Far East Journal of Mathematical Sciences (FJMS)
© 2017 Pushpa Publishing House, Allahabad, India
http://www.pphmj.com
http://dx.doi.org/10.17654/MS101081655
Volume 101, Number 8, 2017, Pages 1655-1670 **ISSN: 0972-0871**

MANUAL AND AUTOMATED MODEL SELECTION PROCEDURES FOR SEEMINGLY UNRELATED REGRESSION EQUATIONS WITH DIFFERENT ESTIMATION METHODS

Nur Azulia Kamarudin and Suzilah Ismail

School of Quantitative Sciences
Universiti Utara Malaysia (UUM)
06010, Sintok, Kedah, Malaysia
e-mail: nurazulia@uum.edu.my
halizus@uum.edu.my

Abstract

Finding a good model can be a hefty task, especially when there are many predictors, thus providing many possible interactions. Effects and interactions in the model need to be looked into too. Therefore, model selection is one way to make this task simpler. Different strategies of selecting the right model had been proposed throughout the years. In this study, 13 selection procedures are compared in terms of their forecasting performances based on root mean square error (RMSE) and geometric root mean square error (GRMSE). Water quality index (WQI) data of a river in Malaysia has been analysed for two-equation and four-equation models of seemingly unrelated regression equations (SURE) model. The procedures were conducted either through manual or automated selection with ordinary least squares (OLS), feasible general least squares (FGLS) or maximum

Received: October 11, 2016; Revised: December 6, 2016; Accepted: December 27, 2016
2010 Mathematics Subject Classification: 62-07.
Keywords and phrases: automated model selection, seemingly unrelated regression equations, maximum likelihood estimation.

likelihood estimation (MLE) method for the final model. All automated manner procedures showed favourable results over manual selections. This proves that one person's knowledge only may not be sufficient to choose the best model. Out of the 13 procedures, SUREMLE-Autometrics has outperformed for both two- and four-equation models with achievement at rank 1 or 2 only. Therefore, MLE is considered as the best estimation method in this model setting.

1. Introduction

Model selection is a procedure to choose an acceptable model as an alternative to constructing a model randomly. The process comprises addition or exclusion of variables until some termination criterion is fulfilled. Misspecification will occur whenever the relevant variables are omitted from the model, the irrelevant included in the model, improper choices of functional form, and the model failed any diagnostic testing (Lv and Liu [10]). All of these specification errors will influence the properties of estimation technique, the quality of inferences, and the accuracy of the forecasting.

Practically, the modelling process begins with an estimation of a model that initially specified by the modeller. Then it is re-specified according to the results of hypothesis testing of single parameters to determine significant variables, or diagnostic checking of model's assumptions. Sometimes the modellers only implement the diagnostic tests for the initial model or the final model. This whole process basically can be done manually or automatically. In manual selection procedure, the decision on how the model should be re-specified is decided by the modellers. Magnus and Morgan [11] criticised that manual modelling may conclude to different end models as a result of difference in views and interests, added with numerous methods used and various ways of researching. Thus, all these will tend to have influence in deciding the right variables together with their measurements.

Difficulty in handling manual selections has led researches to move to a more efficient and faster manner by choosing the model automatically.

Technology evolution to produce more softwares in model selection enables different researches to obtain the same results by following the same algorithm for a given set of data. The work by Hoover and Perez [7] was a pioneer in automated model selection. Krolzig and Hendry [9] then continued Hoover and Perez's effort. They enhanced algorithm of data mining (Hendry and Krolzig [5, 6]) and produced *PcGets*, a program meant for empirical modeller. A more recent automatic model selection program, *Autometrics*, was then introduced. *Autometrics* is a successor of *PcGets* and has been described in Doornik [3]. Ericsson and Kamin [4] who used *PcGets* and *Autometrics*, discovered that the softwares have contributed in robustness and consumed less time compared to manual modelling. In the case of model selection, using *Autometrics* with relatively tight significance levels and bias correction contributed to a successful approach in selecting dynamic equations even when originating from very long lags to prevent excluding relevant variables or dynamics (Castle et al. [1]).

The automated selections are not only limited to single equation models as there are many settings in which single equation models apply to a group of related variables. In these contexts, it makes sense to consider the several models jointly and treat them as a system of equations. The word "*system*" means that the equations are to be considered collectively, instead of individually. Examples of this kind of system include simultaneous equations, vector auto-regression and seemingly unrelated regression equations (SURE) models. This system has the benefit of describing the dynamic composition of the actual procedure since it considers all relationships occurred, i.e., individual equation relationships and interaction of all the relationships. Consequently, further information may be gained from a set of equations compared to the sum of single equations. At the same time, this information can play a big role throughout the analysis, probably by providing more knowledge on the causal relationships and constructions included, apart from making more accurate forecasts (Pindyck and Rubinfeld [12]).

In SURE model particularly, various estimators have been proposed for the estimation of parameters, including the least square estimator and its variant and also iterative estimators. Nevertheless, least square method is widely used following its wide modifications and applications from the basic principles. One prominent estimator in SURE model is feasible generalized least square (FGLS) where the covariance of disturbances is unknown and replaced by a consistent estimator (Zellner [15] and Zellner and Theil [16]). Apart from it, maximum likelihood (ML) has been used to find system estimators (Chotikapanich et al. [2]). In order to implement MLE method in the context of SURE model, repeated measures analysis of two-stage general least squares estimation is used to obtain regression parameters and variance-covariance matrix. The ML estimators of the regression parameters can be obtained by performing the two-stage estimation iteratively. It is through iterative procedure that yields iterative FGLS (IFGLS). This is the type of MLE method which is being used here in this paper. With regard to decide on the best SURE model, this paper investigates both manual and automated selection approaches. Nonetheless, the final models would employ different estimators, which are OLS, FGLS and MLE. This is to find the most suitable estimation method for this model within the algorithms. Therefore, 13 model selection procedures have been put into tests and compared in this analysis.

2. Methods

2.1. Seemingly unrelated regression equations (SURE) model

The SURE model as suggested by Zellner [15], which consists of some equations, is a generalization of a linear regression model. Every equation in the model can be estimated individually albeit the error terms are assumed to be correlated across the equations. The reason is each equation stands by itself with dependent variable and probably different sets of regressors. Henceforth, these equations are ‘seemingly unrelated’. SURE modeling was introduced to serve the purpose of gaining efficiency in estimation by merging information on different equations, and to impose or to test restrictions that involve parameters in different equations.

Assume the system of equations:

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

which can be written in a general form

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

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

2.2. Feasible generalized least squares (FGLS) estimation

The SUR model is a generalization of multivariate regression using a vectorized parameter model. If the covariance matrix Ω is identified, then the model can be estimated with generalized least squares (GLS). Thus, the best linear unbiased estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y} \tag{3}$$

and the covariance matrix of these estimators is

$$V(\hat{\boldsymbol{\beta}}_{GLS}) = (\mathbf{X}'\Omega\mathbf{X})^{-1}. \tag{4}$$

In general, Ω and $\mathbf{u}_i$ are not known and so they have to be estimated. Every equation is estimated by OLS separately and the unbiased estimators for the coefficients of the ith equation are given by

$$\hat{\boldsymbol{\beta}}_{OLS_i} = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{y}_i, \quad i = 1, 2, ..., m \tag{5}$$

and

$$V(\hat{\boldsymbol{\beta}}_{OLS}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \tag{6}$$

The corresponding OLS residuals are given by

$$\hat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_i, \quad i = 1, 2, ..., m. \tag{7}$$

Let $\hat{\Omega}$ be a consistent estimator based on the residuals

$$\hat{\Omega} = \hat{\Sigma} \otimes I \tag{8}$$

$$\text{with } \hat{\Sigma} = [\hat{u}_i]'[\hat{u}_j], \quad i, j = 1, 2, ..., m \tag{9}$$

$$\text{or } \hat{\sigma}_{ij} = \frac{\hat{\mathbf{u}}_i'\hat{\mathbf{u}}_j}{T}, \quad i, j = 1, 2, ..., m, \tag{10}$$

where $\otimes$ denotes the Kronecker product and $\hat{\Sigma}$ is a $M \times M$ matrix based on single equation OLS residuals. Srivastava and Giles [13] referred this estimator as the seemingly unrelated restricted regression (SURR), which yields the following FGLS estimator of β:

$$\hat{\boldsymbol{\beta}}_{FGLS} = (\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}^{-1}\mathbf{y} \tag{11}$$

and the covariance matrix of the estimated parameters is

$$V(\hat{\boldsymbol{\beta}}_{FGLS}) = (\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}. \tag{12}$$

2.3. Maximum likelihood estimation (MLE)

Zellner's FGLS estimator of β as in Subsection 2.2 can be used for computing a new set of residuals leading to a new estimate of Ω. This new estimate is then employed in order to gain new estimates of the regression coefficients β, and so on. Backward and forward iterations between (10) and (11) will produce the ML estimators. Iteration is sustained until convergence is achieved at *k*th round. Let this estimator at the *k*th round be represented by $\hat{\boldsymbol{\beta}}^k_{FGLS}$ or $\hat{\boldsymbol{\beta}}_{ML}$. This method is also known as iterative FGLS (IFGLS):

$$\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}^k_{FGLS} = (\mathbf{X}'\hat{\Omega}^{k^{-1}}\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}^{k^{-1}}\mathbf{y} \tag{13}$$

and the covariance matrix of the estimated parameters is

$$V(\hat{\boldsymbol{\beta}}_{ML}) = V(\hat{\boldsymbol{\beta}}^k_{FGLS}) = (\mathbf{X}'\hat{\Omega}^{k^{-1}}\mathbf{X})^{-1}. \tag{14}$$

2.4. Model selection procedures

Models selection procedures in this paper are based on manual or automated selection with final model estimation methods either by using OLS, FGLS or MLE. *Mine*, *Stepwise* and *Autometrics* use OLS estimation, while *SURE-Autometrics* and *SUREMLE-Autometrics* utilize FGLS and IFGLS, respectively. The automated selection utilizes *Stepwise* or *Autometrics* algorithms.

2.4.1. *Mine*

The manual selections are primarily based on the *p*-values and the final decision to select the model depends on individual's knowledge. In this analysis, *p*-values based on 5% significance level are set to determine the significant or insignificant variables from the general unrestricted models. Variables with high insignificant *p*-values are removed from the model beginning with the highest one. The variable is somehow ignored if the correlation and insignificant *p*-values are also high. Once the variable is eliminated, the standard error is checked for any increase. If exists, then the variable is kept. The variables would be removed as a group if more than one variable are highly insignificant as well as weak correlations persist. The selected model then must succeed for all diagnostic tests. In this paper, this manual selection is also named as *Mine*.

2.4.2. *Stepwise*

Stepwise regression is a well-known algorithm in choosing predictive variables through its three main approaches: (i) forward selection, (ii) backward selection and (iii) bidirectional elimination. Basically, stepwise regression does several multiple regressions. The weakest correlated variable will be eliminated during each regression. Finally, only the related variables that clarify the distribution test are left in the model.

2.4.3. *Autometrics*

On the other hand, *Autometrics* implements a tree search systematically to steer the whole model space. Some strategies such as pruning, bunching, and chopping are executed to drop irrelevant paths and accelerate the

process. *Autometrics* does not only cater for general-to-specific (GETS) approach, but also handles the specific-to-general, which is a contrary approach of GETS. Nonetheless, *Autometrics* only performs individual selection for single model by OLS estimation method. Detailed explanation on the following stages of *Autometrics* can be found in (Doornik [3]):

Stage 1. Specification of initial GUM

The algorithm begins with Stage 1 by specifying the initial generalized unrestricted model (GUM). The IFGLS estimation of initial GUM initializes the whole search procedure. Every equation in SURE model is also estimated by OLS estimation method separately and tested for any misspecifications using the diagnostic tests to check on the contemporaneous correlation errors, normality errors, parameters constancy, autocorrelation, unconditional homoscedasticity and conditional homoscedasticity along with the independence test.

Stage 2. Pre-search reduction

This pre-search reduction is where the algorithm can still operate with or without it. It is added to reduce computational effort since highly insignificant variables are deleted. This stage consists of: (i) encompassing tests to ensure that the simplified model is a valid reduction of the initial system of GUM, (ii) closed lag reduction to test a group of lags from the largest lag downwards and discontinue once a lag cannot be deleted, and (iii) common lag reduction to test all the remaining lags starting from the least significant.

Stage 3. Tree search method

In this stage, the whole spaces of models are generated by the variables in the initial model. Four reduction principles are involved here as below:

(i) Pruning is done when one variable is considered for deletion.

(ii) Bunching is implemented when variables are grouped for deletion instead of one variable at one time.

(iii) Chopping happens when a highly insignificant bunch is eliminated permanently from the search procedure.

(iv) Model contrast principle enables modeler to find out minimum bunch along the path that must be deleted to give a different model.

Stage 4. Tiebreaker

Finally, the tiebreaker stage is applied when there are multiple terminal candidate models to choose the final model.

2.4.4. *SURE-Autometrics*

Yusof and Ismail [14] initiated *SURE-Autometrics*, which is an algorithm for automatic model selection procedures focusing on the multiple equations model of SURE. Multiple equations selection is conducted simultaneously with estimation of FGLS method throughout the process. This algorithm accepts similar operation as its 'parent' algorithm, *Autometrics*.

2.4.5. *SUREMLE-Autometrics*

SUREMLE-Autometrics is proposed here as an alternative in choosing the 'best' model. The development of the *SUREMLE-Autometrics* still adopts the original *SURE-Autometrics* where the similar four stages are involved. As opposed to FGLS estimation method in most SURE models, the original *SURE-Autometrics* algorithm is altered with the use of IFGLS method. This means the new algorithm concentrates on the system estimation which now employs an ML approach, as described in Subsection 2.3. This estimation method is implemented in each stage separately.

Apart from the mentioned selection procedures, there are other selections taken into account in this study for SURE model, which can be classified into their selection manners and the method used for estimation in the final models. Consequently, there are 13 model selection procedures altogether. All procedures involved are summarized here:

(1) *Autometrics* and *Stepwise* are the algorithms for single equation model. Since the model has multiple equations, each is estimated using OLS and individually selected for multiple times. Model selection through *Autometrics* is applied using *PcGive* software and *Stepwise* is employed by IBM SPSS Statistics 21.

(2) *Autometrics-SURE* and *Stepwise-SURE* are procedures that used previous algorithms in the model selection where each equation separately selected with OLS estimation method. However, the final model is estimated using FGLS. Meanwhile, *Autometrics-SUREMLE* and *Stepwise-SUREMLE* estimated the final model using MLE.

(3) *SURE-Autometrics* and *SUREMLE-Autometrics* are the algorithms for automatic model selection procedures focusing on the multiple equations model. The selection of *SURE-Autometrics* is implemented simultaneously with FGLS method of estimation, whereas MLE is embedded in *SUREMLE-Autometrics*.

(4) *Mine*, *Mine-SURE*, *Mine-SUREMLE*, *SURE-Mine* and *SUREMLE-Mine* are non-algorithm model selection procedures which means the selection is a process of trial and error based on personal judgement. The *Mine* and *Mine-SURE* select the equation by equation with OLS estimators. FGLS is used to estimate the final model of *Mine-SURE*, whereas MLE is for *Mine-SUREMLE* estimation. On the other side, *SURE-Mine* used FGLS while *SUREMLE-Mine* employed MLE as methods of estimation with the inspection of variables simultaneously within the model according to the rules above.

3. Results and Discussions

Weekly data of WQI of a river in Malaysia from years 2012 and 2013 has been used as the dependent variable (Y_{it}) in this study. The independent variables (parameters) are dissolved oxygen (DO) (% saturation) (x_{i1t}), dissolved oxygen (DO) (mg/L) (x_{i2t}), biochemical oxygen demand (BOD)

(x_{i3t}), chemical oxygen demand (COD) (x_{i4t}), suspended solids (SS) (x_{i5t}), pH (x_{i6t}), and ammoniacal nitrogen $(NH_3\text{-N})$ (x_{i7t}). These variables will be converted into the sub-indices, which are named SIDO, SIBOD, SICOD, SIAN, SISS and SIPH. These data sets were collected from four sampling stations, namely S6, S7, S8 and S25. Analyses were done on model with four equations and model with two equations. Four sampling stations were represented by four-equation model, whereas two sampling stations indicated two-equation model. Station S6 recorded the highest standard error (SE), followed by station S25. Stations S6 and S25 were then removed for the two-equation model analysis as a result of these high SEs. This is because the values of WQI of these stations are lower compared to the other two, suggesting that the waters around the stations could be more polluted. The nearby free trade industrial zone area could have caused this situation as a result of over-dumped wastes.

This study used analysis in Ismail [8] as a guide in formulating the initial GUMS. The initial GUMS contained 17 explanatory variables: three lags of the dependent variables, seven independent variables and one lag of each independent variable. This is consistent as in autoregressive distributed lag (ADL) model. The first 63 data are used for model estimation and the last five are for model evaluation (i.e., recursive evaluation), which is based on RMSE and GRMSE.

Table 1 lists out model selection procedures according to their selection manners and final model estimation methods. Meanwhile, Tables 2 and 3 exhibit the evaluation results for one, two and three steps ahead forecast of four-equation model for RMSE and GRMSE, respectively. For two-equation model, similar results can be found in Tables 4 and 5. The values are ranked from 1 (the smallest) to 13 (the largest) to indicate forecasting performance.

Table 1. Model selection procedures

Model selection procedures	Selection manners		Final model estimation methods		
	Automated	Manual	OLS	FGLS	MLE
1. *SUREMLE-Autometrics*	/				/
2. *SURE-Autometrics*	/			/	
3. *Autometrics-SUREMLE*	/				/
4. *Autometrics-SURE*	/			/	
5. *Autometrics*	/		/		
6. *Stepwise-SUREMLE*	/				/
7. *Stepwise-SURE*	/			/	
8. *Stepwise*	/		/		
9. *SUREMLE-Mine*		/			/
10. *SURE-Mine*		/		/	
11. *Mine-SUREMLE*		/			/
12. *Mine-SURE*		/		/	
13. *Mine*		/	/		

Table 2. Forecasting performances based on RMSE for four-equation model

Model selection procedures	One-step		Two-steps		Three-steps	
	RMSE	Rank	RMSE	Rank	RMSE	Rank
1. *SUREMLE-Autometrics*	**1.9711**	**1**	**2.1091**	**1**	**2.0877**	**2**
2. *SURE-Autometrics*	2.0653	2	2.1216	2	1.8978	1
3. *Autometrics-SUREMLE*	2.0924	3	2.2155	3	2.1904	4
4. *Autometrics-SURE*	2.0976	4	2.2298	4	2.1794	3
5. *Autometrics*	2.1541	5	2.3054	8	2.2227	5
6. *Stepwise-SUREMLE*	2.2003	8	2.2935	5	2.4210	8
7. *Stepwise-SURE*	2.1829	7	2.2994	6	2.4100	7
8. *Stepwise*	2.1586	6	2.3014	7	2.2334	6
9. *SUREMLE-Mine*	6.2093	13	6.7219	12	7.6793	13
10. *SURE-Mine*	6.1463	11	6.662	10	7.6085	11
11. *Mine-SUREMLE*	6.1490	12	6.7222	13	7.6174	12
12. *Mine-SURE*	6.1100	10	6.6919	11	7.5901	10
13. *Mine*	5.8545	9	6.377	9	7.2203	9

Table 3. Forecasting performances based on GRMSE for four-equation model

Model selection procedures	One-step		Two-steps		Three-steps	
	GRMSE	Rank	GRMSE	Rank	GRMSE	Rank
1. *SUREMLE-Autometrics*	**1.5289**	**1**	**1.5688**	**2**	**1.4889**	**1**
2. *SURE-Autometrics*	1.6192	2	1.6614	5	1.6298	3
3. *Autometrics-SUREMLE*	1.725	5	1.6213	3	1.7231	5
4. *Autometrics-SURE*	1.7913	7	1.6391	4	1.7073	4
5. *Autometrics*	1.6879	4	1.5031	1	1.5623	2
6. *Stepwise-SUREMLE*	1.6485	3	1.6665	6	2.1137	7
7. *Stepwise-SURE*	1.7511	6	1.7572	7	2.1175	8
8. *Stepwise*	1.8593	8	1.9051	8	1.9873	6
9. *SUREMLE-Mine*	4.6312	13	4.8897	13	6.8717	13
10. *SURE-Mine*	4.4602	12	4.7051	12	6.7971	12
11. *Mine-SUREMLE*	4.3046	11	4.5097	11	6.6045	11
12. *Mine-SURE*	4.2184	10	4.3786	10	6.4620	10
13. *Mine*	3.6734	9	3.8663	9	5.8073	9

Table 4. Forecasting performances based on RMSE for two-equation model

Model selection procedures	One-step		Two-steps		Three-steps	
	RMSE	Rank	RMSE	Rank	RMSE	Rank
1. *SUREMLE-Autometrics*	**1.5686**	**1**	**1.7632**	**2**	**1.9941**	**2**
2. *SURE-Autometrics*	1.6902	2	1.7318	1	1.8903	1
3. *Autometrics-SUREMLE*	1.7061	5	1.8655	5	2.1343	5
4. *Autometrics-SURE*	1.698	3	1.8551	4	2.1069	4
5. *Autometrics*	1.7013	4	1.8421	3	2.048	3
6. *Stepwise-SUREMLE*	1.7845	8	1.9615	8	2.2489	8
7. *Stepwise-SURE*	1.7667	7	1.9388	7	2.2142	7
8. *Stepwise*	1.7467	6	1.8981	6	2.1387	6
9. *SUREMLE-Mine*	3.7173	9	4.1178	9	4.749	9
10. *SURE-Mine*	5.9088	13	6.4699	13	7.4393	13
11. *Mine-SUREMLE*	5.6804	12	6.2542	12	7.1988	12
12. *Mine-SURE*	5.4809	11	6.0668	11	6.9918	11
13. *Mine*	5.0674	10	5.6629	10	6.535	10

Table 5. Forecasting performances based on GRMSE for two-equation model

Model selection procedures	One-step		Two-steps		Three-steps	
	GRMSE	Rank	GRMSE	Rank	GRMSE	Rank
1. *SUREMLE-Autometrics*	**1.0869**	**1**	**1.2508**	**2**	**1.6514**	**1**
2. *SURE-Autometrics*	1.4108	8	1.4391	6	1.679	2
3. *Autometrics-SUREMLE*	1.2475	5	1.3257	3	1.8318	5
4. *Autometrics-SURE*	1.2874	7	1.4083	5	1.8082	4
5. *Autometrics*	1.2793	6	1.5031	8	1.7483	3
6. *Stepwise-SUREMLE*	1.1451	2	1.2191	1	1.9614	8
7. *Stepwise-SURE*	1.2159	3	1.3323	4	1.9313	7
8. *Stepwise*	1.2421	4	1.4517	7	1.8586	6
9. *SUREMLE-Mine*	2.1547	9	2.1736	9	3.7733	9
10. *SURE-Mine*	2.9874	13	2.8238	11	5.512	13
11. *Mine-SUREMLE*	2.7599	11	2.6017	10	5.313	12
12. *Mine-SURE*	2.8979	12	3.0103	12	5.1832	11
13. *Mine*	2.6507	10	3.0378	13	4.8197	10

Results for both two- and four-equation models have been consistent with *SUREMLE-Autometrics* which showed high performance compared to other procedures. For the four-equation model, the *SUREMLE-Autometrics* was ranked at 1 for all one, two and three steps-ahead forecasts except for RMSE of three-steps and GRMSE of two-steps. These comparable results were also found for two-equation model. *SUREMLE-Autometrics* again gained top spot for RMSE of one-step and GRMSE of one-step and three-step forecasts. Regardless the change in the number of equations, *SUREMLE-Autometrics*' performance is not much affected in terms on its ability to forecast. Iterative estimation is seen to give advantage to this algorithm.

4. Conclusion

Overall outcomes displayed excellent performance for automated selection. All procedures under this selection manner were positioned from 1 to 8, contrast to the manual selection which failed to show any good performance for all conditions here. Hence, automated model selection using

algorithm not only revealed its superiority through its easiness in handling selections, but also from the high performance shown. In addition, since MLE was embedded in *SUREMLE-Autometrics*, MLE is deemed as the best estimation method in this model setting. This strategy by executing simultaneous selection with MLE method is therefore proven to outclass in this analysis. Therefore, it is recommended that analysis of automated SURE model selection should embed more other MLE methods besides IFGLS, including within *Autometrics* algorithm.

Acknowledgements

The authors would like to gratefully acknowledge the Department of Environment (DOE), Malaysia for providing the data.

The authors also thank the anonymous referees for their valuable suggestions improving the presentation of the manuscript.

References

[1] J. L. Castle, J. Doornik and D. F. Hendry, Evaluating automatic model selection, Journal of Time Series Econometrics 3(1) (2011), 1-33.

[2] D. Chotikapanich, W. E. Griffiths and C. L. Skeels, Sample size requirements for estimation in SUR models, No. 794, 2001, 18 pages.

[3] J. Doornik, Autometrics, J. Castle and N. Shephard, eds., The Methodology and Practice of Econometrics, Oxford University Press, Oxford, 2009.

[4] N. R. Ericsson and S. B. Kamin, Constructive Data Mining: Modeling Argentine Broad Money Demand, J. L. Castle and N. Shephard, eds., The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry, Oxford University Press, Oxford, 2009.

[5] D. F. Hendry and H. Krolzig, Improving on "Data mining reconsidered" by K. D. Hoover and S. J. Perez, Econom. J. 2 (1999), 202-219.

[6] D. F. Hendry and H. Krolzig, The properties of automatic GETS modeling, Economic J. 115(502) (2005), C32-C61.

[7] K. Hoover and S. Perez, Data mining reconsidered: encompassing and the general-to-specific approach to specification search, Econom. J. 2 (1999), 167-191.

[8] S. Ismail, Algorithmic Approaches to Multiple Time Series Forecasting, University of Lancaster, 2005.

[9] H. Krolzig and D. F. Hendry, Computer automation of general-to-specific model selection procedures, J. Econom. Dynam. Control 25 (2001), 831-866.

[10] J. Lv and J. S. Liu, Model selection principles in misspecified models, J. R. Stat. Soc. Ser. B Stat. Methodol. 76(1) (2014), 141-167.

[11] J. R. Magnus and M. S. Morgan, Methodology and Tacit Knowledge, J. Wiley, New York, 1999.

[12] R. S. Pindyck and D. L. Rubinfeld, Econometric Models and Economic Forecasts, 4th ed., Irwin/McGraw-Hill, Boston, Massachusetts, 1998.

[13] V. K. Srivastava and D. E. A. Giles, eds., Seemingly Unrelated Regression Equations Models, Marcel Dekker, Inc., New York, NY, USA, 1987.

[14] N. Yusof and S. Ismail, Independence test in SURE-Autometrics algorithm, International Symposium on Forecasting (ISF), Prague, 2011.

[15] A. Zellner, Estimators for seemingly unrelated regression equations: some exact finite sample results, J. Amer. Statist. Assoc. 58(304) (1963), 977-992.

[16] A. Zellner and H. Theil, Three-stage least squares: simultaneous estimation of simultaneous equations, Econometrica 30(1) (1962), 54-78.

View publication stats