Estimating Regression Coefficients using Weighted Bootstrap with Probability

NORAZAN M. R.[1], HABSHAH MIDI[2] AND A. H. M. R. IMON[3]

[1]Faculty of Computer and Mathematical Sciences,
University Technology MARA,
40450 Shah Alam, Selangor,
MALAYSIA
[2]Laboratory of Applied and Computational Statistics,
Institute for Mathematical Research,
University Putra Malaysia,
43400 Serdang, Selangor,
MALAYSIA
[3]Department of Mathematical Sciences,
Ball State University,
Muncie, IN 47306,
U.S.A.
Email:[1]norazan@tmsk.uitm.edu.my,[2]habshahmidi@gmail.com,[3]imon_ru@yahoo.com

Abstract: In this paper we propose a new Weighted Bootstrap with Probability (WBP). The basic idea of the proposed bootstrap technique is to do re-sampling with probabilities. These probabilities become the control mechanism for getting good estimates when the original data set contain multiple outliers. Numerical examples and simulation study are carried out to evaluate the performance of the WBP estimates as compared to the Bootstrap 1 and Diagnostic-Before Bootstrap estimates. The results of the study signify that the WBP method is more efficient than the other two methods.

Key-Words: - regression, outliers, weighted bootstrap with probability, weighting function

1 Introduction

Bootstrap method is a procedure that can be used to obtain inference such as confidence intervals for the regression coefficient estimates. The bootstrap method proposed by Efron with the basic idea of generating a large number of sub-samples by randomly drawing observations with replacement from the original dataset [4, 5]. 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 practitioners in many areas, as its usage does not rely on the normality assumption. Kun and Yan, for example, did analysis bullwhip effect in supply chain model using bootstrap techniques [9]. An interesting property of the bootstrap method is that it can provide the standard errors of any complicated estimator without requiring any theoretical calculations.

It is now evident that the presence of outliers have an unduly effect on the bootstrap estimates. Outliers are observations that 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. There is a possibility that the bootstrap samples may contain more outliers than the original sample because the bootstrap re-sampling procedure is with replacement [12]. As a consequence, the variance estimates and also the confidence intervals are affected and thus resulting to bootstrap distribution break down. We may use robust estimator to deal with possible outliers, but this may not be enough since robust estimation is expected to perform well only up to a certain percentage of outliers.

In this paper, we propose a modification of the bootstrap procedure proposed by Imon and Ali [12]. The main idea is to form each bootstrap sample by re-sampling with probabilities so that the more outlying observations will have smaller probabilities of selection. We organize this paper as follows – we discuss and summarize several existing bootstrap

procedures in Section 2; in Section 3 we present the newly proposed bootstrap method and examine its performance; and finally, some conclusions are made in Section 4.

2 Some Bootstrap Techniques

In this paper, the application of bootstrap techniques will be applied to multiple linear regression models. These models are considered as they are among the most popular ones and widely used in various areas especially for forecasting or prediction [3, 8, 18].

Let a general linear regression model be in the following form

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

where Y is a (n x 1) vector of continuous response variable, X is a (n x p) data matrix that includes the intercept, β is a (p x 1) vector of unknown parameters to be estimated from the data, and ε is an (n x 1) vector of unobservable random errors, normally and independently distributed with mean zero and constant variance σ^2. For an i^{th} observation, equation (1) can be written as

$$y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon_i \quad (2)$$

It is generally well known that fixed-x re-sampling and random-x re-sampling are the two commonly bootstrapping techniques that usually used for linear regression model [6, 10, 12, 23]. For clarity we summarize the procedures for the two techniques in the following sections.

2.1 Fixed-*x* Re-sampling

In the fixed-x re-sampling, we generate bootstrap replications when the model matrix X is fixed. We treat the fitted values $\hat{y}_i$ from the model as giving the expectation of the response for the bootstrap samples. The fixed-x re-sampling can be summarized as follows:

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

Step 2. Get the residuals $\varepsilon_i = y_i - \hat{y}_i$.

Step 3. Draw ε_i^* from ε_i and attach to $\hat{y}_i$ to get a fixed-x bootstrap value y_i^* where $y_i^* = f(x_i, \hat{\beta}) + \varepsilon_i^*$.

Step 4. Regress the bootstrapped values y_i^* on the fixed X to obtain β^*.

Step 5. Repeat Step 3 and Step 4 for B times to get $\hat{\beta}^{*1}, \ldots, \hat{\beta}^{*B}$

2.2 Random-*x* Re-sampling

On the other hand, the random-x re-sampling offers a different approach of bootstrapping. Assuming that we want to fit a regression model with response y_i and predictors x_i which forms a sample of n observations $z_i' = (y_i, x_i)$. The following summarizes the random-x re-sampling procedure:

Step 1. The bootstrap data $(y_1, x_1)^*, \ldots, (y_n, x_n)^*$ are taken independently with equal probabilities $1/n$ from the original cases $(y_1, x_1), \ldots, (y_n, x_n)$

Step 2. Compute β^* for the bootstrap data set $(y_1, x_1)^*, \ldots, (y_n, x_n)^*$

Step 3. Repeat Step 1 and Step 2 for B times to get $\hat{\beta}^{*1}, \ldots, \hat{\beta}^{*B}$

These two re-sampling methods are also known by other names. Some authors or researchers refer to the fixed-x re-sampling as bootstrapping the residuals of linear regression models or bootstrap 1 method of linear regression model. Meanwhile, the random-x re-sampling is also known as the bootstrapping pairs or case- re-sampling or bootstrap 2 methods of linear regression estimate [6,10, 12, 15, 16].

2.3 Diagnostic-Before Bootstrap

A new way of bootstrapping in linear regression was proposed by Imon and Ali [12]. The method is called Diagnostics-Before Bootstrap. In this procedure, the suspected outliers are identified and omitted from the analysis before performing bootstrap with the remaining set of observations. The bootstrap estimates of parameters involve only good observations. Outliers are identified using robust reweighted least squares (RLS) residuals as proposed by Rousseeuw and Leroy [14]. In order to compute the RLS residuals, a regression line is fitted without the observations identified as outliers by the least median square (LMS) technique [13, 14]

The matrix X and Y are partitioned as follows:

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix}, \; Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix} \tag{3}$$

where R represents the set of cases 'remaining' in the analysis and D as the set of 'deleted' cases. If $\hat{\beta}^{(-D)}$ represents the vector of the estimated parameters after the deletion of d cases, then the Diagnostics-Before Bootstrap can be summarized as follows:

Step 1. Fit a model to the 'remaining' observations to get $\hat{\beta}^{(-D)}$ and the fitted values $\hat{y}_i^{(-D)} = f(x_i, \hat{\beta}^{(-D)})$.

Step 2. Get the residuals $\hat{\varepsilon}_i^{(-D)} = y_i - f(x_i, \hat{\beta}^{(-D)})$.

Step 3. Draw $\varepsilon_i^{*(-D)}$ from $\hat{\varepsilon}_i^{(-D)}$ and attach to $\hat{y}_i^{(-D)}$ to get a fixed-x bootstrap values y_i^* where $y_i^{*(-D)} = f(x_{R\,i}, \hat{\beta}^{(-D)}) + \varepsilon_i^{*(-D)}$.

Step 4. Regress the bootstrapped values $y_i^{*(-D)}$ on the fixed X_R to get $\hat{\beta}^{*(-D)}$.

Step 5. Repeat Step 3 and Step 4 for B times to get $\hat{\beta}^{*1(-D)}, \ldots, \hat{\beta}^{*B(-D)}$.

When outliers are present in our data, both fixed-x and random-x re-sampling methods are expected to breakdown. This can happen as there is no mechanism used to control the presence of outliers in the bootstrap samples produced by these methods. Consequently, the possibility to have bootstrap samples with larger percentage of outliers than that in the original data set is high. The Diagnostics-Before Bootstrap on the other hand accommodates the outliers' influence by first identifying outliers based on the robust re-weighted least squares residuals as proposed in [14] by applying the weight function written in equation (4).

$$w_i = \begin{cases} 0, & \text{if } \text{abs}(r_i) > 2.5\,s \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

The s in equation (4) is the robust scale estimate being defined as

$$s = 1.48268[1 + \{5/(n - p)\}]\sqrt{\text{median}(r_i^2)} \tag{5}$$

where n is the sample size and p is the number of regression coefficients. Each observation would either receive weights "0" or "1" which depends on its outlyingness. Due to its crude weight assignment, the deletion set may not be very accurate and this may possibly affect its bootstrap estimates. Thus, in the next section, a proposed bootstrap method that is expected to accommodate the problem will be presented.

3 Weighted Bootstrap with Probability (WBP)

Many researchers use a mechanism so that the re-sampling plan is not so much affected by the outlying observations. For example, Amado and Pires used an influence function to compute those selection probabilities and applied the procedure to obtain confidence intervals for the univariate location and for the correlation coefficient and selection of variables in two group linear discriminant analysis [1]. Other authors have addressed the problem in slightly different ways for different applications. Stromberg, for example recommended to use a 50% breakdown S-estimate of variability instead of the sample variance for the computation of the bootstrap variance estimate [21]. Robustifying the bootstrap method by applying winsorization for certain L and M estimators was proposed by Singh [20].

Our proposed bootstrap method also attempts to protect the bootstrap procedure against a given number of arbitrary outliers. We propose several modifications on the Diagnostic-Before Bootstrap procedure. Hampel's weighting psi function will be used to determine the weight assigned to each observation. These weights are calculated from the least median squares (LMS) standardized residuals. If we let r_i to represent the LMS residuals (where $i = 1, 2, \ldots, n$), then the standardized LMS residuals $u_i = \dfrac{r_i}{\text{MAD}(r_i)}$. The Hampel's weighting psi function (as shown in equation (7) with tuning constants a = 1.31, b = 2.039, c = 4 is used to compute the weights for all cases of original sample. If w_i denotes the weight for the i^{th} observation, then this weight is defined as

$$w_i(u_i) = \frac{\psi(u_i)}{u_i}. \tag{6}$$

$$\psi_{Hampel}(u) = \begin{cases} u & ,0 \le abs(u) \le a \\ a\ sign(u), & ,a \le abs(u) \le b \\ a(c - abs(u))/(c-b)\ sign(u) & ,b \le abs(u) \le c \\ 0 & ,c \le abs(u) \end{cases} \quad (7)$$

Based on these weights, we expect that outliers in the original sample will receive proper weights according to its outlying ness. We expect that only the very bad outliers will receive weight "0" and be included in deleted set D. To protect the whole procedure against outliers, we propose to do bootstrap re-sampling with probabilities. Thus, the i^{th} observation will get the selection probability of p_i where

$$p_i = \frac{w_i}{\sum_{j=1}^{n} w_j} \quad (8)$$

For $0 \le p_i \le 1$ and $i = 1, 2, \ldots, n$. These probabilities become the control mechanism whereby the bad observations are ascribed less importance than the good ones and thus attributed with lower probabilities for re-sampling.

Assigning probabilities $p_1, p_2, \ldots, p_n$ to $\{(y_1, x_1), (y_2, x_2), \ldots, (y_n x_n)\}$, we are now to present our newly proposed bootstrap method. Our proposed method will be called as the Weighted Bootstrap with Probability (WBP). For simplicity, most of the notations used in [8] are adopted in this paper. We let R represents the set of 'remaining' cases and D represents the set of 'deleted' cases. We propose that the remaining set R should contain observations with $p_i > 0$, thus allowing more observations to be involved in the bootstrapping process. The matrix X and Y are as defined in equation (3).

Let $\hat{\beta}^{(-D)}$ to denote the vector of the estimated parameters after the deletion of d cases and the $\hat{\beta}^{(-D)}$.is estimated by fitting a linear model to the remaining observations only, namely the X_R and the Y_R. The following steps describe the WBP procedure:

Step 1: Fit original data with LMS. Apply Hampel's weighting function to identify outliers based on the LMS residuals. Fit a model to the 'remaining' observations (with $w_i > 0$) to get $\hat{\beta}^{(-D)}$ and the fitted values $\hat{y}_i^{(-D)} = f(x_i, \hat{\beta}^{(-D)})$.

Step 2: Get the residuals $\hat{\varepsilon}_i^{(-D)} = y_i - f(x_i, \hat{\beta}^{(-D)})$.

Step 3: Draw $\varepsilon_i^{*(-D)}$ from $\hat{\varepsilon}_i^{(-D)}$ by re-sampling with probabilities as shown in equation (8). Attach $\varepsilon_i^{*(-D)}$ to $\hat{y}_i^{(-D)}$ to get a fixed-x bootstrap values y_i^* where $y_i^{*(-D)} = f(x_{R\,i}, \hat{\beta}^{(-D)}) + \varepsilon_i^{*(-D)}$.

Step 4: Regress the bootstrapped values $y_i^{*(-D)}$ on the fixed X_R to get $\hat{\beta}^{*(-D)}$.

Step 5: Repeat Step 3 and Step 4 for B times to get $\hat{\beta}^{*1(-D)}, \hat{\beta}^{*2(-D)}, \ldots, \hat{\beta}^{*B(-D)}$.

In this study, re-sampling with probability in Step 3 above was done by making use of the available S-Plus procedure called "sample".

3.1 Examples using real data sets

It is generally known that least squares estimates are very sensitive to the outliers, thus can lead to misleading inference. Similarly, as we mentioned earlier, not all existing bootstrapping techniques can remain efficient when outliers are present. We will assess the goodness of our proposed bootstrap method, compared to the Bootstrap 1 and Diagnostics-Before Bootstrap. Bootstrap 2 is not included as its performance was already found to be very poor [12]. Two real data sets namely the Hawkins-Bradu-Kass data and the Stackloss data that commonly used by other researchers for validating their robust methods, were used as numerical examples. It was reported that the first ten observations in Hawkins-Bradu-Kass data set are outliers [12]. Meanwhile, Stackloss data set consists of 4 outliers [14].

For each bootstrap method, 5000 bootstrap sub-samples were drawn. Least squares estimates for each sub-sample were computed. For simplicity, let the term $\hat{\beta}^{*B}$ represents the estimate of the B^{th} bootstrapped sample and $\hat{\beta}$ is the vector for estimate from the original sample. To check for the stability of the bootstrapped estimates, we constructed 95% confidence intervals for the bootstrapped regression parameters base on the variance of the bootstrapped re-calculated estimates. The 95% standard confidence intervals for β_j is defined as

$$(\hat{\beta}_j \pm z_{0.025}\, s_j^*) \quad (9)$$

where s_j^* is the sample standard deviation of $\hat{\beta}_j^{*B}$. To graphically illustrate the stability of the proposed bootstrap procedure, we also displayed the scatter plots of $\hat{\beta}_j^{*B} - \hat{\beta}_j$ (where $B = 1, 2, \ldots, 5000$ and j = 1, 2, …, p). We expect that a bootstrap procedure is stable when $\hat{\beta}_j^{*B} - \hat{\beta}_j$ is close to zero.

For all the bootstrapping techniques discussed earlier, the estimate of β_j is defined as

$$\hat{\beta}_j^* = \frac{1}{5000}\sum_{B=1}^{5000}\hat{\beta}_j^{*B} = \frac{1}{5000}\sum_{B=1}^{5000}\hat{\beta}_j^{*B(-D)} \quad (10)$$

Hence, the i^{th} residual for each bootstrap method can be written as

$$\hat{\varepsilon}_i^* = y_i - x_i^T\hat{\beta}^* \quad (11)$$

Figure 1 - Figure 3 exhibit the plots of $(\hat{\beta}_0^{*B} - \hat{\beta}_0)$ versus $(\hat{\beta}_1^{*B} - \hat{\beta}_1)$ for Hawkins-Bradu-Kass data. Plots for other regression coefficients are not displayed here due to space constraint, however their results are consistent. These figures show that the proposed bootstrap procedure is the most stable estimates followed by the Bootstrap 1 and the Diagnostic-Before bootstrap estimates. The plot of Figure 3 clearly indicates that the WBP estimates is the most stable even with the presence of multiple outliers, evidenced by the values of the bootstrap biases which are close to zero. On the other hand, the Diagnostic-Before bootstrap and the Bootstrap 1 method fail to provide stable estimates as can be observed from Figure 1 and Figure 2.

Bootstrapped residual estimates for Hawkins-Bradu-Kass data are presented in Table 1. For comparison purpose, we also include least squares (LS) and re-weighted least squares (RLS) residuals. Many authors generally agree that for this data set, the robust RLS can generate estimates that most likely to be very close to the error true values [12, 14]. In this respect, we would expect that the more robust method would be the one with residuals closest to the RLS residuals. From Table 1, it reveals that the WBP method is appreciably the most robust method since its residuals are very close to the RLS residuals. The performance of the OLS and Bootstrap 1 are fairly close and not encouraging. Their residuals are very far from the RLS residuals. It is interesting to note here that both the proposed and the Diagnostic-before bootstrap methods can easily detect and identify all the 10 outliers in the given data set. Unfortunately, both the least squares and Bootstrap 1 method not only fail to identify the correct outliers but suffer masking problem.

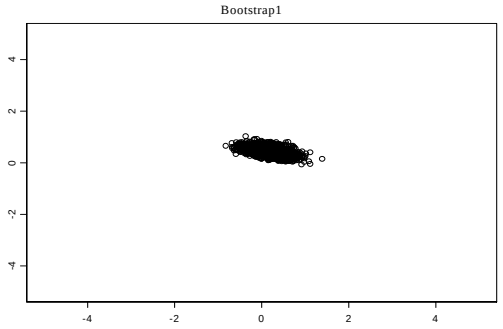

Figure 1: Plots of $(\hat{\beta}_0^{*B} - \hat{\beta}_0)$ versus $(\hat{\beta}_1^{*B} - \hat{\beta}_1)$ for Hawkins-Bradu-Kass data using Bootstrap1 method.

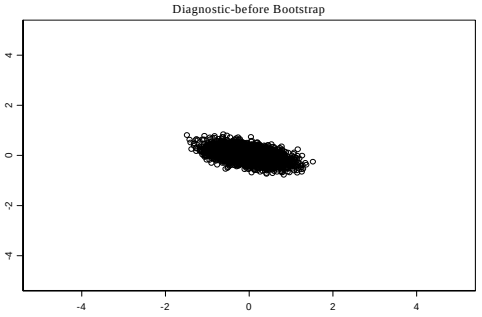

Figure 2: Plots of $(\hat{\beta}_0^{*B} - \hat{\beta}_0)$ versus $(\hat{\beta}_1^{*B} - \hat{\beta}_1)$ for Hawkins-Bradu-Kass using Diagnostic-before Bootstrap method.

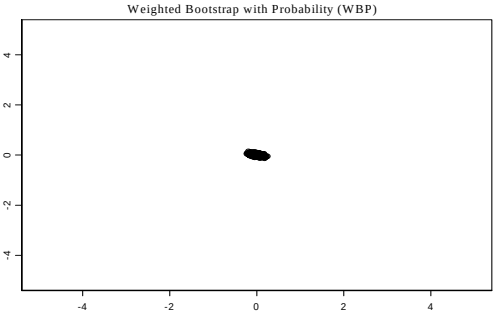

Figure 3: Plots of $(\hat{\beta}_0^{*B} - \hat{\beta}_0)$ versus $(\hat{\beta}_1^{*B} - \hat{\beta}_1)$ for Hawkins-Bradu-Kass using Weighted Bootstrap with Probability method.

Table 1: Bootstrap Residuals for Hawkins-Bradu-Kass Data

i	LS	BOOT1	DIAG	WBP	RLS
1	3.380	3.383	**8.393**	**9.740**	**9.739**
2	3.995	3.998	**8.833**	**10.185**	**10.183**
3	3.003	3.006	**9.057**	**10.405**	**10.405**
4	2.561	2.564	**8.302**	**9.656**	**9.655**
5	3.061	3.065	**8.757**	**10.108**	**10.107**
6	3.436	3.438	**8.652**	**9.997**	**9.996**
7	4.513	4.515	**9.451**	**10.797**	**10.796**
8	3.837	3.840	**9.033**	**10.382**	**10.381**
9	2.709	2.713	**8.414**	**9.768**	**9.767**
10	3.039	3.043	**8.749**	**10.104**	**10.103**
11	**-7.831**	**-7.827**	-1.416	-0.063	-0.064
12	**-9.372**	**-9.367**	-1.555	-0.204	-0.202
13	**-6.118**	**-6.116**	-0.721	0.626	0.623
14	-3.802	-3.803	-1.557	-0.205	-0.215
15	-0.661	-0.661	-1.839	-0.499	-0.503
16	0.867	0.866	-0.878	0.461	0.456
17	0.646	0.647	-1.422	-0.067	-0.073
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
70	0.473	0.474	-0.488	0.866	0.861
71	0.016	0.016	-1.122	0.225	0.221
72	0.138	0.139	-1.419	-0.065	-0.071
73	0.441	0.442	-0.750	0.607	0.602
74	-0.390	-0.389	-2.076	-0.719	-0.725
75	-0.347	-0.345	-0.879	0.478	0.474

Table 2 presents the least squares (LS) and the robust RLS coefficient parameter estimates from the original Stackloss data. Meanwhile, Table 3 - Table 4 illustrate the least squares coefficient parameter estimates of the Stackloss bootstrapped sub-samples. It is worth to mention here that the least squares estimates of the original sample is sensitive to multiple outliers, but not the robust RLS.

We clearly observe that the WBP again repeats its excellent performance. The results of Table 3 indicate that the WBP always outperforms the other two bootstrap methods (see Table 4 and Table 5). The confidence intervals of the WBP estimates signify the narrowest average interval length for all of the regression coefficients. On the other hand, the confidence intervals for the Bootstrap 1 and the Diagnostic-Before Bootstrap give bad results. Their average confident lengths are prominently large.

Table 2: True Coefficient Estimates obtained from the original Stackloss data

Coefficient Parameter	LS	Robust RLS
β_0	-39.920	-37.652
β_1	0.716	0.798
β_2	1.295	0.577
β_3	-0.152	-0.067

Table 3: 95% Confidence Intervals for WBP Bootstrap Estimates using Stackloss Data

Coefficient Parameter	WBP CI	WBP CI Length
β_0	(-44.088, -31.217)	12.472
β_1	(0.705, 0.890)	0.185
β_2	(0.350, 0.804)	0.454
β_3	(-0.150, 0.016)	0.166

Table 4: 95% Confidence Intervals for Diagnostic Bootstrap Estimates using Stackloss Data

Coefficient Parameter	Diagnostic CI	Diagnostic CI Length
β_0	(-64.555, -15.284)	49.270
β_1	(0.360, 1.071)	0.711
β_2	(0.424, 2.166)	1.742
β_3	(-0.470, 0.166)	0.636

Table 5: 95% Confidence Intervals for Bootstrap 1 Estimates using Stackloss Data

Coefficient Parameter	Bootstrap1 CI	Bootstrap1 CI Length
β_0	(-60.230, -19.609)	40.621
β_1	(0.481, 0.951)	0.470
β_2	(0.660, 1.931)	1.271
β_3	(-0.424, 0.119)	0.543

3.2 Examples using simulated data sets

Examples from the real data sets in Section 3.1 have shown that the WBP coefficient estimates are in general found to be the most stable bootstrapped estimates with the shortest confidence interval lengths. In this section we would further investigate the robustness of our proposed bootstrap method by conducting a simulation study. The simulation study was performed using a multiple linear model of three predictors.

Data sets of size n = 20, 40 and 100 with residual outliers α = 10% and 20% were created based on the adapted simulation design used by Sebert et. al [17]. The observations for predictor variables were selected at random from the $U(0, 35)$ distribution. For both good and bad observations, their random errors were generated from $N(0,1)$. All outliers were placed away from the good observations by a distance of 10 standard deviations (standard deviation σ = 1).

We now illustrate the procedure for creating artificial data set for the case of a multiple linear model with a single response and three predictor variables. Our approach is to randomly generate n regression observations. These n observations include the n_c (clean) observation and the n_0 outliers (where $n_0 = \alpha\%$ of n). Thus altogether, we have $n_c + n_0 = n$ observations. The n_c clean observations were generated according to the model

$$y_{i_c} = \beta_0 + \beta_1 x_{1i_c} + \beta_2 x_{2i_c} + \beta_3 x_{3i_c} + \varepsilon_i \qquad (12)$$

where $i = 1, 2, \ldots, n_c$ and $\beta_0 = \ldots = \beta_3 = 5$. The x_{i_c} and ε_i are from $U(0, 35)$ and $N(0,1)$ respectively. The n_o outlying observations were generated from the model

$$y_{i_c} = \beta_0 + \beta_1 \bar{x}_{1i_c} + \beta_2 \bar{x}_{2i_c} + \beta_3 \bar{x}_{3i_c} + y\text{shift} + \varepsilon_i \qquad (13)$$

where $i = 1, 2, \ldots, n_o$. The residual outliers were created when the $y$shift represents the number of standard deviations (in this study is taken to be 10) the outliers are placed away from the good observations. For any contaminated data sets of size n, residuals outliers are placed as the last α% observations.

Using each of the bootstrap method, we generated 1000 bootstrapped random samples. For each k^{th} bootstrapped sample ($k = 1, 2, \ldots, 1000$), a least squares bootstrapped estimate was computed and denoted the k^{th} bootstrapped estimate as $\hat{\beta}_j^{*k}$ ($j = 0, 1, 2, 3$). Based on these re-computed $\hat{\beta}_j^{*k}$, we calculated its standard deviation. The bootstrapped standard errors of the WBP procedure will be compared to the Bootstrap 1 and Diagnostics-Before Bootstrap (DIAG) standard errors.

Table 6 - Table 8 present the bootstrapped estimates using the simulated data. In the contaminated data sets, the bad performance is observed for both the Diagnostic-Before Bootstrap and the Bootstrap 1 estimators. Their bootstrapped coefficient estimates are far from the true value. The WBP method gives very appealing results with the lowest values of standard errors. More serious consequences are observed when we increased the outlier percentage to 20%. Enhancing the percentage of outliers by more than 10% would result to significant increase in the bootstrap standard errors. In other words, we would suggest that generally, the reliability of these bootstrap estimates decreases as the percentages of outliers exceed 10%. This is noticeable in both the Diagnostic-Before bootstrap and the Bootstrap 1 estimates, but no so apparent in the WBP estimates.

Figure 4 – Figure 9 provide density plots for β_3 bootstrapped estimate. The plots graphically represent the summarized performances of the three bootstrap methods for data sets with 10% and 20% outliers. It seems that the estimates from the WBP are not so much affected as compared to those of the Bootstrap 1 and the Diagnostic-Before-Bootstrap methods. The advantages of the WBP over other methods are more apparent in data sets with big sample sizes and for outliers exceeding the level of 10%. In summary, the results from the experiments indicate that the WBP procedure performs well in most of the given situations.

Table 6 : Bootstrapped Estimates with their respective standard errors written in brackets for n =20

Outliers = 10%			
Coefficient Parameter	WBP	BOOTSTRAP1	DIAG
β_0	5.870	16.040	10.155
	(0.292)	(6.678)	(8.202)
β_1	5.001	4.860	5.004
	(0.024)	(0.551)	(0.648)
β_2	4.914	5.307	4.943
	(0.023)	(0.571)	(0.646)
β_3	4.964	4.107	4.983
	(0.022)	(0.499)	(0.606)
Outliers = 20%			
Coefficient Parameter	WBP	BOOTSTRAP1	DIAG
β_0	3.099	41.395	-0.554
	(0.751)	(13.836)	(20.507)
β_1	5.193	2.861	5.226
	(0.058)	(1.105)	(1.578)
β_2	5.053	3.966	5.053
	(0.041)	(0.883)	(1.107)
β_3	4.996	3.429	4.996
	(0.043)	(0.851)	(1.163)

Table 7 : Bootstrapped Estimates with their respective standard errors written in brackets for n =40

Outliers = 10%			
Coefficient Parameter	WBP	BOOTSTRAP1	DIAG
β_0	5.569	13.261	2.442
	(0.547)	(5.695)	(7.304)
β_1	4.965	4.724	4.967
	(0.029)	(0.287)	(0.336)
β_2	4.990	4.356	5.017
	(0.029)	0.300 ()	(0.372)
β_3	5.003	4.966	5.046
	(0.033)	(0.333)	(0.394)
Outliers = 20%			
Coefficient Parameter	WBP	BOOTSTRAP1	DIAG
β_0	5.718	32.694	9.898
	(0.682)	(10.243)	(11.998)
β_1	4.969	4.253	4.982
	(0.032)	(0.514)	(0.578)
β_2	4.980	4.076	4.985
	(0.028)	(0.419)	(0.488)
β_3	4.981	4.378	4.981
	(0.033)	(0.508)	(0.566)

Table 8 : Bootstrapped Estimates with their respective standard errors written in brackets for n =100

Outliers = 10%			
Coefficient Parameter	WBP	BOOTSTRAP1	DIAG
β_0	5.084	19.264	7.665
	(0.276)	(4.688)	(5.269)
β_1	5.024	4.382	5.004
	(0.014)	(0.233)	(0.286)
β_2	4.972	5.061	4.967
	(0.016)	0.257 ()	(0.318)
β_3	4.996	4.405	4.987
	(0.015)	(0.237)	(0.275)
Outliers = 20%			
Coefficient Parameter	WBP	BOOTSTRAP1	DIAG
β_0	4.739	41.903	4.513
	(0.391)	(7.131)	(8.984)
β_1	4.994	4.010	5.006
	(0.021)	(0.398)	(0.489)
β_2	5.036	3.279	5.048
	(0.023)	(0.423)	(0.544)
β_3	5.003	4.155	4.989
	(0.020)	(0.374)	(0.453)

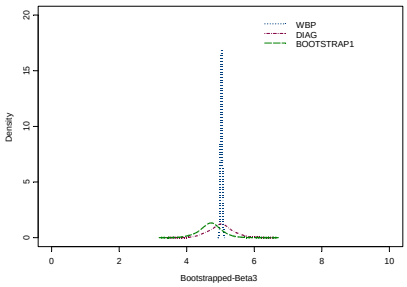

Figure 4: Density plots of bootstrapped coefficient estimates for β_3 using contaminated data set with sample size n=20 and 10% residual outliers

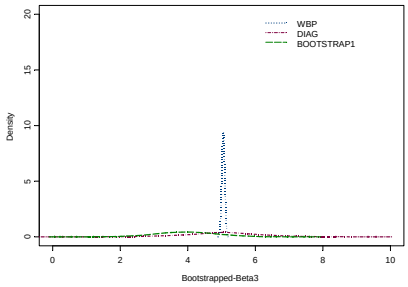

Figure 5: Density plots of bootstrapped coefficient estimates for β_3 using contaminated data set with sample size n=20 and 20% residual outliers

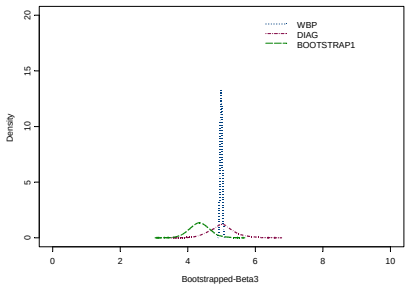

Figure 6: Density plots of bootstrapped coefficient estimates for β_3 using contaminated data set with sample size n=40 and 10% residual outliers

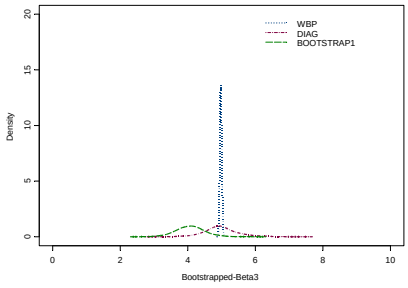

Figure 7: Density plots of bootstrapped coefficient estimates for β_3 using contaminated data set with sample size n=40 and 20% residual outliers

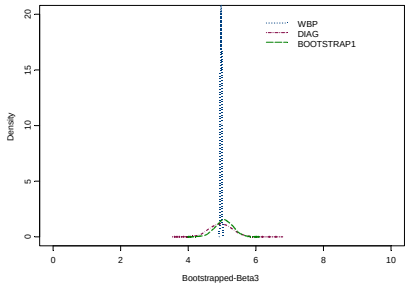

Figure 8: Density plots of bootstrapped coefficient estimates for β_3 using contaminated data set with sample size n=100 and 10% residual outliers

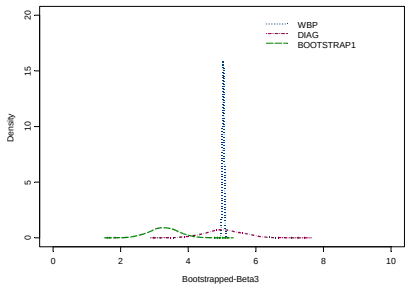

Figure 9: Density plots of bootstrapped coefficient estimates for β_3 using contaminated data set with sample size n=100 and 20% residual outliers

4 Conclusion

In this paper, we propose a new bootstrap method to reduce the effect of outliers on the bootstrap estimates. The numerical studies suggest that Bootsrap 1 performs poorly in the presence of outliers. The Diagnostic-Before bootstrap is more efficient than the Bootstrap 1 but it is not sufficiently robust because it is not very stable and has relatively large confidence interval lengths.

The WBP method consistently outperformed the Bootstrap 1 and Diagnostic-Before bootstrap methods. It emerges that the Hampel's weighting function and re-sampling probability schemes introduced in the WBP procedure help to improve the performance of the bootstrapped estimates. The results of the study clearly indicate that the WBP is the best estimator as it is consistently provides stable estimates, closest residuals to the error true values and shortest average confident length. Hence, it should provide a robust alternative to other existing bootstrap methods.

References:

[1] Amado, C. and Pires, A. M., Robust Bootstrap with Non Random Weights Based on the Influence Function, *Communications in Statistics*, Volume 33, Issue 2, 2004, page 377-396.

[2] Azami, Z., Ibrahim, M., Shahrum, A. and Mohd Sahar, Y. An Evaluation of Test Statistics for Detecting level Change in BL (1, 1, 1, 1) Models. *WSEAS TRANSACTIONS on MAHEMATICS* Issue 2, Volume 7, 2008, page 67-70.

[3] Darmesah, G., Zainodin, H. J., Kamsia, B., and Suriani, H. Multiple Linear Regression in Forecasting the Number of Asthmatics. *WSEAS TRANSACTIONS on INFORMATION SCIENCE and APPLICATIONS* Issue 6, Volume 5, 2008, page 972-977.

[4] Efron, B. (1979). Bootstrap Methods. Another Look at the Jackknife. *Ann. Statist.,*Volume 7, 1979, page 1-26.

[5] Efron B. and Tibshirani R.J., *An Introduction to the Bootstrap*, Chapman Hall, 1998.

[6] Fox, J., *Bootstrapping Regression Models : An Appendix to An R and S-Plus Companion to Applied Regression*. Available on line: http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-bootstrapping.pdf ,2002.

[7] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A., *Robust Statistics: The Approach based on Influence Functions.* John Wiley and Sons, 1986.

[8] Kamsia, B., Zainodin, J., Darmesah, G., Noraini, A. and Amran, A., Effect of Water Parameters on Ephemeroptera Abundance in Telipok River, Sabah Malaysia, *WSEAS TRANSACTIONS on ENVIRONMENT and DEVELOPMENT* Issue 5, Volume 4, 2008, page 447-451.

[9] Kun, L.H., and , Yan, K. C., Analysis Bullwhip Effect in Supply Chain Model by Using Bootstrap Technique, *WSEAS TRANSACTIONS on INFORMATION SCIENCE and APPLICATIONS* Issue 12, Volume 3, 2006.

[10] Midi, H., Bootstrap Methods in a Class of Non-Linear Regression Models. *Pertanika J. Sci & Techno*, 8(2), page 175-189, 2000.

[11] Rahmatullah Imon, A.H.M., Identifying Multiple High Leverage Points in Linear Regression, *Journal of Statistical Studies*, Special Volume, page 207-218, 2002.

[12] Rahmatullah Imon, A.H.M., and Ali, M. M., Bootstrapping Regression Residuals, *Journal of Korean Data & Information Science Society,* 16(3), page 665-682, 2005.

[13] Rousseeuw, P. J., Least Median of squares regression. *Journal of the American Statistical Association,* 79, page 871-880, 1984.

[14] Rousseeuw, P. J. & Leroy, A. M.., *Robust Regression and Outlier Detection,* Wiley, 1987.

[15] Salibian-Barrera, M. and Zamar, R. H., Bootstrapping Robust Estimates of Regression, *Ann. Stat.*, 30(2), page 556-582, 2002.

[16] Salibian-Barrera, M., Bootstrapping MM-Estimates for Linear Regression with Fixed Designs, *Statistics and Probability Letters,* 63, page 259-266, 2006.

[17] Sebert, D.M Montgomery, D. C Rollier, D. A, A Clustering Algorithm for Identifying Multiple Outliers in Linear Regression, *Computational Statistics and Data Analysis*, 27, page 461-484, 1988.

[18] Seung, W. L. and Jun, Y. S., Construction and Operation of Knowledge Base on Intelligent Machine Tools, *WSEAS TRANSACTIONS on SYSTEMS,* Issue 3, Volume 7, 2008, page 148-155.

[19] Shao, J., and Tu, D., *The Jackknife and Bootstrap*, Springer-Verlag, 1995.

[20] Singh, K., Breakdown Theory for Bootstrap Quantiles. *The Annals of Statistics*, 26, page 1719-1732, 1998.

[21] Stromberg, A. J., Robust Covariance Estimates based on Resampling. *Journal of Statistical Planning and Inference*, 57, page 321-334, 1997.

[22] Ulmanis, J., and Kolyshkin, A., The Impact of ICT on the Development of Latvia as a New Member of the EU, *WSEAS TRANSACTIONS on BUSINESS and ECONOMICS*, Issue 10, Volume 4, 2007, page 152-159.

[23] Venables, W.N. and Ripley, B.D., *Modern Applied Statistics with S-Plus*, 3rd edition, Springer-Verlag, 2000.

[24] Willems, G. and Aelst, S. V., Fast and Robust Bootstrap for LTS. *Computational Stat. and data Analysis*, 48, page 703-715, 2005.