Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 12, Number 2 (2016), pp. 1821-1829

http://www.ripublication.com/gjpam.htm

Model Selection Approaches of Water Quality Index Data

Nur Azulia Kamarudin
School of Quantitative Sciences,
Universiti Utara Malaysia (UUM)
06010, Sintok, Malaysia.

Suzilah Ismail
School of Quantitative Sciences,
Universiti Utara Malaysia (UUM)
06010, Sintok, Malaysia.

Abstract

Automatic model selection by using algorithm can avoid huge variability in model specification process compared to manual selection. With the employment of algorithm, the right model selected is then also used for forecasting purposes. In order to select the best model, it is vital to ensure that proper estimation method is chosen in the modelling process. Different estimators have been proposed for the estimation of parameters of a model, including the least square and iterative estimators. This study aims to evaluate the forecasting performances of two algorithms on water quality index (WQI) of a river in Malaysia based on root mean square error (RMSE) and geometric root mean square error (GRMSE). Feasible generalised least squares (FGLS) and iterative maximum likelihood (ML) estimation methods are used in the algorithms, respectively. The results showed that *SUREMLE-Autometrics* has surpassed *SURE-Autometrics*; another simultaneous selection procedure of multiple-equation models. Two individual selections, namely *Autometrics-SUREMLE* and *Autometrics-SURE*, though showed consistency only for GRMSE. All in all, ML estimation is a more appropriate method to be employed in this seemingly unrelated regression equations (SURE) model selection.

AMS subject classification:
Keywords: Automated model selection, multiple equations, maximum likelihood estimation.

1. Introduction

Model selection is a procedure of choosing an adequate model instead of making a random choice of model. It involves the inclusion or removal of variables until some termination criterion is satisfied. Model selection can possibly be done either manually or automatically by using an algorithm. Manual selection however is regarded as uncertain and may even conclude to different end models as a result of difference in views and interests, numerous methods used and various ways of researching [7]. Granger and Hendry [2] believed automated approach can be a formal way to overcome the problem. Algorithm is developed to provide a rule in guiding the researchers to formulate and testing the model while obtaining the same results by following the same algorithm for a given data set. By doing this, it facilitates in diminishing the role of tacit knowledge as well as labour saving through elimination of computational burden especially if there are many potential candidate variables [1].

Hoover and Perez's [3] automated model-selection algorithm had been continued by Krolzig and Hendry [6], who developed a computer program *PcGets*. From there, Doornik and Hendry [1] employed a third-generation algorithm known as *Autometrics*. *PcGets* and *Autometrics* adopted general-to-specific (GETS) concept, where a general model comprises of all variables and is reduced to a simpler model by removing variables with coefficients that are not statistically significant.

With the advancement of *PcGets* and *Autometrics* in automated model selection, Ismail [4] and Yusof and Ismail [8] have come out with automated model selection algorithms for multiple equations, namely SURE-*PcGets* and *SURE-Autometrics* respectively. These two algorithms have exploited the advantages of automated model selection and system of equations in Seemingly Unrelated Regression Equations (SURE) model concept. Nevertheless, the algorithms have only used FGLS estimator in the modelling process. However, Kmenta and Gilbert [5] showed that FGLS is not always efficient in small samples. ML estimator resulted in smaller variance compared to FGLS if there is high correlation between disturbances, but low correlation between the explanatory variables.

There has been insufficient evidence to show any attempts have been made so far to incorporate other estimation methods, such as ML estimation method. It is one of the most attractive estimators due to their asymptotic properties. Hence, it would be interesting to add new estimation method for multiple equations model selection within the GETS concept. In particular, a modification of seemingly unrelated regression automated model selection algorithm, namely *SURE-Autometrics* algorithm is set to take place by employing maximum likelihood estimation. It is then known as *SUREMLE-Autometrics*. All in all, this study is directed towards measuring differences of automated selection procedures between two different approaches, which are FGLS and ML estimation methods through its forecasting performance with the use of error measures. An empirical data set of water quality index (WQI) is analysed here.

2. Methods

2.1. Seemingly Unrelated Regression Equations (SURE)

SURE are multivariate regression models with dependent variables that follow a joint Gaussian distribution. Usually different regressions contain different independent variables and seem "unrelated". However, due to the correlated response variables the regressions are only "seemingly unrelated" and contain valuable information about each other [9]. The SUR model as suggested by Arnold Zellner in 1962, which comprises some equations, is a generalization of a linear regression model. Even though the error terms are assumed to be correlated across the equations, every equation can be estimated individually. This is because each equation stands on its own with dependent variable and possibly different sets of regressors. Hence, these equations are 'seemingly unrelated'.

The motivations of the SUR modelling are to gain efficiency in estimation by combining information on different equations, and to impose or to test restrictions that involve parameters in different equations. Suppose the series of equations are,

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

which can be written in general form,

$$\underset{T\times 1}{y_i} = \underset{T\times k_i}{X_i}\ \underset{k_i\times 1}{\beta_i} + \underset{T\times 1}{u_i}\ , i = 1, 2, \ldots, m \tag{2}$$

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

2.2. Feasible Generalised Least Squares (FGLS)

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

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

and the covariance matrices of these estimators are,

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

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 i^{th} equation in (2.2) are given by,

$$\hat{\beta}_{OLS_i} = \left(X_i'X_i\right)^{-1} X_i'y_i, i = 1, 2, \ldots, m \tag{5}$$

and

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

The corresponding OLS residuals are given by

$$\hat{u}_i = y_i - X_i\hat{\beta}_i, i = 1, 2, \ldots, m \tag{7}$$

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

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

with

$$\widehat{\Sigma} = [\hat{u}_i]'[\hat{u}_j]\ i, j = 1, 2, \ldots, m \tag{9}$$

or

$$\hat{\sigma}_{ij} = \frac{\hat{u}_i'\hat{u}_j}{T} i, j = 1, 2, \ldots, m \tag{10}$$

where $\otimes$ denotes the Kronecker product and $\widehat{\Sigma}$ is a MxM matrix based on single equation OLS residuals. Srivastava and Giles (1987) referred this estimator as the seemingly unrelated restricted regression (SURR), which yields the following FGLS estimator of β,

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

and the covariance matrix of the estimated parameters is,

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

2.3. Maximum Likelihood (ML)

Zellner's FGLS estimator of as in Section 2.2 can be used for calculating a new set of residuals leading to a new estimate of , which later employed for obtaining new estimates of the regression coefficients, and so on. ML estimators are obtained by iterating back and forth between (10) and (11). Iteration is continued until convergence is achieved at k^{th} round. Let this estimator at the k^{th} round be denoted by β^k_{FGLS} or β_{ML}. This method is also known as iterative FGLS (IFGLS).

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

and the covariance matrix of the estimated parameters is

$$V(\beta_{ML}) = V(\beta^k_{FGLS}) = (X'\Omega^{k^{-1}}X)^{-1} \tag{14}$$

2.4. Model Selection Algorithms

SURE-Autometrics as initiated by Yusof and Ismail [8] is an algorithm for automatic model selection procedures focussing on the multiple equations model of SURE. This algorithm adopts similar operation as its 'parent' algorithm, *Autometrics*, where a tree search systematically steers the whole model space. Still, it is computationally incompetent to find all possible models. Therefore, some strategies such as pruning, bunching, and chopping are executed to drop irrelevant paths and accelerate the process. *Autometrics* does not only cater for GETS approach, but also handles the specific-to-general, which is a reverse approach of GETS. Nonetheless, *Autometrics* only performs individual selection for single model by Ordinary Least Squares (OLS) estimation method. Thus, *Autometrics* has been extended to *SURE-Autometrics* to conduct multiple equations selection simultaneously with estimation of FGLS method throughout the process. Meanwhile, *Autometrics-SURE* is a procedure that uses *Autometrics* in model selection where each equation is separately selected with OLS estimation. However, the final model is estimated using FGLS.

In addition, *SUREMLE-Autometrics* and *Autometrics-SUREMLE* are proposed in this paper as alternatives in choosing the 'best' model. *SUREMLE-Autometrics* and *Autometrics-SUREMLE* are modification of *SURE-Autometrics* and *Autometrics-SURE*, respectively. The development of the *SUREMLE-Autometrics* still adopts the original *SURE-Autometrics* where four main stages are involved. Briefly, Stage 1 explains the initial specification of GUM and Stage 2 handles pre-search reduction of the general model. Meanwhile, Stage 3 is on tree search process and Stage 4 is where the final model selection takes place. The modification of this algorithm concentrates on the system estimation which utilizes a ML approach as described in Section 2.3. On the other hand, *Autometrics-SUREMLE* is only different from *Autometrics-SURE* at the final model estimation method. Instead of using FGLS, this algorithm uses the same ML method as in *SUREMLE-Autometrics*. Consequently, there are four model selection procedures taken into account in this study while giving emphasis on SURE model.

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_3N) (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-equation model indicated four sampling stations, while two-equation model represented two sampling stations. This study used Ismail (2005) as a guideline in formulating the initial GUMS. The initial GUMS contained 31 explanatory

Table 1: Model selection procedures and sampling stations

Procedures		**S6**	**S7**	**S8**	**S25**
SUREMLE-Autometrics	$\overline{R}^2$	0.949	0.956	0.948	0.945
	SE	1.945	1.050	1.360	1.767
SURE-Autometrics	$\overline{R}^2$	0.953	0.961	0.953	0.947
	SE	1.870	0.989	1.295	1.737
*Autometrics-*SUREMLE-	$\overline{R}^2$	0.954	0.967	0.947	0.950
	SE	1.910	0.975	1.516	1.657
Autometrics-SURE	$\overline{R}^2$	0.955	0.969	0.949	0.952
	SE	1.901	0.940	1.487	1.621

variables: three lags of the dependent variables (Δy_{it}), seven independent variables (Δx_{ikt}) and three lags of each Δx_{ikt}. This is consistent as in Autoregressive Distributed Lag (ADL) model. The first 63 data is used for model estimation and the last five is for model evaluation (i.e. recursive evaluation), which is based on RMSE and GRMSE.

Table 1 presents the adjusted R square ($\bar{R}^2$) and standard error (SE) based on different model selection algorithm. Station S7 has the highest ($\bar{R}^2$), while stations S8 and S25 have the lowest values. Meanwhile, the SE is recorded highest for station S6, followed by station S25. Due to these high SEs, stations S6 and S25 were removed for the two-equation model analysis. This is because the values of WQI of these stations are lower compared to the other two, meaning that the waters around the stations are more polluted. Waste disposals coming from the nearby Free Trade Industrial Zone area could have contributed to this situation. Table 2 and 3 exhibit the evaluation results for one, two and three steps ahead forecast of four-equation model, while Table 4 and 5 show similar findings for two-equation model. The values are ranked from 1 (the smallest) to 4 (the largest) to indicate forecasting performance.

In the analysis of four equations, all selection procedures have maintained steady performance for both RMSE and GRMSE. *SUREMLE-Autometrics* has been ranked at 1 for all one, two and three step-ahead forecasts. In the meantime, mixed results are seen for two-equation model. Nonetheless, *SUREMLE-Autometrics* still tops the ranks in general, regardless of number of equations. With reference to individual selection, *Autometrics-SUREMLE* and *Autometrics-SURE* were unable to reach rank 1 at any settings. Unlike their counterparts, they have managed to show consistency only for GRMSE. The overall outcome reveals that *SUREMLE-Autometrics* is the 'best' approach with rank 1 in all conditions, except for GRMSE of one-step forecast in two-equation model with rank 2. By adopting ML estimation method in the algorithm, the 'best' model can be chosen

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**
SUREMLE-Autometrics	1.24	1	1.42	1	1.52	1
SURE-Autometrics	2.16	4	2.27	4	2.11	4
Autometrics-SUREMLE	1.75	3	1.88	2	2.00	3
Autometrics-SURE	1.74	2	1.90	3	1.72	2

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**
SUREMLE-Autometrics	0.50	1	0.73	1	0.93	1
SURE-Autometrics	1.09	2	1.32	2	1.03	2
Autometrics-SUREMLE	1.59	4	1.70	4	1.64	4
Autometrics-SURE	1.47	3	1.59	3	1.42	3

simultaneously from multiple equations for SURE model. This means that ML estimation is found to be more applicable instead of FGLS in this study.

4. Conclusion

This paper demonstrates that *SUREMLE-Autometrics* outclassed other model selection procedures in both four and two equations model. This proves that number of equations in a model do not affect the superiority of *SUREMLE-Autometrics*' forecasting performance. It also suggests that the use of ML estimation should be given more attention in multiple equations model selection such as in SURE model, particularly in an automatic

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**
SUREMLE-Autometrics	1.11	1	1.45	1	1.46	1
SURE-Autometrics	1.29	2	1.91	4	2.08	4
Autometrics-SUREMLE	1.57	3	1.65	2	1.63	2
Autometrics-SURE	1.59	4	1.66	3	1.63	2

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**
SUREMLE-Autometrics	0.85	2	1.24	1	1.25	1
SURE-Autometrics	0.62	1	1.36	2	1.43	4
Autometrics-SUREMLE	1.21	3	1.39	3	1.41	3
Autometrics-SURE	1.23	4	1.41	4	1.39	2

setting. In addition, a proper choice of estimation method can aid researchers and practitioners in finding the 'best' parsimonious model from a very general model and use the chosen model in forecasting purposes.

Acknowledgements

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

References

[1] Doornik, J. & Hendry, D.F., 2007, Empirical Econometric Modelling using PcGive 12: Volume 1., Timberlake Consultants Ltd, London.

[2] Granger, C.W.J. & Hendry, D.F., 2005, "A Dialogue Concerning a New Instrument for Econometric Modeling. Econometric Theory," 21(01), pp. 278–297.

[3] Hoover, K. & Perez, S., 1999, "Data mining reconsidered: encompassing and the general-to-specific approach to specification search," The Econometrics Journal, 2, pp. 167–191.

[4] Ismail, S., 2005, "Algorithmic Approaches to Multiple Time Series Forecasting," Ph.D. thesis, University of Lancaster, Lancaster.

[5] Kmenta, J. & Gilbert, R.F., 1968, "Small Sample Properties of Alternative Estimators of Seemingly Unrelated Regressions," Journal of the American Statistical Association, 63(324), pp. 1180–1200.

[6] Krolzig, H. & Hendry, D.F., 2001, "Computer automation of general-to-specific model selection procedures," Journal of Economic Dynamics & Control, 25, pp. 831–866.

[7] Magnus, J.R. & Morgan, M.S., 1999, Methodology & tacit knowledge, J. Wiley, NY.

[8] Yusof, N. & Ismail, S., 2011, "Independence Test in SURE-Autometrics Algorithm," International Symposium on Forecasting (ISF). Prague.

[9] Zellner, A., 1962, "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias," Journal of the American Statistical Association, 57(298), pp. 348–368.