Australian Journal of Basic and Applied Sciences, 5(3): 317-321, 2011
ISSN 1991-8178

New Weighted Information Criteria to Select the True Regression Model

Ali Hussein Al-Marshadi

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

Abstract: This article considers the analysis of multiple linear regressions (MLR) that is used frequently in practice. We propose new weighted information criteria (WIC) that could be used to guide the selection of the "true" regression model for different sample size. Usually, weighted information criterion is calculated by summing weighted different selection criteria when the weights of the weighted information criterion are determined heuristically. In this study, we used simulation study to compare two new weighted information criteria with other seven model selection criteria in terms of their ability to identify the "true" model. The comparison of the nine model selection criteria was in terms of their percentage of number of times that they identify the "true" model. The simulation results indicate that overall, the first proposed weighted information criterion (PWIC-I) showed very good performance over all where it provided the second best performance after SBC criterion. The main result of our article is that we recommend considering the new weighted information criterion (PWIC-I) as a reliably criterion to identify the "true" model.

Key words: Multiple Linear Regression; Information Criteria; Bootstrap Procedure; principal component Procedure

INTRODUCTION

Regression is a tool that allows researcher 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:

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

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 $N(0, \sigma^2)$; $i = 1, ..., n$

We are interested in selecting the suitable regression model. In general, this is what SAS procedures, PROC GLM, PROC REG, and PROC AUTOREG, are set up to do, (SAS Institute Inc., 2008). In practice many researchers recommend considering all possible regression models that can be constructed of all the available variables to select the true model among them using some information criterion, (Neter, J. *et al*., 1996). A lot of efforts are usually needed to decide what the suitable model of the data is. Statisticians often use information criteria such as Akaike's Information Criterion (AIC), (Akaike, H., 1969), Sawa's Bayesian Information Criterion (BIC), (Judge G. G. *et al*., 1980; Sawa, T., 1978), Schwarz's Bayes Information Criteria (SBC), (Schwarz, G., 1978), Amemiya's Prediction Criterion (PC), (Judge G. G. *et al*., 1980; Amemiya, T., 1976; Amemiya, T., 1985), Final Prediction Error (JP), (Judge G. G. *et al*., 1980; Hocking R. R., 1976), Estimated Mean Square Error of Prediction (GMSEP), (Hocking R. R., 1976), and SP Statistic (SP), (Hocking R. R., 1976) to guide the selection of the true model, (SAS Institute Inc., 2008; Neter, J. *et al*., 1996). Many studies have proposed either new or modified criteria to be used to select the true model. New different approaches have been proposed in the literature in order to select the true model. One of these approaches is to establishing a new information criterion by summing weighted different selection criteria when the weights of the weighted information criterion are determined heuristically, (Egrioglu *et al*., 2008). Egrioglu *et al*. proposed weighted information criterion (WIC) when the weights of his weighted information criterion were determined intuitively, (Egrioglu *et al*., 2008). Aladag *et al*. improved the WIC criteria by optimizing the weights used as coefficients in the WIC criterion and called it the adaptive weighted information criteria (AWIC), (Aladag *et al*., 2010).

Our research objective is proposing two new weighted information criteria that could be used to guide the selection of the true regression model. Also, our research objective involves comparing the new weighted

Corresponding Author: Ali Hussein Al-Marshadi, Department of Statistics, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia,
E-mail: aalmarshadi@kau.edu.sa

information criteria with seven well-known model selection criteria in terms of their ability to identify the true model.

Methodology:

The REG procedure of the SAS system is a standard tool for fitting data with multiple linear regression models, (SAS Institute Inc., 2008). In REG procedure, users find the following seven model selection criteria available, which give users tools can be used to select an appropriate regression model. The seven model selection criteria are, (SAS Institute Inc., 2008):

1. Akaike's Information Criterion (AIC), (Akaike, H., 1969),
2. Sawa's Bayesian Information Criterion (BIC), (Judge G. G. *et al*., 1980; Sawa, T., 1978),
3. Schwarz's Bayes Information Criteria (SBC), (Schwarz, G., 1978),
4. Amemiya's Prediction Criteria[4,7,8] (PC), (Judge G. G. *et al*., 1980; Amemiya, T., 1976; Amemiya, T., 1985),
5. Final Prediction Error (JP), (Judge G. G. *et al*., 1980; Hocking R. R., 1976),
6. Estimated Mean Square Error of Prediction (GMSEP), (Hocking R. R., 1976), and
7. SP Statistics (SP), (Hocking R. R., 1976).

Our study concerns with comparing the two new weighted information criteria to the previous seven information criteria in terms of their ability to identify the true model.

The two new weighted information criteria involves using the bootstrap technique, (Efron, B., 1983; Efron, B., 1986), and the principal component analysis, (Khattree, R., and Naik N. D., 2000) as tools to determine the weights of the new weighted information criteria with respect to the data that are analyzed. The idea of using the bootstrap in improving the performance of a rule of model selection was introduced by Efron, (Efron, B., 1983; Efron, B., 1986), and is extensively discussed by Efron and Tibshirani, (Efron, B., and Tibshirani, R. J., 1993).

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

Let the observation vector O_{i} is defined as follows: $O_{\mathrm{i}} = \left[Y_i \quad X_{i1} \quad ... \quad X_{i,p-1}\right]'$, where $i = 1,2,...,n.$.

1. Generate the bootstrap sample on case-by-case using the observed data (original sample) *i.e.*, based on resampling from $(O_1, O_2, ..., O_n)$. The 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 when the bootstrap sample size is equal to the original sample size (Efron, B., and Tibshirani, R. J., 1993).

2. Fit the all possible regression models, which we would like to select the true model from them, to the bootstrap data, thereby obtaining the bootstrap AIC*, BIC*, SBC*, PC*, JP*, GMSEP*, and SP* for each model.

3. Repeat steps (1) and (2) (R) times.

4. Statisticians often use the previous collection of information criteria to guide the selection of the true model such as selecting the model with the smallest value of the information criteria, (SAS Institute Inc., 2008; Neter, J. *et al*., 1996). We will follow the same rule in our new weighted criteria. We have the advantage that each information criteria has (R) replication values result of the bootstrapping of the observed data (from step (1), (2), and (3)). In other ward, we have seven bootstrapping samples of size R for each candidate model one sample for each information criterion (AIC*, BIC*, SBC*, PC*, JP*, GMSEP*, and SP*). To use this advantage, we propose using the principal component analysis, to determine the weights of the two new weighted information criteria as follow: Suppose $\sum$ is the variance-covariance matrix of 7 variables AIC*, BIC*, SBC*, PC*, JP*, GMSEP*, and SP* for any candidate model. The total variance of these variables is defined as $tr\sum$ (the trace of $\sum$), The first principal component of 7 by 1 vector $\mathrm{X}=\left(AIC^{*} \quad BIC^{*} \quad SBC^{*} \quad PC^{*} \quad JP^{*} \quad GMSEP^{*} \quad SP^{*}\right)'$ is a linear combination: $\mathrm{a}_1'\mathrm{x}=a_{11}AIC^{*}+a_{12}BIC^{*}+a_{13}SBC^{*}+a_{14}PC^{*}+a_{15}JP^{*}+a_{16}GMSEP^{*}+a_{17}SP^{*}$, where $\mathrm{a}_1=\left(a_{11} \quad a_{12} \quad a_{13} \quad a_{14} \quad a_{15} \quad a_{16} \quad a_{17}\right)'$, with $\mathrm{a}_1'\mathrm{a}_1=1$ and such that $\mathrm{var}(\mathrm{a}_1'\mathrm{x})$ is the maximum among all linear combination of x, with the coefficient vector having unit length. Thus the first principal component so obtained accounts for the maximum variation. Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_7 \geq 0$ be the eigenvalues and $\mathrm{a}_1, ..., \mathrm{a}_7$ be the corresponding eigenvectors of $\sum$. Then $u_1=\mathrm{a}_1'\mathrm{x}$, $u_2=\mathrm{a}_2'\mathrm{x}$, ..., $u_7=\mathrm{a}_7'\mathrm{x}$ are the first, second,…, 7^{th} principal components of x. Furthermore, $\mathrm{var}(u_1)=\lambda_1, ..., \mathrm{var}(u_7)=\lambda_7$. Also, the correlation coefficient between each variable of 7 variables AIC*, BIC*, SBC*, PC*, JP*, GMSEP*, and

SP*, and the j^{th} principal component u_j has some nice interpretation. For example the correlation coefficient between the variable AIC*, and the first principal component u_1 is given by

$$corr(AIC^*,u_1) = a_{11}\sqrt{\frac{\lambda_1}{\operatorname{var}(AIC^*)}} \Rightarrow a_{11} = \frac{corr(AIC^*,u_1)\sqrt{\operatorname{var}(AIC^*)}}{\sqrt{\lambda_1}} \Rightarrow$$

$$\frac{1}{a_{11}} = \frac{\sqrt{\lambda_1}}{corr(AIC^*,u_1)\sqrt{\operatorname{var}(AIC^*)}};$$

Since the variables (AIC*, BIC*, SBC*, PC*, JP*, GMSEP*, and SP*) with coefficients of larger magnitude in the first principal component have larger contribution to that component, we suggest the weights of the first new weighted information criterion to be calculated for each candidate model separately using the inverse of each coefficient of the first principal component as a weight for the corresponding information criterion (variable) of the seven information criteria (AIC, BIC, SBC, PC, JP, GMSEP, and SP) (variables) that compose the first new weighted information criterion, when the seven bootstrapping samples of the variables (AIC*, BIC*, SBC*, PC*, JP*, GMSEP*, and SP*) for the corresponding model were used in the principal component analysis to determine the coefficients of the first principal component. After the weights were determined for each candidate model, we calculate the first new weighted information criterion for each candidate model using its corresponding original data of the seven information criteria (AIC, BIC, SBC, PC, JP, GMSEP, and SP) (variables) and its corresponding weights in the following way:

$$\text{PWIC-I} = \frac{1}{a_{11}}(AIC) + \frac{1}{a_{12}}(BIC) + \frac{1}{a_{13}}(SBC) + \frac{1}{a_{14}}(PC)$$
$$+\frac{1}{a_{15}}(JP) + \frac{1}{a_{16}}(GMSEP) + \frac{1}{a_{17}}(SP).$$

where a_{1i} $(i = 1, 2, ..., 7)$ represents the coefficients for the first principal component. Also, we will consider second new weighted information criterion with the same setup as the first new weighted information criterion when we substitute its weights with the first principal component coefficients instead of the inverse of each coefficient of the first principal component in the following way:

$$\text{PWIC-II} = a_{11}(AIC) + a_{12}(BIC) + a_{13}(SBC) + a_{14}(PC)$$
$$+a_{15}(JP) + a_{16}(GMSEP) + a_{17}(SP).$$

where a_{1i} $(i = 1, 2, ..., 7)$ represents the coefficients for the first principal component.

The Simulation Study:

A simulation study of PROC REG's regression model analysis of data was conducted to compare the two new weighted information criteria with the well-known seven model selection criteria in terms of their percentage of number of times that they identify the true model.

Normal data were generated according to all possible regression models that can be constructed of three independent variables X_1, X_2, X_3, (total of 7 models). These regression models are special cases of model (1) with known regression parameters $(\beta_0 = 2, \beta_1 = 3, \beta_2 = 4, \beta_3 = 5)$. There were 14 scenarios to generate data involving two different sample sizes ($n =$ 50, and 100 observations) with all the possible regression models. The independent variables, X_1, X_2, X_3 were drawn from normal distributions with $\mu = 0$ and $\sigma^2 = 4$. The error term of the model was drawn from normal distribution with $\mu = 0$ and $\sigma^2 = 9$. For each scenario, we simulated 500 datasets. SAS code was written to generate the datasets according to the described models using the SAS/IML (SAS Institute Inc., 2008). The algorithm of our approach was applied to each one of the 500 generated data sets with each possible model for each one of the nine information criteria in order to compare their performance. We close this section by commenting on how to choose the number of bootstrap samples R (i.e. the number of times the observed data was bootstrapped) used in the evaluation of the new approach. As R increases, the results of the new weighted information criteria stabilize. Although, choosing a value of R which is smaller than the sample size may result in inaccurate results, choosing a value of R which is larger than the sample size will be wasting of computational time. The values of 50, and 100 were chosen for R to be equal to the two cases of the sample sizes considered in the simulation study.

Results:

Table 1 summarizes results of the percentage of number of times that each criterion selects the true regression model from all possible regression models, when n=50, and R=50. Table 2 summarizes results of the percentage of number of times that each criterion selects the true regression model from all possible regression models, when n=100, and R=100.

The first new weighted information criterion (PWIC-I) shows outstanding performance over all except when it is compared with the SBC criterion in both sample sizes considered. In general, as expected, the performance of most the information criteria improved with increasing sample size n.

Table 1: The Percentage of number of times that each criterion selects the true regression model from the all possible regression models when n=50, and R=50.

The right model	The percent of success (%)								
	PWIC I	PWIC II	AIC	BIC	SBC	PC	JP	GMSEP	SP
X1	77.40	75.60	64.80	71.40	88.80	64.80	64.80	66.80	66.80
X2	78.60	77.40	67.40	74.00	87.80	67.40	67.40	70.40	70.40
X3	78.20	76.40	66.00	73.00	88.20	66.00	66.00	67.80	67.80
X1,X2	88.00	87.80	82.20	86.80	94.40	82.60	82.60	84.20	84.20
X1,X3	88.20	87.20	80.00	85.60	94.20	80.20	80.20	81.60	81.60
X2,X3	88.00	87.80	83.20	86.80	93.00	83.20	83.20	84.20	84.20
X1,X2,X3	100	100	100	100	100	100	100	100	100
Overall percent of success	85.486	84.6	77.657	82.514	92.343	77.743	77.743	79.286	79.286

Table 2: The Percentage of number of times that each criterion selects the true regression model from the all possible regression models when n=100, and R=100.

The right model	The percent of success (%)								
	PWIC I	PWIC II	AIC	BIC	SBC	PC	JP	GMSEP	SP
X1	79.80	78.00	68.80	71.60	92.20	68.80	68.80	69.80	69.80
X2	80.40	77.60	68.00	70.80	92.40	68.00	68.00	69.20	69.20
X3	81.40	80.20	70.00	71.80	92.40	70.00	70.00	71.00	71.00
X1,X2	90.00	86.60	81.80	83.60	97.20	81.80	81.80	83.20	83.20
X1,X3	87.80	88.60	83.80	84.80	95.20	83.80	83.80	84.00	84.00
X2,X3	90.00	89.40	83.40	84.80	96.20	83.40	83.40	84.00	84.00
X1,X2,X3	100	100	100	100	100	100	100	100	100
Overall percent of success	87.057	85.771	79.4	81.057	95.086	79.4	79.4	80.171	80.171

Conclusion:

In our simulation, we considered multiple linear regressions, looking at the performance of two new weighted information criteria for selecting the suitable regression model with two different sample sizes. Overall, the first new weighted criterion (PWIC-I) provided the second best guide to select the suitable model among the nine compared criteria. Thus, the first new weighted criterion (PWIC-I) can be recommended to be considered as reliable criterion.

REFERENCES

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

Aladag C.H., E. Egrioglu, S. Gunay and M.A. Basaran, 2010. Improving weighted information criterion by using optimization, Journal of computational and applied mathematics, 233: 2683-2687.

Amemiya, T., 1985. Advanced Econometrics, Cambridge: Harvard University Press.

Amemiya, T., 1976. Estimation in Nonlinear Simultaneous Equation Models, Paper presented at Institut National de La Statistique et Des Etudes Ecnomiques, Paris,March 10 and published in Malinvaued, E. (ed.), Cahiers Du SeminarireD'econometrie, 19.

Efron, B. and R.J. Tibshirani, 1993. Introduction to the Bootstrap, New York: Chapman and Hall.

Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. J. Amer. Statist. Assoc., 78: 316-331.

Efron, B., 1986. How biased is the apparent error rate of a prediction rule?, J. Amer. Statist. Assoc., 81: 416-470.

Egrioglu, E., C.H. Aladag and S. Gunay, 2008. A new model selection strategy in artificial neural networks, Applied mathematics and computation, 195: 591-597.

Hocking, R.R., 1976. The analysis and selection of variables in linear regression, Biometrics, 32: 1-49.

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

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

Neter, J., M.H. Kutner, C.J. Nachtsheim and W. Wasserman, 1996. Applied Linear Regression Model, (Third Edition). Richard D. Irwin, Inc., Chicago.

SAS Institute Inc, 2008.SAS Online Doc 9.13. 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, Annals of Statistics, 6: 461-464.