JKAU: Sci., **Vol. 21** No. 2, pp: 197-212 (2009 A.D. / 1430 A.H.); DOI: 10.4197 / Sci. 21-2.2

Bootstrap Simulation Procedure Applied to the Selection of the Multiple Linear Regressions

Ali Hussein Al-Marshadi

Department of Statistics, Faculty of Science,
King Abdulaziz University, Jeddah, Saudi Arabia
AALMarshadi@kau.edu

Abstract. This article considers the analysis of multiple linear regressions (MLR) that is used frequently in practice. We propose new approach could be used to guide the selection of the "true" regression model for different sample size in both cases of existing and not existing of multicollinearity, first-order autocorrelation, and heteroscedasticity. We used simulation study to compare eight model selection criteria in terms of their ability to identify the "true" model with the help of the new approach. The comparison of the eight model selection criteria was in terms of their percentage of number of times that they identify the "true" model with the help of the new approach. The simulation results indicate that overall, the new proposed approach showed very good performance with all the eight model selection criteria where the GMSEP, JP, and SP criteria provided the best performance for all the cases. The main result of our article is that we recommend using the new proposed approach with GMSEP, or JP, or SP criteria as a standard procedure to identify the "true" model.

Keywords: Multiple Linear Regression; Information Criteria; Bootstrap Procedure; MCB Procedure.

1. 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:

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

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

We are interested in estimating these, βs, in order to have an equation for predicting a future Y from the associated Xs. The usual way of estimating the βs is the method referred to as ordinary least squares in which we estimate the β_k parameters by b_k that minimize the error sum of squares $SSE = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_{i1} + ... + b_{p-1} X_{i,p-1}))^2$. In general, this is what SAS procedures, PROC GLM, PROC REG, and PROC AUTOREG[1], are set up to do. 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[2]. 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)[3], Sawa's Bayesian Information Criterion (BIC)[4,5], Schwarz's Bayes Information Criteria (SBC)[6], Amemiya's Prediction Criterion (PC)[4,7,8], Final Prediction Error (JP)[9,4], Estimated Mean Square Error of Prediction (GMSEP)[9], and SP Statistic (SP)[9] to guide the selection of the true model[1,2]. Many studies have proposed either new or modified criteria to be used to select the true model. Recently, empirical study was conducted to illustrate the behavior of well-known information criteria in selecting the true regression model for different sample size, intercorrelations, and intracorrelations[10]. Unfortunately, these criteria have low percentage of selecting the true model as 19% of times. Also, he concluded that the sample size, intercorrelations, and intracorrelations have no significant effect on the performance of the information criteria.

Our research objective is proposing and evaluating new approach could be used to guide the selection of the true regression model. Also, our research objective involves comparing eight model selection criteria in terms of their ability to identify the true model with the help of the new approach.

2. Methodology

The REG procedure of the SAS system[1] is a standard tool for fitting data with multiple linear regression models. One of the main reasons that the REG procedure of the SAS system is very popular is the fact that it is a general-purpose procedure for regression. 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[1]:

1. Akaike's Information Criterion[3] (AIC)
2. Sawa's Bayesian Information Criterion[4,5] (BIC)
3. Schwarz's Bayes Information Criteria[6] (SBC)
4. Amemiya's Prediction Criteria[4,7,8] (PC)
5. Final Prediction Error[9,4] (JP).
6. Estimated Mean Square Error of Prediction[9] (GMSEP) and
7. SP Statistics[9] (SP).

One more model selection criteria will be presented that is equal to the average of the previous seven model selection criteria and it will be called Average Information Criterion (AVIC7). Our study concerns with comparing the eight information criteria in terms of their ability to identify the true model with the help of the new approach.

The new approach involves using the bootstrap technique[11,12] and the Multiple Comparisons with the Best (MCB) procedure[13] as tools to help the eight information criterion in identifying the right regression model. The idea of the new approach can be justified and applied in a very general context, one which includes the selection of the true regression model[14]. The idea of using the bootstrap to improve the performance of a model selection rule was introduced by Efron[11,12], and is extensively discussed by Efron and Tibshirani[15].

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

Let the observation vector O_i is defined as follows:

$$O_i = \begin{bmatrix} Y_i & X_{i1} & \dots & X_{i,p-1} \end{bmatrix}', \text{ where } i = 1,2,...,n..$$

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). The properties of the bootstrap when the bootstrap sample size is equal to the original sample size are discussed by Efron and Tibshirani [15].

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*, SP*, and AVIC7* for each model.

3. Repeat steps (1) and (2) (**W**) 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 [1,2]. We will follow the same rule in our approach, but we have the advantage that each information criteria has (**W**) replication values result of the bootstrapping of the observed data (from step (1), (2), and (3)). To make use of this advantage, we propose using MCB procedure [13] to pick the winners (*i.e.* selecting the best set of models or single model if possible), when we consider the bootstrap replicates of the information criteria, which is produced by each of the model, as groups.

3. The Simulation Study

A simulation study of PROC REG's regression model analysis of data was conducted to compare the eight model selection criteria with the new approach in terms of their percentage of number of times that they identify the true model alone.

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 98 scenarios to generate data involving one case where multicollinearity was induced into the data and the other case where no multicollinearity was induced into the data, one case where first-order autocorrelation was induced into the data and the other case where no autocorrelation was induced into the data, one case where heteroscedasticity was induced into the data and the other case where no heteroscedasticity was induced into the data, and three different sample sizes ($n = 15$, 21, and 50 observations) with all the possible regression models (total of 7 models). The independent variables, X_1, X_2, X_3 were drawn from normal distributions with $\mu = 0$ and $\sigma^2 = 4$. In the case where the multicollinearity was induced into the data, we set the correlation, ρ, between X_1 and X_2 to 0.70 using multivariate technique[15]. The error term of the model was drawn from normal distribution with $\mu = 0$ and $\sigma^2 = 9$. In the case where the first-order autocorrelation was induced into the data, the error term of the model was drawn from normal distribution with $\mu = 0$ and $\sigma^2 = 9$ such that $\varepsilon_i = \rho\varepsilon_{i-1} + u_i$; $u_i \sim i.i.d.\, N(0, \sigma^2)$; $i = 1, 2, ..., n$. we set the correlation, ρ, equal to 0.90 using multivariate technique[16]. In the case where the heteroscedasticity was induced into the data, the error term of the model was drawn from normal distribution with $\mu = 0$ and the error variance not being constant over all cases which increase as cases increases such that $i = \sigma_i^2$; $i = 1, 2, ..., n$. For each scenario, we simulated 1000 datasets. SAS (Version 9.1) SAS/IML[1] code was written to generate the datasets according to the described models. The algorithm of our approach was applied to each one of the 1000 generated data sets with each possible model (total of 7 models) for each one of the eight information criteria in order to compare their performance with the new approach. We close this section by commenting on how to choose the number of bootstrap samples **W** (*i.e.* the number of times the observed data was bootstrapped) used in the evaluation of the new approach. As **W** increases, the results of the new approach stabilize. Although, choosing a value of **W** which is too small may result in inaccurate results, choosing a value of **W** which is too large will be wasting of computational time. The values of 10 and

15 were chosen for **W,** since smaller values seemed to marginally diminish the number of correct model selections with the new approach while larger values did not significantly improve the performance of the new approach. The objective of implanting MCB procedure[13] in our new approach is to select models into a subset with a probability of correct selection $p(\text{correct selection})=(1-\alpha)$ that the "best" model is included in the subset where the subset could be single model if possible. The percentage of number of times that the MCB procedure[13] selects the right model alone was reported.

4. Results

Due to space limitations, we present only part of the total simulation results of the 98 scenarios. The complete results are available from the author upon request. Table 1 summarizes results of the percentage of number of times that the procedure selects the true regression model alone from all possible regression models (total of 7 models) for all the eight criteria, when n = 15, and **W** = 10. Table 2 summarizes results of the percentage of number of times that the procedure selects the true regression model alone from all possible regression models (total of 7 models) for all the eight criteria, when multicollinearity was induced into the data, and n = 15, and **W** = 10. Table 3 summarizes results of the percentage of number of times that the procedure selects the true regression model alone from all possible regression models (total of 7 models) for all the eight criteria, when n = 21, and **W** = 10. Table 4 summarizes results of the percentage of number of times that the procedure selects the true regression model alone from all possible regression models (total of 7 models) for all the eight criteria, when multicollinearity was induced into the data, and n = 21, and **W** = 10. The comparisons of Table 1 with Table 2 and Table 3 with Table 4 reveal no significant effect for the multicollinearity on the performance of the eight criteria with the propose approach. Therefore, we restrict our attention to the case where no multicollinearity exists in the data.

Table 5 summarizes results of the percentage of number of times that the procedure selects the true regression model alone from all possible regression models (total of 7 models) for all the eight criteria, when n = 15, and **W** = 15. Table 6 summarizes results of the percentage of number of times that the procedure selects the true regression model alone from

all possible regression models (total of 7 models) for all the eight criteria, when n = 21, and **W** = 15. Table 7 summarizes results of the percentage of number of times that the procedure selects the true regression model alone from all possible regression models (total of 7 models) for all the eight criteria, when n = 50, and **W** = 10. Table 8 summarizes results of the percentage of number of times that the procedure selects the true regression model alone from all possible regression models (total of 7 models) for all the eight criteria, when n = 50, and **W** = 15. The comparisons of Table 1 with Table 5 and Table 3 with Table 6 reveal no significant improving for the performance of the new approach with increasing the value of **W** from 10 to 15. Therefore, we restrict our attention to the case of **W** = 10 on the rest of the simulation.

Table 1. The percentage of number of times that the procedure selects the true regression model alone from the all possible regression models for all the eight criteria when n = 15, W = 10, and (nominal Type I error = 0.05).

The right model	The eight criteria							
	AIC	**BIC**	**SBC**	**PC**	**JP**	**GMSEP**	**SP**	**AVIC7**
X1	86.7%	97.1%	91.3%	96.8%	99.1%	99.5%	99.5%	97.3%
X2	86.7%	96.9%	91.5%	99.1%	99.6%	99.8%	99.8%	98.7%
X3	85.7%	96.1%	89.7%	99.5%	100%	100%	100%	99.3%
X1,x2	96.4%	99.8%	97.1%	100%	100%	100%	100%	100%
X1,X3	96.1%	99.6%	97.1%	99.9%	100%	100%	100%	100%
X2,X3	96.1%	100%	96.2%	100%	100%	100%	100%	100%
X1,X2,X3	100%	100%	100%	100%	100%	100%	100%	100%

Table 2. The percentage of number of times that the procedure selects the true regression model alone from the all possible regression models for all the eight criteria when multicollinearity was induced, n = 15, W = 10, and (nominal Type I error = 0.05).

The right model	The eight criteria							
	AIC	**BIC**	**SBC**	**PC**	**JP**	**GMSEP**	**SP**	**AVIC7**
X1	86.3%	96.7%	91.6%	97.8%	99.2%	99.3%	99.3%	97.2%
X2	86.7%	95.9%	91.0%	99.2%	99.8%	99.8%	99.8%	98.1%
X3	85.1%	96.4%	90.3%	99.7%	100%	100%	100%	99.5%
X1,x2	96.6%	99.7%	97.7%	100%	100%	100%	100%	99.9%
X1,X3	96.0%	99.4%	96.8%	100%	100%	100%	100%	99.9%
X2,X3	95.5%	99.6%	96.8%	100%	100%	100%	100%	99.9%
X1,X2,X3	100%	100%	100%	100%	100%	100%	100%	100%

Table 3. The percentage of number of times that the procedure selects the true regression model alone from the all possible regression models for all the eight criteria when n = 21, W = 10, and (nominal Type I error = 0.05).

The right model	The eight criteria							
	AIC	BIC	SBC	PC	JP	GMSEP	SP	AVIC7
X1	93.5%	97.6%	96.5%	98.7%	100%	99.5%	99.5%	98.2%
X2	92.6%	97.9%	96.5%	99.9%	100%	100%	100%	99.2%
X3	93.3%	97.6%	96.4%	99.7%	100%	100%	100%	99.5%
X1,x2	98.7%	99.5%	98.9%	100%	100%	100%	100%	99.8%
X1,X3	98.4%	99.6%	99.0%	100%	100%	100%	100%	100%
X2,X3	98.5%	100%	98.8%	100%	100%	100%	100%	100%
X1,X2,X3	100%	100%	100%	100%	100%	100%	100%	100%

Table 4. The percentage of number of times that the procedure selects the true regression model alone from the all possible regression models for all the eight criteria when multicollinearity was induced, n = 21, W = 10, and (nominal Type I error = 0.05).

The right model	The eight criteria							
	AIC	BIC	SBC	PC	JP	GMSEP	SP	AVIC7
X1	94.7%	98.0%	97.3%	98.7%	99.7%	99.7%	99.7%	98.2%
X2	94.0%	98.2%	96.8%	99.4%	100%	100%	100%	99.2%
X3	94.2%	98.2%	97.0%	100%	100%	100%	100%	100%
X1,x2	98.8%	99.4%	98.8%	100%	100%	100%	100%	100%
X1,X3	98.4%	99.6%	99.0%	100%	100%	100%	100%	100%
X2,X3	98.2%	99.6%	98.9%	100%	100%	100%	100%	99.9%
X1,X2,X3	100%	100%	100%	100%	100%	100%	100%	100%

Table 5. The percentage of number of times that the procedure selects the true regression model alone from the all possible regression models for all the eight criteria when n = 15, W = 15, and (nominal Type I error = 0.05).

The right model	The eight criteria							
	AIC	BIC	SBC	PC	JP	GMSEP	SP	AVIC7
X1	82.1%	95.2%	88.3%	95.6%	99.0%	99.5%	99.5%	94.6%
X2	79.8%	94.7%	87.6%	97.5%	99.5%	99.6%	99.6%	97.1%
X3	80.2%	94.6%	86.5%	99.1%	100%	100%	100%	98.9%
X1,x2	93.9%	99.1%	95.2%	100%	100%	100%	100%	99.8%
X1,X3	94.1%	98.8%	95.4%	100%	100%	100%	100%	100%
X2,X3	93.0%	98.4%	94.6%	100%	100%	100%	100%	100%
X1,X2,X3	100%	100%	100%	100%	100%	100%	100%	100%

Table 6. The percentage of number of times that the procedure selects the true regression model alone from the all possible regression models for all the eight criteria when n = 21, W = 15, and (nominal Type I error = 0.05).

The right model	The eight criteria							
	AIC	BIC	SBC	PC	JP	GMSEP	SP	AVIC7
X1	91.2%	96.4%	95.2%	97.9%	99.5%	99.6%	99.6%	97.8%
X2	89.8%	96.1%	95.0%	99.4%	100%	100%	100%	98.6%
X3	87.6%	96.3%	93.5%	99.6%	100%	100%	100%	99.4%
X1,x2	98.0%	99.7%	98.6%	100%	100%	100%	100%	99.9%
X1,X3	97.0%	99.4%	98.3%	100%	100%	100%	100%	100%
X2,X3	97.4%	99.5%	98.2%	100%	100%	100%	100%	99.9%
X1,X2,X3	100%	100%	100%	100%	100%	100%	100%	100%

Table 7. The Percentage of number of times that the procedure selects the true regression model alone from the all possible regression models for all the eight criteria when n=50, W=10, and (nominal Type I error=0.05).

The right model	The eight criteria							
	AIC	BIC	SBC	PC	JP	GMSEP	SP	AVIC7
X1	98.8%	99.7%	99.9%	99.7%	100%	100%	100%	99.9%
X2	98.5%	99.2%	99.6%	99.8%	100%	100%	100%	99.8%
X3	99.3%	99.7%	99.8%	100%	100%	100%	100%	100%
X1,x2	99.6%	99.7%	100%	100%	100%	100%	100%	99.9%
X1,X3	99.9%	100%	100%	100%	100%	100%	100%	100%
X2,X3	99.7%	100%	100%	100%	100%	100%	100%	100%
X1,X2,X3	100%	100%	100%	100%	100%	100%	100%	100%

Table 8. The percentage of number of times that the procedure selects the true regression model alone from the all possible regression models for all the eight criteria when n = 50, W = 15, and (nominal Type I error = 0.05).

The right model	The eight criteria							
	AIC	BIC	SBC	PC	JP	GMSEP	SP	AVIC7
X1	98.1%	99.0%	99.5%	99.4	100%	100%	100%	99.5%
X2	96.8%	98.5%	99.5%	100%	100%	100%	100%	99.7%
X3	98.4%	99.5%	99.7%	100%	100%	100%	100%	100%
X1,x2	99.4%	99.5%	99.6%	100%	100%	100%	100%	99.9%
X1,X3	99.8%	99.9%	100%	100%	100%	100%	100%	100%
X2,X3	99.5%	99.9%	99.9%	100%	100%	100%	100%	100%
X1,X2,X3	100%	100%	100%	100%	100%	100%	100%	100%

Table 9 summarizes results of the percentage of number of times that the procedure selects the true regression model alone from all possible regression models (total of 7 models) for all the eight criteria, when first-order autocorrelation was induced into the data, and n = 15, and **W** = 10. Table 10 summarizes results of the percentage of number of times that the procedure selects the true regression model alone from all possible regression models (total of 7 models) for all the eight criteria, when first-order autocorrelation was induced into the data, and n = 21, and **W** = 10. Table 11 summarizes results of the percentage of number of times that the procedure selects the true regression model alone from all possible regression models (total of 7 models) for all the eight criteria, when first-order autocorrelation was induced into the data, and n = 50, and **W** = 10. The comparisons of Table 1 with Table 9, Table 3 with Table 10, and Table 7 with Table 11 reveal no significant effect for the first-order autocorrelation on the performance of the eight criteria with the propose approach.

Table 9. The percentage of number of times that the procedure selects the true regression model alone from the all possible regression models for all the eight criteria when first-order autocorrelation was induced, n = 15, W = 10, and (nominal Type I error = 0.05).

The right model	The eight criteria							
	AIC	BIC	SBC	PC	JP	GMSEP	SP	AVIC7
X1	87.5%	97.5%	92.6%	93.9%	97.9%	98.7%	98.7%	96.3%
X2	85.8%	96.4%	91.7%	97.3%	99.7%	99.7%	99.7%	97.1%
X3	84.5%	97.8%	90.0%	98.2%	100%	100%	100%	98.8%
X1,x2	96.1%	99.5%	96.9%	99.6%	99.9%	100%	100%	99.6%
X1,X3	95.5%	99.2%	97.2%	99.6%	99.6%	99.6%	99.6%	99.5%
X2,X3	97.8%	100%	98.4%	100%	100%	100%	100%	100%
X1,X2,X3	100%	100%	100%	100%	100%	100%	100%	100%

Table 10. The percentage of number of times that the procedure selects the true regression model alone from the all possible regression models for all the eight criteria when first-order autocorrelation was induced, n = 21, W = 10, and (nominal Type I error = 0.05).

The right model	The eight criteria							
	AIC	**BIC**	**SBC**	**PC**	**JP**	**GMSEP**	**SP**	**AVIC7**
X1	93.7%	97.9%	96.9%	97.2%	99.5%	99.6%	99.6%	98.8%
X2	90.9%	96.6%	95.7%	98.0%	100%	100%	100%	98.6%
X3	93.3%	97.4%	96.5%	98.6%	99.8%	99.8%	99.8%	99.2%
X1,x2	98.4%	99.2%	98.8%	99.7%	99.9%	99.9%	99.9%	99.7%
X1,X3	98.1%	99.8%	99.0%	100%	100%	100%	100%	99.9%
X2,X3	98.6%	99.6%	99.1%	100%	99.9%	99.9%	99.9%	99.8%
X1,X2,X3	100%	100%	100%	100%	100%	100%	100%	100%

Table 11. The percentage of number of times that the procedure selects the true regression model alone from the all possible regression models for all the eight criteria when first-order autocorrelation was induced, n = 50, W = 10, and (nominal Type I error = 0.05).

The right model	The eight criteria							
	AIC	**BIC**	**SBC**	**PC**	**JP**	**GMSEP**	**SP**	**AVIC7**
X1	98.3%	99.2%	99.9%	98.6%	99.7%	99.8%	99.8%	99.3%
X2	98.2%	98.7%	99.5%	99.0%	100%	100%	100%	99.6%
X3	98.4%	98.8%	99.5%	99.6%	100%	100%	100%	99.7%
X1,x2	99.6%	99.8%	99.8%	100%	100%	100%	100%	99.9%
X1,X3	99.8%	100%	100%	100%	100%	100%	100%	100%
X2,X3	99.5%	99.6%	100%	100%	100%	100%	100%	100%
X1,X2,X3	100%	100%	100%	100%	100%	100%	100%	100%

Table 12 summarizes results of the percentage of number of times that the procedure selects the true regression model alone from all possible regression models (total of 7 models) for all the eight criteria, when heteroscedasticity was induced into the data, and n = 15, and **W** = 10. Table 13 summarizes results of the percentage of number of times that the procedure selects the true regression model alone from all possible regression models (total of 7 models) for all the eight criteria, when heteroscedasticity was induced into the data, and n = 21, and **W** = 10. Table 14 summarizes results of the percentage of number of times that the procedure selects the true regression model alone from all

possible regression models (total of 7 models) for all the eight criteria, when heteroscedasticity was induced into the data, and n = 50, and **W =** 10. The comparisons of Table 1 with Table 12, Table 3 with Table 13, and Table 7 with Table 14 reveal no significant effect for the heteroscedasticity on the performance of the eight criteria with the propose approach.

Table 12. The percentage of number of times that the procedure selects the true regression model alone from the all possible regression models for all the eight criteria when heteroscedasticity was induced, n = 15, W = 10, and (nominal Type I error = 0.05).

The right model	The eight criteria							
	AIC	BIC	SBC	PC	JP	GMSEP	SP	AVIC7
X1	87.5%	97.1%	92.7%	99.1%	99.9%	99.9%	99.9%	97.5%
X2	87.8%	96.4%	91.5%	99.1%	99.8%	99.8%	99.8%	98.5%
X3	85.4%	97.7%	89.7%	99.7%	100%	100%	100%	99.8%
X1,x2	96.3%	99.6%	97.8%	99.9%	100%	100%	100%	99.7%
X1,X3	95.9%	99.7%	97.6%	100%	100%	100%	100%	99.9%
X2,X3	97.4%	99.7%	97.8%	100%	100%	100%	100%	100%
X1,X2,X3	100%	100%	100%	100%	100%	100%	100%	100%

Table 13. The percentage of number of times that the procedure selects the true regression model alone from the all possible regression models for all the eight criteria when heteroscedasticity was induced, n = 21, W = 10, and (nominal Type I error = 0.05).

The right model	The eight criteria							
	AIC	BIC	SBC	PC	JP	GMSEP	SP	AVIC7
X1	94.2%	97.4%	97.0%	98.5%	99.7%	99.7%	99.7%	98.5%
X2	93.2%	97.1%	96.0%	99.5%	100%	100%	100%	98.7%
X3	96.4%	98.8%	98.7%	99.7%	99.8%	99.8%	99.8%	99.8%
X1,x2	97.4%	99.6%	98.2%	99.9%	100%	100%	100%	99.9%
X1,X3	97.9%	99.8%	98.7%	100%	100%	100%	100%	100%
X2,X3	98.4%	99.3%	99.0%	100%	100%	100%	100%	100%
X1,X2,X3	100%	100%	100%	100%	100%	100%	100%	100%

Table 14. The percentage of number of times that the procedure selects the true regression model alone from the all possible regression models for all the eight criteria when heteroscedasticity was induced, n = 50, W = 10, and (nominal Type I error = 0.05).

The right model	The eight criteria							
	AIC	BIC	SBC	PC	JP	GMSEP	SP	AVIC7
X1	99.6%	99.6%	99.9%	99.7%	100%	100%	100%	99.9%
X2	98.7%	98.7%	99.6%	99.0%	100%	100%	100%	99.6%
X3	99.2%	99.2%	99.6%	99.9%	100%	100%	100%	99.8%
X1,x2	99.9%	99.9%	99.9%	100%	100%	100%	100%	100%
X1,X3	99.8%	99.8%	99.9%	100%	100%	100%	100%	100%
X2,X3	99.8%	99.8%	99.9%	100%	100%	100%	100%	100%
X1,X2,X3	100%	100%	100%	100%	100%	100%	100%	100%

Although the new approach shows very good performance over all with all the criteria for all the cases, it was outstanding with GMSEP[9], JP[9,4], and SP[9] criteria. Also, as expected, the performance of the new approach improved with increasing sample size n. Finally, we applied the proposed approach to real data for a studying 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 15 contains the data obtained in the study. The researcher decided to fit 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[17]. We will apply the proposed approach to select the best model or the best subset of models for this data considering three predictor variables X1, X2, and a cross-product interaction effect X12. Table 16 shows the mean and the standard deviation of the three criteria GMSEP[9], JP[9,4], and SP[9] when **W** = 5 for the four considered models. The MCB procedure[13] selects the best model as the one that has been selected by the researcher but with a cross-product interaction effect X12, with $\alpha = 0.7$.

Table 15. Data for power cells case example.

Cell (i)	1	2	3	4	5	6	7	8	9	10	11
Number of Cycles (Y)	150	86	49	288	157	131	184	109	279	235	224
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	10	10	20	20	20	20	20	30	30	30

Table 16. The mean and the standard deviation of the criteria for the considered models when W = 5.

The considered models	The criteria					
	GMSEP		JP		SP	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
X1	5126.81283	1377.49815	4936.93087	1326.47970	522.175381	140.300737
X2	3190.37903	2454.08272	3072.21684	2363.19076	324.946012	249.952869
X1,x2	1023.67651	742.93274	928.89164	674.14267	104.263348	75.669075
X1,X2,X12	746.49992	500.28423	622.08327	416.90353	76.032399	50.954875

5. Conclusion

In our simulation, we considered multiple linear regressions, looking at the performance of the new proposed approach for selecting the suitable regression model with different cases. Overall, the new approach provided the best guide to select the suitable model. The new approach showed outstanding performance with $GMSEP^{[1]}$, $JP^{[1,2]}$, and $SP^{[1]}$ criteria. Thus, this new approach can be recommended to be used with one of the three mentioned criteria. Note for users of the propose approach: if the MCB procedure suggested the best subset of models contains more than one model, we recommend selecting the true model as the one with a small subset of predictors since the examination of simulation results showed that in this case the other models are overfitted models, *i.e.* model that contains the predictors of the true model, plus any additional predictors. The main result of our article is that the three criteria $GMSEP^{[1]}$, $JP^{[1,2]}$, and $SP^{[1]}$ criteria are competitive in term of their ability to identifying the right model with the help of the new proposed approach even under the violation of regression assumptions such as heteroscedastic regression model, autocorrelated error model of regression, and multicollinearity.

References

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

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

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

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

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

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

[7] **Amemiya, T.,** 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, no. 19 (1976).

[8] **Amemiya, T.,** *Advanced Econometrics*, Cambridge: Harvard University Press (1985).

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

[10] **Al-Subaihi, A. Ali,** A Monte Carlo study of univariate variable selection criteria, *Pak. J. Statist*., **23** (1)**:** 65-81 (2007).

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

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

[13] **Hsu, J.C.,** Constrained simultaneous confidence intervals for multiple comparisons with the best, *Annals of Statistics*, **12:** 1136-1144 (1984).

[14] **AL-Marshadi Ali Hussein**, The new approach to guide the selection of the covariance structure in mixed model, *Research Journal of Medicine and Medical Sciences*, **2** (2)**:** 88-97 (2007).

[15] **Efron, B.** and **Tibshirani, R.J.,** *Introduction to the Bootstrap*, New York: Chapman and Hall (1993).

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

[17] **Neter J., Kutner, H.M.** and **Wasserman, W.,** *Applied Linear Regression Models***,** Richard D. Irwin, (1990).

استخدام أسلوب البوت استراب بطريقة المحاكاة في اختيار أفضل نموذج انحدار خطي متعدد

علي حسين المرشدي

قسم الإحصاء، كلية العلوم، جامعة الملك عبدالعزيز

جدة – المملكة العربية السعودية

المستخلص. تهتم هذه الورقة العلمية بدراسة واحده من أهم الطرق الإحصائية من حيث كثرة استخداماتها في مجالات العلوم، وهي طريقة الانحدار الخطي المتعدد. تتركز دراستنا حول مشكلة عادةً ما تواجة الباحثين في المجالات التطبيقية، وهي تحديد النموذج المناسب. ولقد تم في هذه الورقة تقديم أسلوب جديد قد يساعد في تحديد النموذج المناسب في حالة وجود وعدم وجود مشكلة ال **multicollinearity**، وال **first-order autocorrelation**، وال **heteroscedasticity** مع حجم عينات مختلفة. هدف الدراسة هو مقارنة أداء ثمانية من المقايس التي تستخدم في تحديد النموذج المناسب بمساعدة الأسلوب الجديد. وقد أوضحت النتائج أنه يمكننا أن نوصي باستخدام الأسلوب الجديد مع المقايس **GMSEP, JP, and SP** كطريقة أساسية في تحديد النموذج المناسب.