European Journal of Scientific Research
ISSN 1450-216X Vol.32 No.3 (2009), pp.277-287
© EuroJournals Publishing, Inc. 2009
http://www.eurojournals.com/ejsr.htm

Dynamic Robust Bootstrap Method Based on LTS Estimators

Habshah Midi
Laboratory of Applied and Computational Statistics
Institute For Mathematical Research [INSPEM]
University Putra Malaysia
43400 UPM Serdang, Selangor Malaysia
E-mail: habshahmidi@gmail.com

Hassan S. Uraibi
Laboratory of Applied and Computational Statistics
Institute For Mathematical Research [INSPEM]
University Putra Malaysia
43400 UPM Serdang, Selangor Malaysia

Bashar A. Talib
Laboratory of Applied and Computational Statistics
Institute For Mathematical Research [INSPEM]
University Putra Malaysia
43400 UPM Serdang, Selangor Malaysia

Abstract

The applications of bootstrap methods in regression analysis have drawn much attention to the statistics practitioners because of some practical reasons. In order to make reliable inferences about the parameters of a model, require that the parameter estimates are normally distributed. Nevertheless, in real situations, many estimates are not normal and the use of bootstrap method is more appropriate as it does not rely on the normality assumption. It is now evident that the presence of outliers have an unduly effect on the bootstrap estimates. There is a possibility that the bootstrap samples may contain more outliers than the original sample. In this paper, we propose a robust bootstrap algorithm based on Least Trimmed of Squares (LTS) estimator which will be unaffected in the presence of outliers. We call this method Dynamic Robust Bootstrap-LTS based (DRBLTS) because here we have employed the LTS estimator in the modified bootstrap algorithm. The performance of the DRBLTS is evaluated by real data sets and simulation study. The numerical examples indicate that the DRBLTS is more efficient than the other methods.

Keywords: Bootstrap samples, Outliers, LTS, Bias estimation and RMSE

1. Introduction

The Ordinary Least Squares (OLS) method is the most popular technique in statistics and it is often use to estimate the parameters of a model because of tradition and ease of computation. According to Gauss-Marcov Theorem, the OLS estimators, in the class of unbiased linear estimators, have minimum variance that is they are best linear unbiased estimator (BLUE). Nonetheless, the OLS estimates are easily affected by the presence of outliers and will produce inaccurate estimates (see Huber (1973), Rousseeuw and Leroy (2003) and Maronna et. al (2006)). Outliers are observations which are markedly different from the bulk of the data or from the pattern set by the majority of the observations. In a regression problem, observations corresponding to excessively large residuals are treated as outliers. According to Hampel et. al (1986), the existence of 1-10% outliers in a routine data is rather rule than exceptions. Midi et. al (2009) pointed out that the detection of outliers is crucial due to their responsibility for misleading conclusion about the fitting of multiple linear regression model, causing multicollinearity problems, masking and swamping of outliers. Hampel (1971) pointed out that even one single outlier can have an arbitrary large effect on the OLS estimates. In this connection, he introduced a breakdown point (*BP*) as the smallest percentage of outliers that can cause an estimator to take an arbitrary large value. The robustness of each estimator is measured by the BP. An estimator becomes more robust as the value of *BP* increases. The BP of the OLS estimator is 0% which implies that it can be easily affected by a single outlier. As an alternative a robust methods which are much less affected by outliers are put forward (see Huber (1973), Chatterjee and Hadi(1988), Barnett and Lewis (1994), Rousseeuw and Van Driessen(1999), Rousseeuw and Leroy(2003) and Maronna et. al(2006)). However, most robust methods are relatively difficult and computationally complicated.

A better approach is to use a bootstrap method which was introduced by Efron (1979) with the basic idea of generating a large number of sub-samples by randomly drawing observations with replacement from the original dataset. These sub-samples are then being termed as bootstrap samples and are used to recalculate the estimates of the regression coefficients. Bootstrap method has been successful in attracting statistics practitioners as its usage does not rely on the normality assumption. An interesting feature of the bootstrap method is that it can provide the standard errors of any complicated estimator without requiring any theoretical calculations. These interesting properties of the bootstrap method have to be traded off with computational cost and time. There are considerable papers that deal with bootstrap methods (see Efron and Tibshiriani(1986), Efron and Tibshiriani(1993)). The classical bootstrap technique usually based on the OLS estimates. It is now evident that the presence of outliers could make a great deal of damage to the bootstrap inferential procedure (see Liu(1988), Barrera and Zamar (2002), Imon and Ali(2005) and Willems and Aelts (2005). The reason for inaccurate conclusion is that we suspect that there is a possibility that the bootstrap samples may contain more outliers than the original samples(see Barrera and Zamar(2002) and Willems and Van Aelts(2005)). The problem becomes worse when the percentage of outliers is more than the breakdown point (BP). In this situation the model structure may change and will affect the bootstrap estimates. By ignoring the outliers in the bootstrap samples and analyzing data using the classical bootstrap method may produce sub-optimal or even invalid inferential statements and inaccurate predictions. Unfortunately, many statistics practitioners are not aware of this consequence.

In this paper, we propose a robust bootstrap algorithm that we called Dynamic Robust Bootstrap LTS-based method that combined the bootstrapping algorithm and the Least Trimmed of Squares (LTS) estimator. First, we attempt to develop a mechanism to detect the percentage of outliers, denoted as α .The advantage of knowing the α value in each bootstrap sample is that the LTS will trim the correct number of outliers if real outliers exist. Any bootstrap samples with α value greater than the BP will be deleted and a new bootstrap sample will be generated until we get the desired number of bootstrap replications, which is referred as B. The main idea is to identify the percentage of outliers in each bootstrap sample and use the LTS to estimate the model parameters of each bootstrap sample. We anticipate that the DRBLTS will provide more accurate results as the LTS will trim the exact percentage of outliers and any bootstrap sample having values of α greater than the BP will be eliminated from the bootstrap samples and will be replaced with relatively ‘good’ samples. In this paper, a ‘good bootstrap’ sample is the one in which the percentage of outliers in each bootstrap sample is less than the BP of the LTS estimator.

2. Material and Method

In regression setting, there are two different ways of conducting bootstrapping; namely the random X-resampling and the fixed X-resampling which is also refer as bootstrapping the residuals. The later is the most popular technique of bootstrapping in linear regression. In this paper we will consider this technique in the multiple linear regression model with additive error terms

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

where y is the $n \ x \ 1$ vector of observed values for the response variable and X is the $n \ x \ p$ matrix of observed values for the k explanatory variables. The vector β is an unknown $p \ \text{x} \ 1$ vector of regression coefficients and ε is the $n \ \text{x} \ 1$ vector of error terms which is assumed to be independent, identically and normally distributed with mean 0 and constant variance, σ^2.

2.1. Bootstrap Based on the OLS (BOLS)

The OLS estimates of β is given by $\hat{\beta} = (X^T X)^{-1} X^T Y$. Since its origination, the classical bootstrap relies on the OLS estimates to acquire the residuals of the original data. The bootstrap samples are then obtained by re-sampling the residuals from the original regression. We will summarize this algorithm as follows;

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

Step 2: Get the residuals ε_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^* is drawn from ε_i and attached to $\hat{y}_i$ to get a fixed- x bootstrap values y_i^* where $y_i^* = f(x_i, \hat{\beta}_{ols}) + \varepsilon_i^*$.

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

Step 5: Repeat steps 3 and 4 for B times to get $\hat{\beta}_{ols}^{*1}, \ldots, \hat{\beta}_{ols}^{*B}$ where B is the bootstrap replications.

2.2. Robust Bootstrap Based on LTS (RBLTS)

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. To overcome this problem, we propose to modify the BOLS algorithm by substituting the OLS with a robust estimator with high BP. In this paper, we consider the LTS estimator which was proposed by Rousseeuew (1984) and have a high BP which is equal to [{(n-p)/2+1}/n]. $\hat{\beta}_{LTS}$ is obtained by minimizing $\sum_{i=1}^{h} \hat{\varepsilon}_{(i)}^2$ where $\hat{\varepsilon}_{(i)}$ is the i-th ordered residual. This technique trims a certain percentage of outliers (α) in the data. For a trimming percentage of α, Rousseeeuw and Leroy (1987) suggested choosing $h = [n/2] + [(p+1)/2]$ where p is the number of parameters. The advantage of using the LTS is that we can control the level of trimming which depend on the suspected percentage of outliers. If we suspect the data contains nearly 10% outliers then the LTS will trim 10% of that outliers from the data. It is important to note here, that in the LTS routine in S-PLUS, one may choose the default α value or specify the exact α value based on the percentage of outliers in the original data. The former and the later LTS bootstrapping methods are referred as RBLTS2 and RBLTS1, respectively. The main difference between the two bootstrap methods is that the RBLTS2 uses the default value of α which is equal to 0.10 while RBLTS1 utilizes α values specified in the original sample. The following algorithm describes the RBLTS procedure;

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

Step 2: Obtain the residuals ε_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^* is drawn from ε_i and attached to $\hat{y}_i$ to get a fixed- x bootstrap values y_i^* where y_i^*= $f(x_i, \hat{\beta}_{Lts})$ + ε_i^*.

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

Step 5: Repeat steps 3 and 4 for B times to get $\hat{\beta}^{*1},\ldots,\hat{\beta}^{*B}$ where B is the bootstrap replications.

It is important to point out that some statistics practitioners are not aware that there is a possibility that some bootstrap samples have percentage of outliers greater than the percentage of outliers in the original data. If this problem is not treated properly, the bootstrap estimates will be affected. In this respect, The RBLTS1 and RBLTS2 have some shortcomings. In the minimizing technique, in some occasions these two estimators may over trim or down trim the proportion of outliers in the bootstrap samples because this techniques do not have a detection algorithm to check the percentage of outliers in each bootstrap sample. Consequently, they may over trim or down trim the percentage of outliers in each bootstrap sample, unnecessarily. For better understanding, the explanations of the said weaknesses are illustrated with example. For this purpose we will consider the Stackloss Data (see Rousseeuw and Leroy (2003)). This data set with three independent variables contains 21 observations has been extensively analyzed by several authors (see Atkinson (1985), Rousseew and Leroy(2003) and Midi et. al (2009)). They reported that this data set has four outliers (cases 1,3,4 and 21) or 19.2 % outliers. Let us first focus our attention to the RBLTS1 in which the LTS algorithm in S-Plus Routine specifies the exact α value of the original data. For illustration, 21 observations are sampled with replacement from the original sample. We repeated for 1000 bootstrap samples and record the observed number of bootstrap samples out of 1000 which have certain percentage of outliers. In this example, any bootstrap sample which has percentage of outliers in the range of say, 0<$\alpha \leq$ 5, 5<$\alpha \leq$ 10, 10<$\alpha \leq$ 15, 15<$\alpha \leq$ 20, 20<$\alpha \leq$ 25, 25<$\alpha \leq$ 30, will be declared to have percentage of outliers as 5%, 10%, 15%, 20%, 25%, 30%, respectively. In this respect, the percentage of outliers in the original data is 20% which is equivalent to 5. Table 1 illustrates the number of outliers (*NO*), number of bootstrap samples (*NBS*), observed number of outliers (*ONO*), number of trimmed observations (*NT*) and the Difference between *NT* and *ONO*, denoted as Diff.

Positive Diff indicates the number of clean observations that are trimmed unnecessarily (over trimmed clean observations) while negative Diff indicates the number of remaining outliers that still exist after the trimmings (down trim the number of outliers).

Value of Diff equals to 0 suggests that the LTS trimmed the exact or the true number of outliers. This is the desired situation.

For illustrations, we refer to the second row of Table 1 which corresponds to *NO*=2 and *NBS*=75. There are 75 bootstrap samples out of 1,000, each containing 2 outliers.

The number of outliers in 75 samples is given by;

ONO= NO x NBS = 2x75=150

For RBLTS1, the trimming depends on the number of outliers in the original sample where in this example equals to 5. In this respect, the number of trimmed observations is given by *NT*= 5x*NBS*=5x75=375.

Diff=NT-ONO=375-150=+225

For RBLTS2, the trimming depends on the default value of alpha of the S-Plus Routine which is equals to 0.10 where in this example equals to 3. In this respect, the number of trimmed observations is given by *NT*= 3x*NBS*=3x268=804.

Diff=NT-ONO=804-536=+268

The breakdown point (*BP*) for this data is [{(n-p)/2+1}/n]= [(21-4)/2+1]/21= 47.5% and the number of outliers corresponds to *BP*= *BP*x21$\approx$10.

Table 1: Some Results of RBLTS1 and RBLTS2 bootstrap re-samples of Stackloss data

N=21 B=1000 **Outliers %(NO)**	**RBLTS1**				**RBLTST2**			
	NBS	***ONO***	***NT***	*Diff*	*NBS*	***ONO***	***NT***	***Diff***
0%	13	0	65	+65	125	0	375	+375
5%=2	75	150	375	+225	268	536	804	+268
10%=3	153	459	765	+306	268	804	804	0
15%=4	193	772	965	+193	205	820	615	-205
20%=5	213	1065	1065	0	91	455	273	-182
25%=6	170	1020	850	-170	31	186	93	-93
30%=7	108	756	540	-216	12	84	36	-48
35%=8	53	424	265	-159	0	0	0	0
40%=9	10	90	50	-40	0	0	0	0
BP=10	10	100	5	-50	0	0	0	0
>BP (at least 11)	2	>=22	10	>=(-12)	0	0	0	0
Total	**1000**	**>=4858**	**5000**	**>= (+142)**	**1000**	**2885**	**3000**	**+115**

Based on Table 1, we observe that many bootstrap samples have percentage of outliers in each sample greater than the percentage of outliers in the original data. The results also reveal that RBLTS1 and RBLTS2 will over trim (unnecessary trimming of good observations) and down trim (do not trim all outliers) the number of outliers in each bootstrap sample. These are indicated by the positive and negative values of the Diff. In addition to that, several bootstrap samples contain percentage of outliers in each bootstrap sample which is larger than the number of outliers associate with the BP values.

The same process was repeated for several sets of 1,000 bootstrap samples and other real data and due to space limitations, the results are not reported here. However, the results are consistent where we encounter many bootstrap samples with percentage of outliers in each sample larger than the percentage of outliers of the original data. Similarly, several samples contain percentage of outliers in each sample larger than the BP value.

This illustrations suggest that there is still problem when applying the RBLTS1 and RBLTS2. In order to rectify this problem, it is necessary to identify the exact number of outliers in each bootstrap sample so that LTS will trim the exact or correct number of outliers.

2.3. Dynamic Robust Bootstrap for LTS [DRBLTS]

The RBLTS may be a good alternative to the BOLS if the percentage of outliers in each bootstrap sample is equal to the percentage of outliers in the original data. Nevertheless, in real situation the proportion of outliers in the bootstrap samples can be higher than that in the original data. It is now evident that the bootstrap estimates can be adversely affected by outliers (Hampel, 1971). These situations are not desirable because they might produce inefficient results. An attempt has been made to make the RBLTS estimates more robust. We propose to modify the RBLTS procedure by first identifying the exact number of outliers in the original data and hence specified the appropriate value of α thus obtained. Once the value of α is determined, use the LTS with the specified α value to estimate the parameters of the model. Consequently, compute the residuals and identified the residuals as outliers if the absolute value of the standardized residuals are larger than three. Just like the previous two bootstrapping methods, the bootstrap samples are then taken at random with replacement from the original data. It is important to note here, that at this stage, those bootstrap samples which contain the percentages of outliers in each sample larger than the BP will be omitted and replaced with a new sample. The same process is repeated until the desired bootstrap iterations are obtained. We called this method as the Dynamic Robust Bootstrap based on LTS (DRBLTS) and expect it is more robust than other methods discussed in this paper. We summarized the DRBLTS as follows;

Step 1: Identify the exact number of outliers in the data by Least Median of Squares (see Rousseeuw and Leroy (2003)). Consequently α is determined.

Step 2: Fit the LTS to the original sample of observations to get $\hat{\beta}_{LTS}$ and the fitted values $\hat{y}_i$ = $f(x_i, \hat{\beta})$. Here in the S-PLUS routine we specify the exact value of α found in Step 1 for the LTS algorithm.

Step 3. 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 4: Draw n bootstrap random sample with replacement, that is ε_i^* is drawn from ε_i and attached to $\hat{y}_i$ to get a fixed-x bootstrap values y_i^* where y_i^* = $f(x_i, \hat{\beta}) + \varepsilon_i^*$. At this step, we built a dynamic subroutine program for the detection of outliers based on the standardized residuals. This program has the ability to identify a certain percentage of outliers in each bootstrap sample.

Step 5: Fit the LTS to the bootstrapped values y_i^* on the fixed X to obtain $\hat{\beta}^*$. The percentage of outliers that should be trimmed depend on step 4.

Step 6: Repeat steps 3, 4 and 5 for B times to get $\hat{\beta}^{*1}, \ldots, \hat{\beta}^{*B}$ where B is the bootstrap replications. Any bootstrap sample which has percentage of outliers larger than *BP* will be deleted and will not be counted as bootstrap sample and will be replaced with a new sample. The percentage of outliers in each bootstrap sample is determined from Step 4.

According to Imon and Ali(2005), 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.

2.4. Assessment of the Bootstrap Methods

The performance of the four methods are evaluated based on the bias and RMSE. A ‘good’ method is the one which has the smallest bias and smallest RMSE. The biases and the RMSE’s of the four bootstrap methods can be computed by employing the following formula. Let us first illustrate the computation of the BOLS bias, variance, MSE and RMSE. The calculation of other estimates is the same, just substitute the BOLS with the desired estimator, such as the BRLTS1,BRLTS2 and DRBLTS in the corresponding formula.

The BOLS estimate of β is given by $\hat{\beta}_{(bols)} = \frac{\sum_{b=i}^{B} \hat{\beta}_{ols}^{*b}}{B}$ which yielded the bootstrap bias = $\hat{\beta}_{(bols)} - \hat{\beta}_{ols}$. The bootstrap variance are obtained by taking the diagonal values of the covariance matrix $Cov(\hat{\beta}_{(bols)}) = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\beta}^{*b} - \hat{\beta}_{bols})(\hat{\beta}^{*b} - \hat{\beta}_{bols})^T$. The mean-squared error (MSE) is given by $MSE(\hat{\beta}_{bols}) = (bias)^2 + \mathrm{var}(\hat{\beta}_{bols})$. Consequently, the root mean squared error is given by

$$\sqrt{MSE(\hat{\beta}_{bols})}$$

3. Results and Discussion

In this section, several numerical examples and some simulation studies are presented to illustrate the performance of the four estimators.

3.1. Numerical Examples

Several well known data sets in robust regression are presented to compare the performance of the BOLS, BLTS1, BLTS2 and DRBLTS. Comparison between the estimators are based on the bootstrap bias and RMSE. All computations were done on windows with professional edition using S-PLUS @6.2.

3.1.1. Hawkins, Bradu and Kass [1984]

Rousseeuw and Leroy (2003) 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 fit a linear model as follows:

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

Table 2: Average, bias and RMSE of bootstrap estimates of Hawkins Data

Methods	$\bar{\hat{\beta}}_0$	$\bar{\hat{\beta}}_1$	$\bar{\hat{\beta}}_2$	$\bar{\hat{\beta}}_3$	*bias* $(\hat{\beta}_0)$	*bias* $(\hat{\beta}_1)$	*bias* $(\hat{\beta}_2)$	*bias* $(\hat{\beta}_3)$	*RMSE* $(\hat{\beta}_0)$	*RMSE* $(\hat{\beta}_1)$	*RMSE* $(\hat{\beta}_2)$	*RMSE* $(\hat{\beta}_3)$
BOLS	-1.01	0.71	0.00	0.00	-2.01	-0.29	-1.00	-1.00	2.01	0.29	1.00	1.00
RBLTS1	-0.190	0.14	0.05	-0.08	0.045	0.00	0.02	0.00	0.05	0.00	0.02	0.00
RBLTS2	-1.006	0.16	0.20	0.18	-0.771	0.02	0.16	0.25	0.78	0.02	0.16	0.25
DRBLTS	**-0.228**	**0.14**	**0.04**	**-0.07**	**0.010**	**0.00**	**0.00**	**0.00**	**0.01**	**0.00**	**0.000**	**0.00**

The results in Table 2 show that the BOLS has the largest bias and RMSE compared to other estimators. Although the bias and the RMSE of the RBLTS1 and RBLTS2 are relatively smaller than the BOLS, their performances are inferior than the DRBLTS. It is evident from the results that the DRBLTS has the least bias and RMSE, followed by RBLTS1 and RBLTS2.

3.1.2. Stackloss Data

It is a well known data set presented by Brownlee (1965). The data describe the operation of plant for the Oxidation of ammonia to nitric acid and consist of 21 four-dimensional observations. This data presents the stackloss [y] corresponding to three predictor variables namely the rate of operation [x1], the cooling water inlet temperature [x2], and the acid concentration [x3]. As already been mentioned, some researchers reported that this data contains four outliers. Some researchers claimed that this data contains five outliers.

We fit a linear model as follows:

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

Table 3: Average, bias and RMSE of bootstrap estimates of Stackloss Data

Methods	$\bar{\hat{\beta}}_0$	$\bar{\hat{\beta}}_1$	$\bar{\hat{\beta}}_2$	$\bar{\hat{\beta}}_3$	*bias* $(\hat{\beta}_0)$	*bias* $(\hat{\beta}_1)$	*bias* $(\hat{\beta}_2)$	*bias* $(\hat{\beta}_3)$	*RMSE* $(\hat{\beta}_0)$	*RMSE* $(\hat{\beta}_1)$	*RMSE* $(\hat{\beta}_2)$	*RMSE* $(\hat{\beta}_3)$
BOLS	-43.39	1.00	0.01	0.00	-6.19	0.13	-0.41	0.08	6.19	0.13	0.41	0.08
RBLTS1	-38.20	0.81	0.64	-0.08	-1.00	-0.06	0.22	0.00	1.00	0.06	0.22	0.00
RBLTS2	-40.16	0.82	0.73	-0.08	-2.96	-0.05	0.31	0.00	2.96	0.05	0.31	0.00
DRBLTS	**-36.90**	**0.87**	**0.40**	**-0.08**	**0.30**	**0.00**	**-0.02**	**0.00**	**0.30**	**0.00**	**0.02**	**0.00**

Let us now focus to the results in Table 3 for Stackloss data. As can be expected, similar results are obtained for Stackloss data which is known to have several outliers. We can see that the bias and RMSE of the DRBLTS are remarkably the smallest among the four estimators. From these results, it seems that the BOLS is the least efficient estimator with the RBLTS2 being the next least efficient estimator. Again, it can be seen that the DRBLTS consistently has the least bias and RMSE compared to other estimators. Clearly from the these two examples suggest that the BOLS is easily affected by outliers. We have not pursued the analysis of the examples to a final conclusion, but a reasonable interpretation up to this point is that the DRBLTS is the least affected estimator.

3.1.3. Simulation Study

A simulation study is presented to further assess the performance of the DRBLTS estimator. We consider the similar model use by Riadth et. al (2002) in his simulation study. First we generate 25 observations according to linear relation

$$y = 2 + 0.7x_1^{(i)} + 0.5x_2^{(i)} + \varepsilon_i$$

where

$$x_1^{(i)} \sim N(0.6,25)$$

$$x_2^{(i)} \sim N(-0.1,0.81)$$

ε_i is drawn from normal distribution, that is $\varepsilon_{\mathrm{i}}\ \square\ \mathrm{N}(0, 0.04)$. Then contamination of the data was commenced. At each step, one ‘good’ residual was deleted and replaced with a bad data point. The contaminated residuals were generated from normal distribution, that is $\varepsilon_{\mathrm{i}}\ \square\ \mathrm{N}(10, 9)$.

Table 4-6 present the biases and RMSE’s of the four methods with varying sample size 25, 50, 100 and 500. The results are for B=500. For the clean data (with no outliers), all four methods are fairly close to each other with respect to the bias and the RMSE. However, as the percentage of outliers increases, the BOLS immediately affected by outliers. The biases and RMSEs of the BOLS is the largest among the four estimators. The performance of the RBLTS2 is slightly better than the RBLTS1 up to 10% outliers. It is interesting to point out that the RBLTS1 is slightly better than the RBLTS2 for the case with slightly above 10% outliers.

On the other hand, the RMSE of the DRBLTS estimates is consistently the smallest as the percentage of outliers increases. From these results, it seems that the BOLS is very sensitive to the presence of outliers, with the RBLTS1 being the next most sensitive followed by the RBLTS2 when the percentage of outliers is up to 10%. Otherwise, the RBLTS2 is more sensitive to outliers than RBLTS1. The DRBLTS is hardly affected by the outliers, as shown by the values of the biases and RMSE which are consistently the smallest. The DRBLTS estimates emerge to be conspicuously more efficient than the other estimators. The results seem to be consistent in all 500 trials and each sample, size 25, 50, 100, 500.

Table 4: Average, bias and RMSE of bootstrap estimates of simulation data when n=25

Outliers	Methods	$\bar{\hat{\beta}}_0$	$\bar{\hat{\beta}}_1$	$\bar{\hat{\beta}}_2$	$Bias(\hat{\beta}_0)$	$Bias(\hat{\beta}_1)$	$Bias(\hat{\beta}_2)$	$RMSE(\hat{\beta}_0)$	$RMSE(\hat{\beta}_1)$	$RMSE$ $(\hat{\beta}_2)$
0%	BOLS	2.019	0.697	0.561	0.019	-0.003	0.061	0.019	0.003	0.061
	RBLTS1	2.015	0.698	0.561	0.015	-0.002	0.061	0.015	0.002	0.061
	RBLTS2	1.986	0.700	0.563	-0.014	0.00	0.063	0.014	0.00	0.063
	DRBLTS	1.983	0.699	0.555	-0.017	-0.001	0.055	0.017	0.001	0.055
5%	BOLS	2.750	0.760	0.435	0.750	0.060	-0.064	0.750	0.060	0.064
	RBLTS1	2.267	0.724	0.368	0.267	0.024	-0.132	0.267	0.024	0.132
	RBLTS2	2.110	0.709	0.489	0.110	0.009	-0.011	0.110	0.009	0.011
	DRBLTS	1.987	0.698	0.546	-0.013	-0.002	0.046	0.013	0.002	0.046
10%	BOLS	3.713	0.672	0.784	1.713	-0.028	0.284	1.713	0.028	0.284
	RBLTS1	2.374	0.720	0.474	0.374	0.020	-0.026	0.374	0.020	0.026
	RBLTS2	2.338	0.709	0.470	0.338	0.009	-0.030	0.338	0.009	0.030
	DRBLTS	1.973	0.701	0.545	-0.027	0.002	0.045	0.027	0.002	0.045
15%	BOLS	3.985	0.804	1.086	1.985	0.104	0.586	1.985	0.104	0.586
	RBLTS1	2.294	0.783	0.600	0.294	0.083	0.100	0.294	0.083	0.100
	RBLTS2	2.533	0.804	0.663	0.533	0.104	0.163	0.533	0.104	0.163
	DRBLTS	1.988	0.702	0.536	-0.012	0.002	0.036	0.012	0.002	0.036
20%	BOLS	4.300	0.656	0.597	2.300	-0.044	0.097	2.300	0.044	0.097
	RBLTS1	2.260	0.742	0.446	0.260	0.042	-0.054	0.260	0.042	0.054
	RBLTS2	2.711	0.722	0.314	0.711	0.022	-0.186	0.711	0.022	0.186
	DRBLTS	1.985	0.706	0.562	-0.015	0.006	0.062	0.015	0.006	0.062

Table 5: Average, bias and RMSE of bootstrap estimates of simulation data when n=50

Outliers	Methods	$\bar{\hat{\beta}}_0$	$\bar{\hat{\beta}}_1$	$\bar{\hat{\beta}}_2$	$Bias(\hat{\beta}_0)$	$Bias(\hat{\beta}_1)$	$Bias(\hat{\beta}_2)$	$RMSE(\hat{\beta}_0)$	$RMSE(\hat{\beta}_1)$	$RMSE$ $(\hat{\beta}_2)$
0%	BOLS	1.968	0.699	0.462	-0.032	-0.001	-0.038	0.032	0.001	0.038
	RBLTS1	1.971	0.700	0.460	-0.029	-0.001	-0.040	0.029	0.001	0.040
	RBLTS2	1.964	0.701	0.473	-0.036	0.002	-0.027	0.036	0.002	0.027
	DRBLTS	1.967	0.700	0.461	-0.032	-0.001	-0.039	0.322	0.001	0.039
5%	BOLS	1.968	0.699	0.462	-0.032	-0.001	-0.038	0.032	0.001	0.038
	RBLTS1	1.971	0.700	0.460	-0.029	-0.001	-0.040	0.029	0.001	0.040
	RBLTS2	1.964	0.701	0.473	-0.036	0.002	-0.027	0.036	0.002	0.027
	DRBLTS	1.967	0.700	0.461	-0.032	-0.001	-0.039	0.322	0.001	0.039
10%	BOLS	2.592	0.674	0.601	0.592	-0.026	0.101	0.592	0.026	0.101
	RBLTS1	2.150	0.691	0.477	0.150	-0.009	-0.023	0.150	0.009	0.023
	RBLTS2	1.991	0.695	0.470	-0.009	-0.004	-0.030	0.009	0.004	0.030
	DRBLTS	1.958	0.700	0.453	0.042	0.000	0.047	0.042	0.000	0.047
15%	BOLS	3.042	0.650	0.784	1.042	-0.051	0.284	1.042	0.051	0.284
	RBLTS1	2.156	0.684	0.510	0.156	-0.016	0.009	0.156	0.016	0.009
	RBLTS2	2.170	0.676	0.540	0.170	-0.024	0.040	0.170	0.024	0.040
	DRBLTS	1.954	0.700	0.462	-0.046	0.000	-0.038	0.046	0.000	0.038
20%	BOLS	3.789	0.607	0.487	1.789	-0.093	-0.012	1.789	0.093	0.012
	RBLTS1	2.190	0.684	0.531	0.190	-0.016	0.031	0.190	0.016	0.031
	RBLTS2	2.685	0.660	0.664	0.685	-0.040	0.164	0.685	0.040	0.164
	DRBLTS	1.939	0.701	0.466	-0.061	0.001	-0.034	0.061	0.001	0.034

Table 6: Average, bias and RMSE of bootstrap estimates of simulation data when n=100

Outliers	**Methods**	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$Bias(\hat{\beta}_0)$	$Bias(\hat{\beta}_1)$	$Bias(\hat{\beta}_2)$	$RMSE(\hat{\beta}_0)$	$RMSE(\hat{\beta}_1)$	$RMSE(\hat{\beta}_2)$
0%	BOLS	1.993	0.701	0.531	-0.007	0.001	0.031	0.007	0.001	0.031
	RBLTS1	1.993	0.701	0.533	-0.007	0.001	0.033	0.007	0.001	0.033
	RBLTS2	1.994	0.701	0.527	-0.006	0.001	0.027	0.006	0.001	0.027
	DRBLTS	1.991	0.701	0.531	-0.009	0.001	0.031	0.009	0.001	0.031
5%	BOLS	2.571	0.741	1.055	0.571	0.041	0.555	0.571	0.041	0.555
	RBLTS1	2.051	0.706	0.564	0.061	0.006	0.064	0.061	0.006	0.064
	RBLTS2	1.993	0.701	0.525	-0.007	0.001	0.025	0.007	0.001	0.025
	DRBLTS	1.992	0.701	0.540	-0.007	0.001	0.040	0.007	0.001	0.040
10%	BOLS	3.014	0.751	0.986	1.014	0.051	0.486	1.014	0.051	0.486
	RBLTS1	2.080	0.711	0.548	0.080	0.010	0.048	0.080	0.010	0.048
	RBLTS2	2.080	0.710	0.535	0.080	0.010	0.035	0.080	0.010	0.035
	DRBLTS	1.996	0.703	0.532	-0.004	0.003	0.032	0.004	0.003	0.032
15%	BOLS	3.492	0.796	1.085	1.492	0.096	0.585	1.492	0.096	0.585
	RBLTS1	2.076	0.718	0.496	0.076	0.017	-0.004	0.076	0.017	0.004
	RBLTS2	2.331	0.750	0.575	0.331	0.050	0.075	0.331	0.050	0.075
	DRBLTS	1.997	0.703	0.529	-0.003	0.003	0.029	0.003	0.003	0.029
20%	BOLS	3.911	0.808	1.148	1.911	0.108	0.648	1.911	0.108	0.648
	RBLTS1	2.083	0.703	0.478	0.083	0.003	-0.022	0.083	0.003	0.022
	RBLTS2	2.718	0.751	0.697	0.718	0.051	0.197	0.718	0.051	0.197
	DRBLTS	2.004	0.700	0.524	0.004	0.000	0.024	0.004	0.000	0.024

Table7: Average, bias and RMSE of bootstrap estimates of simulation data when n=500

Outliers	**Methods**	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$Bias(\hat{\beta}_0)$	$Bias(\hat{\beta}_1)$	$Bias(\hat{\beta}_2)$	$RMSE(\hat{\beta}_0)$	$RMSE(\hat{\beta}_1)$	$RMSE(\hat{\beta}_2)$
0%	BOLS	2.010	0.700	0.492	0.010	0.000	-0.008	0.010	0.000	0.008
	RBLTS1	2.010	0.700	0.492	0.010	0.000	-0.008	0.010	0.000	0.008
	RBLTS2	2.012	0.701	0.470	0.012	0.001	-0.030	0.012	0.001	0.030
	DRBLTS	2.012	0.700	0.492	0.012	0.000	-0.008	0.012	0.000	0.008
5%	BOLS	2.488	0.710	0.351	0.488	0.007	-0.149	0.488	0.007	0.149
	RBLTS1	2.040	0.702	0.501	0.040	0.002	0.001	0.040	0.002	0.001
	RBLTS2	2.016	0.701	0.479	0.016	0.002	-0.021	0.016	0.002	0.021
	DRBLTS	2.013	0.701	0.489	0.013	0.001	-0.011	0.013	0.001	0.011
10%	BOLS	2.995	0.724	0.315	0.996	0.024	-0.185	0.996	0.024	0.185
	RBLTS1	2.047	0.704	0.489	0.047	0.004	-0.011	0.047	0.004	0.011
	RBLTS2	2.040	0.703	0.488	0.040	0.003	0.012	0.040	0.003	0.012
	DRBLTS	2.015	0.700	0.492	0.015	0.000	-0.008	0.015	0.000	0.008
15%	BOLS	3.538	0.683	0.297	1.538	-0.017	-0.203	1.538	0.017	0.203
	RBLTS1	2.051	0.704	0.500	0.051	0.004	0.001	0.051	0.004	0.001
	RBLTS2	2.333	0.718	0.492	0.333	0.018	-0.008	0.333	0.018	0.008
	DRBLTS	2.014	0.702	0.495	0.014	0.002	-0.005	0.014	0.002	0.005
20%	BOLS	4.062	0.671	0.431	2.062	-0.029	-0.096	2.062	0.029	0.069
	RBLTS1	2.050	0.706	0.497	0.050	0.006	-0.003	0.050	0.006	0.003
	RBLTS2	2.818	0.693	0.347	0.818	-0.007	-0.173	0.818	0.007	0.153
	DRBLTS	2.014	0.700	0.502	0.014	0.001	0.002	0.014	0.001	0.002

4. Conclusion

The empirical studies suggest that the BOLS is the better choice than the other three estimators for a cleaned data. Nonetheless, its performance was inferior to the RBLTS1, RBLTS2 and DRBLTS when contamination occurred in the data. The results seem to suggest that the DRBLTS is the most efficient bootstrap estimator when outliers are presence in the data. Hence, it should provide robust alternative to the classical bootstrap method.

References

[1] Amado, C. and Pires, A.M. 2004. " Robust bootstrap with non random weights based on the influence function", Communications in Statistics, *Simulation and Computation*, 33, pp.377-396.

[2] Atkinson, A.C. 1985. "Plots, Transformations, and Regression", Oxford: Oxford University Press. Oxford,

[3] Barnett V. and Lewis T. 1994. "Outliers in Statistical Data", 3rd ed., New York: John Wiley.

[4] Brownlee, K.A. 1965. "Statistical Theory and Methodology in Science and Engineering", 2nd Edn.,New York: John Wiley and Sons.

[5] Efron, B. and Tibshiriani, R. 1986. "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *J. Stat. Sci*., 1, pp.54-77.

[6] Efron,B. and Tibshiriani, R.J. 1993. "An Introduction to the Bootstrap",6th ed., Chapman and Hall.

[7] Chatterjee, S. and Hadi, A. 1988. "Sensitivity analysis in Linear Regression", 1st ed., NewYork: John Wiley.

[8] Hampel, F.R., 1971. "A general Definition of Qualitative robustness, *The Annals of Statistics., 42,pp.1887-1896.*

[9] Hampel, F,R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A., 1986 " Robust Statistics:The Approach Based on Influence Functions", New York: John Wiely and Sons.

[10] Huber, P. J., 1973. "Robust Regression: Asymptotic, Conjectures and Monte Carlo", *The Annals of Statistics*, 1, pp.799-821.

[11] Imon, A.H.M.R, and Ali.M.M, 2005. "Bootstrapping regression residuals", *J.Korean Data Inform.Sci*.Soc.,16, 665-682.

[12] Liu,Y,R., 1988. "Bootstrap procedures under some Non i.i.d Models", *Ann. Statist*., 16,pp.1696-1708.

[13] Matias, S.B. and Zamar, R.H. 2002. "Botstrapping Robust Estimates of Regression", *J. Ann. Statist*., 30, pp.556-582.

[14] Maronna,R.A, Martin, R.D. and Yohai,V.J. 2006. " Robust Statistics Theory and Methods", 1st ed., New York: Wiely and Sons

[15] Midi, H, Ramli, N.M and Imon, A.H.M.R. 2009. "The performance o f diagnostic-robust generaliozed potential approach for the identification of multiple high leverage points in linear regression", *Journal of Applied Statistics*. To appear in 36(5):1-15.

[16] Riadh, K, Cottrell,L and Vigneron,V. 2002. "Bootstrap for Neural Model Selection",*Neurocomputing J*., 48, pp. 175-183.

[17] Rousseeuw P. J. and Leroy. M.A ., 2003. "Robust Regression and Outlier detection", Illustrated Edn., New York:Wiley-IEEE.

[18] Rousseeuw P. J. and Van Driessen K., 1999. "Computing LTS regression for Large Data Sets", *Data Mining Knowledge Discovery* 12, pp.29-45.

[19] Willems,G and Aelst, S.V. 2005. " Fast and Robust Bootstrap for LTS", *Computational Statistics & Data Analysis.,* 48,pp.703-715.