Journal of Statistical and Econometric Methods, vol.3, no.4, 2014, 21-31
ISSN: 2241-0384 (print), 2241-0376 (online)
Scienpress Ltd, 2014

Challenge the Approach of Collaboration of Statistical Methods in Selecting the Correct Multiple Linear Regressions with Violation of some Assumptions

Ali Hussein Al-Marshadi[1] and Abdullah H. Al-Harbey[2]

Abstract

The analysis of multiple linear regressions (MLR) is investigated in the current study. MLR is applied frequently in different field of applied studies. In the current study simulation technique is used to evaluate the approach of "The Approach of Collaboration of Statistical Methods in Selecting the Correct Regressions (ACSMSCR)" by AL-Marshadi, (2014) with its two options in terms of its ability to identify the right regression model under some model assumptions violations with two different sample sizes. The evaluation is in terms of the percentage of number of times of success in identifying the right model. The simulation results indicate that the approach of (ACSMSCR) provided very good choice to select the right model even under the considered model assumptions violations where the second option provided the best performance for both the

[1] Department of Statistics, King Abdulaziz University. E-mail: AALMarshadi@kau.edu
[2] Department of Statistics, King Abdulaziz University. E-mail: AHarbey@yahoo.com

Article Info: *Received* : September 6, 2014. *Revised* : October 29, 2014.
Published online : December 27, 2014.

sample sizes. The main result of the current study is that we suggest using the approach of (ACSMSCR) with the second option as a reliable procedure to select the right model even under the considered model assumptions violations.

Mathematics Subject Classification: Statistics

Keywords: Multiple Linear Regression; Multicollinearity; Heteroscedasticity; First-order autocorrelation; Information Criteria; Bootstrap Procedure; Clustering Procedure

1 Introduction

Researchers usually use the technique of Regression Analysis to model the relationship between a response variable Y, and some explanatory variable usually denoted X_k. In general form, the statistical model of multiple linear regressions (MLR) is:

$$Y_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} + \varepsilon_i , \quad (1)$$

Where:

$\beta_0, \beta_1, ..., \beta_{p-1}$ are the unknown parameters

$X_{i1}, ..., X_{i,p-1}$ are the explanatory variables

ε_i are independent $N(0, \sigma^2)$; $i = 1, ..., n$ (SAS Institute Inc., 2004; John *et al*., 1996).

Deciding what the right model for the observed data is the most critical part of the regression analysis. In practice most of researchers suggest considering all possible combinations of predictor variables to construct all combinations of regression models in order to select the right model among all combinations of regression models using information criterion (SAS Institute Inc., 2004; John *et al*., 1996; AL-Marshadi, 2014). Plenty of studies have provided either new

information criteria or modified one to be used in selecting the right model (Akaike, 1969; Judge *et al*., 1980; Sawa, 1978; Schwarz, 1978; Amemiya, 1976; 1983; Hocking, 1976; AL-Marshadi, 2014).

The objective of the current study is evaluating the approach of "The Approach of Collaboration of Statistical Methods in Selecting the Correct Regressions (ACSMSCR)" by AL-Marshadi, (2014) in selecting the right model under three violations of the model assumptions with its two options. The three violations of the model assumptions that were considered in the current study are multicollinearity, heteroscedasticity, and first-order autocorrelation error.

2 Methodology

The REG procedure of the SAS system is frequently used for analyzing data with multiple linear regression models. In REG procedure, the following seven model selection criteria are available, which can be used to select the right regression model (SAS Institute Inc., 2004).

- Akaike's Information Criterion (AIC) (Akaike, 1969),
- Sawa's Bayesian Information Criterion (BIC) (Judge *et al*., 1980; Sawa, 1978),
- Schwarz's Bayes Information Criteria (SBC) (Schwarz, 1978),
- Amemiya's Prediction Criteria (PC) (Judge *et al*., 1980; Amemiya, 1976; 1983),
- Final Prediction Error (JP) (Hocking, 1976; Judge *et al*., 1980),
- Estimated Mean Square Error of Prediction (GMSEP) (Hocking, 1976), and
- SP Statistics (SP) (Hocking, 1976).

The approach of (ACSMSCR) involves using the bootstrap technique (Efron, 1983; 1986), and Hierarchical Clustering Methods with two options of distance measures, Ward's Minimum Variance Approach, and Single Linkage Approach

(Khattree and Naik, 2000) to help the previous seven information criteria to select the right regression model (AL-Marshadi, 2014). The approach of (ACSMSCR) showed excellent performance when the model assumptions are satisfied (AL-Marshadi, 2014). In current study the approach of (ACSMSCR) is used to select the right regression model under three violations of the model assumptions that are multicollinearity, heteroscedasticity, and first-order autocorrelation error with its two options.

3 The Simulation Study

A simulation study of PROC REG's regression model analysis of data was carried out to evaluate the approach of (ACSMSCR), under three violations of the model assumptions with its two options, in terms of its percentage of number of times of success in selecting the right model.

The initial setup of the simulation study is quite similar to the setup used in AL-Marshadi, (2014) which is described as following:

Data follow Normal distributions were generated according to all possible regression models, (K=7 models) that can be constructed of three predictor variables, X_1, X_2, X_3. These regression models are special cases of model (1) when the parameters were set up equal to $(\beta_0 = 2, \beta_1 = 3, \beta_2 = 4, \beta_3 = 5)$. There were 42 scenarios to generate data involving three cases of assumptions violations. First case where multicollinearity was inducted into the data, second case where heteroscedasticity was inducted into the data, and third case where first-order autocorrelation error was inducted into the data. Two different sample sizes ($n = 50$, and 100 observations) for all possible combinations of regression models are considered in the current study. The predictor variables, X_1, X_2, X_3 were generated from normal distributions with, $\mu = 0$, and $\sigma^2 = 4$. The error term of

the model was generated from normal distribution with, $\mu = 0$, and $\sigma^2 = 9$. When the multicollinearity was inducted into the data, the covariance, $\sigma_{x_1 x_2}$, between X_1 and X_2 was set up equal to 0.70. When the heteroscedasticity was inducted into the data, the error term of the model was generated from normal distribution with $\mu = 0$ and inconstant variance, such that, $i = \sigma_i^2$; $i = 1, 2, ..., n$. When the first-order autocorrelation was inducted into the data, the error term of the model was generated from correlated normal distribution such that, $\varepsilon_i = \rho \varepsilon_{i-1} + u_i$; $u_i \sim i.i.d.\ N(0, \sigma^2)$; $i = 1, 2, ..., n$, and $\sigma^2 = 9$. Where the correlation, ρ, set up equal to 0.90. For each scenario, 5000 datasets were simulated using SAS/IML (SAS Institute Inc., 2004) according to the described models. The algorithm of (ACSMSCR) approach was applied to each one of the 5000 generated data sets with each possible model (AL-Marshadi, 2014). The percentage of successes in selecting the right model is reported for the two options of the approach.

4 Main Results

Table 1 summarizes results of the percentage of successes in selecting the right regression model from all possible regression models with the two options of the approach of (ACSMSCR), when multicollinearity was inducted into the data, n=50, and **W**=10 (where **W** is equal to the number of the bootstrap samples) . Table 2 summarizes results of the percentage of successes in selecting the right regression model from all possible regression models with the two options of the approach of (ACSMSCR), when multicollinearity was inducted into the data, n=100, and **W**=10.

Table 3 summarizes results of the percentage of successes in selecting the right regression model from all possible regression models with the two options of the

approach of (ACSMSCR), when heteroscedasticity was inducted into the data, n=50, and **W=**10. Table 4 summarizes results of the percentage of successes in selecting the right regression model from all possible regression models with the two options of the approach of (ACSMSCR),when heteroscedasticity was inducted into the data, n=100, and **W=**10.

Table 5 summarizes results of the percentage of successes in selecting the right regression model from all possible regression models with the two options of the approach of (ACSMSCR), when first-order autocorrelation was inducted into the data, n=50, and **W=**10. Table 6 summarizes results of the percentage of successes in selecting the right regression model from all possible regression models with the two options of the approach of (ACSMSCR), when first-order autocorrelation error was inducted into the data, n=100, and **W=**10.

Table 1: The Percentage of number of times that the procedure selects the right regression model from the all possible regression models with the two options when multicollinearity was inducted into the data, n=50, and W=10.

The correct model	The cluster of the best set of models	The percent of success	
		The Word option	The Single option
X1	X1,X1X2,X1X3,X1X2X3	100%	100%
X2	X2,X1X2,X2X3,X1X2X3	100%	100%
X3	X3,X1X3,X2X3,X1X2X3	100%	100%
X1, X2	X1X2,X1X2X3	84.18%	86.20%
X1, X3	X1X3,X1X2X3	33.64%	65.58%
X2, X3	X2X3,X1X2X3	92.84%	96.04%
X1, X2, X3	X1X2X3	31.32%	99.80%
Overall percent of success		77.43%	92.52%

Table 2: The Percentage of number of times that the procedure selects the right regression model from the all possible regression models with the two options when multicollinearity was inducted into the data, n=100, and W=10.

The correct model	The cluster of the best set of models	The percent of success	
		The Word option	The Single option
X1	X1,X1X2,X1X3,X1X2X3	100%	100%
X2	X2,X1X2,X2X3,X1X2X3	100%	100%
X3	X3,X1X3,X2X3,X1X2X3	100%	100%
X1, X2	X1X2,X1X2X3	98.22%	99.78%
X1, X3	X1X3,X1X2X3	80.06%	93.06%
X2, X3	X2X3,X1X2X3	99.96%	100%
X1, X2, X3	X1X2X3	72.32%	99.82%
Overall percent of success		92.94%	98.95%

Table 3: The Percentage of number of times that the procedure selects the right regression model from the all possible regression models with the two options when heteroscedasticity was inducted into the data, n=50, and W=10.

The correct model	The cluster of the best set of models	The percent of success	
		The Word option	The Single option
X1	X1,X1X2,X1X3,X1X2X3	99.98%	99.98%
X2	X2,X1X2,X2X3,X1X2X3	100%	100%
X3	X3,X1X3,X2X3,X1X2X3	100%	100%
X1, X2	X1X2,X1X2X3	60.92%	66.42%
X1, X3	X1X3,X1X2X3	23.16%	22.46%
X2, X3	X2X3,X1X2X3	72.98%	76.00%
X1, X2, X3	X1X2X3	10.12%	58.70%
Overall percent of success		66.74%	74.79%

Table 4: The Percentage of number of times that the procedure selects the right regression model from the all possible regression models with the two options when heteroscedasticity was inducted into the data, n=100, and W=10.

The correct model	The cluster of the best set of models	The percent of success	
		The Word option	The Single option
X1	X1,X1X2,X1X3,X1X2X3	100%	99.98%
X2	X2,X1X2,X2X3,X1X2X3	100%	100%
X3	X3,X1X3,X2X3,X1X2X3	100%	100%
X1, X2	X1X2,X1X2X3	59.18%	71.90%
X1, X3	X1X3,X1X2X3	1.22%	32.20%
X2, X3	X2X3,X1X2X3	77.80%	87.18%
X1, X2, X3	X1X2X3	9.50%	78.98%
Overall percent of success		63.96%	81.46%

Table 5: The Percentage of number of times that the procedure selects the right regression model from the all possible regression models with the two options when first-order autocorrelation was inducted into the data, n=50, and W=10.

The correct model	The cluster of the best set of models	The percent of success	
		The Word option	The Single option
X1	X1,X1X2,X1X3,X1X2X3	99.74%	99.50%
X2	X2,X1X2,X2X3,X1X2X3	99.98%	99.82%
X3	X3,X1X3,X2X3,X1X2X3	100%	99.98%
X1, X2	X1X2,X1X2X3	59.12%	68.46%
X1, X3	X1X3,X1X2X3	5.38%	32.12%
X2, X3	X2X3,X1X2X3	70.64%	78.78%
X1, X2, X3	X1X2X3	9.74%	55.34%
Overall percent of success		63.51%	76.29%

Table 6: The Percentage of number of times that the procedure selects the right regression model from the all possible regression models with the two options when first-order autocorrelation was inducted into the data, n=100, and W=10.

The correct model	The cluster of the best set of models	The percent of success	
		The Word option	The Single option
X1	X1,X1X2,X1X3,X1X2X3	97.84%	99.98%
X2	X2,X1X2,X2X3,X1X2X3	100%	100%
X3	X3,X1X3,X2X3,X1X2X3	100%	100%
X1, X2	X1X2,X1X2X3	68.80%	80.04%
X1, X3	X1X3,X1X2X3	23.88%	41.30%
X2, X3	X2X3,X1X2X3	85.24%	97.60%
X1, X2, X3	X1X2X3	17.34%	77.80%
Overall percent of success		70.44%	85.25%

5 Conclusion

In our simulation, multiple linear regressions were considered when the model suffers from three violations of model assumptions, looking at the performance of the approach of (ACSMSCR) for selecting the right regression model with its two options and two different sample sizes. Comparing Table 1 and Table 2 in the current study to Table 1 and Table 2 in the study of AL-Marshadi, (2014) respectively shows that the violation of multicollinearity assumption has no significant effect on the performance of (ACSMSCR) method with its two options for the two sample size, where the single option shows the best performance for both the sample sizes. Also, comparing Table 3, and Table 4, in the current study and Table 5, and Table 6 in the current study to Table 1 and Table 2 in the study of AL-Marshadi, (2014) respectively shows that the violation

of heteroscedasticity and first-order autocorrelation assumptions, respectively have some little effect on the performance of the approach of (ACSMSCR) with its two options for the two sample sizes, where the single option shows the best performance for both sample sizes.

In general, the increasing of the sample size has a significant effect in the performance of the approach of (ACSMSCR). Finally, we can say that the approach of (ACSMSCR) provided very good choice to be considered in selecting of the right model even in the existing of the considered violation in the model assumptions.

ACKNOWLEDGEMENTS. The author is thankful to the referee and Editors for useful comments which improve the quality of paper.

References

[1] AL-Marshadi, Ali Hussein, Collaboration of Statastical Methods in Selecting the Correct Multiple Linear Regressions, *American Journal of Biostatistics*, **4**, (2014), 29-33.

[2] SAS Institute, Inc., SAS*/STAT User's Guide SAS OnlineDoc 9.1.2.*, Cary NC: SAS Institute Inc., 2004.

[3] John, N., K.H. Michael and W. William, *Applied Linear Regression Models*, 3rd Edn., Richard D. Irwin, Inc., Chicago, 1996.

[4] Akaike, H., Fitting autoregressive models for prediction, *Ann. Inst. Statist. Math.*, **21**, (1969), 243-247.

[5] Judge, G.G., W.E. Griffiths, R.C. Hill and T. Lee, *Theory and Practice of Econometrics*, New York: Wiley, 1980.

[6] Sawa, T., Information criteria for discriminating among alternative regression models, *Econometrica*, **46**, (1978), 1273-1291.

[7] Schwarz, G., Estimating the dimension of a model, *Ann. Statistics*, **6**, (1978), 461-464.

[8] Amemiya, T., Estimation in Nonlinear Simultaneous Equation Models, *Paper presented at Institut National de La Statistique et Des Etudes Ecnomiques, Malinvaued*, E. (Ed.), Paris, Cahiers Du SeminarireD'econometrie, no 19, (1976).

[9] Amemiya, T., *Advanced Econometrics*, Cambridge: Harvard University Press, 1983.

[10] Hocking, R.R., The analysis and selection of variables in linear regression. *Biometrics*, **32**, (1976), 1-49.

[11] Efron, B., Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Am. Statist. Assoc.*, **78**, (1983), 316-331.

[12] Efron, B., How biased is the apparent error rate of a prediction rule? *J. Am. Statist. Assoc.*, **81**, (1986), 416-470.

[13] Khattree, R. and N.D. Naik, *Multivariate Data Reduction and Discrimination with SAS Software*, SAS Institute Inc.Cary NC, USA, 2000.