American Journal of Applied Sciences 6 (6): 1191-1198, 2009
ISSN 1546-9239

Linear Regression Model Selection Based on Robust Bootstrapping Technique

[1]Hassan S. Uraibi, [1]Habshah Midi, [2]Bashar A. Talib and [3]Jabar H. Yousif
[12]Laboratory of Applied and Computational Statistics,
Institute for Mathematical Research
University Putra Malaysia, 43400 UPM, Serdang Selangor, Malaysia
[3]Faculty of Computing and Information Technology,
Sohar University, Sultanate of Oman

Abstract: Problem statement: Bootstrap approach had introduced new advancement in modeling and model evaluation. It was a computer intensive method that can replace theoretical formulation with extensive use of computer. The Ordinary Least Squares (OLS) method often used to estimate the parameters of the regression models in the bootstrap procedure. Unfortunately, many statistics practitioners are not aware of the fact that the OLS method can be adversely affected by the existence of outliers. As an alternative, a robust method was put forward to overcome this problem. The existence of outliers in the original sample may create problem to the classical bootstrapping estimates. There was possibility that the bootstrap samples may contain more outliers than the original dataset, since the bootstrap re-sampling is with replacement. Consequently, the outliers will have an unduly effect on the classical bootstrap mean and standard deviation. **Approach:** In this study, we proposed to use a robust bootstrapping method which was less sensitive to outliers. In the robust bootstrapping procedure, we proposed to replace the classical bootstrap mean and standard deviation with robust location and robust scale estimates. A number of numerical examples were carried out to assess the performance of the proposed method. **Results:** The results suggested that the robust bootstrap method was more efficient than the classical bootstrap. **Conclusion/Recommendations:** In the presence of outliers in the dataset, we recommend using the robust bootstrap procedure as its estimates are more reliable.

Key words: Bootstrap, outliers, robust location, robust standard deviation

INTRODUCTION

Model selection is an important subject in the areas of scientific research, especially in regression predictions. Riadh *et al*.[1] proposed utilizing the bootstrap techniques for model selection. Bootstrap method which was introduced by[2] is a very attractive method because it can be utilized without relying any assumptions on the underlying population. It is a computer intensive method that can replaced theoretical formulation with extensive use of computer. There are considerable papers related to bootstrap method[3-7]. Despite the good properties of the bootstrap method, it suffers numerical instability when outliers are present in the data. The bootstrap distribution might be a very poor estimator of the distribution of the regression estimates because the proportion of outliers in the bootstrap samples can be higher than that in the original data set[4]. Most of the bootstrap techniques use the Ordinary Least Squares (OLS) procedures to estimate the parameters of the model. It is well known that the OLS is extremely sensitive to outliers and will produce inaccurate estimates[8]. In this study, we propose using robust method in which the final solutions are not easily affected by outliers.

MATERIALS AND METHODS

Classical bootstrap based on the fixed-x re-sampling: Consider the general multiple linear regression model with additive error terms:

$$y = X\beta + \varepsilon \quad (1)$$

Where:

y = The n×1 vector of observed values for the response variable
X = The n×p matrix of observed values for the m explanatory variables

The vector β is an unknown p×1 vector of regression coefficients and ε is the n×1 vector of error terms which is assumed to be independent, identically

Corresponding Author: Hassan S. Uraibi, Laboratory of Applied and Computational Statistics

and normally distributed with mean 0 and constant variance, σ^2. In regression setting, there are two different ways of conducting bootstrapping; namely the Random-x Re-Sampling and the fixed-x Re-Sampling which is also refer as bootstrapping the residuals. Riadh *et al*.[1] use the random-x Re-Sampling together with the OLS method in their bootstrap algorithm. In this study, the fixed-x Re-Sampling technique with OLS method is adopted. We call this estimator the Classical Bootstrap fixed-x Resampling Method (CBRM).

The CBRM procedure as enumerated by Efron and Tibshirani[3] is summarized as follows:

Step 1: Fit the OLS to the original sample of observations to get $\hat{\beta}$ and the fitted values $\hat{y}_i = f(x_i, \hat{\beta})$.

Step 2: Obtain the residuals $\varepsilon_i = y_i - \hat{y}_i$ and giving probability 1/n for each ε_i value.

Step 3: Draw n bootstrap random sample with replacement, that is ε_i^b is drawn from ε_i and attached to $\hat{y}_i$ to get a fixed-x bootstrap values y_i^b where $y_i^b = f(x_i, \hat{\beta}) + \varepsilon_i^b$.

Step 4: Fit the OLS to the bootstrapped values y_i^b on the fixed X to obtain $\hat{\beta}^b$.

Step 5: Repeat steps 3 and 4 for B times to get $\hat{\beta}^{b1},\ldots,\hat{\beta}^{bB}$ where B is the bootstrap replications.

According to Imon and Ali[5], there is no general agreement among statisticians on the number of the replications needed in bootstrap. B can be as small as 25, but for estimating standard errors, B is usually in the range of 25-250. They point out that for bootstrap confidence intervals, a much larger values of B is required which normally taken to be in the range of 500-10,000.

Riadh *et al*.[1] pointed out that the bootstrap standard deviation can be estimated as follows:

$$\hat{\sigma}_{boot} = \left[\frac{1}{B-1}\sum_{b=1}^{B}(MSR^{(b)} - \mu_{boot})^2\right]^{\frac{1}{2}} \quad (2)$$

where, MSR is the mean squared residual denoted as:

$$MSR = \sum_{i=1}^{n}\left(e_i^b\right)^2 / n \quad (3)$$

and

$$e^b = (y_i - \hat{y}_i^b), \quad b = 1,2,\ldots\ldots,B \quad i = 1,2,\ldots\ldots,n$$

$$\mu_{boot} = \frac{1}{B}\sum_{b=1}^{B}(MSR^{(b)} \quad b = 1,2,\ldots\ldots,B \quad (4)$$

The drawback of using the classical standard deviation and the classical mean in estimating the bootstrap scale and location in Eq. 2 and 4 is that it is very sensitive to outliers. As an alternative, a robust location and scale estimates which are less affected by outliers are proposed. The robust bootstrap location and robust scale estimates are given by (5) and (6) as follows:

$$Med_{boot} = Median(MSR^{(b)}) \quad (5)$$

$$MAD_{boot} = Med\,|\,MSR^{(b)} - Med_{boot}\,|\,/0.6745, b = 1,2,\ldots B \quad (6)$$

Robust Bootstrap Based on the Fixed-x Resampling (RBRM): Unfortunately, many researchers are not aware that the performance of the OLS can be very poor when the data set for which one often makes a normal assumption, has a heavy-tailed distribution which may arise as a result of outliers. Even with single outlier can have an arbitrarily large effect on the OLS estimates[8]. It is now evident that the bootstrap estimates can be adversely affected by outliers, because the proportion of outliers in the bootstrap samples can be higher than that in the original data[4]. These situations are not desirable because they might produce misleading results. An attempt has been made to make the bootstrap estimates more efficient. We propose to modify the CBRM procedure by using some logical procedure with robust Least Trimmed Squares (LTS) estimator, so that outliers have less influence on the parameter estimates. We call this estimator as Robust Bootstrap fixed-x Re-Sampling Method (RBRM).We summarized the RBRM as follows:

Step 1: Fit the LTS to the original sample of observations to get $\hat{\beta}$ and the fitted values $\hat{y}_i = f(x_i, \hat{\beta})$.

Step 2: Obtain the residuals $\varepsilon_i = y_i - \hat{y}_i$ and giving probability 1/n for each ε_i value. Standardized the residuals and identify them as outliers if the absolute value of the standardized residuals larger than three.

Step 3: Draw n bootstrap random sample with replacement, that is ε_i^b is drawn from ε_i and attached to get a fixed-x bootstrap values y_i^b where $y_i^b = f(x_i, \hat{\beta}) + \varepsilon_i^b$.

At this step, we built a dynamic subroutine program for the detection of outliers. This program has the ability to identify a certain percentage of outliers in each bootstrap sample.

Step 4: Fit the LTS to the bootstrapped values y_i^b on the fixed X to obtain $\hat{\beta}^b$. The percentage of outlier that should be trimmed depend step 2.

Step 5: Repeat steps 3 and 4 for B times to get $\hat{\beta}^{b1},\ldots,\hat{\beta}^{bB}$ where B is the bootstrap replications.

The bootstrap scale and location estimates in Eq. 2 and 4 are based on Mean Squared Residuals which is sensitive to outliers. We propose to replace the Mean Squared Residual (MSR) with a more robust measure that is the Median Squared Residual (RMSR). The propose robust bootstrap location and robust bootstrap scale estimates are as follows:

$$RMed_{boot} = Median(RMSR^{(b)}) \quad (7)$$

$$RMAD_{boot} = Med\,|\,RMSR^{(b)} - RMed_{boot}\,|\,/0.6745 \quad (8)$$

where, RMSR is the Median Squared Residual and for each observation i, i = 1,2,…….., n and for each b = 1, 2,…, B; compute:

$$RMSR^{(b)} = Median(e_i^{*(b)})^2 \quad (9)$$

$$e_i^{*(b)} = (y_i - \hat{y}_i^{(b)}) \quad (10)$$

We also would like to compare the performance of (7) and (8) with the classical formulation of bootstrap standard deviation and location but based on Median Squared Residuals instead of Mean Squared Residuals. These measures are given by:

$$\hat{\sigma}_{boot}(\hat{\theta}^*) = \left[\frac{1}{B-1}\sum_{b=1}^{B}(RMSR^{(b)} - \mu_{boot})^2\right]^{\frac{1}{2}} \quad (11)$$

$$\mu_{boot} = \frac{1}{B}\sum_{b=1}^{B}(RMSR^{(b)}) \quad (12)$$

The RBRM procedures commences with estimating the robust regression parameters using LTS method which trim some of the values from both size. This means that, some values from the data which are labeled as outliers are deleted. In this situation $\hat{\beta}_{LTS}$, will be either larger or smaller than $\hat{\beta}$. In step 2, outliers might be present and it can be the candidate to be selected in Step 3. Since we consider sampling with replacement, each outlier might be chosen more than once. Consequently, there is a possibility that a bootstrap samples may contain more outliers than the original sample. We try to overcome this problem by determining the alpha value based on the percentage of outliers in the bootstrap resamples which are detected in Step 3. In this respect, we develop a dynamic detection subroutine program that can detect the proportion of outliers in each bootstrap resample. Step 4 of RBRM includes the computation of y bootstap by using the LTS based on the first three logical steps. The LTS is expected to be more reliable than the OLS when outliers are present in the data, because it is based on robust method which is not sensitive to outliers. As mentioned earlier, the number of outliers that should be trimmed in the LTS procedure depends on the alpha value that correspond to the percentage of outliers detected. In this way, the effect of outliers is reduced. According to Riadh *et al*.[1], the best model to be selected among several models, is the one which has the smallest value of location and scale estimates or the minimum scale estimate.

RESULTS

Several well known data sets in robust regression are presented to compare the performance of the CBRM and the RBRM procedures. Comparisons between the estimators are based on their bootstrap locations and scales estimates. We have performed many examples and due to space constraints, we include only three real examples and one simulated data. The conclusions of other results are consistent and are not presented due to space limitations. All computations are done by using S-Plus®6.2 for windows with Professional Edition.

Hawkins, Bradu and Kass Data: Hawkin *et al*.[8] constructed an artificial three-predictor data set containing 75 observations with 10 outliers in both of the spaces (cases 1-10), 4 outliers in the X-space (cases 11-14) and 61 low leverage inliers (cases 15-75). Most of the single case deletion identification methods fail to identify the outliers in Y-space though some of them point out cases 11-14 as outliers in the Y-space.

We consider four models:

$$M1: P = 3: Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \qquad \text{True}$$

$$M2: P = 2: Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

$$M3: P = 2: Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \varepsilon_i$$

$$M4 \rightarrow: P = 2: Y_i = \beta_0 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

Table 1: CBRM results of Hawkins data

	Mean		Median	
Models	μ_{boot}	$\hat{\sigma}_{boot}$	med_{boot}	MAD_{boot}
M1	0.3154	0.0346	0.3133	0.0303
M2	0.2975	0.0399	0.2937	0.0372
M3	0.3472	0.0438	0.3474	0.0360
M4	0.3224	0.0395	0.3179	0.0368

Table 2: RBRM results of Hawkins data

	Mean		Median	
Models	μ_{boot}	$\hat{\sigma}_{boot}$	med_{boot}	MAD_{boot}
M1	5.1866	0.3155	5.1053	0.2280
M2	7.0047	0.7206	6.8892	0.6631
M3	5.7588	0.4730	5.6654	0.4079
M4	5.0593	0.2247	5.0036	0.1371

Table 3: CBRM results of Stackloos data

	Mean		Median	
Models	μ_{boot}	$\hat{\sigma}_{boot}$	med_{boot}	MAD_{boot}
M1	10.26	1.25	10.009	1.0425
M2	10.33	1.12	10.075	0.9359
M3	16.83	1.91	16.351	1.4574
M4	25.85	15.7	25.164	2.0667

Table 1 and 2 show the estimated bootstrap location and scale estimates based on CBRM and RBRM procedures.

Stackloss data[8]: The Stackloss data is a well known data set which is presented by Brownlee[9]: The data describe the operation of plant for the Oxidation of ammonia to nitric acid and consist of 21 four-dimensional observations. The Stackloss (y) is related to the rate of operation (x1), the cooling water inlet temperature (x2) and the acid concentration (x3). Most robust statistics researchers concluded that observations 1, 2, 3 and 21 were outliers.

We consider four models:

$$M1: P = 3: Y_i = \beta_0 + \beta_1 X_{1i} + \beta_1 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \quad \text{True}$$

$$M2: P = 2: Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

$$M3: P = 2: Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \varepsilon_i$$

$$M4 \rightarrow: P = 2: Y_i = \beta_0 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

Table 3 and 4 show the bootstrap location and scale estimates of the Stackloss data based on CBRM and RBRM procedures.

Table 4: RBRM results of Stackloos data

	Mean		Median	
Models	μ_{boot}	$\hat{\sigma}_{boot}$	med_{boot}	MAD_{boot}
M1	0.9293	1.6751	1.1717	0.4196
M2	0.6609	0.5061	1.0605	0.4684
M3	1.3661	1.1465	2.0388	0.5655
M4	4.6048	6.0732	6.8214	1.7793

Coleman data[8]: This data which was studied by Coleman *et al.*[10] contains information on 20 schools from the Mid-Atlantic and new England states. Mosteller and Tukey[11] analyzed this data with measurements of five independent variables. The previous study refer observations 3, 17 and 18 as outliers[8].

We consider fifteen models as follow:

$$M1: P = 5: Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i \quad \text{True}$$

$$M2: P = 4: Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$$

$$M3: P = 4: Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon_i$$

$$M4: p = 4: Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

$$M5: P = 4: Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

$$M6: P = 4: Y_i = \beta_0 + \beta_2 X_{2i} + \beta_3 X_{3i} + + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

$$M7: P = 3: Y_i = \beta_0 + \beta_2 X_{2i} + + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

$$M8: P = 3: Y_i = \beta_0 + \beta_2 X_{2i} + + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$$

$$M9: p = 3: Y_i = \beta_0 + \beta_3 X_{3i} + + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

$$M10: P = 3: Y_i = \beta_0 + \beta_1 X_{1i} + + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$$

$$M11: P = 3: Y_i = \beta_0 + \beta_1 X_{1i} + + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

$$M12: P = 3: Y_i = \beta_0 + \beta_1 X_{1i} + + \beta_2 X_{2i} + \beta_5 X_{5i} + \varepsilon_i$$

$$M13: P = 3: Y_i = \beta_0 + \beta_1 X_{1i} + + \beta_2 X_{2i} + \beta_4 X_{4i} + \varepsilon_i$$

$$M14: P = 3: Y_i = \beta_0 + \beta_1 X_{1i} + + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

$$M15: P = 3: Y_i = \beta_0 + \beta_2 X_{2i} + + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$$

Table 5 and 6 show the results of CBRM and RBRM of the Coleman data.

Table 5: CBRM results of Coleman data

	Mean		Median	
Models	μ_{boot}	$\hat{\sigma}_{boot}$	med_{boot}	MAD_{boot}
M1	264.79	407.86	114.31	155.05
M2	1083.70	614.98	1004.40	585.65
M3	910.05	938.52	662.40	741.07
M4	6314.00	4238.10	5568.10	4299.70
M5	696.64	651.45	537.50	527.02
M6	605.49	684.37	402.30	490.25
M7	8010.50	4690.30	7303.80	4549.20
M8	291.17	277.38	225.70	249.09
M9	1161.50	873.58	995.40	712.73
M10	1093.60	634.07	1019.40	588.30
M11	21783.40	6146.60	21262.40	6069.00
M12	1797.40	2013.60	1042.90	1292.80
M13	4843.40	2489.50	4502.70	2305.60
M14	178.50	196.57	105.05	126.48
M15	1801.50	904.77	1668.20	821.28

Table 6: RBRM results of Coleman data

	Mean		Median	
Models	μ_{boot}	$\hat{\sigma}_{boot}$	med_{boot}	MAD_{boot}
M1	0.4296	0.2362	0.4036	0.1607
M2	0.9157	0.8943	0.7362	0.3878
M3	1.5126	1.2118	1.2296	0.6643
M4	5.1409	3.1416	4.5042	1.8494
M5	0.9022	0.5792	0.8250	0.3828
M6	0.5587	0.4421	0.4472	0.1964
M7	5.9627	2.8438	5.7292	2.7066
M8	266.06	269.86	228.71	155.20
M9	1.0832	0.5817	1.0361	0.4596
M10	0.9274	0.5748	0.8587	0.3959
M11	6.4584	3.5923	6.3301	2.0415
M12	7.4923	3.5314	7.4160	2.4681
M13	4.9836	3.7238	4.2552	2.0028
M14	2.0063	0.9448	1.9293	0.8440
M15	1.2564	1.3637	1.0154	0.5616

Simulation study: A simulation study similar to that of Riadh *et al*.[1] is presented to assess the performance of the RBRM procedure. Consider the problem of fitting a linear model:

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_p X_p + \varepsilon_i$$

$$B_0 = (x_1^{(i)}, x_1^{(i)}, y_i) \quad i = 1, 2, ..., 500$$

In this study, we simulate a data set by putting:

$$x_1^{(i)} \sim N(0.6, 25)$$
$$x_2^{(i)} \sim N(-0.1, 0.81)$$
$$y = 2 + 0.7x_1^{(i)} + 0.5x_2^{(i)} + \varepsilon_i$$

where, ε_i is a random variable which possesses the distribution N(0, 0.04).

Table 7: CBRM results of simulated data

		Mean B = 500		Median B = 500	
Out.	M	μ_{boot}	$\hat{\sigma}_{boot}$	med_{boot}	MAD_{boot}
5%	M1	0.024	0.001	0.024	0.001
	M2	0.112	0.004	0.111	0.004
	M3	5.908	0.221	5.923	0.227
10%	M1	0.029	0.002	0.029	0.002
	M2	0.132	0.003	0.132	0.003
	M3	7.200	0.253	7.145	0.233
15%	M1	0.097	0.011	0.097	0.012
	M2	0.203	0.014	0.202	0.015
	M3	8.845	0.412	8.767	0.420
20%	M1	0.147	0.112	0.124	0.040
	M2	0.241	0.090	0.206	0.056
	M3	10.108	0.494	10.110	0.508

Table 8: RBRM results of simulated data

		Mean B = 500		Median B = 500	
Out.	M	μ_{boot}	$\hat{\sigma}_{boot}$	med_{boot}	MAD_{boot}
5%	M1	5.512	0.026	5.505	0.020
	M2	5.870	0.021	5.863	0.015
	M3	19.369	0.079	19.343	0.058
10%	M1	10.829	0.053	10.816	0.044
	M2	11.092	0.043	11.080	0.030
	M3	25.266	0.102	25.230	0.071
15%	M1	14.891	0.072	14.871	0.056
	M2	15.087	0.058	15.068	0.039
	M3	28.933	0.116	28.896	0.089
20%	M1	18.596	0.089	18.573	0.075
	M2	18.746	0.078	18.720	0.053
	M3	30.283	0.111	30.251	0.088

Then we started to contaminate the residuals. At each step, one 'good' residual was deleted and replaced with contaminated residual. The contaminated residual were generated as N (10, 9). We consider 5, 10, 15 and 20% contaminated residuals and three models:

- Model1 $M_1 : P = 2 : Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ True
- Model2 $M_2 : P = 1 : Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$
- Model3 $M_3 : P = 1 : Y_i = \beta_0 + \beta_2 X_{2i} + \varepsilon_i$

Table 7 and 8 show the results of CBRM and RBRM procedures. Graphical displays are used to explain why a particular model is selected. We only present the results for Model 1-3 of the simulated data at 5% outliers due to space limitations. The residuals plot before the bootstrap procedure is shown in Fig. 1. Figure 2-4 shown the box-plot of the MSR boot for Model 1-3 while Fig. 5-7 exemplified the box-plot of the RMSR for Model 1-3.

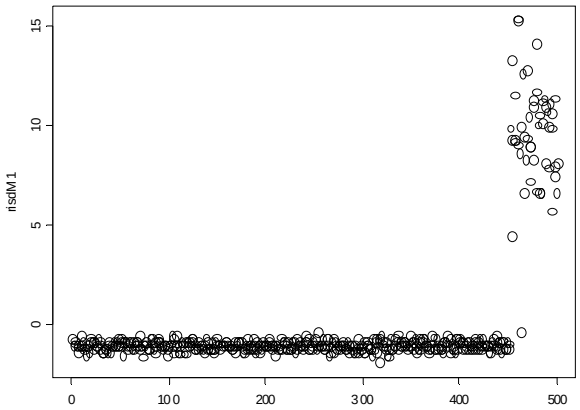

Fig. 1: Residuals before bootstrap

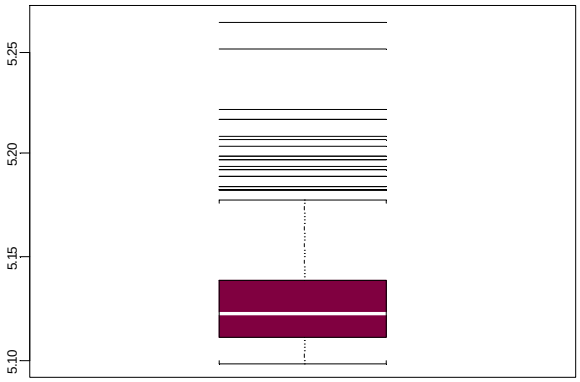

Fig. 2: The Box-plot for the MSR boot M1

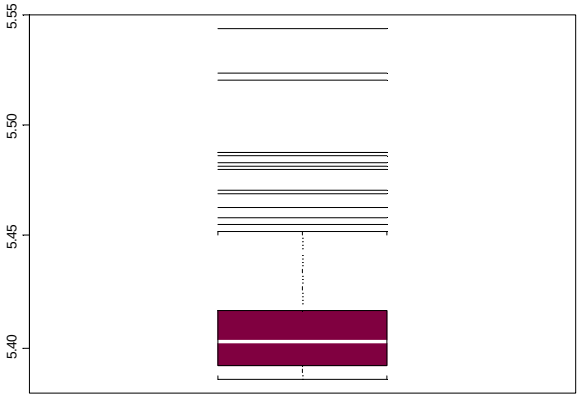

Fig. 3: The Box-plot for the MSR boot M2

DISCUSION

Let us first focus our attention to the results of the Hawkin's data for the CBRM procedure, presented in Table 1. Among the 4 models considered, the bootstrap location and scale estimate of Model 4 is the smallest.

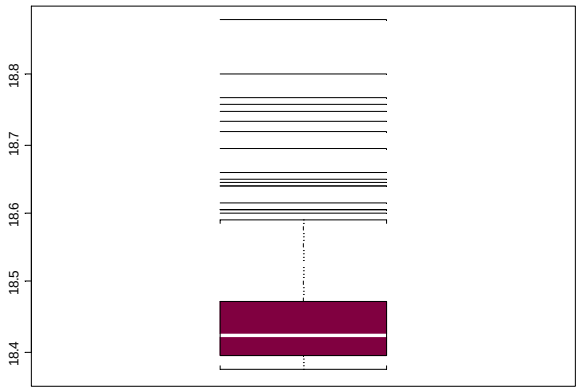

Fig. 4: The Box-plot for the MSR boot M3

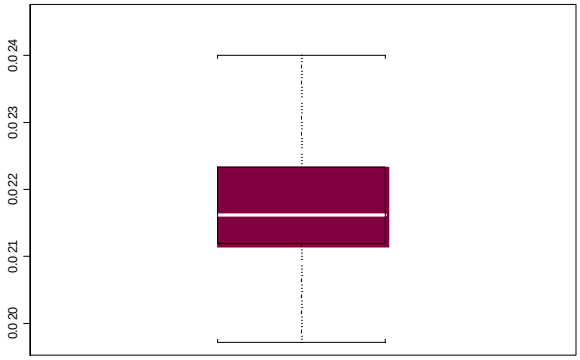

Fig. 5: The Box-plot for the RMSR for M1

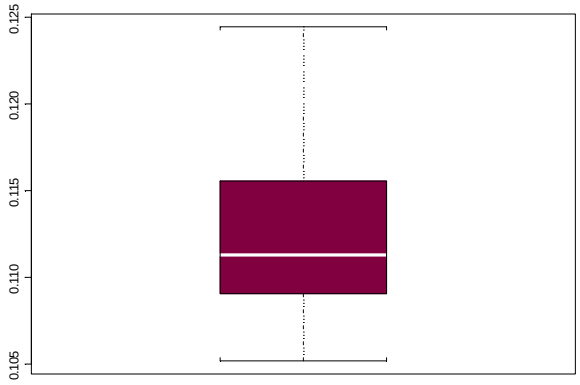

Fig. 6: The Box-plot for the RMSR for M2

It is important to note that the scale estimate-median based is smaller than the scale estimates-mean based. This indicates that the formulation of scale estimates based on median is more efficient than when based on mean. In this respect, the CBRM suggests that Model 4 is the best model. However, the results of Table 2 of the RBRM procedure signify that Model 1 is the best

model. It can be seen that the scale estimate for Model 1 which is based on median is the smallest among the four models. It is interesting to note here that the overall results indicate that the scale estimate-median based of the RBRM procedure is the smallest. Thus, the RBRM based on median has increased the efficiency of the estimates.

It can be observed from Table 3 and 4 of the Stackloss data that the scale estimates based on median is more efficient than when based on mean for both CBRM and RBRM procedures. Similarly, the RBRM-median based has the least value of the scale estimates. The CBRM indicates that Model 2 is the best model while the RBRM suggest that Model 1 is the best model. Nonetheless, the model selection based on RBRM-median based is more efficient and more reliable. These are indicated by its location and scale estimates which are the smallest among the models considered.

By looking at Table 5 of the Coleman data reveals that Model 14 of the CBRM is the best model, evident by the smallest value of the location and scale estimates. In fact, the location and scale estimate of Model 14 which is based on median is smaller than when based on mean. The results of RBRM in Table 6 signify that Model 1's location and scale estimates is the smallest among the 15 models considered. For this model, the RBRM median-based is more efficient than the RBRM mean-based. These are indicated by its location and scale estimates which are smaller than the RBRM mean based.

The results of the simulated data in Table 7 shows that Model 2 is the best model for all outlier percentage levels because the scale estimate of Model 2 is the smallest compared to other models. Nonetheless, the RBRM results of Table 8 suggest that Model 1 is the best model.

Similar to that of the Hawkin, Stackloss and Coleman data, the RBRM median-based is more efficient than the RBRM mean-based. In fact the scale estimates of the RBRM median-based are remarkably smaller than the RBRM mean-based for all outlier percentage levels. From the results of the simulation study indicates that the RBRM median-based is more efficient and reliable procedure.

Here, we would like to explain further why Model 2 is selected by considering only at 5% outliers due to space constraint. By looking at Fig. 1, it is obvious that there are 5% outliers in the residuals before the bootstrap is employed. It can be seen from Fig. 2-4 that the number of outliers of the MSR in Fig. 2 equal to 18 while only 14 in Fig. 3.

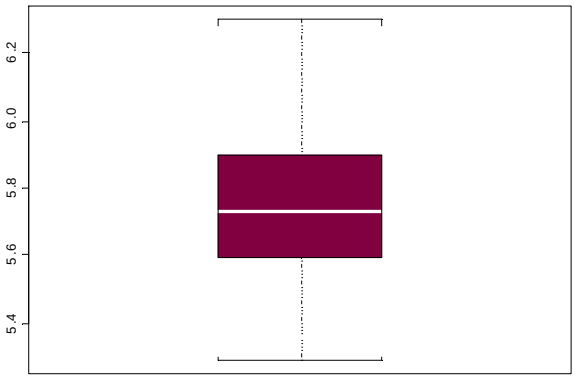

Fig. 7: The Box-plot for the RMSR for M3

The median of the MSR in Fig. 4 is very large compared to Fig. 2 and 3. Among the three models in Fig. 2-4, the CBRM chooses Model 2 as the best model because the proportion of outliers of the MSR is less than the other two models.

On the other hand, the RBRM select M1 as the best model. By comparing Fig. 5-7 with Fig. 2-4, it can be seen that there is no outlier in the distribution of the Median Squared Residuals when we employed RBRM method, while apparent outliers are seen in the distribution of the Mean Squared Residuals, when the CBRM are employed. In this situation, the RBRM has an attractive feature. Among the three models being considered, the RMSR bootstrap resample of Model 1 is more efficient as it is more compact in the central region compared to the other two models. In this situation, Model 1 is recommended as the RMSR is more consistent and more efficient.

CONCLUSION

In this study, we propose a new robust bootstrap method for model selection criteria. The proposed bootstrap method attempts to overcome the problem of having more outliers in the bootstrap samples than the original data set. The RBRM procedure develops a dynamic subroutine program that is capable of detecting certain percentage of outliers in the data. The results indicate that the RBRM consistently outperformed the CBRM procedure. It emerges that the best model selected always corresponds to the RBRM-median based that has the least bootstrap scale estimate. Hence, utilizing the RBRM median-based in the model selection, can improve substantially the accuracy and the efficiency of the estimates. Thus the RBRM median-based is more reliable for linear regression model selection.

REFERENCES

1. Kallel, R., M. Cottrell and V. Vigneron, 2002. Bootstrap for neural model selection. J. Neurocomput., 48: 175-183. DOI: 10.1016/S0925-2312(01)00650-6
2. Efron, B. and R.J. Tibshiriani, 1994. An Introduction to the Bootstrap. 6th Edn., CRC Press, Tyloe and Francis Group, USA., ISBN: 9780412042317, pp: 456.
3. Efron, B. and R. Tibshiriani, 1986. Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. J. Stat. Sci., 1: 54-77. DOI:10.1214/ss/1177013815
4. Barrera, M.S. and R.H. Zamar, 2002. Botstrapping robust estimates of regression. J. Ann. Stat., 30: 556-582. DOI:10.1214/aos/1021379865
5. Imon, A.H.M.R and Ali.M.M, 2005. Bootstrapping regression residuals. J. Korean Data Inform. Sci. Soc., 16: 665-682. http://www.yeskisti.net/yesKISTI/InfoSearch/LinkingImgView.jsp?koi=KISTI1.1003/JNL.JAKO200522941493176&dbt=JAKO
6. Liu, Y.R., 1988. Bootstrap procedures under some Non i.i.d Models. Ann. Stat., 16: 1696-1708. DOI: 10.1214/aos/1176351062
7. Sahinler, S. and D. Topuz, 2007. Bootstrap and jackknife resampling algorithms for estimation of regression parameters. J. Applied Quant. Methods, 2: 188-199. http://jaqm.ro/issues/volume-2,issue-2/pdfs/sahinler_topuz.pdf
8. Rousseeuw, P.J. and L.M. Annick, 2003. Robust Regression and Outlier Detection. Illustrated Edn., Wiley-IEEE, ISBN: 0471488550, pp: 360.
9. Brownlee, K.A., 1965. Statistical Theory and Methodology in Science and Engineering. 2nd Edn., John Wiely and Sons, New York, ISBN: 0471113557, pp: 596.
10. Coleman, J., *et al*., 1979. Equality of Educational Opportunity. Reprent Edn., Ayer Co Pub. ISBN: 0405120885, pp: 737.
11. Mosteller, F. and J.W. Tukey, 1977. Data Analysis and Regression: A Second Course in Statistics. Illustrated Edn. Addison-Wesely, Reading, MA., ISBN: 020104854X, pp: 588.