Journal of Modern Applied Statistical Methods

Volume 8 | Issue 2 Article 14

11-1-2009

Multiple Search Paths and the General-To-Specific Methodology

Paul Turner
Loughborough University, United Kingdom, p.m.turner@lboro.ac.uk

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 505-510

Multiple Search Paths and the General-To-Specific Methodology

Paul Turner
Loughborough University
United Kingdom

Increased interest in computer automation of the general-to-specific methodology has resulted from research by Hoover and Perez (1999) and Krolzig and Hendry (2001). This article presents simulation results for a multiple search path algorithm that has better properties than those generated by a single search path. The most noticeable improvements occur when the data contain unit roots.

Key words: Multiple search paths, general-to-specific, Monte Carlo simulation.

Introduction

The general-to-specific methodology introduced by Davidson, et al. (1978) and discussed by Gilbert (1986) is now a well established part of the technical toolkit of applied econometricians. The idea of this approach is to begin with a deliberately over-parameterized model, examine its properties (particularly those of the residuals) to ensure that it is data congruent and then to progressively simplify the model to obtain a parsimonious specification. Arguably, the main advantage of this approach is that, provided the original over-parameterized model is data congruent, tests of restrictions are always conducted against a statistically well specified alternative model. This contrasts with the alternative specific-to-general approach in which the alternative model is frequently badly specified, thereby invalidating the testing procedure.

A typical situation facing a modeler can be illustrated as follows. The modeler begins with a general model of the form which relates two variables of interest y and x which follow a dynamic relationship disturbed by a random error u:

$$y_t = \beta_0 + \sum_{i=0}^{4} \beta_{i+1} x_{t-i} + \sum_{i=1}^{4} \beta_{5+i} y_{t-i} + u_t \qquad (1)$$

Economic theory suggests that an equilibrium relationship exists between the variables y and x. However, theory typically indicates little about the short-run dynamic relationship between the variables. Therefore, beginning with (1)[1], exclusion restrictions for the right-hand side variables are tested and these are progressively eliminated until either the null $H_0 : \beta_i = 0$ is rejected or the model begins to show signs of misspecification in the form of serial correlation in the residuals, heteroscedasticity, non-normality etc. When a parsimonious specification is found then the model is often re-written in a more convenient form such as the error-correction representation.

One of the problems which arises with the general-to-specific methodology is the search path involved in moving from the general model (1) to a parsimonious specification is not unique. Typically, the general model contains a large number of highly co-linear variables. Exclusion of large numbers of variables at an early stage is a dangerous strategy since variables that may be insignificant in the most general model may become significant as other co-linear variables are excluded. Most advocates of this methodology therefore recommend proceeding gradually, eliminating a few variables at each stage of the specification search, until the final specification is obtained. However, the number of possible search paths

Paul Turner is Reader in Economics in the School of Business and Economics, Loughborough University. Email him at: P.M.Turner@lboro.ac.uk

may become large even in a relatively small model. Suppose, for example, that the true model requires a total of n restrictions on the most general model. It follows that there are $n!$ separate search paths which involve the elimination of one variable at each stage and which will succeed in getting from the general model to the final, correct specification. If the elimination of several variables at once is allowed then the number of search paths increases still further.

Another problem arising within the general-to-specific methodology is that there is always the chance of making a Type II error during the process of the specification search and a variable which should be present in the final specification is eliminated at some stage. The result of this is typically that other variables, which ideally would have been excluded in the final specification, are retained as proxies for the missing variable. The resulting final specification is therefore over-parameterized. It is difficult to identify cases such as this from real world data where the investigator does not have the luxury of knowledge of the data generation process. However, it is straightforward to demonstrate this phenomenon using Monte Carlo analysis of artificial data sets.

In the early years of general-to-specific analysis it was argued that the only solution to the problems discussed above was to rely on the skill and knowledge of the investigator. For example, Gilbert (1986) argued the following:

> How should the econometrician set about discovering congruent simplifications of the general representation of the DGP …. Scientific discovery is necessarily an innovative and imaginative process, and cannot be automated. (p.295)

However, more recent research by Hoover and Perez (1999), Hendry and Krolzig (2001) and Krolzig and Hendry (2001) has suggested that automatic computer search algorithms can be effective in detecting a well specified econometric model using the now established 'general-to-specific' methodology. This has been facilitated by the introduction of the PC-GETS computer package which will automatically conduct a specification search to obtain the best data congruent model based on a given data set.

The purpose of this paper is to investigate the properties of a simple automatic search algorithm in uncovering a correctly specified parsimonious model from an initially overparameterized model. The algorithm works by estimating multiple search paths and choosing the final specification which minimizes the Schwartz criterion. This is compared with a naïve search algorithm in which the least significant variable in the regression is successively eliminated until all remaining variables are significant at a pre-determined level.

Methodology

The main problem encountered in conducting multiple search paths is the number of possible search paths that might be legitimately investigated. For example, consider a model in which the final specification involves twelve exclusion restrictions relative to the original model (not an unusual situation when working with quarterly data). In this case there are $12! = 479,001,600$ possible search paths involving the progressive elimination of one variable at each stage. Therefore, even with the power of modern computing, consideration of every possible search path is simply not an option. However, the situation is not as impossible as it may first appear. Many search paths will eventually converge on the same final specification and the problem is simply to ensure that enough are tried so as to maximize the chance of obtaining the correct specification. The pseudo-code below sets out the algorithm used in this research to achieve this.

```
FOR j = 1 to R, where R is a predetermined number of iterations.
    REPEAT UNTIL $\left|t_{\hat{\beta}_i}\right| > t_c$ where $t_c$ is a predetermined critical value for all $i = 1,..,N$ where $N$ is the number of variables included in the equation.

        Estimate equation.
        FOR each variable in the model examine $\left|t_{\hat{\beta}_i}\right|$. IF $\left|t_{\hat{\beta}_i}\right| < t_c$ AND
        $\gamma > 0.5$ where $\gamma$ is a random drawing from a uniform distribution with the
        interval $[0,1]$ THEN eliminate associated variable and re-estimate equation. ELSE IF $\gamma < 0.5$ THEN
        retain variable.
        IF $\left|t_{\hat{\beta}_i}\right| > t_c$ for all $i$ then STOP and record the variables included in the
        equation as well as the value of the Schwartz criterion. Otherwise go back to previous step.
    FOR j = 1 to R, compare the value of the Schwartz criterion for each final specification and choose the
    specification with the lowest value
```

The data generation process takes the form of the familiar partial adjustment model. This formulation is consistent with a cost minimization process in which agents minimize a quadratic cost function which includes costs of adjustment as well as costs of being away from equilibrium. The equation used to generate the data takes the form:

$$\begin{aligned} y_t &= 0.5x_t + 0.25y_{t-1} + u_t \\ t &: 1,...,T \end{aligned} \tag{2}$$

where $u_t : t = 1,...,T$ are iid standard normal random variables. The *x* variable is generated in two alternative ways. In the first $x_t : t = 1,...,T$ are also iid standard normal random variables with $\operatorname{cov}(x_t, u_t) = 0$. In the second, $x_t = x_{t-1} + \varepsilon_t$ where $\varepsilon_t : t = 1,...,T$ are iid standard normal variables with $\operatorname{cov}(x_t, \varepsilon_t) = 0$. Thus in case 1 the relationship is one between stationary variables while, in case 2, it is between $I(1)$ variables.

Using (2) to generate the data and (1) as the starting point for a specification search, the search algorithm discussed above is applied as well as the naïve search algorithm of simply eliminating the least significant variable at each stage of the search process. Ten thousand specification searches[2] are carried out using seeded pseudo-random numbers generated by the EViews regression package and the results of each search are classified according to the classification set out by Hoover and Perez (1999) as shown below:

A: Final model = True Model
B: True Model $\subset$ Final Model and $\hat{\sigma}_{Final} < \hat{\sigma}_{True}$
C: True Model $\subset$ Final Model and $\hat{\sigma}_{Final} > \hat{\sigma}_{True}$
D: True Model $\not\subset$ Final Model and $\hat{\sigma}_{Final} < \hat{\sigma}_{True}$
E: True Model $\not\subset$ Final Model and $\hat{\sigma}_{Final} > \hat{\sigma}_{True}$

Thus the final specification is classified as to whether it matches the true model (case A), contains all the variables included in the true model and has a lower standard error (case B), contains all the variables included in the true model but has a higher standard error (case C), omits at least one variable from the true model but has a lower standard error (case D) or omits at least one variable from the true model and has a higher standard error (case E).

Results

Table 1 presents the results for the multiple search path algorithm when the data are stationary. In all cases $R = 100$, that is 100 different specification searches were carried out and the equation with the lowest Schwartz criterion[3] was chosen. Examination of Table 1 indicates that both the sample size and the choice of critical value used in the specification search are important factors. If the sample size is small $T = 100$ then $t_c = t_c^{5\%}$ performs better than $t_c = t_c^{1\%}$ value in terms of identifying the true model more often (case A) and avoiding the elimination of variables that should be present in the true model (case E). However, as the sample size increases, this situation is reversed and in large samples with $T = 500$ then $t_c = t_c^{1\%}$ performs much better than $t_c = t_c^{5\%}$. Note that case C is never observed in any of the simulations carried out.

Does the multiple search path algorithm offer any gains over a naïve specification search? Examination of the results in Table 2 suggests that this is the case. In all cases the multiple search path algorithm identifies the true model more often. Moreover, as the sample size gets large, the frequency with which the multiple search path algorithm identifies the true model appears to be converging towards 100% with $t_c = t_c^{1\%}$. This is not the case for the naïve algorithm in which, with the same specification, the true model was identified in only 67.6% of the simulations.

Next, the effects of working with non-stationary data was considered. Here the x variable is generated as a random walk series with the implication that the y variable also contains a unit root. However, the specification of an equilibrium relationship between the variables ensures that they are co-integrated. This means that it is still reasonable to conduct a specification search in levels of the series even though individually each series contains a unit root. The results for the multiple search path algorithm are given in Table 3.

The results from Table 3 are very similar to those for non-stationary data shown in Table 1. The actual percentages differ slightly but the general pattern remains the same. If the sample size is small then $t_c = t_c^{5\%}$ performs better than $t_c = t_c^{1\%}$. However, as the sample size gets larger, this situation is reversed with case A converging towards 100% (when $t_c = t_c^{1\%}$) as the sample size becomes large.

Table 1: Multiple Search Paths General-To-Specific
($y_t = 0.5x_t + 0.25y_{t-1} + u_t$, x and u are independently generated *iid* processes)

	T=100		T=200		T=500	
Classification	5% Nominal Size	1% Nominal Size	5% Nominal Size	1% Nominal Size	5% Nominal Size	1% Nominal Size
A	52.4	48.2	76.4	83.3	80.9	93.0
B	15.2	4.0	17.3	5.7	19.1	6.9
C	0.0	0.0	0.0	0.0	0.0	0.0
D	14.7	10.8	2.8	2.5	0.0	0.0
E	17.7	37.0	3.5	8.5	0.0	0.0

Table 2: Single Search Path General-To-Specific
($y_t = 0.5x_t + 0.25y_{t-1} + u_t$, x and u are independently generated *iid* processes)

	T=100		T=200		T=500	
Classification	5% Nominal Size	1% Nominal Size	5% Nominal Size	1% Nominal Size	5% Nominal Size	1% Nominal Size
A	36.5	34.2	49.4	60.3	51.4	67.6
B	18.5	3.3	27.5	6.0	29.4	7.1
C	0.0	0.0	0.0	0.0	0.0	0.0
D	24.3	17.4	12.2	14.9	9.6	12.9
E	20.7	45.1	10.9	18.8	9.6	12.4

Finally, the multiple search path algorithm is contrasted with the naïve algorithm for the case of non-stationary data. The results for the naïve algorithm are shown in Table 4. These indicate that the naïve algorithm performs extremely badly when applied to non-stationary data. Case A is achieved in at best one quarter of the simulations, even with a large sample $T = 500$ and irrespective of the critical value employed. This suggests that the real value of a multiple search path algorithm may lie in its application to the modeling of non-stationary series. Since this is very often the case with econometric model building, it suggests that the approach may have considerable practical value.

Conclusion

In this article the use of a multiple search path algorithm for the general-to-specific approach to econometric analysis has been investigated. It has been shown that this algorithm has significant advantages over a naïve approach to specification searches. Moreover the relative advantage of this approach increases when dealing with non-stationary data. Since non-stationary data is the norm rather than the exception in econometric model building, it is arguable that a multiple search path approach offers real advantages to the applied econometrician.

Table 3: Multiple Search Paths General-To-Specific
($y_t = 0.5x_t + 0.25y_{t-1} + u_t$, *x* is a random walk process and *u* is a stationary *iid* process)

	T=100		T=200		T=500	
Classification	5% Nominal Size	1% Nominal Size	5% Nominal Size	1% Nominal Size	5% Nominal Size	1% Nominal Size
A	54.4	49.8	81.2	85.2	88.9	94.7
B	10.1	2.7	11.5	4.4	11.1	5.2
C	0.0	0.0	0.0	0.0	0.0	0.0
D	19.8	14.7	4.7	4.7	0.0	0.1
E	15.8	32.9	2.6	5.7	0.0	0.0

Table 4: Single Search Path General-To-Specific
($y_t = 0.5x_t + 0.25y_{t-1} + u_t$, *x* is a random walk process and *u* is a stationary *iid* process)

	T=100		T=200		T=500	
Classification	5% Nominal Size	1% Nominal Size	5% Nominal Size	1% Nominal Size	5% Nominal Size	1% Nominal Size
A	17.2	16.4	20.9	27.8	14.8	21.5
B	16.7	2.3	29.8	3.9	33.3	7.5
C	0.0	0.0	0.0	0.0	0.0	0.0
D	39.3	32.3	25.5	32.4	26.6	36.9
E	26.8	49.0	23.8	35.9	25.3	34.1

Notes

[1] The lag length in equation (1) is set at 4 for illustrative purposes only. This is often the case when dealing with quarterly data but alternative lag lengths are frequently employed for data with different frequencies.

[2] The specification searches were carried out using an EViews program which is available from the author on request.

[3] In fact, examination of the results indicates that many different specification search paths converge on the true model. The problem is not one of picking a single search path which gives the correct result but rather one of avoiding rogue search paths which give the wrong result.

References

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

Gilbert, C. L. (1986). Professor Hendry's econometric methodology, *Oxford Bulletin of Economics and Statistics*, *48*, 283-307.

Hendry, D.F., & Krolzig, H. (2001). New developments in automatic general-to-specific modelling. In Stigum, B. (Ed.) *Econometrics and the philosophy of economics*. Cambridge, MA: MIT Press.

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

Krolzig, H., & Hendry, D. (2001). Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control*, *25*, 831-866.