American Journal of Biostatistics 4 (2): 29-33, 2014
ISSN: 1948-9889

doi:10.3844/ajbssp.2014.29.33 Published Online 4(2) 2014 (http://www.thescipub.com/ajbs.toc)

COLLABORATION OF STATISTICAL METHODS IN SELECTING THE CORRECT MULTIPLE LINEAR REGRESSIONS

Ali Hussein Al-Marshadi

Department of Statistics, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

Received 2014-07-12; Revised 2014-08-20; Accepted 2014-09-11

ABSTRACT

This article considers the analysis of Multiple Linear Regressions (MLRs) that are essential statistical method for the analysis of medical data in various fields of medical research like prognostic studies, epidemiological risk factor studies, experimental studies, diagnostic studies and observational studies. An approach is used in this article to select the "true" regression model with different sample sizes. We used the simulation study to evaluate the approach in terms of its ability to identify the "true" model with two options of distance measures: Ward's Minimum Variance Approach and the Single Linkage Approach. The comparison of the two options performed was in terms of their percentage of the number of times that they identify the "true" model. The simulation results indicate that overall, the approach exhibited excellent performance, where the second option providing the best performance for the two sample sizes considered. The primary result of our article is that we recommend using the approach with the second option as a standard procedure to select the "true" model.

Keywords: Multiple Linear Regression, Information Criteria, Bootstrap Procedure, Clustering Procedure

1. INTRODUCTION

Regression is a tool that allows researchers to model the relationship between a response variable Y and a number of explanatory variable, usually denoted X_k. In general form, the statistical model of Multiple Linear Regressions (MLRs) is:

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

Where:

$\beta_0, \beta_1,\ldots,\beta_{p-1}$ = The unknown parameters
$X_{i1},\ldots,X_{i,p-1}$ = The explanatory variables
ε_i = Independent $N(0,\sigma^2)$; i = 1,…,n (SAS, 2004; John *et al*., 1996)

The primary concern in the analysis of the use of regressions is concentrated on determining the suitable model of the data. In practice, many researchers recommend considering all possible combinations of independent variables used to construct the regression models to select the true model among them using some information criterion (SAS, 2004; John *et al*., 1996). Statisticians often use information criteria, such as 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 Criterion (PC) (Judge *et al*., 1980; Amemiya, 1976; 1985), final Prediction Error (JP) (Hocking, 1976; Judge *et al*., 1980), Estimated Mean Square Error of Prediction (GMSEP) (Hocking, 1976) and SP Statistic (SP) (Hocking, 1976), to select the true model (SAS, 2004; Neter *et al*., 1990). Many studies have proposed the use of either new or modified information criteria to select the true model (Akaike, 1969; Judge *et al*., 1980; Sawa, 1978; Schwarz, 1978; Amemiya, 1976; 1985; Hocking, 1976; AL-Marshadi, 2011).

Our research objective is to apply and evaluate the idea of the ACSMSCCS approach (AL-Marshadi, 2014) in

selecting the true regression model. The evaluation involves a comparison of two options of the approach in terms of the ability of each option to identify the true model.

2. METHODOLOGY

The REG procedure of the SAS system is a standard tool for analyzing data with multiple linear regression models. In the REG procedure, we can find the following seven model selection criteria available, which can be used to select an appropriate regression model (SAS, 2004). The seven model selection criteria are:

- 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; 1985)
- 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 involves using the bootstrap technique (Efron, 1983; 1986) and Hierarchical Clustering Methods with two options of distance measures of Ward's Minimum Variance Approach and the Single Linkage Approach (Khattree and Naik, 2000) as tools to accommodate the effort of the seven information criteria in identifying the correct regression model. This approach showed excellent performance in different context (AL-Marshadi, 2014). The general idea of using the bootstrap technique to improve the performance of a model selection rule was introduced by Efron (1983; 1986) and was extensively discussed by Efron and Tibshirani (1993).

In the context of the multiple linear regression models described in Equation (1), the algorithm for using the parametric bootstrap technique in the approach can be outlined as follows:

Let the observation vector O_i be defined as follows: $O_i = [Y_i X_{i1} \ldots X_{i,p-1}]$, where i = 1, 2,…,*n*.

Generate the W bootstrap samples on a case-by-case basis using the observed data i.e., based on resampling from (O_i, O_2,…., O_n). Each bootstrap sample size is taken to be the same as the size of the observed sample (i.e., *n*). Efron and Tibshirani discussed the properties of the bootstrap technique when the bootstrap sample size is equal to the original sample size (Efron and Tibshirani, 1993).

Fit all the possible regression models using the considered independent variables (K model), from which we wish to select the true model, to the W bootstrap samples, thereby obtaining the bootstrap AIC*, BIC*, SBC*, PC*, JP*, GMSEP* and SP* for each model from the W bootstrap samples.

Statisticians often use the previous collection of information criteria to select the true model, such as selecting the model with the smallest value of the information criteria (SAS, 2004; Neter *et al*., 1990). We will follow a different rule in the approach. The bootstrapping of the observed data provides us the advantage that, for each model and each information criteria, we have (W) replication values (from steps (1) and (2)). To make use of this advantage, we propose using the averages of each information criteria for each model separately in the approach to construct a random vector that follows a 7-dimensional multivariate normal distribution.

$$\left[\overline{AIC}^{*} \quad \overline{BIC}^{*} \quad \overline{SBC}^{*} \quad \overline{PC}^{*} \quad \overline{JP}^{*} \quad \overline{GMSEP}^{*} \quad \overline{SP}^{*}\right]_{Model-i}$$

$$; i = 1, 2, ..., K$$

To briefly justify that the random vector follows a 7-dimensional multivariate normal distribution, let us consider each model separately and assume that each average of the information criteria approximately follows normal distribution according to the central limit theorem. Additionally, those averages of the information criteria for each model are assumed to be correlated. Therefore, we can consider the averages of the information criteria of each model to be a random vector that follows a 7-dimensional multivariate normal distribution. In this stage, the Clustering method will play the main role of the approach by clustering all possible regression models to two clusters, with one of them being determined to be the cluster of the best set of models. The cluster of the best set of models will be determined according to the cluster that includes the general model (the full model). Next, the best model will be the simplest model in the cluster of the best set of models. We refer to the approach in the context of the regression models as the Approach of Collaboration of Statistical Methods in Selecting the Correct Regressions (ACSMSCR).

3. THE SIMULATION STUDY

A simulation study of PROC REG's regression model analysis of data was conducted to evaluate the

approach in terms of its percentage of times that it identifies the true model.

The setup of the simulation study is quite similar to the setup used in (AL-Marshadi, 2011) which is described briefly as following:

Normal data were generated according to all possible regression models, (K = 7 models) that can be constructed using three independent variables X_1, X_2, X_3. These regression models are special cases of model (1) with known regression parameters (β_0 = 2, β_1 = 3, β_2 = 4, β_3 = 5). There were 14 scenarios to generate data involving two different sample sizes (n = 50 and n = 100 observations) with all the possible regression models. The independent variables X_1, X_2, X_3 were generated from the normal distributions with μ = 0 and σ^2 = 4. The error term of the model was drawn from the normal distribution with μ = 0 and σ^2 = 9. For each scenario, we simulated 5000 datasets. The SAS/IML (SAS, 2004) code was written to generate the datasets according to the described models. The algorithm of the approach was applied to each one of the 5000 generated data sets with each possible model. The percentage of times that the approach selects the right model was reported for both options of the distance measures: Ward's Minimum Variance Approach and the Single Linkage Approach (Khattree and Naik, 2000).

4. RESULTS

Table 1 summarizes results of the percentage of times that the procedure selects the true regression model from all possible regression models for the approach with the two options, when n = 50 and W = 10. **Table 2** summarizes the results of the percentage of times that the procedure selects the true regression model from all the possible regression models for the approach with the two options, when n = 100 and W = 10. Overall, the approach provided significant improvement in term of the percentage of success in selecting the right model which can be seen by compare the performance of the approach to the performance of each of the information criteria considered with the approach. The performance of each of the information criteria considered with the approach can be seen in (AL-Marshadi, 2011) for the performance comparison.

Finally, the approach is applied on real data that study the effects of the charge rate and temperature on the life of a new type of power cell in a preliminary small-scale experiment. The charge rate (X1) was controlled at three levels (0.6, 1.0 and 1.4 amperes) and the ambient temperature (X2) was controlled at three levels (10, 20 and 30°C). Factors pertaining to the discharge of the power cell were held at fixed levels. The life of the power cell (Y) was measured in terms of the number of discharge-charge cycles that a power cell underwent before it failed. **Table 3** contains the data obtained in the study. The researcher decided to fit the first order model in terms of X1 and X2, without a cross-product interaction effect X12, to this initial small-scale study data after detailed analysis (John *et al*., 1996). The approach was applied to select the best model for these data considering all possible regression models, (K = 3 models) that can be constructed of the two predictor variables of X1 and X2. **Table 4** describes the result of the approach when W = 5 for the three considered models. The approach (ACSMSCR) selected the best model as the one that was selected by the researcher with the both options of Word and Single.

Table 1. The percentage of times that the procedure selects the true regression model from the all possible regression models for the approach with the two options when 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	100.00	100.00
X2	X2,X1X2,X2X3,X1X2X3	100.00	100.00
X3	X3,X1X3,X2X3,X1X2X3	100.00	100.00
X1, X2	X1X2,X1X2X3	87.66	94.20
X1, X3	X1X3,X1X2X3	47.42	66.66
X2, X3	X2X3,X1X2X3	92.86	96.06
X1, X2, X3	X1X2X3	40.22	91.46
Overall percent of success		81.17%	92.63

Table 2. The percentage of times that the procedure selects the true regression model from the all possible regression models for the approach with the two options when n = 100 and W = 10

		The percent of success	
The correct model	The cluster of the best set of models	The word option (%)	The single option (%)
X1	X1,X1X2,X1X3,X1X2X3	100.00	100.00
X2	X2,X1X2,X2X3,X1X2X3	100.00	100.00
X3	X3,X1X3,X2X3,X1X2X3	100.00	100.00
X1, X2	X1X2,X1X2X3	98.96	99.80
X1, X3	X1X3,X1X2X3	82.22	93.72
X2, X3	X2X3,X1X2X3	100.00	100.00
X1, X2, X3	X1X2X3	78.52	99.72
Overall percent of success		94.24%	99.03

Table 3. Data for the power cells example

Cell (i)	1	2	3	4	5	6	7	8	9	10	11
Number of cycles (Y)	150.0	86.0	49.0	288.0	157.0	131.0	184.0	109.0	279.0	235.0	224.0
Charge rate (X1)	0.6	1.0	1.4	0.6	1.0	1.0	1.0	1.4	0.6	1.0	1.4
Temperature (X2)	10.0	10.0	10.0	20.0	20.0	20.0	20.0	20.0	30.0	30.0	30.0

Table 4. The result of the approach for the power cells example when W = 5 for the three considered models

	The cluster number	
The considered models	The word option	The single option
X1	1	1
X2	1	1
X1, X2	2	2
The selected model	X1,X2	X1,X2

5. CONCLUSION

In our simulation, we performed a multiple linear regressions analysis to apply and evaluate the idea of the ACSMSCCS approach (AL-Marshadi, 2014) for selecting the suitable regression model with two options and different sample sizes. Overall, the approach provided excellent results in selecting the true model, with the second option providing the best performance for the two sample sizes that were considered.

6. REFERENCES

Akaike, H., 1969. Fitting autoregressive models for prediction. Ann. Inst. Stat. Math., 21: 243-247.

AL-Marshadi, A.H., 2011. New weighted information criteria to select the true regression model. Aus. J. Basic Applied Sci., 5: 317-321.

AL-Marshadi, A.H., 2014. Selecting the covariance structure in mixed model using statistical methods calibration. J. Math. Stat., 10: 309-315. DOI: 10.3844/jmssp.2014.111.116

Amemiya, T., 1976. 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.

Amemiya, T., 1985. Advanced Econometrics. 1st Edn., Harvard University Press, Cambridge, ISBN-10: 0674005600, pp: 521.

Efron, B. and R.J. Tibshirani, 1993. An Introduction to the Bootstrap. 1st Edn., CRC Press, ISBN-10: 0412042312, pp: 456.

Efron, B., 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. J. Am. Stat. Assoc., 78: 316-331. DOI: 10.1080/01621459.1983.10477973

Efron, B., 1986. How biased is the apparent error rate of a prediction rule? J. Am. Stat. Assoc., 81: 416-470. DOI: 10.1080/01621459.1986.10478291

Hocking, R.R., 1976. A biometrics invited paper. the analysis and selection of variables in linear regression. Biometrics, 32: 1-49.

John, N., K.H. Michael and W. William, 1996. Applied Linear Regression Models, 3rd Edn., Richard D. Irwin, Inc., Chicago. ISBN-10: 0072955678

Neter, J., W. Wasserman and M.H. Kutner, 1990. Applied Linear Regression Models. 3rd Edn., Irwin, Homewood, ISBN-10: 025608338X, pp: 1179.

Judge, G.G., W.E. Griffiths, R.C. Hill and T. Lee, 1980. The Theory and Practice of Econometrics. 1st Edn., John Wiley and Sons Canada, Limited, New York, ISBN-10: 0471087548, pp: 810.

Khattree, R. and N.D. Naik, 2000. Multivariate Data Reduction and Discrimination with SAS Software. 1st Edn., SAS Institute, ISBN-10: 1580256961, pp: 574.

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

Sawa, T., 1978. Information criteria for discriminating among alternative regression models. Econometrica, 46: 1273-1291.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat., 6: 461-464.