Some Observations on Automatic model Selection

Jurgen A. Doornik [*]
Institute for New Economic Thinking at the Oxford Martin School
University of Oxford, UK

December 9, 2013

Preliminary and incomplete; not ready for citation yet

Abstract

The foundations of automated general-to-specific (Gets) modelling are reviewed. The algorithms under consideration are formulated, and the salient differences highlighted. Additional motivations are provided for the procedural decisions that have been made. We show that estimating all possible models, while not possible in practice, can be useful in clarifying some outstanding issues (at the moment this issue is not fully decided yet).

1 Introduction

2do: Introduction

2 Selective historical background

In the last two decades, automatic modelling software has become more popular. Aided by faster computers, algorithms can be implemented and investigated using Monte Carlo methods.

Interest in strategies for modelling goes further back. Efroymson (1960) proposed stepwise regression, which would have been attractive for its computational simplicity (many coding optimizations employed in those days are not really worth the effort today). Hamaker (1962) discusses the implementation of forward selection and backward elimination, and notes that forward selection can break down when correlations between variables are high. Beale, Kendall, and Mantel (1967) consider finding the

[*] Thanks

best subset of a certain size. Cox and Snell (1974) and Hocking (1976) provide a brief overview of the early methods of variable selection.

From the offset, there was a debate about the best method for variable selection. Mantel (1970) argues that backward elimination (or 'step-down') should be preferred over forward selection. One argument put forward in favour of backward elimination is reduced computational effort, but this is unclear, and now unimportant anyway. The second is again related to the situation where correlations mask the significance of a single variable. Beale (1970) counters that the advantage is exaggerated, because, when backward elimination drops a variable, it can not re-enter anymore at a later stage. this path dependence can be a problem for stepwise regression as well: if two almost equally significant variables swap place, a different final model may be found.

The early selection literature was primarily focussed on goodness-of-fit, while working under severe computational constraints. There seems to be no concern about the quality of the final model, or any distinction between good and bad models.

Around the same time, the econometric literature was more interested in the development of mis-specification (or diagnostic) tests: when inference in a linear regression model is based on independently normal error terms, those assumptions can and should be tested as a matter of course. E.g., the influential empirical modelling of consumption in the UK in Davidson, Hendry, Srba, and Yeo (1978) uses mis-specification testing for all specifications considered. Such tests became an important aspect of econometric software, see Hendry (1986) and Ericsson, Campos, and Tran (1990). Relatedly, the concept of what makes a good model was developed in the work of David Hendry and colleagues, culminating in the concept of *congruence*, see e.g. Gilbert (1986) and Hendry and Nielsen (2007, Ch.20). A model is congruent if:

(1) it is consistent with theory;
(2) its formulation is logically consistent, i.e. the left and right-hand side are coherent;
(3) it is based on valid conditioning;
(4) it has constant parameters;
(5) it satisfies the stochastic assumptions;
(6) and finally, it encompasses rival models.

The first three are largely up to the researcher, and recipes for modelling cannot offer much help.[1] Next, parameters should be constant inside the estimation sample, but ideally also out-of-sample. One approach is to hold back data for this purpose (in other contexts referred to as the 'test' data, after estimation using the 'training' set). However, when data is scarce, such as in macro-economic modelling, it is not clear if anything is gained from this (see Hendry and Krolzig, 2004a). Also, in time-series modelling it is often unattractive to hold back the most recent period. Item (5) can be verified through the mis-specification tests referred to above. Because modelling starts from the general, with the initial set of variables determined by the researcher, the general model is one such rival model, assuming it can be estimated. So the selected model should, as a minimal requirement, be a valid reduction of this *general unrestricted model* (GUM). Ideally, it is not dominated by another model that can be selected from the GUM. A rival model obtained from another source may result in an extension of the information set.

[1] But note Hendry and Santos (2010).

A general-to-specific methodology (*Gets*) was proposed to find a congruent model. The claim of Gilbert (1986, p.295), that scientific discovery, and by extension *Gets*, cannot be automated, is at least partially overturned by later developments: Hoover and Perez (1999) pioneered an automated version of *Gets* and studied its properties. This was subsequently extended by Hendry and Krolzig (1999) and Doornik (2009).

There is a large literature on model selection, partially falling under the label of statistical learning or data mining, see e.g. Hastie, Tibshirani, and Friedman (2009). We restrict ourselves to only a few methods that can also be used with continuous data in a time-series setting.

3 A fundamental difference

Imagine the following situation. A student approaches you with a pressing question: 'I have a set of variables G, from which I selected this model M. But I may have made a mistake. Could you tell me if this is the $\{$*stepwise, backward elimination, Lasso, information criterion, ...*$\}$ solution?' Then you have no alternative but to rerun the procedure from scratch to answer the question. So, while the procedure itself may be quick, it is quite cumbersome to verify the results. Each entails that there is a unique, but different, solution.

Now let the student ask the question: is M the Gets model? This question is easily answered, provided the student gives some additional information:

1. the likelihood of G,
2. the likelihood of M,
3. diagnostic tests of M.

If this checks out fine, the model can be accepted as satisfactory. This is based on items (4)–(6) of congruence. We could have paid some more attention to constancy and valid conditioning. There could be other models that we prefer (there is not necessarily a unique Gets model), or we may wish to suggests improvements to G. Despite this, the question was quite readily answered.

The importance of this distinction does not seem to be sufficiently appreciated in the literature. By classifying models as acceptable and non-acceptable, we limit what we are looking for in a model selection procedure. By requesting that the GUM G satisfies the criteria, we ensure that there is a solution. While finding all Gets models is generally too costly, finding several to choose from is quite feasible.

When stepwise regression, backward elimination, and Lasso (Efron, Hastie, Johnstone, and Tibshirani, 2004) each yield a different model, what does that entail? Is one better than the other, or are they all equally good? We can run horse races (e.g. the forecasting comparison in Kock and Teräsvirta, 2014), but these results tend not to generalize to other settings or time periods. The notion of congruence can, of course be applied here too: we can even be lucky that all found models are acceptible to us. If not, there is no obvious route to recovery. We return to this issue below. Another approach is to ignore all this, choose one method, and proceed with the selected model.[2] However, the results can be quite fragile, making it difficult to convince others of the value of the model (unless the choosen method happens to be *en vogue*).

[2] One further approach is not to select at all.

It is not suggested that G should consist of anything and anything ever thought of, say a dump of all the information on the internet. A researcher should restrict herself to a relevant, and potentially useful information set. This can be a subject of much debate, but, once determined, automatic model selection methods can save much research time. Even stepwise regression requires formulation of G: the candidate set has to be determined before pressing a button on the computer. As a consequence, we belief that we study of automated model selection is a well-defined problem, and the tools are statistical analysis and Monte Carlo experimentation. Appeals to metaphysical principles, as in McAleer (2005), serve only to obfuscate.

4 Data mining: good or bad?

The economics discipline is unique in that the term 'data mining' tends to be used to in a pejorative manner. As Hoover (2013, p.53) points out, outside economics, when people mine for something, they try to harvest a valuable material (gold, coal, etc.). In computer science, data mining is considered by some to be a novel activity: the preface, written in 2001, of Hastie, Tibshirani, and Friedman (2009) defines the new field of data mining as extracting patterns and trends from vast amounts of data. The distinguishing feature here is the search in large databases.

Sargan (2001a, p.159), writing in 1973, adopts the definition of data mining as '*A model which has been fitted to data over some sample period is found to have a significantly worse error variance than it should when used to predict in a later period.*'

Lovell (1983, p.1), in his paper entitled 'Data Mining', describes it as follows: '*When a data miner uncovers t-statistics that appear significant at the 0.05 level by running a large number of alternative regressions on the same body of data, the probability of a Type I error of rejecting the null hypothesis when it is true is much greater than the claimed 5%.*'

Sargan's notion is that of overfitting. This is not a logical consequence of model selection: if we are so conservative that we almost always select the empty model, there will not be any overfitting. However, it is an issue we need to be aware of, particularly if model selection is based on some form of maximization of in-sample fit.

Lovell (1983)'s statement is stronger, but we shall show that it (1) depends on the hypothesis being considered, (2) depends on test procedure that is used, and, (3) when it is considered an issue, can be avoided.

To clarify the issue, we specify the example of Lovell (1983, Section II) as a Monte Carlo experiment. The data generation process (DGP) is given by

$$\begin{array}{rcll} y_t & = & \sum_{i=1}^{K}\beta_i x_{i,t}+\epsilon_t, & \epsilon_t\sim \mathsf{IN}\left[0,1\right],\ t=1,...,T, \\ \mathbf{x}_t & = & (x_{1,t},...,x_{K,t})', & \mathbf{x}_t\sim \mathsf{IN}_K\left[\mathbf{0},\mathbf{I}_K\right]. \end{array} \tag{1}$$

The general unrestricted model (GUM) consists of the K variables and an intercept:

$$y_t \ = \ \gamma_0^F+\sum_{i=1}^{10}\gamma_i x_{i,t}+u_t \quad u_t\sim \mathsf{IN}\left[0,\sigma_u^2\right]. \tag{2}$$

The superscript F on γ_0 indicates that the intercept is forced in all models (although omitting the intercept throughout would not have a material impact).

The test of $H_0 : \gamma_0 = ... = \gamma_K = 0$ in (2), under the assumption that the null hypothesis is true, i.e. all the β_i are zero, has the standard central F distribution. Similarly, the test for $\gamma_i = 0$ when it is true has the expected central Student-t distribution.

Geary (1967) analyzes the simplified case, where the regressors are orthogonal, so $\mathbf{X}'\mathbf{X}$ is a diagonal matrix, $\mathbf{X}' = (\mathbf{x}_1...\mathbf{x}_T)$. Moreover, σ^2 is known, so the t-values are independent draws from $\mathsf{N}[0, 1]$ when the true coefficients are zero. This is the setting for the remainder of thissection. It is easily seen that, when using a p-value of α and corresponding critical value c_α, the probability that all absolute t-statistics are less than c_α is given by $(1 - \alpha)^K$. In other words, when $K = 20$ and $\alpha = 0.05$, one will find one variable significant on average using a critical value of two, independently of the sample size.

The procedure investigated in Lovell (1983, Section II) is that of selecting the two most significant variables from the candidate set of size K. The null hypothesis that no variables matter is rejected when any of the selected variables are significant, i.e non-rejection occurs when both coefficients are deemed insignificant. In the simplified orthogonal setting, this amounts to estimating (2) and selecting those two variables with the highest t-values (they need not be significant). If any of these is significant, the null hypothesis is rejected. If we denote the ordered t-values as:

$$|t_{(K)}| \geq |t_{(K-1)}| \geq ... \geq |t_{(1)}|,$$

and the corresponding variables as $x_{(K)}, x_{(K-1)}, ...$, then the procedure is:

$$\boxed{\begin{array}{ll} P_1 : & \text{Estimate (2)} \\ & \text{Reject } H_0 : \gamma_0 = ... = \gamma_K = 0 \text{ if } |t_{(K)}| > c^*_\alpha. \end{array}}$$

We would expect that every modeller, data miner or not, would know that c^*_α should not be taken from the normal or t-distribution: we are working now with order statistics.

As an example we take $K = 10$ variables to select from, and set $c^*_\alpha = 1.96$ (i.e. adopting a standard normal distribution). Then for procedure P_1, the probability to reject H_0 when it is true equals one minus the probability of finding no regressor significant, which is $1 - 0.95^{10} = 0.4$. Geary (1967, Table 1) shows that we should have used 2.8 as the critical value. In other words, we should have used a 99.5% quantile from the normal distribution to obtain a size of 5% for test P_1:

$$(1 - 0.05) \approx (1 - 0.005)^{10}.$$

How can we reconcile the value 0.4 with a claimed 'true significance level' of 0.226 in Table 1 of Lovell? The answer is that we cannot. The same table shows the 'true significance level' for $K = 2$ to be 5%. But the footnote shows that it is derived as a tautology, namely the value of $\widehat{\alpha}$ that solves:[3]

$$(1 - \alpha)^2 = (1 - \widehat{\alpha})^2.$$

[3] The entries for $K > 2$ are the answers to the following question: which $\widehat{\alpha}$ would give us the same rejection frequency as that found when using α for $K = 2$?

The actual probability to reject H_0 using normality for each t-statistic (without selection), also given in the Lovell's Table 1, equals $1 - 0.95^2 = 0.1$. It is not the claimed 5%. The selection is actually a red herring here: in the simplified setup it does not matter if we select or not. However, the number of variable matters, because K determines the distribution of the order statistic used in P_1. Only for $K = 1$ can we use a 5% critical value of 1.96.

One could argue that this is just a cosmetic issue: the problem could have been avoided had the setup been restricted to selecting only one regressor. On the other hand, a researcher wishing to test H_0 should not do it using P_1, but indeed look at the F-test of all coefficients in (2).

Another method of testing H_0 is to require both coefficients to be significant:

```
$P_2$ :  Estimate (2)
         Reject $H_0 : \gamma_0 = ... = \gamma_K = 0$ if $|t_{(K-1)}| > c^*_\alpha$.
```

Geary (1967, Table 1) gives a 5% critical of 2.07 for $K = 10$, which is not far from 1.96. For smaller K the procedure is undersized, but, again, the distribution depends on K.

Now consider the following model selection procedure:

```
$P_3$ :  Estimate (2)
         Estimate all possible submodels, keeping only those with $Z^2(k) \leq c_\alpha$
         Keep only those models in the selected pool that are ‘minimal’
         If all models are rejected, return the GUM
         Reject $H_0 : \gamma_0 ... \gamma_K = 0$ if one or more non-empty model is selected.
```

This needs some further explanation. In the second step we estimate $2^K - 1$ models, and reject those that that are significant at 5% in an F-test on all coefficients (ignoring the intercept, which is always forced in the models). The test statistic is denoted $Z^2(k)$. We may also use a likelihood-ratio test and a $\chi^2(k)$ distribution. The third step is a logical step: all redundant models are removed from the models that survive the tests, leaving one or more final model. Here a model is redundant if a subset of it is also a model in the surviving pool.

P_3 can be analyzed under the null that all variables are irrelevant. The test of the empty model is rejected with probability α. If that is the case either the GUM or some sub model(s) survive, and H_0 is rejected. Otherwise, the empty model is accepted (with probability $1 - \alpha$). In that case, all other models are redundant, because every model nests the empty model. Here we have a procedure that estimates all possible models, which surely qualifies as a ‘large number of alternative regressions on the same body of data’. However, the size of the joint test on all coefficients is actually α.

There is one issue though that remains: it is mathematically impossible in the setup that we considered to control both the size of $H_0 : \gamma_0 = ... = \gamma_K = 0$ at $\alpha = 5\%$ and that of $H_0^* : \gamma_i = 0$ at $\alpha^* = 5\%$. If we adopt the former, then, because of orthogonality and known σ, we find that α^* is defined through:

$$(1 - \alpha^*)^K = 1 - \alpha.$$

As already noted above, for $K = 10$, $\alpha = 5\%$ corresponds to $\alpha^* = 0.5\%$, and $\alpha^* = 5\%$ corresponds to $\alpha = 40\%$.

Because H_0 is intrinsically of less interest, we prefer a method that controls α^*. A model selection method that controls neither seems undesirable. This defines our aims when allowing for correlated regressors.

Table 2 reports a small Monte Carlo experiment, to show that estimation of σ, and independence rather than orthogonality, does not change the results obtained in this section. The DGP is (1) with $\beta_1 = ...\beta_K = 0$; $x_{1,t}, ..., x_{K,t}$ provides the candidate regressor set, so the GUM is (2). The selection methods that are used are S_1 and S_3, the model selection versions of P_1 and P_3, as well as stepwise regression and backward elimination. S_1 is simply selecting significant regressors:

```
$S_1$ :  Estimate (2)
         Select all regressors that have $|t_i| > c_\alpha$.
```

In the case of S_3, the Bayesian information criterion[4] is used as tie breaker:

```
$S_3$ :  Estimate (2)
         Estimate all possible submodels, keeping only those with $Z^2(k) \leq c_\alpha$
         Keep only those models in the selected pool that are ‘minimal’
         If all models are rejected, return the GUM
         Otherwise return the model with the smallest BIC.
```

The test used in S_3 is the likelihood-ratio test with a χ^2 distribution. We set the sample size to $T = 200$ and use $M = 10000$ replications.

The first column with results in Table 2 is labelled ‘gauge’. This is the retention rate of irrelevant variables in the final model. The gauge is averaged over all regressors, but, because they are exchangeable, it also applies to each individual regressor. The gauge can therefore be interpreted as the type I error of the test $H_0 : \gamma_i = 0$. The column with $\widehat{\sigma}$ reports the average residual standard error, which is unity in the DGP, and always close to one here – there is no sign of overfitting, so no data mining in the sense of Denis Sargan.

The last three columns relate to the average model size: the percentage of models that is empty ($k = 0$), the percentage that has one regressor ($k = 1$), and the remainder. One hundred minus the percentage of empty models is the type I error of the test for $H_0 : \gamma_1 = ... = \gamma_K = 0$.

Returning to the gauge, we see that for all methods except S_3, the empirical rejection frequencies are equal to the nominal frequency of 5%. These three methods are all very similar, which is to be expected in an independent design. They all have a type I error for $H_0 : \gamma_1 = ... = \gamma_K = 0$ that corresponds with the theory, e.g. about 40% when $K = 10$. As noted, this is not the objective of these selection methods.

S_3 is very different. Here the type I error for $H_0 : \gamma_1 = ... = \gamma_K = 0$ is close to α, ranging from 5.1% to 6% in the table. The price that logically must be paid is that the type I error for $H_0 : \gamma_i = 0$ gets smaller as K grows. We should note that S_3 is

[4] Also called Schwarz criterion (SC). BIC is defined as $(-2\widehat{\ell} + k \log T)/T$, where $\widehat{\ell}$ is the estimated log-likelihood, and k the number of parameters in the model.

	Gauge	$\widehat{\sigma}$	$k=0$	$k=1$	$k\geq 2$
			$K=1$		
S_1	0.050	0.9984	95.0%	5.0%	0%
S_3	0.051	0.9984	94.9%	5.1%	0%
Stepwise	0.050	0.9984	95.0%	5.0%	0%
Backward	0.050	0.9984	95.0%	5.0%	0%
			$K=2$		
S_1	0.049	0.9977	90.4%	9.4%	0.2%
S_3	0.028	0.9981	94.7%	5.1%	0.2%
Stepwise	0.049	0.9977	90.4%	9.4%	0.2%
Backward	0.049	0.9977	90.4%	9.4%	0.2%
			$K=10$		
S_1	0.050	0.9935	60.6%	30.1%	9.2%
S_3	0.007	0.9979	94.0%	5.3%	0.7%
Stepwise	0.051	0.9933	59.4%	32.0%	8.6%
Backward	0.052	0.9931	58.9%	31.8%	9.3%

Table 1 Type I errors of some model selection procedures at nominal significance level $\alpha = 5\%$ and with K variables in the GUM. All variables are irrelevant in the DGP, $T = 200, M = 10^5$.

not feasible in general: our brute force method of estimating 2^K models can be taken to about $K = 15$.

****** ends here ******

T	Gauge	$\widehat{\sigma}$	$k=0$	$k=1$	$k\geq 2$
			S_1		
10	0.048	0.9504	91.4%	7.7%	0.9%
20	0.051	0.9765	90.3%	9.2%	0.4%
50	0.049	0.9898	90.6%	9.1%	0.4%
1000	0.051	0.9997	90.1%	9.7%	0.2%
100000	0.052	1.0000	89.9%	9.8%	0.3%
			min. BIC		
10	0.201	0.9170	65.3%	28.5%	6.2%
20	0.118	0.9688	79.6%	18.5%	1.9%
50	0.054	0.9895	89.7%	9.9%	0.4%
1000	0.008	0.9998	98.4%	1.6%	0%
100000	0.0007	1.0000	99.9%	0.1%	0%

Table 2 Gauges of S_1 at $\alpha = 5\%$, and model selection by BIC. $K = 2$ variables in the GUM, all irrelevant in the DGP, $M = 10^5$.

5 Asymptotic properties

A model selection procedure is consistent if it finds the correct model with probability approaching certainty as the sample size goes to infinity. In that case, inference based on the selected model is asymptotically identical to that of the true model (Pötscher, 1991, Lemma 1).

There are many consistent model selection procedures. For example, selecting the model with the smallest BIC is consistent, whereas using AIC $= (-2\widehat{\ell} + 2k)/T$ instead is not (AIC leads to larger models). Information criteria can be expressed as a likelihood-ratio test with critical value depending on sample size and number of regressors.

The distinction made in the previous section was also addressed in Sargan (2001b), showing that S_1 can be made consistent by shrinking the significance level sufficiently quickly as the sample size grows.

Consistency requires that the correct model is within the set being searched.

6 Finding Gets models: multipath search

The core of the multipath search algorithm adopted by Hoover and Perez (1999) and Hendry and Krolzig (2006) can be described as follows:

1. Specify the GUM and choose a significance level α.
2. Estimate the GUM, and abort if it fails diagnostic testing.
3. Order the variables according to their squared t-values, most insignificant first:
   $$t^2_{(1)}, t^2_{(2)}, ..., t^2_{(k)}.$$
   Let m be the cut-off, such that $t^2_{(m+1)} \geq c^2_\alpha$, so there are m insignificant variables in the GUM.
4. for $i = 1, ..., m$: delete variable (i), re-estimate, and follow the backward elimination path, stopping when
   (a) the next marginal variable is significant (standard termination of backward elimination);
   (b) the model fails the reduction test against the GUM (encompassing or backtesting failure);
   (c) the model fails diagnostic testing (diagnostic tracking failure);
   A failure means that the previous model is a terminal candidate (with one or more insignificant variables), otherwise it is the current model.
5. Remove duplicate terminal candidates to find the first set of terminal models.

If the m insignificant variables are also insignificant in each path that is followed, then there is one terminal model with those m removed, which is found after $1+m(m-1)$ model estimations.

The first terminal model that is found, in the absence of backtesting and diagnostic tracking is the backward elimination model.

7 Finding Gets models: tree search

The algorithm at the heart of Doornik (2009) is different, but tries to achieve the same aims more efficiently:

[description needed]

Again, in its basic form, without backtesting and diagnostic tracking, *Autometrics* finds the backward elimination model first.

8 Finding Gets models: pre-search

A pre-search to remove highly insignificant variables prior to the application of the search algorithm, is sometimes adopted to reduce the time of searching for terminal models. This notion was introduced by Hendry and Krolzig (2006) in their *PcGets* software, and can be seen as backward elimination without re-estimation:

1. Order the variables according to their squared t-values, most insignificant first:
$$t^2_{(1)} \leq t^2_{(2)} \leq ... \leq t^2_{(k)}.$$
Let m be the cut-off, such that $t^2_{(m+1)} \geq c^2_\alpha$, so there are m insignificant variables in the GUM.
2. for $i = 1, ..., m$: if variables $(1)..(i)$ are not a valid reduction of the GUM, set $\overline{m} = i - 1$ and stop.
3. Remove variables $1, .., \overline{m}$.

9 Properties of selection

- Measuring success
- Distribution of post-selection estimator (bias correction)
- Consistent selection (oracle nonsense)

10 Properties of automated Gets

To study the properties of automated Gets, simplified versions of the algorithms given above (which are already simplifications) could provide useful starting points.

One concern that has frequently been raised, is that the algorithms use excessive testing, to such an extent that the final model is meaningless and inference distorted. (bit vague) However, to some extent, the method of discovery may not matter: the questions asked by the student above can be answered, irregardless of how the student found the model. Indeed, if there were a known DGP, and the student had found it, it would surely be perverse to reject it because of excessive 'data mining'.

Hendry and Krolzig (2004b) state that, when all regressors are mutually orthogonal, and the set of variables is more general than the DGP, model selection can be based on selecting all those variables with a squared t-value below a specified critical value (say using $\alpha = 5\%$). (provided $k << T$) In such a world, only one decision is required, based upon a single estimation. In that case, stepwise regression and backward elimination would be the same (allowing for some small-sample fluctuations at the margin). Similarly, because the independent t-tests contain the same information as the joint F-tests against the GUM (or likelihood-ratio tests), the multi-path algorithm (without diagnostic testing) is also the same. In that case all estimations are redundant, all further testing a duplication, and there is no impact of testing beyond the single decision.

We then claim that all testing in the path searched is to improve t-testing in the presence of non-orthogonality. I feel that this can be substantiated for Autometrics, but perhaps not for PcGets.

11 Properties of 1-cut selection

Castle, Doornik, and Hendry (2011) label selection on the basis of ordered t-values as '1-cut selection'. Hendry and Krolzig (2003) derive the probabilities of selecting one or more variables when none matter (the 'gauge'), assuming orthogonality, and book(2014) shows that 1-cut selection is consistent.

12 Properties of backward elimination

References

Beale, E. M. L. (1970). Note on procedures for variable selection in multiple regression. *Technometrics 12*, 909–914.

Beale, E. M. L., M. G. Kendall, and N. Mantel (1967). The discarding of variables in multivariate analysis. *Biometrika 54*, 357–366.

Castle, J. L., J. A. Doornik, and D. F. Hendry (2011). Evaluating automatic model selection. *Journal of Time Series Econometrics 3*, 884–889. doi:10.2202/1941-1928.1097.

Cox, D. R. and E. J. Snell (1974). The choice of variables in observational studies. *Applied Statistics 23*, 51–59.

Davidson, J. E. H., D. F. Hendry, F. Srba, and J. S. Yeo (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., *Econometrics: Alchemy or Science?* Oxford: Blackwell Publishers, 1993, and Oxford University Press, 2000; and in Campos, J., Ericsson, N.R. and Hendry, D.F. (eds.), *General to Specific Modelling*. Edward Elgar, 2005.

Doornik, J. A. (2009). Autometrics. In J. L. Castle and N. Shephard (Eds.), *The Methodology and Practice of Econometrics: Festschrift in Honour of David F. Hendry*. Oxford: Oxford University Press.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*, 407–499.

Efroymson, M. A. (1960). Multiple regression analysis. In A. Ralston and H. S. Wilf (Eds.), *Mathematical Meethods for Digital Computers*. New York: Wiley.

Ericsson, N. R., J. Campos, and H.-A. Tran (1990). PC-GIVE and David Hendry's econometric methodology. *Revista de Econometria 10*, 7–117.

Geary, R. C. (1967). Ex post determination of significance in multivariate regression when the independent variables are orthogonal. *Journal of the Royal Statistical Society B 29*, 154–161.

Gilbert, C. L. (1986). Professor Hendry's econometric methodology. *Oxford Bulletin of Economics and Statistics 48*, 283–307. Reprinted in Granger, C. W. J. (ed.) (1990), *Modelling Economic Series*. Oxford: Clarendon Press.

Hamaker, H. C. (1962). On multiple regression analysis. *Statistica Neerlandica 16*, 31–56.

Hastie, T. J., R. J. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Predictions*. New York: Springer.

Hendry, D. F. (1986). Using PC-GIVE in econometrics teaching. *Oxford Bulletin of Economics and Statistics 48*, 87–98.

Hendry, D. F. and H.-M. Krolzig (1999). Improving on 'Data mining reconsidered' by K.D. Hoover and S.J. Perez. *Econometrics Journal 2*, 202–219. Reprinted in J. Campos, N.R. Ericsson and D.F. Hendry (eds.), *General to Specific Modelling*. Edward Elgar, 2005.

Hendry, D. F. and H.-M. Krolzig (2003). New developments in automatic general-to-specific modelling. In B. P. Stigum (Ed.), *Econometrics and the Philosophy of Economics*, pp. 379–419. Princeton: Princeton University Press.

Hendry, D. F. and H.-M. Krolzig (2004a). Sub-sample model selection procedures in general-to-specific modelling. In R. Becker and S. Hurn (Eds.), *Contemporary Issues in Economics and Econometrics: Theory and Application*, pp. 53–74. Cheltenham: Edward Elgar.

Hendry, D. F. and H.-M. Krolzig (2004b). We ran one regression. *Oxford Bulletin of Economics and Statistics 66*, 799–810.

Hendry, D. F. and H.-M. Krolzig (2006). *Automatic Econometric Model Selection using PcGets* (2nd ed.). London: Timberlake Consultants Press.

Hendry, D. F. and B. Nielsen (2007). *Econometric Modeling: A Likelihood Approach*. Princeton: Princeton University Press.

Hendry, D. F. and C. Santos (2010). An automatic test of super exogeneity. In M. W. Watson, T. Bollerslev, and J. Russell (Eds.), *Volatility and Time Series Econometrics*, pp. 164–193. Oxford: Oxford University Press.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics 32*, 1–49.

Hoover, K. D. (2013). The role of hypothesis testing in the molding of econometric models. *Erasmus Journal for Philosophy and Economics 6*, 43–65.

Hoover, K. D. and S. J. Perez (1999). Data mining reconsidered: Encompassing and the general-to-specific approach to specification search. *Econometrics Journal 2*, 167–191.

Kock, A. B. and T. Teräsvirta (2014). Forecasting performance of three automated modelling techniques during the economic crisis 2007-2009. *International Journal of Forecasting forthcoming*.

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

Mantel, N. (1970). Why stepdown procedures in variable selection. *Technometrics 128*, 621–625.

McAleer, M. (2005). Automated inference and learning in modeling financial volatility. *Econometric Theory 21*, 232–261.

Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory 7*, 163–185.

Sargan, J. D. (2001a). Model building and data mining. *Econometric Reviews 20*, 159–170. Written and presented in 1973.

Sargan, J. D. (2001b). Model building and data mining. Technical report. Written and presented in 1981.