ASSESSING THE SIMULATION PERFORMANCES OF MULTIPLE MODEL SELECTION ALGORITHM

Norhayati Yusof[1], Suzilah Ismail[2], and T Zalizam T Muda[3]

[1]*Universiti Utara Malaysia, Malaysia, norhayati@uum.edu.my*
[2]*Universiti Utara Malaysia, Malaysia, halizus@uum.edu.my*
[3]*Universiti Utara Malaysia, Malaysia, zalizam@uum.edu.my*

ABSTRACT. The *Autometrics* is an algorithm for single equation model selection. It is a hybrid method which combines expanding and contracting search techniques. In this study, the algorithm is extended for multiple equations modelling known as *SURE-Autometrics*. The aim of this paper is to assess the performance of the extended algorithm using various simulation experiment conditions. The capability of the algorithm in finding the true specification of multiple models is measured by the percentage of simulation outcomes. Overall results show that the algorithm has performed well for a model with two equations. The findings also indicated that the number of variables in the true models affect the algorithm performances. Hence, this study suggests improvement on the algorithm development for future research.

Keywords: algorithm, SURE-Autometrics, Autometrics, seemingly unrelated regressions, feasible generalized least squares

INTRODUCTION

Generally, the modelling process is ambiguously explained by the expert modellers due to tacit knowledge. This knowledge only can be learned through research experiences which will be difficult for practitioners who are usually non-experts and no statistical background. Thus, an automatic modelling has become increasingly important tool for model building process. According to Hendry and Doornik (2014), the automatic modeller (i.e., algorithm) would be able to find a better model with additional information than the human modeller by discovering more than one possible models. Researchers also agreed with this new approach after revisited their previous studies and re-modelling the data (Doornik, 2009; Ericsson & Kamin, 2009; Hendry & Krolzig, 1999). Hence, this paper emphasizes on the algorithm that is developed for a seemingly unrelated regression equations (SURE) model. The development is based on a general-to-specific modelling approach using the search strategy adapted from *Autometrics* algorithm (Doornik, 2008, 2009). The *Autometrics* is not applicable for a multiple equations model such as SURE, as it is only suitable for single equation modelling. Hence, the algorithm is named *SURE-Autometrics* algorithm (Yusof & Ismail, 2014). Our focus is on the performance of the *SURE-Autometrics* with respect to its ability of finding the true model specification using Monte-Carlo simulation.

MULTIPLE MODELS SELECTION ALGORITHM

The properties and performances of the *Autometrics* were extensively reviewed in literatures (see among others, Castle, Doornik, & Hendry, 2011; Castle, Qin, & Robert Reed, 2013; Hendry & Doornik, 2014; Hoover & Perez, 2004). The algorithm was developed by combining the expanding and contracting search techniques. The expanding technique can also be called as specific-to-general, which starts from an empty model and adding variables until some termination criterion is satisfied. Regularly, the termination is based on a measure of penalized fit or marginal significance. In contrast, the contracting technique begins at the other end where variables are reduced from an initial model that comprised of all variables until a termination criterion is reached. Hence, the technique is generally known as general-to-specific (GETS).

The algorithm is fully described in Doornik and Hendry (2007), and Doornik (2008, 2009). Basically, it aims to improve the computational efficiency in searching the best model from the general unrestricted model (GUM). Thus, the algorithm uses a tree search method by implementing systematic strategies such as pruning, bunching and chopping in order to cut off irrelevant path and speed up the discovery of best model. Figure 1 shows example of the process where the GUM consists of four variables. The resulting tree is a unique representation of the model space. Precisely, all possible models would be estimated if moving from left to the right, and top to the bottom. Moreover, the search is done iteratively by using model contrast so it can seek more possible models. Therefore, the algorithm employs both expanding and contracting techniques.

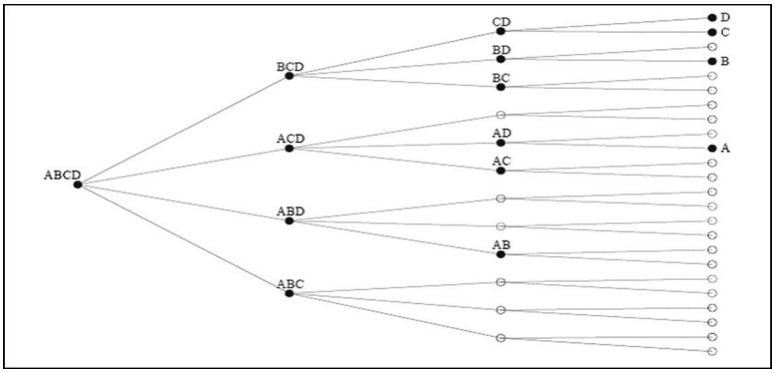

Figure 1. Search Strategy in Autometrics Algorithm

Meanwhile, the seemingly unrelated regression equations (SURE) model consists of several single equations that are related through the disturbances amongst equations. The series of equations are specified as follows,

$$\begin{aligned} y_{1t} &= \beta_{11}x_{1t,1} + \beta_{12}x_{1t,2} + \ldots + \beta_{1k_1}x_{1t,k_1} + u_{1t} \\ y_{2t} &= \beta_{21}x_{2t,1} + \beta_{22}x_{2t,2} + \ldots + \beta_{2k_1}x_{2t,k_2} + u_{2t} \\ &\vdots \\ y_{mt} &= \beta_{m1}x_{mt,1} + \beta_{m2}x_{mt,2} + \ldots + \beta_{mk_1}x_{mt,k_m} + u_{mt} \end{aligned} \tag{1}$$

which can be written in general form,

$$\underset{T\times 1}{\mathbf{y}_i} = \underset{T\times k_i}{\mathbf{X}_i}\ \underset{k_i\times 1}{\boldsymbol{\beta}_i} + \underset{T\times 1}{\mathbf{u}_i} \qquad i = 1,2,\ldots,m \tag{2}$$

where $\mathbf{y}_i$ is vector of T identically distributed observations for each random variable, $\mathbf{X}_i$ is a non-stochastic matrix of fixed variables of rank k_i, $\boldsymbol{\beta}_i$ is vector of unknown coefficients, and u_i is a vector of disturbances.

Therefore, estimation using feasible generalized least squares (FGLS) is more efficient than ordinary least squares (OLS) which is appropriate for single equation modelling. This model has wide range of applications mostly arise in economic, financial, and sociological modelling (Fildes, Wei, & Ismail, 2011; Srivastava & Giles, 1987; Zellner, 1962). It can also be applied to other areas such as human genetics (Verzilli, Stallard, & Whittaker, 2005) and behavioural science (Fernandez, Smith, & Wenger, 2007; Schwartz, 2006).

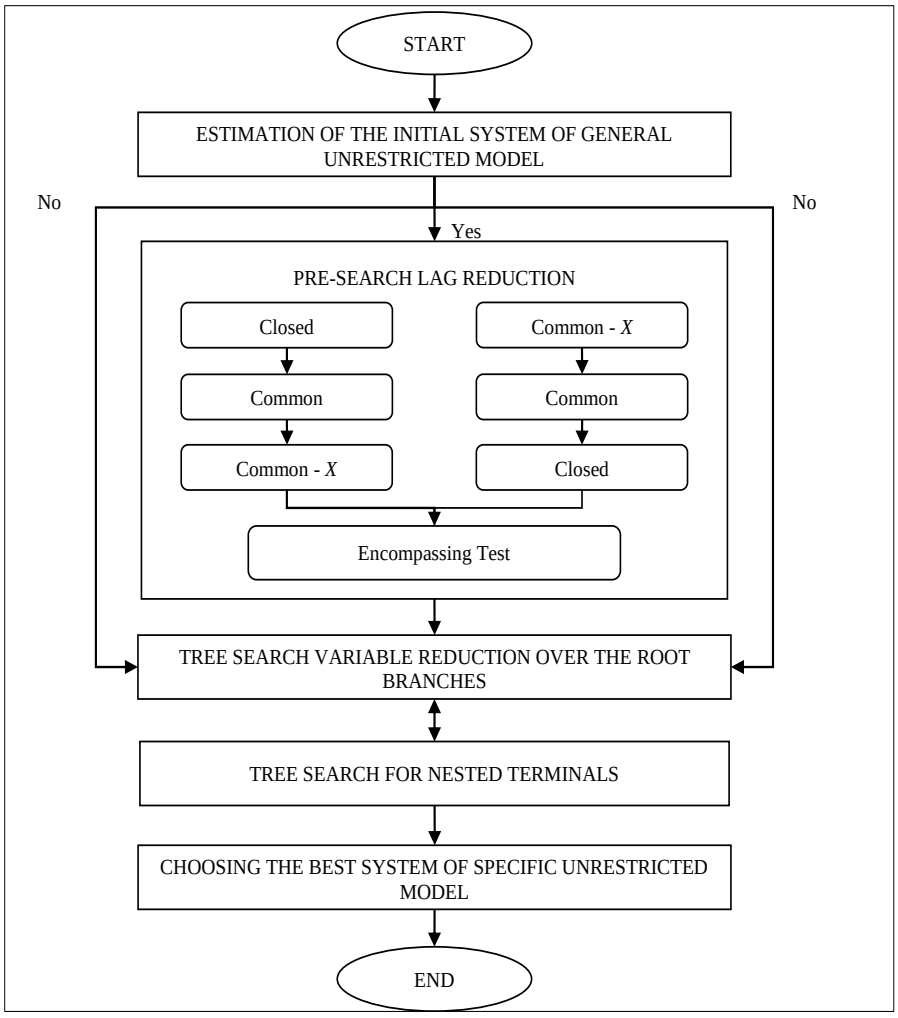

Figure 2. *SURE-Autometrics* Algorithm Framework

One of the properties in *Autometrics* is each single equation should be congruent. Hence, the *SURE-Autometrics* is developed by maintaining the search method in *Autometrics* and the

OLS method of estimation is replaced by FGLS method. It means that the model selection processes are done simultaneously. As indicated in Figure 2, the algorithm framework consists of five phases. The first phase deals with the formulation of an initial specification of the multiple equations of GUMs, and then followed by the second phase which focuses on pre-search reduction process. In this phase, the highest insignificant variables are deleted to reduce the models complexities in the previous phase. Third phase is the tree search procedure of finding the simplified GUMs. The fourth phase is to make sure the search is iterative which will result in multiple numbers of models that survived the reduction processes. These survived models known as terminal models. The final phase will deal with these terminals where an information criterion is used to select the final models.

SIMULATION ANALYSIS AND FINDINGS

In this paper, we demonstrate the performance of *SURE-Autometrics* for a model of two equations. The experimental frames require a formulation of several SURE models to be the true models specification. The true models were generated based on evaluation study of *Autometrics* (Doornik, 2009). The simulation analysis involves 120 combinations of experiment conditions as shown in Table 1, where each analysis has 100 simulated replications of each experiment that are designed in order to test the performances of *SURE-Autometrics*. The artificial data were simulated depending on five SURE models with true specification, three levels of correlation error, two sets of initial GUMs and two sample sizes. The table also shows two different setting of significance level in the algorithm.

Table 1. Summaries of Experimental Conditions

Condition of experiment	Level
1. True models specification	S1: $y_{1t} = 0.0230 + 0.0293\varepsilon_{1t}$ $y_{2t} = 0.0182 + 0.0240\varepsilon_{2t}$ S2: $y_{1t} = 0.0087 + 0.6170y_{1t-1} + 0.0229\varepsilon_{1t}$ $y_{2t} = 0.0058 + 0.6825y_{2t-1} + 0.0173\varepsilon_{2t}$ S3: $y_{1t} = 0.0078 + 0.6340y_{1t-1} + 0.3685x_{11t} - 0.3020x_{11t-1} + 0.0201\varepsilon_{1t}$ $y_{21t} = 0.0060 + 0.6915y_{2t-1} + 0.2811x_{12t} - 0.2224x_{12t-1} + 0.0151\varepsilon_{2t}$ S4: $y_{1t} = 0.0049 + 0.5966y_{1t-1} + 0.4820x_{21t} - 0.2072x_{21t-1} + 0.0221\varepsilon_{1t}$ $y_{21t} = 0.0028 + 0.6517y_{2t-1} + 0.1273x_{22t} + 0.1053x_{22t-1} + 0.0171\varepsilon_{2t}$ S5: $y_{1t} = 0.0049 + 0.6154y_{1t-1} + 0.3376x_{11t} - 0.2881x_{11t-1} + 0.3429x_{21t} - 0.1237x_{21t-1} + 0.0197\varepsilon_{1t}$ $y_{2t} = 0.0038 + 0.6720y_{2t-1} + 0.2742x_{12t} - 0.2268x_{12t-1} + 0.0278x_{22t} + 0.1377x_{22t-1} + 0.0149\varepsilon_{2t}$
2. Strength of correlation disturbances	Weak, $\rho = 0.2$, Moderate, $\rho = 0.6$, Strong, $\rho = 0.9$
3. Initial GUMs	Small, k = at most 18 irrelevant variables Large, k = at most 39 irrelevant variables
4. Sample sizes	Small, $n = 73$ Large, $n = 146$
5. Significance level	$\alpha = 0.05$ $\alpha = 0.01$

The first model, S1 can be referred as an empty model whereas S2 consists of the first lag of dependent variable. Model S3 and S4 are similar but have different independent variables. While the last model, S5 combines the variables from S3 and S4. Subsequently, numerous

irrelevant variables were added to these true models during the first phase of the algorithm. The performances were measured by calculating the percentages of the final models selected by *SURE-Autometrics* similar to the true models, since the data-generating process is known. Our aim is to have a substantial high percentage of these outcomes.

Overall results suggest that the performances are almost similar regardless of different level of correlation strength amongst the two equations. Hence, Table 2 summarizes the percentages of simulation outcomes for the strongest correlation disturbances. On average, the percentages were at least 80% for all except one experimental condition. The condition resulted in lowest percentages for both sample sizes that is below 71%. It was from S5 which received 34 irrelevant variables and the search procedure was administered at 1% level of significance. The results also revealed that 96% of the simulation of initial GUMs contained 18 irrelevant variables with large sample sizes and administered at 5% significance level able to achieve the S1 model. Additionally, percentages from a model who received small number of irrelevant variables were considerably higher compared to large number of irrelevant variables.

Table 2. Percentages of simulation outcomes

		$\alpha = 5\%$		$\alpha = 1\%$	
True SURE model	Sample sizes, *n*	*k* = at most 39 irrelevant variables	*k* = at most 18 irrelevant variables	*k* = at most 39 irrelevant variables	*k* = at most 18 irrelevant variables
S1	146	89	***96***	85	85
	73	88	91	83	83
S2	146	88	91	88	85
	73	85	89	80	84
S3	146	83	88	89	88
	73	82	84	82	85
S4	146	87	92	80	89
	73	84	89	79	84
S5	146	83	90	***70***	84
	73	79	87	***65***	80

CONCLUSION

In general, the *SURE-Autometrics* is able to achieve the true model specification with high percentages of simulation outcomes for multiple models with two equations. However, the number of variables in the true models appears to affect the algorithm performances. This can be seen from the results where there is high percentage in finding S1 as compared to low achievement in finding S5. The situation occurs due to assessment procedure. Since the model consists of multiple equations, the percentages are counted if all equations were similar to the true model. The final models may consists only one equation that is similar, and this state might be difficult for a true model such as S5 that have more independent variables compared to other models. Hence, a new assessment method can be developed in future study to overcome this problem. A parallel search strategy can also be implemented in the algorithm development as an attempt to improve the computational efficiency since it involves multiple equations.

ACKNOWLEDGMENTS

The authors are grateful for the financial support received from the Ministry of Education Malaysia under the Fundamental Research Grant Scheme (FRGS) and Universiti Utara Malaysia.

REFERENCES

Castle, J. L., Doornik, J. A., & Hendry, D. F. (2011). Evaluating automatic model selection. *Journal of Time Series Econometrics*, *3*(1), 33.

Castle, J. L., Qin, X., & Robert Reed, W. (2013). Using model selection algorithms to obtain reliable coefficient estimates. *Journal of Economic Surveys*, *27*(2), 269–296. doi:10.1111/j.1467-6419.2011.00704.x

Doornik, J. A. (2008). Encompassing and automatic model selection. *Oxford Bulletin of Economics and Statistics*, *70*, 915–925. doi:10.1111/j.1468-0084.2008.00536.x

Doornik, J. A. (2009). Autometrics. In J. L. Castle & N. Shephard (Eds.), *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry* (pp. 88–121). New York: Oxford University Press.

Doornik, J. A., & Hendry, D. F. (2007). *Empirical Econometric Modelling using PcGive 12: Volume 1*. London: Timberlake Consultants Ltd.

Ericsson, N. R., & Kamin, S. B. (2009). Constructive Data Mining: Modeling Argentine Broad Money Demand. In J. L. Castle & N. Shephard (Eds.), *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*. New York: Oxford University Press.

Fernandez, S., Smith, C. R., & Wenger, J. B. (2007). Employment, privatization, and managerial choice: Does contracting out reduce public sector employment? *Journal of Policy Analysis and Management*, *26*, 57–77.

Fildes, R., Wei, Y., & Ismail, S. (2011). Evaluating the forecasting performance of econometric models of air passenger traffic flows using multiple error measures. *International Journal of Forecasting*, *27*(3), 902–922. doi:10.1016/j.ijforecast.2009.06.002

Hendry, D. F., & Doornik, J. A. (2014). *Empirical Model Discovery and Theory Evaluation: Automatic Selection Methods in Econometrics*. MIT Press.

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

Hoover, K. D., & Perez, S. J. (2004). Truth and Robustness in Cross-country Growth Regressions. *Oxford Bulletin of Economics and Statistics*, *66*(5), 765–798.

Schwartz, J. (2006). Family structure as a source of female and male homicide in the United States. *Homicide Studies*, *10*(4), 253–278.

Srivastava, V. K., & Giles, D. E. A. (1987). Seemingly Unrelated Regression Equations Models: Estimation and Inference. *Statistics: Textbooks and Monographs*. New York: Marcel Dekker, Inc.

Verzilli, C. J., Stallard, N., & Whittaker, J. C. (2005). Bayesian modelling of multivariate quantitative traits using seemingly unrelated regressions. *Genetic Epidemiology*, *38*, 313–325. doi:10.1002/gepi.20072

Yusof, N., & Ismail, S. (2014). Lag variables reduction in multiple models selection algorithm. In *International Conference on the Analysis and Mathematical Applications in Engineering and Science* (pp. 169–173). Sarawak, Malaysia.

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association*, *57*(298), 348–368.