ASSESSING THE SIMULATION PERFORMANCES OF MULTIPLE MODEL SELECTION ALGORITHM

Norhayati Yusof¹, Suzilah Ismail², and T Zalizam T Muda³

¹Universiti Utara Malaysia, Malaysia, norhayati@uum.edu.my
²Universiti Utara Malaysia, Malaysia, halizus@uum.edu.my
³Universiti 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](image)

**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,

\[
y_{1t} = \beta_{11}x_{1t,1} + \beta_{12}x_{1t,2} + \ldots + \beta_{1k}x_{1t,k} + u_{1t}
\]

\[
y_{2t} = \beta_{21}x_{2t,1} + \beta_{22}x_{2t,2} + \ldots + \beta_{2k}x_{2t,k} + u_{2t}
\]

\[
\vdots
\]

\[
y_{mt} = \beta_{m1}x_{mt,1} + \beta_{m2}x_{mt,2} + \ldots + \beta_{mk}x_{mt,k} + u_{mt}
\]

(1)
which can be written in general form,

\[ y_i = X_i \beta_i + u_i \quad i = 1, 2, \ldots, m \]

where \( y_i \) is vector of \( T \) identically distributed observations for each random variable, \( X_i \) is a non-stochastic matrix of fixed variables of rank \( k_i \), \( \beta_i \) 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>
<thead>
<tr>
<th>Condition of experiment</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. True models specification</td>
<td>S1: ( y_1 = 0.0230 + 0.0293 \varepsilon ) &lt;br&gt; ( y_2 = 0.0182 + 0.0240 \varepsilon ) &lt;br&gt; S2: ( y_1 = 0.0087 + 0.6170 x_{1t-1} + 0.0229 \varepsilon ) &lt;br&gt; ( y_2 = 0.0058 + 0.6825 x_{2t-1} + 0.0173 \varepsilon ) &lt;br&gt; S3: ( y_1 = 0.0078 + 0.6340 x_{1t-1} + 0.3685 x_{2t-1} + 0.3020 x_{3t-1} + 0.0201 \varepsilon ) &lt;br&gt; ( y_2 = 0.0060 + 0.6915 x_{2t-1} + 0.2811 x_{3t-1} - 0.2224 x_{2t-2} + 0.0151 \varepsilon ) &lt;br&gt; S4: ( y_1 = 0.0049 + 0.5966 x_{1t-1} + 0.4820 x_{2t-1} - 0.2072 x_{3t-1} + 0.0221 \varepsilon ) &lt;br&gt; ( y_2 = 0.0028 + 0.6517 x_{2t-1} + 0.1273 x_{3t-1} + 0.1053 x_{2t-2} + 0.0171 \varepsilon ) &lt;br&gt; S5: ( y_1 = 0.0049 + 0.6154 x_{1t-1} + 0.3376 \varepsilon_{1t-1} - 0.2881 x_{2t-1} + 0.3429 x_{2t-1} - 0.1237 x_{2t-2} + 0.0197 \varepsilon ) &lt;br&gt; ( y_2 = 0.0038 + 0.6720 x_{2t-1} + 0.2742 x_{2t-2} - 0.2268 x_{2t-1} + 0.2078 x_{2t} + 0.1377 x_{2t-2} + 0.0149 \varepsilon )</td>
</tr>
<tr>
<td>2. Strength of correlation disturbances</td>
<td>Weak, ( \rho = 0.2 ) &lt;br&gt; Moderate, ( \rho = 0.6 ) &lt;br&gt; Strong, ( \rho = 0.9 )</td>
</tr>
<tr>
<td>3. Initial GUMs</td>
<td>Small, ( k = \text{at most 18 irrelevant variables} ) &lt;br&gt; Large, ( k = \text{at most 39 irrelevant variables} )</td>
</tr>
<tr>
<td>4. Sample sizes</td>
<td>Small, ( n = 73 ) &lt;br&gt; Large, ( n = 146 )</td>
</tr>
<tr>
<td>5. Significance level</td>
<td>( \alpha = 0.05 ) &lt;br&gt; ( \alpha = 0.01 )</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>True SURE model</th>
<th>Sample sizes, n</th>
<th>$\alpha = 5%$</th>
<th>$\alpha = 1%$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k = \text{at most } 39$</td>
<td>$k = \text{at most } 18$</td>
<td>$k = \text{at most } 39$</td>
</tr>
<tr>
<td>S1</td>
<td>146</td>
<td>89</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>73</td>
<td>88</td>
<td>91</td>
</tr>
<tr>
<td>S2</td>
<td>146</td>
<td>88</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>73</td>
<td>85</td>
<td>89</td>
</tr>
<tr>
<td>S3</td>
<td>146</td>
<td>83</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>73</td>
<td>82</td>
<td>84</td>
</tr>
<tr>
<td>S4</td>
<td>146</td>
<td>87</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>73</td>
<td>84</td>
<td>89</td>
</tr>
<tr>
<td>S5</td>
<td>146</td>
<td>83</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>73</td>
<td>79</td>
<td>87</td>
</tr>
</tbody>
</table>

**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


