder uses the default determined in (4.1). The second (*prune* = 2) is chopping only: one regressor at a time is considered for deletion (at $p_a$) and removed permanently from the search procedure when its significance is above $p_c$. The next case (*prune* = 3) is bunching only: regressors are grouped for deletion, if the whole group fails jointly, it is shrunk until the remaining group succeeds. The fourth case extends this by permanent removal (chopping) if the group's significance is above $p_c$. The aggressive case (*prune* = 5) differs from cases 3 and 4 in how the inital candidate group is determined.

The bunching in cases 3 and 4 works as follows. If a regressor's p-value is above $p_a$, it is considered for deletion. If it also is above $p_b \equiv p_b^*$ (1), bunching is started, adding variables as long as the *smallest p-value in the bunch* is above $p_b^*(k_b)$:

$$p_b^*(k_b) = p_b^{1/2}\left[1 - \left(1 - p_b^{1/2}\right)^{k_b}\right] ,$$

with *kb* the size of the bunch. The following table gives some numerical examples:

| $p_a$ | 0.25 | 0.1 | 0.05 | 0.01 | 0.001 |
|---|---|---|---|---|---|
| $p_b = p_b^*(1) = p_c$ | 0.353 | 0.178 | 0.112 | 0.050 | 0.016 |
| $p_b^*(2)$ | 0.497 | 0.281 | 0.186 | 0.089 | 0.030 |
| $p_b^*(3)$ | 0.555 | 0.340 | 0.236 | 0.119 | 0.042 |
| $p_b^*(5)$ | 0.588 | 0.394 | 0.291 | 0.161 | 0.062 |
| $p_b^*(10)$ | 0.595 | 0.420 | 0.329 | 0.206 | 0.093 |

Of course, for the actual bunch to be deleted, its joint p-value must be above $p_a$ (in addition to satisfying backtesting and diagnostic testing conditions).

In the aggressive case variables are added to the bunch as long as their individual significance is above $p_b$. The bunches are therefore allowed to be larger.

Figure 4.5 shows that pruning values 1–4 are almost identical in terms of size and power. There is a considerable difference in speed: *prune* = 4 takes about 2/3 of the time of single step deletion (*prune* = 1). Only at $T = 60$ is this

end p.103

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

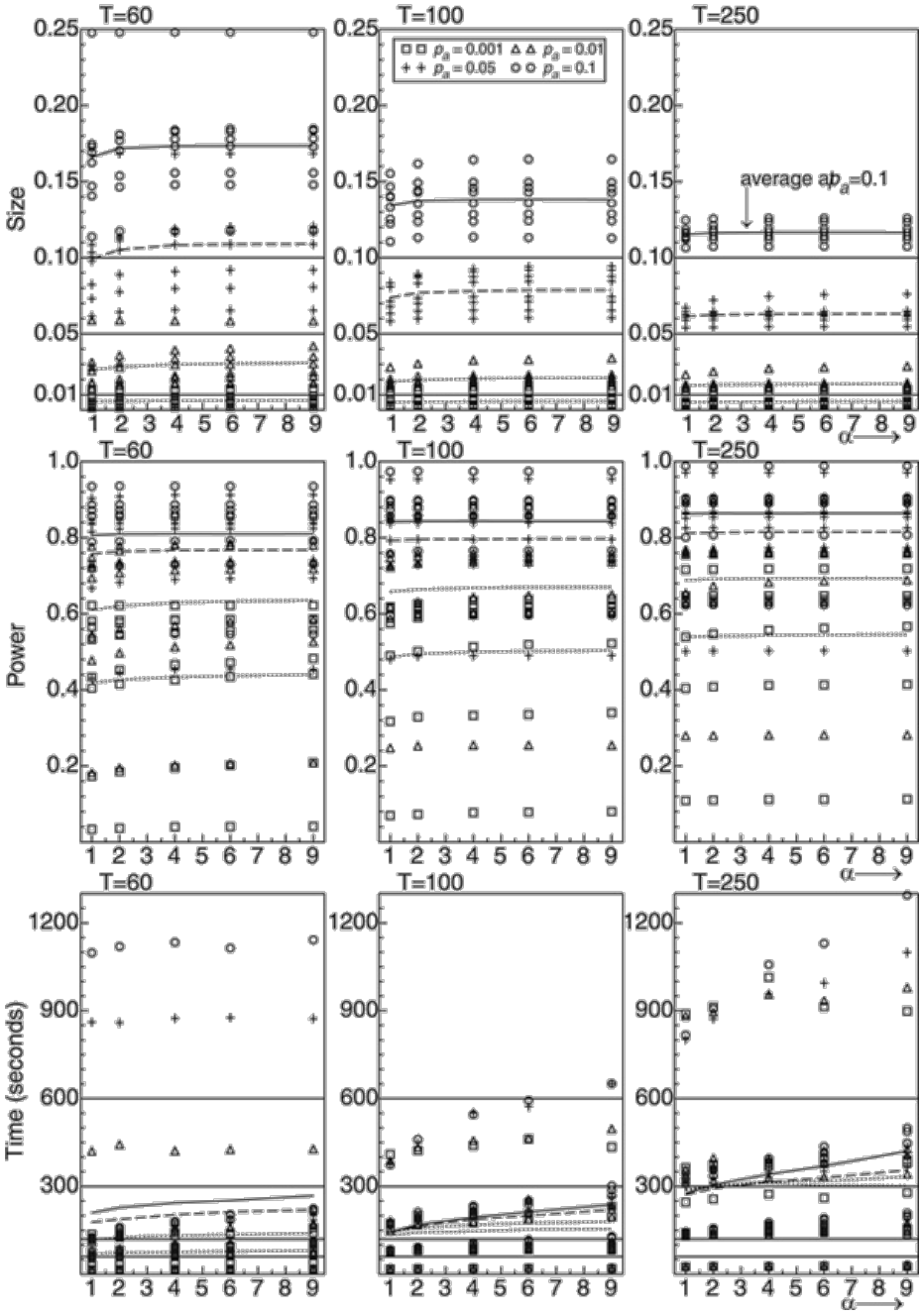


**Fig. 4.5. Tree search method: size (top three), power (middle three), and time (bottom three) for five pruning methods: individual results and averages for each $p_a$ (squares for $p_a$ = 0.001, triangles for $p_a$ = 0.01, plusses for $p_a$ = 0.05, circles for $p_a$ = 0.1). Method 1: prune only, 2: prune & chop, 3: prune & bunch, 4: prune & bunch & chop, 5: aggressive.**

end p.104

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

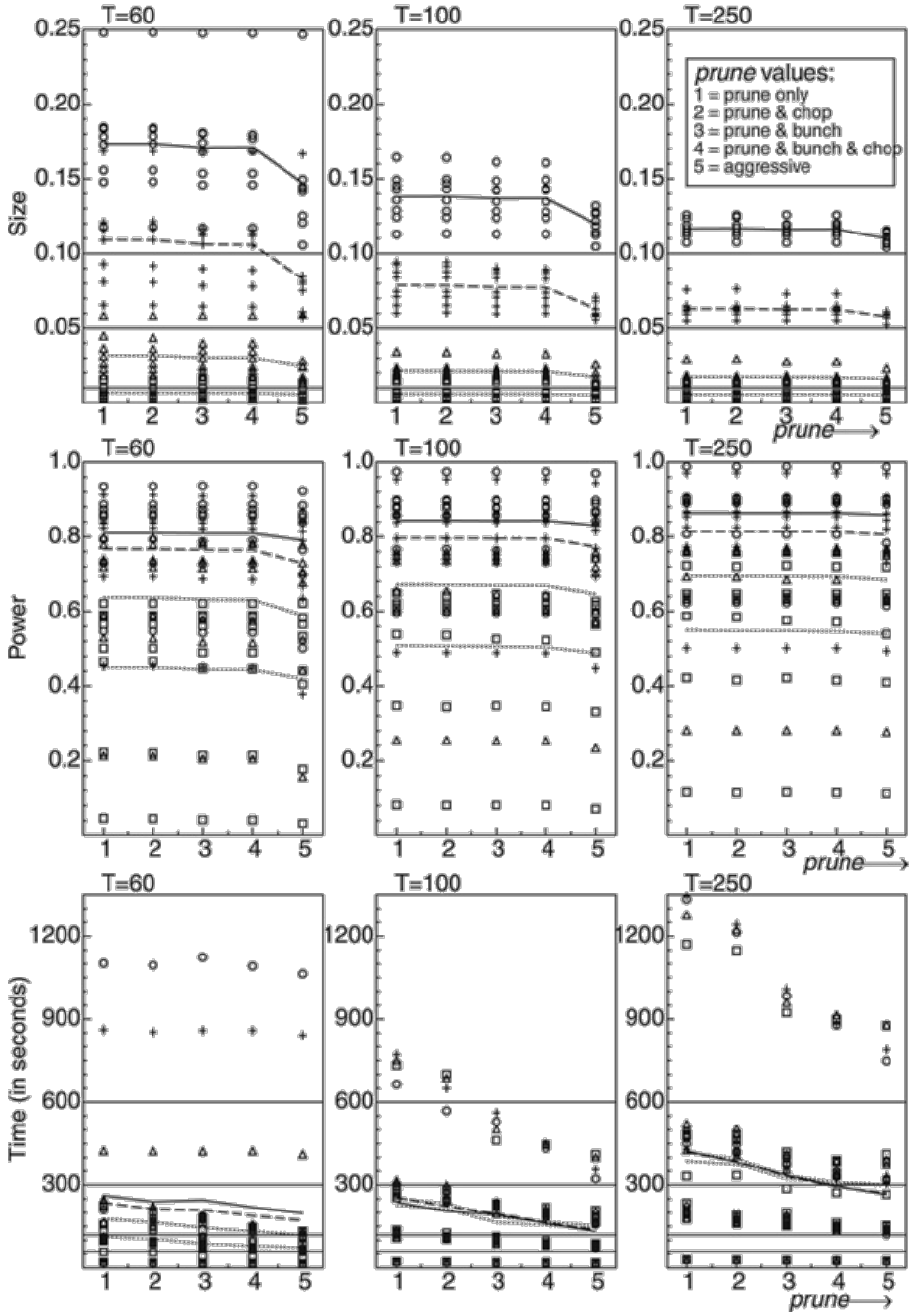


**Fig. 4.6. Backtesting: size (top three) and power (bottom three) for three backtesting methods: individual results and averages for each $p_a$ (squares for $p_a = 0.001$, triangles for $p_a = 0.01$, plusses for $p_a = 0.05$, circles for $p_a = 0.1$). Backtest mode 0: none, 1: w.r.t. initial GUM, 2: w.r.t. GUM0, 3: w.r.t. current GUM.**

end p.105

less pronounced, mainly because the largest experiment is not much affected. Among these, we have a clear preference for *prune* = 4.

Aggressive pruning provides an additional speed-up, but at lower size and power. This will not be chosen as default.

### 4.4.4 Backtesting

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

The three backtesting modes are considered with the bunching p-value set according to (4.1), and *prune* = 4, the defaults determined so far. Figure 4.6 shows that no backtesting at all (value 0 on the *x*-axis) is disastrous: the size

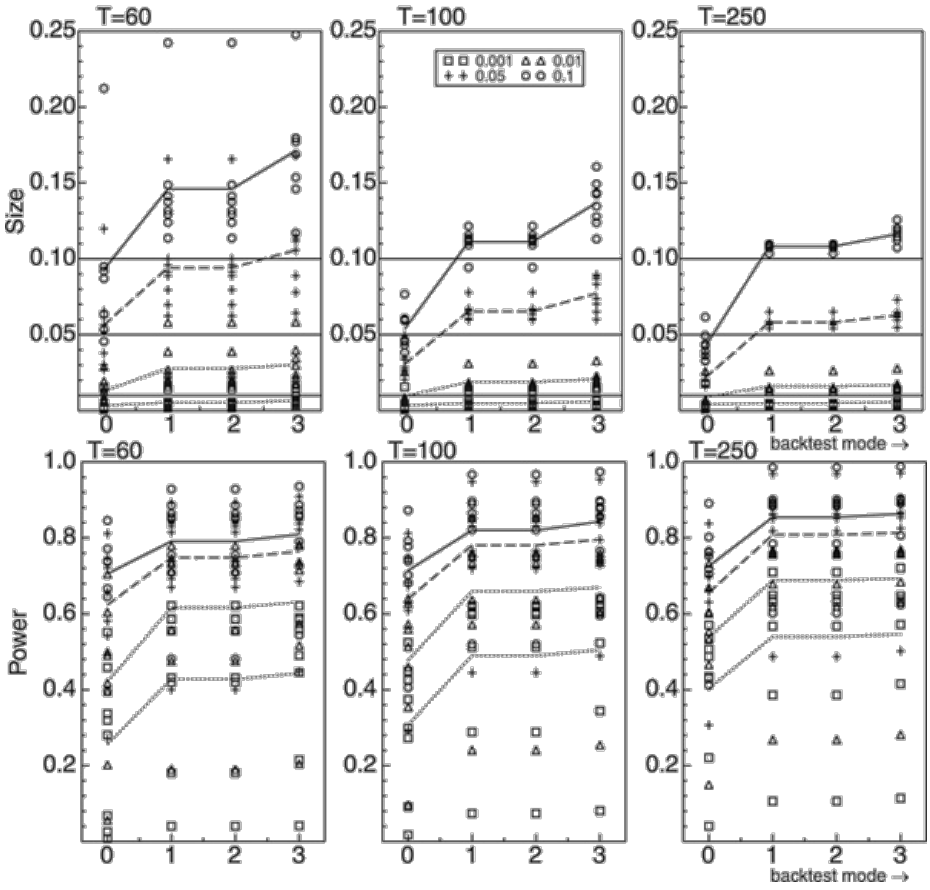


**Fig. 4.7. Presearch: size (top three), power (middle three), and time (bottom three) for three presearch settings: individual results and averages for each $p_a$ (squares for $p_a = 0.001$, triangles for $p_a = 0.01$, plusses for $p_a = 0.05$, circles for $p_a = 0.1$). Presearch method 0: none, 1: lags (two passes), 2: lags (three passes), 3: lags and variables.**

end p.106

collapses as the sample size grows, while the power is much lower too. This applies *a fortiori* to backwards selection, also see Doornik (2008).

The other backtesting modes are: (1) backtesting with respect to the original GUM (value 1 on the *x*-axis) and (3) backtesting with respect to the most recent GUM. Surprisingly, this makes a difference in terms of size, but perhaps less so in terms of power, at least at the smaller sample sizes. Backtesting w.r.t. the original GUM also results in smaller dispersion of size between experiments. The timings of the experiments are unaffected by the choice of backtesting.

At this stage, backtesting modes 2 and 3 are identical. However, presearch adds another possible mode of backtesting.

### 4.4.5 Presearch

Presearch adds a new GUM before the tree search algorithm commences. The initial GUM is the GUM before any presearch is performed. GUM 0 is the GUM after the presearch.

Lag reduction groups the regressors by lag. Any variables that have no lags at all are excluded in the Autometrics implementation, while seasonals are not considered to be lagged. So, for example, constant, trend, and seasonals are not part of the regressors that are tested at lag zero. Also, lag presearch is skipped altogether if there are no lags in the model.

Three methods of presearch lag reduction and one method of variable reduction are considered:

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

- *Closed lag reduction tests* lags from the largest lag downwards, stopping as soon as a lag cannot be deleted. The following steps are taken:

  Collect all regressors at lag $m$, then if:

  1. no individual p-value is below max{ $p^*_{p,1}$ (kp), ps}, with $p_s$ defined in Table 4.2,

     (4.2) $$p^*_{p,1}(k_p) = 1 - (1 - p_{p,1})^{k_p},$$

     and where $k_p$ is the number of regressors involved (so the marginal p-value decreases as $k_p$ grows), and
  2. their joint p-value in the most recent model is above $p_p$, and
  3. backtesting with respect to initial GUM and diagnostic testing conditions are satisfied,

  then delete all these regressors, set this to the current model, and continue with lag $m - 1$.
- *Common lag reduction* tests all remaining lags starting from the least significant.

end p.107

  Determine the joint significance of each lag in the starting model (ie the model after common lag reduction), and order the lags with the most insignificant first. Then if:

  1. their joint p-value in the starting model is above $p_p$, and
  2. no individual p-value is below max{ $p^*_{p,1}$ ($k_p$), $p_s$ }, and
  3. their joint p-value in the most recent model is above $p_p$, and
  4. backtesting with respect to initial GUM and diagnostic testing conditions are satisfied,

  then delete all these regressors, set this to the current model, and continue with the next lag.
- *Common X-lag reduction* is the same as the previous method, but now the lagged dependent variables are excluded from the lag reduction.
- *Variable reduction* is similar to common lag reduction, but focusing on variables rather than lags:
  Determine the joint significance of variable in the starting model (ie the model after lag presearch), and order the variables with the most insignificant first. Then if:
  1. the joint p-value in the starting model is above $p_p$,
  2. no individual p-value is below max{ $p^*_{p,1}$ ($k_p$), $p_s$ }, and
  3. backtesting with respect to initial GUM and diagnostic testing conditions are satisfied,

  then delete all the regressors belonging to this variable.

Four levels of presearches are considered in the simulations:

0 no research,

1 lag reduction (closed followed by common lag reduction),
  a 'Model' 1 is determined by two lag presearches, first closed, then common.
  b 'Model' 2 is determined by running the two lag presearches in opposite order.
  c Remove any failed model.
  d Form the union of the remaining models (usually both survive the previous step), which is now GUM 0.

2 extended lag reduction (closed followed by two passes of common lag reduction),
  a 'Model' 1 is determined by three lag presearches, first closed, then common, then common on *X*'s only.

end p.108

  b 'Model' 2 is determined by running the three lag presearches in opposite order.
  c Remove any failed model.
  d Form the union of the two models, which is now GUM 0.

3 lag reduction (closed, common) followed by variable reduction.
  a 'Model' 1 is determined by two lag presearches, first closed, then common, followed by variable reduction.
  b 'Model' 2 is determined by running variable presearch, common, and closed lag presearches.
  c Remove any failed model.
  d Form the union of the two models, which is now GUM 0.

The presearch cut-off p-value are set to $p_{p,1} = p_c$ and $p_p = f(p_a^{0.8}, 5) \approx 5p_a^{0.8}$:

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

| $p_a$ | 0.25 | 0.1 | 0.05 | 0.01 | 0.001 |
|---|---|---|---|---|---|
| $p_p$ | 0.711 | 0.485 | 0.334 | 0.114 | 0.020 |
| $p^*_{p,1}(1) = p_c$ | 0.353 | 0.178 | 0.112 | 0.050 | 0.016 |

These were chosen to be conservative with presearch removal, and no other values were tried (but could lead to better or worse results).

Figure 4.7 gives the results for the four presearch settings. Note that 3 out of 8 experiments are static, and are unaffected by the presearch lag reduction. Lag reduction is beneficial here: the size is reduced at no cost in power, while time is also substantially reduced.

Subsequent variable reduction is less clear cut, because, while power improves, the size also deteriorates. These effects are offsetting. For example, comparing presearch methods 2 and 3 at $T = 250$, the average size at $p_a = 0.001$ is halfway between the average sizes for $p_a = 0.001$ and 0.01 using method 2. But so is the power: the same power gain can be achieved using method 2 by raising $p_a$ . Note also that the size dispersion increases. At $T = 250$ and $p_a = 0.01$ the size for lag presearch ranges from around 0.010 to 0.017 for methods 1 and 2, but increases to 0.018–0.057 when variable presearch is used.

To study the effect of lag presearch more closely, we repeat experiments 4–8 (see Table 4.3), but now adding all variables up to lag 2 in the GUM. Denoting the new experiments with an asterisk, then experiments 4*, 5*, 6* have 33 irrelevant variables, and 7* has 15. Finally, 8* has 64 irrelevant variables, so we omit $T = 60$. The results for these extended experiments are in Figure 4.8. The advantages of lag presearch are accentuated. Subsequent variable reduction is again ineffective, although the increase in size dispersion is now only visible at $T = 100$.

end p.109

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

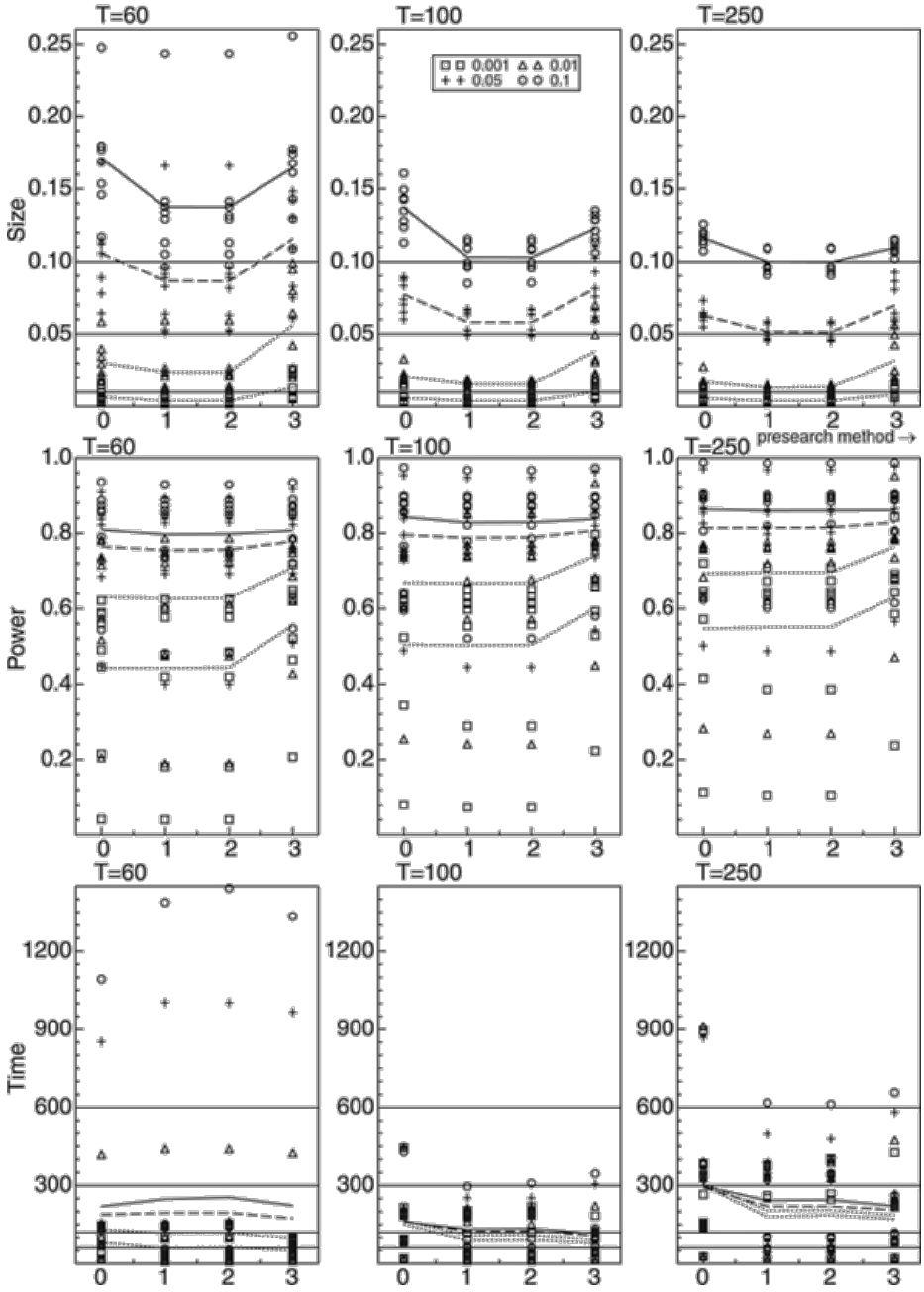


**Fig. 4.8. Presearch*: size (left), power (right) for three presearch settings in extended experiments: individual results and averages for each $p_a$ (squares for $p_a = 0.001$, triangles for $p_a = 0.01$, plusses for $p_a = 0.05$, circles for $p_a = 0.1$). Presearch method 0: none, 1: lags (two passes), 2: lags (three passes), 3: lags and variables.**

end p.110

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

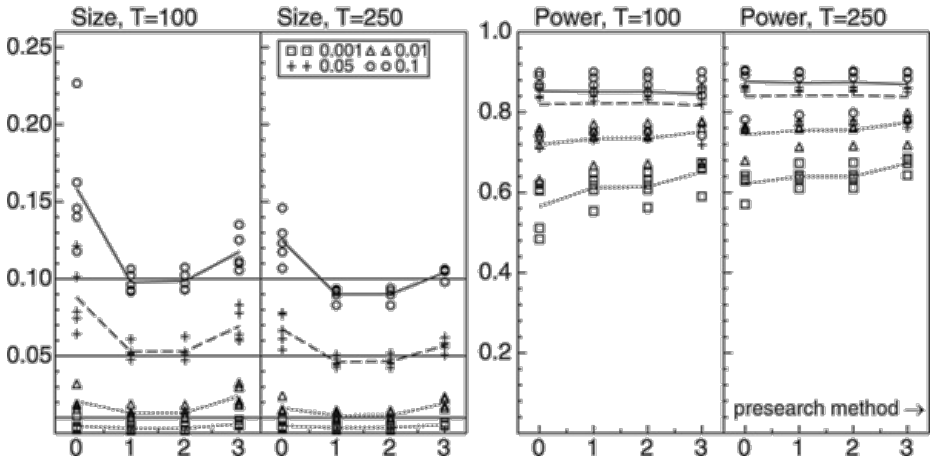


**Fig. 4.9. Backtesting after presearch*: size (left), power (right) for two backtest settings after presearch lag reduction: individual results and averages for each $p_a$ (squares for $p_a = 0.001$, triangles for $p_a = 0.01$, plusses for $p_a = 0.05$, circles for $p_a = 0.1$). Backtest mode 1: with respect to initial GUM, 2: with respect to GUM0, 3: with respect to current GUM.**

The conclusion is that presearch lag reduction is beneficial and should be switched on by default (Autometrics adopts method 2, which uses three passes of lag reduction). Variable reduction, on the other hand, is ambiguous at best (but other methods of presearch variable reduction may behave better).

### 4.4.6 Backtesting after Presearch

When presearch is used, three GUMs could be used for backtesting:

1. initial GUM, prior to presearch,
2. GUM 0, after presearch but before further reduction,
3. current GUM, the union of final models after a reduction pass.

The additional mode 3 is investigated in the larger experiments 4* to 8*, which have all variables up to lag 2 in the GUM. Figure 4.9 shows that it is almost identical to backtesting with respect to GUM 0. Otherwise, mode 1 leads again to lower and more concentrated size for roughly the same power.

The conclusions related to backtesting remain unaffected by the presearch: the lower, less dispersed size, with small impact on power, leads us to prefer backtesting with respect to GUM 0.

### 4.4.7 Defaults after Calibration

Table 4.4 summarizes the default settings for the Autometrics algorithm that were inferred from the simulation results in this section.

end p.111

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

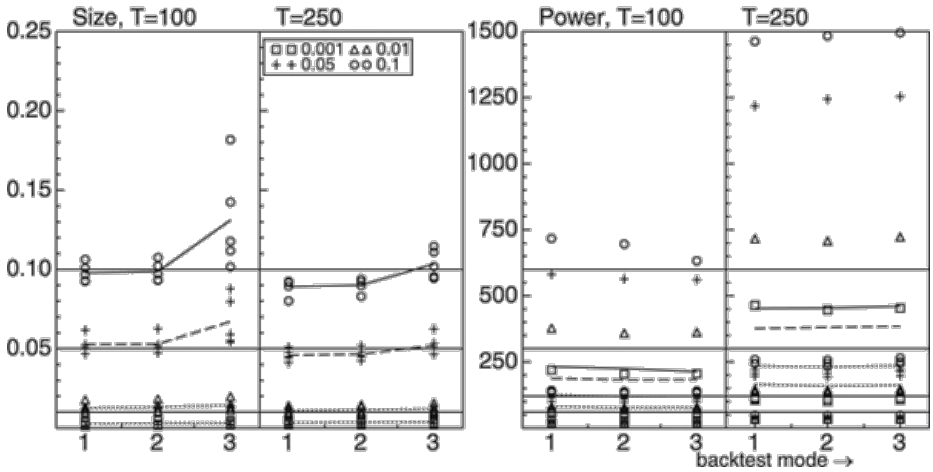


**Fig. 4.10. Tiebreaker: size (top three), power (bottom three) for the first GUM, the final GUM, and 5 tiebreaker settings: individual results and averages for each $p_a$ (squares for $p_a = 0.001$, triangles for $p_a = 0.01$, plusses for $p_a = 0.05$, circles for $p_a = 0.1$). Tiebreaker modes listed in Table 4.5.**

## 4.5 Impact of Other Design Aspects

### 4.5.1 Termination: Iteration and Tiebreaker

Autometrics finishes iterating when the new GUM (ie the union of the final models) is the same as the previous GUM. It is common at that stage to find multiple final candidate models. These models are all valid reductions of the initial GUM, and any could be chosen by the modeller. In an automated setting, as well as Monte Carlo experiments, a choice must be made by the program. Following PcGets, the default is the Schwarz criterion, but other choices can be made. This ranges from the best fitting model, via AIC, HQ, SC to the smallest model; we label these 2, ..., 6, where the ordering is in increasing penalty of model size, as in Table 4.5. There may well be multiple smallest models among the final candidates, but ties among the other criteria are exceedingly unlikely.

**Table 4.4. Design aspects that are determined through calibration**

| | | |
|---|---|---|
| *Bunch and chop factor* | $\alpha$ | $\alpha = \max\left\{\frac{1}{2}p_a^{1/2}, p_a^{3/4}\right\}/pa$ |
| | | see(1). |
| *Tree search method* | 4 | prune & bunch & chop; |
| *Backtesting* | 2 | GUM0 backtest; |
| *Presearch* | 2 | extended lag reduction. |

end p.112

**Table 4.5. GUM and tiebreaker modes used in Figure 4.10**
0 GUM 1 (ie GUM after one iteration of the algorithm)
1 Final GUM
2 Best fitting final model (highest likelihood)
3 Final model with smallest AIC (Akaike Information Criterion)
4 Final model with smallest HQ (Hannan–Quinn Information Criterion)
5 Final model with smallest SC (Schwarz Criterion)
6 Smallest final model (best fitting smallest if there is more than one)

Figure 4.10 plots size and power for the standard set of experiments of the previous section. It also includes the first GUM of the reduction (ie the union of models after one round), labelled 0, and the final GUM, labelled 1. The initial GUM has a size close to one, and a power of one. Figure 4.10 shows the progress made from that starting point via the first GUM to the final GUM. Because backtesting is with respect to GUM 0, there is no difference between GUM 1 and the final GUM: terminal candidates remain unchanged after GUM 1 because the backtesting criteria are unchanged (but additional terminal candidates may still be found).

The final GUM is always oversized, remaining so at $T = 250$ (but not asymptotically). The tiebreakers are ordered in decreasing model size, with the best fit being the least stringent. Choosing the best fitting terminal model has roughly a similar impact on size and power. But then, moving from best fit to SC, the size keeps going down with little impact on power,

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

motivating our choice of SC as the default tiebreaker (although, when forecasting is the main objective, another choice may be preferred). Choosing the smallest model impacts power as well as size. These are largely small sample issues, as the plots suggest.

### 4.5.2 Squeezing

Squeezing allows for a single reduction to be tried at a smaller p-value than $p_a$. The subsequent branch is accepted only if, jointly with one or more subsequent variables, the reduction is valid at $p_a$. So squeezing will not result in an invalid reduction, but allows for the case when two or more variables are highly correlated so that they are jointly insignificant, even though the individual terms appear significant. Whether squeezing is on or off has no impact on any of the simulations in this chapter.

### 4.5.3 Diagnostic Testing

Diagnostic testing has no effect on size and power in these sets of experiments. We need to find a different set of experiments to show the impact of diagnostic testing. However, in terms of speed, the effect is dramatic: delayed diagnostic

end p.113

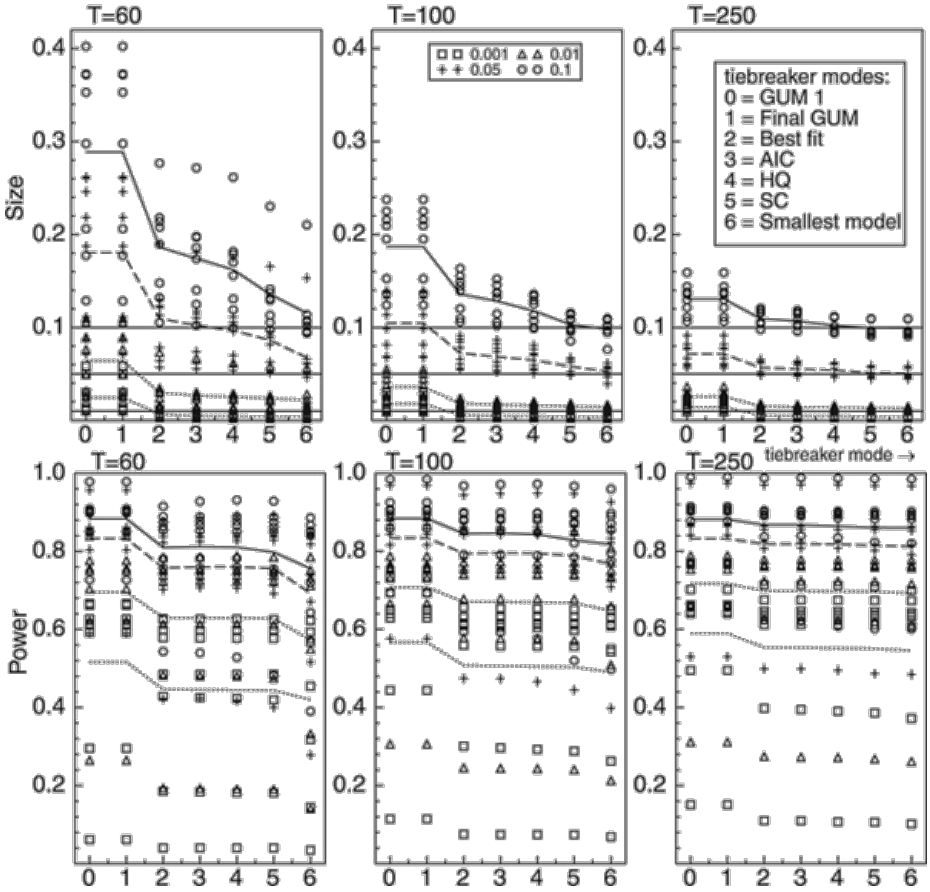


**Fig. 4.11. Comparison of PcGets (dotted lines) and Autometrics (solid lines). Horizontal axis: ρ ranging from 0 to 0.9; percentages on vertical axis.**

testing is 2.5 times faster than always diagnostic testing (in experiment 2 of Table 4.3, where the t-value is 3, using $T = 100$ and $T = 250$).

## 4.6 Comparison to Hoover–Perez and PcGets

Hendry and Krolzig (2003) discuss the shortcomings of some of the experiments used by Hoover and Perez (1999) (denoted HP1 to HP9, *op.cit.*, Table 3). In particular, the variables that matter are either highly significant (*t*-values of 5, 8, 10, 12), or

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

insignificant ($t$-value less than 1). The former are perhaps too easy, while the latter are almost impossible to detect on statistical grounds. Indeed, in some experiments stepwise regression fares just as well as PcGets

end p.114

or Autometrics. The exception is experiment HP8, where stepwise regression is dramatically worse. As pointed out by Hendry and Krolzig (2003), in most of these experiments it is beneficial to set the size very tightly. This will not affect power much, but greatly help with finding the DGP.

Here we focus on HP2, HP7, HP8, and HP9. Hendry and Krolzig (2003) choose HP2 and HP7 because all relevant variables are significant. [12]

[12] Hendry and Krolzig (2003) should also have included HP8. There is a mistake in Table 3 of Hoover and Perez (1999): the coefficient on $x3$ of model 8' should be 0.0345 (computed as 0.75 × 0.046). Model 7' also has a mistake: 6.73 should be 6.44. The code of Hoover and Perez (1999) shows that they used the version with autoregressive errors in the simulations, which are therefore unaffected by these typos.

We add HP9 because of the size problems reported by Hoover and Perez (1999).

Note again that Hoover and Perez (1999) define the size of the Monte Carlo experiments in relation to *significant* variables only: the rate at which falsely significant variables are retained in the final model. I follow Hendry and Krolzig (1999) in defining the size for *all* falsely included variables, which is always higher than the Hoover and Perez (1999) definition.

Table 4.6 compares the Hoover and Perez (1999) and PcGets results with Autometrics. The former are taken from their published tables, so use their definition of size. All other results report size over all irelevant variables; this slightly favours Hoover and Perez (1999). Also note that in the Hoover and Perez (1999) results the constant is always included in the model, but excluded from the evaluation. In the PcGets and Autometrics experiments the constant is kept free, and treated as the other variables. Again the difference is small. Because Hoover and Perez (1999) do not implement a presearch, the PcGets and Autometrics results in Table 4.6 do not either. The PcGets results are for an older version of the software, but, because almost all later changes are in the presearch, these results are still deemed relevant. Autometrics is run without presearch, and with pruning set to 1 (so no bunching and chopping—the latter could be seen as a form of embedded presearch).

Table 4.6 shows that the tree search of Autometrics improves on the multiple-path search algorithms: it is able to deliver consistent performance across experiments without presearch. Although somewhat oversized, the empirical size is closer to the nominal size. At first sight, Autometrics does not do as well in finding the DGP in HP8, but this is entirely a result of the larger empirical size, as discussed above. [13]

[13] And as Autometrics quick reduction and default presearch confirms, see Table 4.7.

Table 4.7 shows the results with presearch switched on, using the default settings for PcGets and Autometrics. It includes the Autometrics results using 'quick mode' (aggressive bunching and reduced effort). The presearch in PcGets has a much more dramatic effect than with Autometrics. The lower score of Autometrics in terms of the percentage of experiments in which the

end p.115

**Table 4.6. Comparision of three model selection algorithms without presearch**

| | HP2 | | | HP7 | | | HP8 | | HP9 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | HP | PcGets* | Auto* | HP | PcGets* | Auto* | HP | Auto* | HP | PcGets* | Auto* |
| | | | | | | | | | | 1% nominal size | |
| Size (%) | 5.7 | 2.4 | 2.0 | 3.0 | 2.4 | 2.1 | 0.9 | 2.1 | 3.2 | 2.5 | 2.1 |
| Power (%) | 100.0 | 100.0 | 100.0 | 94.0 | 99.9 | 99.8 | 99.9 | 100.0 | 57.3 | 61.9 | 60.9 |
| DGP (%) | 0.8 | 60.2 | 62.0 | 24.6 | 59.0 | 61.7 | 78.0 | 63.8 | 0.8 | 0.0 | 0.1 |
| 5% nominal size | | | | | | | | | | | |
| Size (%) | 10.7 | 10.7 | 6.3 | 8.2 | 10.2 | 6.6 | 3.7 | 6.5 | 8.5 | 10.4 | 6.5 |
| Power (%) | 100.0 | 100.0 | 100.0 | 96.7 | 99.9 | 99.9 | 100.0 | 100.0 | 60.4 | 66.2 | 63.0 |
| DGP (%) | 0.0 | 8.4 | 11.9 | 4.0 | 4.0 | 11.1 | 31.6 | 13.0 | 1.2 | 0.0 | 0.5 |
| 10% nominal size | | | | | | | | | | | |
| Size (%) | 16.2 | — | 10.0 | 14.2 | — | 10.1 | 10.6 | 10.0 | 14.1 | — | 10.1 |
| Power (%) | 100.0 | — | 100.0 | 96.9 | — | 99.9 | 100.0 | 100.0 | 62.5 | — | 64.6 |
| DGP (%) | 0.0 | — | 2.3 | 0.2 | — | 1.8 | 7.6 | 1.8 | 0.4 | — | 0.0 |

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

Source HP: Hoover and Perez (1999) Tables 4, 6, 7.

Source PcGets*: Hendry and Krolzig (1999) Table 2, no presearch.

Auto*: Autometrics using default settings, but with presearch, bunching and chopping switched off.

DGP is % of experiments in which the DGP is exactly found.

Size HP: % falsely significant variables; others: % falsely included variables.

Power: % of DGP variables included.

DGP is found is commensurate with the higher empirical size. The exception is HP2 at 5%, where PcGets has higher size and a higher score on DGP found.

Next, we consider a set of experiments with a better range of t-values, namely those reported under section 5.1 of Hendry and Krolzig (2005). These are smaller versions of experiments 4, 5, and 6 in Table 4.3. Now the constant is

**Table 4.7. Comparision of two model selection algorithms with presearch**

| | HP2 | | | HP7 | | | HP8 | | HP9 |
|---|---|---|---|---|---|---|---|---|---|
| | PcGets | Auto | Quick | PcGets | Auto | Quick | Auto | Quick | Auto |
| 1% nominal size | | | | | | | | | |
| Size (%) | 0.9 | 1.3 | 0.9 | 1.0 | 1.7 | 1.0 | 1.6 | 0.8 | 1.7 |
| Power (%) | 100.0 | 100.0 | 100.0 | 99.8 | 99.3 | 99.2 | 100.0 | 100.0 | 60.4 |
| DGP (%) | 81.0 | 75.0 | 81.1 | 80.8 | 67.3 | 80.8 | 69.8 | 82.5 | 0.1 |
| 5% nominal size | | | | | | | | | |
| Size (%) | 5.5 | 5.2 | 3.9 | 5.4 | 6.0 | 4.2 | 5.9 | 3.9 | 6.1 |
| Power (%) | 100.0 | 100.0 | 100.0 | 99.8 | 99.8 | 99.6 | 100.0 | 100.0 | 62.6 |
| DGP (%) | 34.5 | 27.2 | 39.8 | 34.7 | 15.6 | 39.5 | 18.6 | 41.3 | 0.5 |

Source PcGets: Hendry and Krolzig (2003) Table 6.

Auto: Autometrics using default settings.

Quick: Autometrics using quick reduction which is more aggressive.

Additional notes, see Table 4.6.

end p.116

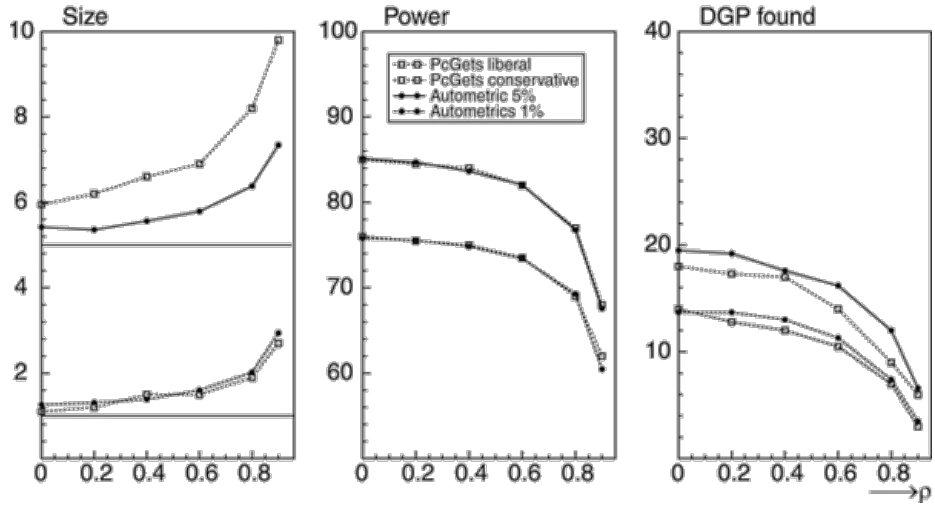


**Fig. 5.1. Mean historical and DCC correlations.**

free and $k_1 = 5$, so there are 22 regressors of which 17 are irrelevant. Figure 4.11 gives the results (the PcGets outcomes have been taken from Figure 10 of Hendry and Krolzig, 2005). Remember that the liberal strategy of PcGets is aimed at 5% and conservative at 1%. The figure shows that the size behaviour of Autometrics is similar at 1% and 5%, the size is less sensitive to $\rho$, and closer to the nominal size at 5%.

Castle and Hendry (2007) run a most interesting experiment to study co- linearity caused by nonlinear transformations of the data (such as using age and age-squared in a regression). They find that, with a nonzero mean, the colinearity can be substantial. The DGP is as follows (now using $N$ to denote the sample size in this cross-section example):

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

$$y_i = \beta x_i + u_i,\ \beta = N^{-1/2}5,\ u_i \sim IN(0,1),$$
$$x_i = \mu + v_i, \quad v_i \sim IN(0,1),$$

with $\beta$ chosen to give $x_i$ a $t$-value of 5 in the DGP. The GUM contains three variables, two of which are nuisance:

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \epsilon_i,\ \epsilon_i \sim IN\left(0, \sigma_\epsilon^2\right).$$

Two values of $\mu$ are used (0 and 10) at two sample sizes (100 and 1, 000). Table 4.8 compares PcGets and Autometrics for this experiment, showing that Autometrics does dramatically better. Some closer investigation with PcGets shows that it seems to be that its presearch gets in the way.

end p.117

**Table 4.8. Comparision of two model selection algorithms in a cross-section case, *M* = 10 000 at a 1% nominal size (conservative strategy).**

| | $\mu = 0, N = 100$ | | $\mu = 10, N = 100$ | | $\mu = 0, N = 1000$ | | $\mu = 10, N = 1000$ | |
|---|---|---|---|---|---|---|---|---|
| | PcGets | Auto | PcGets | Auto | PcGets | Auto | PcGets | Auto |
| Size (%) | 2 | 2 | 43 | 5 | 2 | 2 | 43 | 2 |
| Power (%) | 98 | 99 | 58 | 97 | 98 | 99 | 58 | 98 |

Source PcGets: Castle and Hendry (2007) Figure 1.
Auto: Autometrics using default settings.

## 4.7 BHS revisited (again)

The empirical application of Autometrics is illustrated using the model for real US money demand of Baba *et al.* (1992). Hendry and Krolzig (1999) tabulate the GUM, which has 38 coefficients: real money in logs, $m-p$, up to lag 6, GNP in logs, $y$, up to lag 6, 3 interest rates up to lag 1 (20-year treasury bills, $R_{20}$, 1-month bills, $R_1$, and M2 instruments, $R_{ma}$ ), 2 current account rates up to lag 2 ($R_{na}$ , $R_{sna}$ ), a volatility measure up to lag 1 ($V$), and an interaction between volatility and the interest spread up to lag 2 ($SV$). Finally, there is a dummy which is 1 in 1980(2) and $-1$ in 1980(3) and an intercept.

Following Hendry and Krolzig (1999), the selection is run at $p_a = 0.01$. The Autometrics presearch on lags removes only three variables: $y_{t-5}$, $\Delta p_{t-5}$, $y_{t-6}$. This gives a GUM 0 with 35 coefficients. The first search through the model space is in branches that do not already have a terminal, yielding 4 terminal candidate models (including the final model). Next the union contrast yields another 6 terminals (the number is restricted). GUM 1 is the union of these 10 terminals with 29 coefficients.

In the next round, Autometrics finds an additional 8 contrasting terminals, taking the total to 18. PcGets is reported to have found only 2 terminals, illustrating how different the search procedures can be. [14]

[14] Hendry and Krolzig (1999) use an early version of PcGets. The presearch removes 16 variables from the initial GUM, leaving 11 insignificant in GUM 0, and 11 paths to explore. A more recent PcGets (version 1.12 from 2002) has the more elaborate 5-step presearch. This removes 20 variables, leaving 18 significant variables (at 1%) in GUM 0. As a consequence, there are no paths to explore. So GUM 0 is the final model, corresponding to terminal 5 of Autometrics. This happens to be the terminal with the lowest AIC and HQ.

The first terminal found by Autometrics is the final model selected by PcGets. However, Autometrics chooses the second terminal because it has a smaller SC (and a better fit, because it also has 16 coefficients), see Table 4.9. The difference is very small though. The final model has one insignificant coefficient, which survives because of backtesting against GUM 0.

end p.118

**Table 4.9. GUM 1 and first two terminals found by Autometrics**

| | GUM 1 | Terminal 1 | | | Terminal 2 | | |
|---|---|---|---|---|---|---|---|
| | p-value | coeff. | (s.e.) | p-value | coeff. | (s.e.) | p-value |
| $(m-p)_{t-1}$ | 0.0000 | 1.115 | (0.056) | 0.0000 | 1.259 | (0.063) | 0.0000 |
| $(m-p)_{t-2}$ | 0.0026 | −0.419 | (0.062) | 0.0000 | −0.564 | (0.069) | 0.0000 |
| $(m-p)_{t-3}$ | 0.9531 | — | — | — | — | — | — |
| $(m-p)_{t-4}$ | 0.2143 | — | — | — | — | — | — |
| $(m-p)_{t-5}$ | 0.0000 | 0.260 | (0.061) | 0.0000 | 0.294 | (0.062) | 0.0000 |

Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $(m-p)_{t-6}$ | 0.0022 | −0.134 | (0.052) | 0.0121 | −0.148 | (0.053) | 0.0060 |
| $y_t$ | 0.0000 | 0.113 | (0.008) | 0.0000 | 0.103 | (0.009) | 0.0000 |
| $y_{t-3}$ | 0.0101 | — | — | — | — | — | — |
| $y_{t-4}$ | 0.0522 | — | — | — | — | — | — |
| $\Delta p_t$ | 0.0000 | −0.845 | (0.105) | 0.0000 | −0.842 | (0.111) | 0.0000 |
| $\Delta p_{t-1}$ | 0.2905 | — | — | — | 0.250 | (0.128) | 0.0534 |
| $\Delta p_{t-2}$ | 0.0038 | −0.408 | (0.112) | 0.0005 | −0.411 | (0.114) | 0.0005 |
| $\Delta p_{t-3}$ | 0.0246 | −0.335 | (0.114) | 0.0042 | −0.406 | (0.114) | 0.0006 |
| $\Delta p_{t-1}$ | 0.0378 | — | — | — | — | — | — |
| $V_t$ | 0.1430 | 0.701 | (0.080) | 0.0000 | — | — | — |
| $V_{t-1}$ | 0.5509 | — | — | — | 0.663 | (0.080) | 0.0000 |
| $R_{20}$ | 0.0000 | −0.751 | (0.065) | 0.0000 | −0.830 | (0.083) | 0.0000 |
| $R_{20,\,t-1}$ | 0.0083 | — | — | — | — | — | — |
| $R_{1,\,t}$ | 0.0038 | — | — | — | — | — | — |
| $R_{ma,\,t}$ | 0.0032 | — | — | — | — | — | — |
| $R_{ma,\,t-1}$ | 0.0005 | — | — | — | 0.134 | (0.048) | 0.0064 |
| $R_{na,\,t}$ | 0.0189 | 1.187 | (0.384) | 0.0026 | 0.241 | (0.070) | 0.0008 |
| $R_{na,\,t-1}$ | 0.0330 | −1.857 | (0.660) | 0.0059 | — | — | — |
| $R_{na,\,t-2}$ | 0.0521 | 0.902 | (0.349) | 0.0111 | — | — | — |
| $R_{sna,\,t}$ | 0.1502 | — | — | — | — | — | — |
| $D80_t$ | 0.0001 | 0.017 | (0.003) | 0.0000 | 0.019 | (0.003) | 0.0000 |
| $SV_{t-1}$ | 0.0286 | — | — | — | — | — | — |
| $SV_{t-2}$ | 0.0000 | −13.03 | (1.705) | 0.0000 | −11.74 | (1.825) | 0.0000 |
| 1 | 0.0000 | 0.240 | (0.022) | 0.0000 | 0.214 | (0.020) | 0.0000 |
| $\hat{\sigma}$ | | 0.4319% | | | 0.4316% | | |
| SC | -7.3415 | −7.5347 | | | −7.5361 | | |

The content of all 18 terminals is represented in Table 4.10. The remaining terminals differ largely in which lags of $m_p$, $y$, $\Delta p$, $V$, and $SV$ enter the model. In addition, they illustrate that it is difficult to distinguish between the many interest rate variables (and their lags). This is probably no surprise, but no previous technology revealed this so clearly to the empirical modeller.

## 4.8 Conclusions

Autometrics is a new algorithm for model selection within the general-to- specific framework. It follows on the work by Hoover and Perez (1999) and Hendry and Krolzig (2005). The multiple path search is replaced by a tree search, which improves the small sample behaviour considerably without

end p.119

**Table 4.10. Eighteen terminals found by Autometrics; in the last column ++ denotes a variable in all terminals, + in twelve or more terminals**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $(m-p)_{t-1}$ | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | ++ |
| $(m-p)_{t-2}$ | * | * | * | * | * | * | — | — | * | * | — | * | * | * | * | * | * | * | + |
| $(m-p)_{t-3}$ | — | — | — | — | — | — | * | * | — | — | * | — | * | — | * | — | — | * | |
| $(m-p)_{t-4}$ | — | — | — | * | — | — | — | — | — | — | — | * | — | * | — | * | * | * | |
| $(m-p)_{t-5}$ | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | ++ |
| $(m-p)_{t-6}$ | * | * | * | * | — | — | — | * | * | — | * | * | * | * | * | * | * | * | + |
| $Y_t$ | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | ++ |
| $Y_{t-3}$ | — | — | — | — | — | — | — | — | * | — | — | — | * | — | * | — | — | — | |
| $Y_{t-4}$ | — | — | — | — | — | — | — | — | * | — | — | — | * | — | * | — | — | — | |
| $\Delta p_t$ | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | ++ |
| $\Delta p_{t-1}$ | — | * | — | — | — | — | — | — | — | — | — | * | * | — | * | — | — | — | |
| $\Delta p_{t-2}$ | * | * | * | * | * | * | — | — | * | * | * | * | * | * | * | * | * | * | + |
| $\Delta p_{t-3}$ | * | * | * | — | * | * | * | * | * | * | * | * | — | — | — | — | — | * | + |

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\Delta p_{t-4}$ | — | — | — | * | — | — | * | * | — | — | — | — | — | — | — | * | — | — | * |
| $V_t$ | * | — | * | * | * | * | * | * | — | — | * | — | — | * | * | * | * | — | + |
| $V_{t-1}$ | — | * | — | — | — | — | — | — | * | * | — | * | * | — | — | — | — | * | |
| $R_{20}$ | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | ++ |
| $R_{20,\,t-1}$ | — | — | * | — | * | * | * | * | — | * | * | — | — | * | — | — | * | — | |
| $R_{1,\,t}$ | — | — | * | — | * | * | * | * | — | * | * | * | * | * | * | — | * | — | + |
| $R_{ma,\,t}$ | — | — | * | — | * | * | * | * | — | — | — | * | * | * | * | — | * | — | |
| $R_{ma,\,t-1}$ | — | * | * | — | * | * | * | * | — | * | * | * | * | * | * | * | * | — | + |
| $R_{na,\,t}$ | * | * | — | * | — | * | * | * | * | * | * | — | — | * | · | * | — | * | + |
| $R_{na,\,t-1}$ | * | — | — | * | — | — | * | * | * | * | * | — | — | — | — | * | — | * | |
| $R_{na,\,t-2}$ | * | — | — | * | * | — | * | * | * | * | * | * | * | — | — | * | * | * | + |
| $R_{sna,\,t}$ | — | — | * | — | — | — | — | — | — | — | — | — | — | * | * | — | — | — | |
| $D80_t$ | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | ++ |
| $SV_{t-1}$ | — | — | * | — | * | * | * | * | — | — | * | — | — | * | — | — | * | — | |
| $SV_{t-2}$ | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | ++ |
| 1 | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | ++ |
| k | 16 | 16 | 19 | 17 | 18 | 18 | 20 | 20 | 18 | 18 | 20 | 18 | 21 | 19 | 22 | 17 | 19 | 18 | |

end p.120

requiring the addition of presearches to the algorithm. Additional proposed improvements include delayed diagnostic testing, and using the unique tree representation to implement union and terminal contrasts. Another interesting finding is the somewhat better performance of backtesting with respect to the initial GUM (or the GUM after presearch), thus effectively removing the encompassing step from the iterations (because it already takes place inside the tree search).

**References**

Baba, Y., Hendry, D. F., and Starr, R. M. (1992). The demand for M1 in the U.S.A., 1960–1988. *Review of Economic Studies*, **59**, 25–61. Link

Castle, J. L. and Hendry, D. F. (2007). Extending the boundaries of automatic selection: Non-linear models. Mimeo, Department of Economics, Oxford University.

Doornik, J. A. (2007). *Object-Oriented Matrix Programming Using Ox*, 6th edition. London: Timberlake Consultants Press.

——(2008). Encompassing and automatic model selection. Oxford Bulletin of Economics and Statistics, **70**, 915–925. Link

Hendry, D. F. (2000). *Econometrics: Alchemy or Science ?* Oxford: Oxford University Press. Link OSO X-Reference

——and Doornik, J. A. (2007). *Empirical Econometric Modelling Using PcGive: Volume I*, 5th edition. London: Timberlake Consultants Press.

——and Krolzig, H.-M. (1999). Improving on 'Data mining reconsidered' by K. D. Hoover and S. J. Perez. *Econometrics Journal*, **2**, 202–219. Link

—— ——(2001). *Automatic Econometric Model Selection*. London: Timberlake Consultants Press.

—— ——(2003). New developments in automatic general-to-specific modelling. In Stigum, B. P. (ed.), *Econometrics and the Philosophy of Economics*, pp. 379–419. Princeton: Princeton University Press.

—— ——(2004). Sub-sample model selection procedures in general-to-specific modelling. In Becker, R. and Hurn, S. (eds.), *Contemporary Issues in Economics and Econometrics: Theory and Application*, pp. 53–74. Cheltenham: Edward Elgar.

—— ——(2005). The properties of automatic Gets modelling. *Economic Journal*, **115**, C32–C61.

——and Nielsen, B. (2007). *Econometric Modeling: A Likelihood Approach*. Princeton: Princeton University 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**, 167–191. Link

**The Methodology and Practice of Econometrics ,A Festschrift in Honour of David F. Hendry**
**Castle, Jennifer** (Editor), British Academy Postdoctoral Fellow in Economics
**Shephard, Neil** (Editor), Professor of Economics at Oxford University and Research Director of the University's Oxford-Man Institute
Print publication date: 2009, Published to Oxford Scholarship Online: September 2009
Print ISBN-13: 978-0-19-923719-7, doi:10.1093/acprof:oso/9780199237197.001.0001

Hoover, K. D. and Perez, S. J. (2004). Truth and robustness in cross-country growth regressions. *Oxford Bulletin of Economics and Statistics*, **66**, 765–798. Link

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

Lynch, A. W. and Vital-Ahuja, T. (1998). Can subsample evidence alleviate the data-snooping problem ? A comparison to the maximal $R_2$ cutoff test. Discussion Paper, Stern Business School, New York University.

end p.121