ROBUSTNESS OF BOOTSTRAP ESTIMATORS TO INFLUENTIAL OBSERVATIONS

Tayo P. Ogundunmade, Adedayo A. Adepoju

Department of Statistics, University of Ibadan, Ibadan, Nigeria

Corresponding Author: Tayo P. Ogundunmade, ogundunmadetayo@yahoo.com

ABSTRACT: The existence of outliers in the original sample may create problem to the classical bootstrapping estimates. There is 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. This study examined the performance of three bootstrap estimators namely: Case Bootstrapping, Fixed-X Bootstrapping and Residual Resampling method under different levels of outliers. The objective was to determine which of these bootstrap methods is resistant to the presence of outliers in the data. Three levels of outliers; 5%, 10% and 20% were considered and injected into sample sizes, N = 20, 30, 50, and 100 each replicated 1000 and 5000 times respectively. The performances of the bootstrap methods were evaluated using the mean, standard error, absolute bias, mean square error and the root mean square error. The results showed that the Residual resampling Bootstrap performed better than the other two estimators.

KEYWORDS: bootstrapping, Fixed-X bootstrapping, case bootstrapping, Fixed-X Bootstrapping.

1.0 INTRODUCTION

Bootstrapping is a nonparametric approach to statistical inference that substitutes computation for more traditional distributional assumptions and asymptotic results. Bootstrapping offers a number of advantages: the bootstrap is quite general, although there are some cases in which it fails. Because it does not require distributional assumptions (such as normally distributed errors), the bootstrap can provide more accurate inferences when the data are not well behaved or when the sample size is small.

It is possible to apply the bootstrap to statistics with sampling distributions that are difficult to derive, even asymptotically. It is relatively simple to apply the bootstrap to complex data-collection plans (such as stratified and clustered samples). Bootstrapping uses the sample data to estimate relevant characteristics of the population.

The sampling distribution of a statistic is then constructed empirically by resampling from the sample. The resampling procedure is designed to parallel the process by which sample observations were drawn from the population. For example, if the data represent an independent random sample of size n (or a simple random sample of size n from a much larger population), then each bootstrap sample selects n observations with replacement from the original sample.

The main concern with bootstrapping is how to generate a bootstrap distribution close to the true distribution of the original sample. To do that, the bootstrap resampling should appropriately mimic the "true" data generating process (DGP) that produced the "true" dataset. In the context of repeated measurement data and mixed-effects modelling, the bootstrap should therefore respect the true DGP with the repeated measures within a subject and handle two levels of variability: between-subject variability and residual variability. The classical bootstrap methods developed in simple linear regression should be modified to take into account the characteristics of mixed-effects models. Resampling random effects may be coupled with resampling residuals. The case bootstrap can be combined with the residual bootstrap. The performance of these approaches is, however, not well studied in the presence of outliers.

Uraibi et. al ([U+09]) studied linear regression model selection based on robust bootstrapping technique which was less sensitive to outliers. In the robust bootstrapping procedure, they replaced 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. Their results suggested that the robust bootstrap method was more efficient than the classical bootstrap. Alamgir et. al ([AA13]) introduced a new bootstrap procedure, called “split sample bootstrap” to handle outliers. The proposed bootstrap procedure gave bootstrap estimates having smaller standard errors resulting in narrower confidence intervals of the estimates. Zahari et. al ([ZRM14]) examined bootstrapped parameter estimation in ridge regression with multicollinearity and multiple outliers. Their study employed the fixed - x resampling technique for robust ridge regression. The proposed method was

expected to reduce the effect of the problems to the estimation results by producing lower standard error values as compared to the existing methods.
They proposed an alternative method that is resistant to the presence of simultaneous problems of multicollinearity and multiple outliers, using a combination of bootstrapping technique and robust methods in ridge regression. We propose a slight modification from what have been found in the literatures, which is to combine the bootstrapping method and robust ridge regression. We would like to examine the superiority of the proposed method as an enhancement to the existing methods when handling simultaneous problems of multiple outliers and multicollinearity. The results of the study showed that the proposed technique produced better parameter estimates with lower standard error values.
The aim of this study is to examine the performances of three bootstrap approaches; case bootstrap, Fixed-X bootstrap and residual resampling bootstrap methods when applied to linear models in the presence of outliers. The concept of bootstrap methods is discussed in Section 2. The design of the simulation is described in Section 3. The results of the simulation studies are discussed in Section 4. while section 5 summarises and concludes the study.

2.0 METHODOLOGY

Bootstrapping is the practice of estimating properties of an estimator by measuring those properties when sampling from an approximating distribution. In the case where a set of observations can be assumed to be from an independent and identically distributed population, this can be implemented by constructing a number of resamples with replacement, of the observed dataset and of equal size of the observed dataset. There are different types of bootstrapping schemes like case resampling, Bayesian bootstrap, smooth bootstrap, resampling residuals, wild bootstrap, block bootstrap for time series etc. In this study, we shall compare the performance of case bootstrapping, residual resampling and fixed -X resampling in the presence of outlier.
Let us consider the general regression model:

$$Y = X\beta + e \qquad (1)$$

Where:
Y is a (n ×1) vector of responses,
X is a (n × k) data matrix containing values of the predictors ($k = p + 1$),
the parameter vector β is (k × 1) (including the intercept) which is to be estimated from the observed data and
e is a (n × 1) vector of unobservable random error terms.
The bootstrap procedures are discussed in brief in the following section.

2.1 Residual Resampling

This employs the resampling residual technique. It involves first fitting the model and bootstrapping the residuals with OLS. Classical bootstrap approach with OLS (bootstrapping the residual)

- fit the model and obtain the residuals as $\hat{e} = y_i - g_i(\hat{\beta})$
- place probability $1/n$ at each $\hat{e}_i$ and sample the residual independently with replacement from $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_n$.
- generate bootstrap residuals e_i^* for $i = 1,2,\dots,n$
- then have a bootstrap data set $y_i^* = g_i(\hat{\beta}) + e_i^* \ \ for\ i = 1,2,\dots,n$
- for each bootstrap data set, we obtain $\hat{\beta}^*$
- the procedure is repeated B times and the covariance matrix for $\hat{\beta}$ is estimated; = $\frac{1}{B-1}\sum_{j=1}^{B}(\hat{\beta}_j^* - \hat{\beta}_*^*)(\hat{\beta}_j^* - \hat{\beta}_*^*)^T$,
- where $\hat{\beta}_j^*$ is the bootstrap estimate from the j^{th} bootstrap sample and

$$\hat{\beta}_*^* = \frac{1}{B}\sum_{j=1}^{B}\hat{\beta}_j^*$$

2.2 Case Bootstrapping

This method consists of resampling with replacement the entire subjects, that is, the joint vector of design variables and corresponding responses (X_i, Z_i, Y_i) from the original data before modelling. It is also called the paired bootstrap. This procedure omits the second step of resampling the observations inside each subject. However, it is the most obvious way to do bootstrapping and makes no assumptions on the model.
This method resamples first the entire subjects with replacement. The individual residuals are then resampled with replacement globally from the residual distribution of the original simulated dataset or individually from the residual distribution of new subjects obtained after bootstrapping in the first step. The bootstrap sample is obtained as follows:

- Fit the model to the data then calculate the residuals
- Draw N entire subjects (X_i, Z_i, Y_i) with replacement from (X_i, Z_i, Y_i) in the original data and keep their predictions from model fitting and their corresponding residuals. The new subject has $n_i^{\bullet}$ observations.

- Draw the residuals with replacement globally from all residuals of the original data or individually from each new subject.
- Generate the bootstrap responses.

2.3 Fixed–X Bootstrapping

Let B be the number of bootstrap samples to be drawn from the original dataset, a general bootstrap algorithm is:

- Generate a bootstrap sample by resampling from the data and/or from the estimated model.
- Obtain the estimates for all parameters of the model for the bootstrap sample
- Repeat the two steps above B times to obtain the bootstrap distribution of parameter estimates and then compute mean and standard deviation.

3.0 SIMULATION PROCESS

The main task is the generation of stochastic dependent Y_t which is used in estimating the parameters of the model.
To achieve this, the following have to be assumed

1. Values of the predetermined variable X
2. Values of the parameter, $b_0 = 0.2, b_1 = 0.5$

The simulation of the error term U_t is generated from a normal distribution with mean 0 and variance 0.25. We then proceed as follows:

a. The sample size N is specified as N = 20, 40, 80 and 100.
b. For all cases X are generated from a uniform (0,1)
c. Different levels of outlier to be considered are 5%, 10% and 20% for each sample size.

3.1 Criteria for evaluating the performance of the estimators

For comparison and assessment of the performance of the estimators, the following criteria are used in this study.

(i) Mean values of the parameter estimates.
(ii) Absolute Bias
(iii) Standard deviation
(iv) Mean square error
(v) Root Mean Square Error (RMSE)

4.0 RESULT AND DISCUSSION

Tables 1, 2 and 3 show the results of Case Bootstrapping, Residual Resampling Bootstrapping, and Fixed-X Bootstrapping respectively. For all of the bootstrap approaches, the mean estimate, standard deviation, absolute bias, mean square, and the root mean square are displayed for 1000 and 5000 boots at N = 20, 40, 80 and 100 respectively. The results show that Case Bootstrapping performs better than the other two estimators. Using the absolute bias result comparison, Case bootstrapping produces the least estimates, followed by Residual Resampling Bootstrapping and lastly Fixed-X Bootstrapping.
Tables 1, shows the results for the behaviour of Case Bootstrapping, using mean estimates, standard error, absolute bias, mean square error and the root mean square error at different sample sizes and percentage outliers for 1000 and 5000 runs respectively.
In general, the absolute bias, MSE and RMSE decrease as the number of runs increase from 1000 to 5000. It is worth noting that similar results are obtained at N = 20 for β_0 and β_1 when 10% and 20% of outliers are injected into the data. Also at N = 40, the estimates obtained for β_0 and β_1 at 10% and 20% outliers are the same.
For larger samples, that is, N = 80 and N = 100, the estimates of β_1 decrease with increasing percentage outlier.
As the sample size increases, the standard deviation of all the parameter estimates decreases in an ordered manner. The values of the mean square error did not really show a regular pattern. In all cases, as the number of bootstrap increases from 1000 to 5000 the standard deviation exhibits a minor fluctuation of values, respectively. On other hand, decrease in the absolute bias as the number of bootstrap increases is also not consistent.
Unlike the case bootstrap method, the Residual Resampling method does not show any consistent pattern. In most cases, least values of the standard deviation, absolute bias and MSE are produced when sample is 20 at 5% for the booting.

Table 1: Case Bootstrapping Result

Sample size	Coefficient		Case Bootstrapping				
			Estimate	Standard Deviation	Abs	MSE	RMSE
			Boot = 1000				
n=20	5%	β_0	0.174584	0.070785	0.025416	0.005657	0.07521
		β_1	0.557119	0.119506	0.057119	0.017544	0.132455
	10%	β_0	0.186731	0.061868	0.013269	0.004004	0.063275
		β_1	0.509989	0.125977	0.009989	0.01597	0.126373
	20%	β_0	0.186731	0.061868	0.013269	0.004004	0.063275
		β_1	0.509989	0.125977	0.009989	0.01597	0.126373
n=40	5%	β_0	0.186731	0.061868	0.013269	0.004004	0.063275
		β_1	0.509989	0.125977	0.009989	0.01597	0.126373
	10%	β_0	0.186731	0.061868	0.013269	0.004004	0.063275
		β_1	0.509989	0.125977	0.009989	0.01597	0.126373
	20%	β_0	0.1889	0.062137	0.0111	0.003984	0.063121
		β_1	0.510985	0.11678	0.010985	0.013758	0.117295
n=80	5%	β_0	0.1889	0.062137	0.0111	0.003984	0.063121
		β_1	0.510985	0.11678	0.010985	0.013758	0.117295
	10%	β_0	0.178835	1.01E-03	0.021165	0.000449	0.021189
		β_1	0.50000	8.89E-02	0.0000	0.007906	0.088914
	20%	β_0	0.201554	0.029146	0.001554	0.000852	0.029188
		β_1	0.499958	0.000456	4.19E-05	2.1E-07	0.000458
n=100	5%	β_0	0.20081	0.023445	0.00081	0.00055	0.023459
		β_1	0.500216	0.006531	0.000216	4.27E-05	0.006534
	10%	β_0	0.200589	0.023884	0.000589	0.000571	0.023892
		β_1	0.499994	0.000769	6E-06	5.91E-07	0.000769
	20%	β_0	0.201489	0.024894	0.001489	0.000622	0.024938
		β_1	0.499917	0.000552	8.29E-05	3.11E-07	0.000558
			Boot = 5000				
n=20	5%	β_0	0.174244	0.069066	0.025756	0.005433	0.073712
		β_1	0.560171	0.123811	0.060171	0.01895	0.137658
	10%	β_0	0.187046	0.061399	0.012954	0.003938	0.062751
		β_1	0.509316	0.124013	0.009316	0.015466	0.124363
	20%	β_0	0.187046	0.061399	0.012954	0.003938	0.062751
		β_1	0.509316	0.124013	0.009316	0.015466	0.124363
n=40	5%	β_0	0.187046	0.061399	0.012954	0.003938	0.062751
		β_1	0.509316	0.124013	0.009316	0.015466	0.124363
	10%	β_0	0.186222	0.062062	0.013778	0.004041	0.063573
		β_1	0.510724	0.120946	0.010724	0.014743	0.12142
	20%	β_0	0.186228	0.061509	0.013772	0.003973	0.063032
		β_1	0.509816	0.121364	0.009816	0.014826	0.121761
n=80	5%	β_0	0.186228	0.061509	0.013772	0.003973	0.063032
		β_1	0.509816	0.121364	0.009816	0.014826	0.121761
	10%	β_0	0.178835	1.02E-03	0.021165	0.000449	0.02119
		β_1	0.50000	8.92E-02	0.00000	0.007948	0.089152
	20%	β_0	0.200363	0.029414	0.000363	0.000865	0.029416
		β_1	0.499964	0.000464	3.65E-05	2.16E-07	0.000465
n=100	5%	β_0	0.200554	0.024037	0.000554	0.000578	0.024044
		β_1	0.499907	0.004198	9.32E-05	1.76E-05	0.004199
	10%	β_0	0.200275	0.024827	0.000274	0.000616	0.024828
		β_1	0.500012	0.000759	1.15E-05	5.77E-07	0.000759
	20%	β_0	0.200611	0.025193	0.000611	0.000635	0.025201
		β_1	0.499958	0.00053	4.18E-05	2.83E-07	0.000532

Table 2: Fixed-X Bootstrapping Result

Sample size	Coefficient		Fixed-X Bootstrapping				
			Estimate	Standard Deviation	Abs	MSE	RMSE
			Boot = 1000				
n=20	5%	β_0	0.208001	0.320359	0.008001	0.102694	0.320459
		β_1	0.499642	0.104939	0.000358	0.011012	0.10494
	10%	β_0	0.205784	0.284855	0.005784	0.081176	0.284914
		β_1	0.494241	0.290573	0.005759	0.084466	0.29063
	20%	β_0	0.188428	0.24711	0.011572	0.061197	0.247381
		β_1	0.49877	0.079335	0.001231	0.006296	0.079345
n=40	5%	β_0	0.194585	0.19802	0.005415	0.039241	0.198094
		β_1	0.508301	0.224644	0.008301	0.050534	0.224797
	10%	β_0	0.196885	0.173185	0.003115	0.030003	0.173213
		β_1	0.504233	0.076575	0.004233	0.005882	0.076692
	20%	β_0	0.195837	0.178302	0.004163	0.031809	0.178351
		β_1	0.500015	0.003829	1.51E-05	1.47E-05	0.003829
n=80	5%	β_0	0.193054	0.122546	0.006946	0.015066	0.122743
		β_1	0.503791	0.05023	0.003791	0.002537	0.050372
	10%	β_0	0.195275	0.120985	0.004726	0.01466	0.121077
		β_1	0.500008	0.003555	8.1E-06	1.26E-05	0.003555
	20%	β_0	0.196425	0.124019	0.003575	0.015394	0.124071
		β_1	0.500136	0.002352	0.000136	5.55E-06	0.002356
n=100	5%	β_0	0.199746	0.101794	0.000254	0.010362	0.101794
		β_1	0.499695	0.017752	0.000305	0.000315	0.017754
	10%	β_0	0.197337	0.104249	0.002663	0.010875	0.104283
		β_1	0.500187	0.00382	0.000187	1.46E-05	0.003825
	20%	β_0	0.200204	0.108451	0.000204	0.011762	0.108451
		β_1	0.4999	0.002397	0.0001	5.76E-06	0.002399
			Boot = 5000				
n=20	5%	β_0	0.191682	0.314228	0.008318	0.098809	0.314338
		β_1	0.506536	0.492664	0.006536	0.24276	0.492707
	10%	β_0	0.197646	0.273785	0.002354	0.074964	0.273795
		β_1	0.508421	0.301482	0.008421	0.090962	0.301599
	20%	β_0	0.194241	0.251296	0.005759	0.063183	0.251362
		β_1	0.500791	0.086807	0.000791	0.007536	0.086811
n=40	5%	β_0	0.196738	0.192096	0.003262	0.036911	0.192123
		β_1	0.504998	0.21193	0.004998	0.044939	0.211988
	10%	β_0	0.198471	0.170082	0.001529	0.02893	0.170089
		β_1	0.500844	0.080481	0.000844	0.006478	0.080485
	20%	β_0	0.198603	0.173837	0.001397	0.030221	0.173843
		β_1	0.500063	0.003802	6.27E-05	1.45E-05	0.003803
n=80	5%	β_0	0.199659	0.500252	0.000341	0.250252	0.500252
		β_1	0.500252	0.057293	0.000252	0.003283	0.057293
	10%	β_0	0.199987	0.117718	1.33E-05	0.013858	0.117718
		β_1	0.499996	0.003649	4.1E-06	1.33E-05	0.003649
	20%	β_0	0.201438	0.121553	0.001437	0.014777	0.121561
		β_1	0.499966	0.002405	3.43E-05	5.78E-06	0.002405
n=100	5%	β_0	0.198627	0.102957	0.001373	0.010602	0.102966
		β_1	0.500763	0.025261	0.000763	0.000639	0.025273
	10%	β_0	0.199018	0.104278	0.000982	0.010875	0.104283
		β_1	0.500004	0.003842	4E-06	1.48E-05	0.003842
	20%	β_0	0.198915	0.108158	0.001085	0.011699	0.108163
		β_1	0.50001	0.002379	9.5E-06	5.66E-06	0.002379

Table 3: Residual Resampling Bootstrapping Result

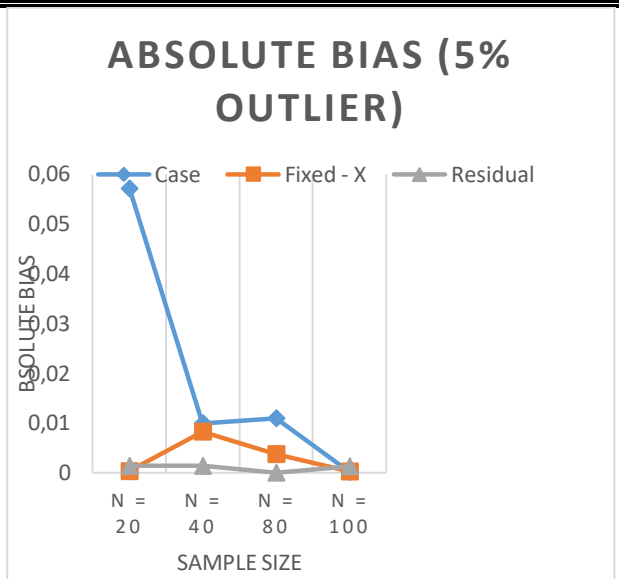
Sample size	Coefficient		Residual Resampling Bootstrapping				
			Estimate	Standard Deviation	Abs	MSE	RMSE
			Boot = 1000				
n=20	5%	β_0	0.18547	0.041369	0.01453	0.001923	0.043847
		β_1	0.50141	0.001884	0.00141	5.54E-06	0.002353
	10%	β_0	0.176592	0.043201	0.023408	0.002414	0.049135
		β_1	0.500914	0.000855	0.000914	1.57E-06	0.001251
	20%	β_0	0.160875	0.043165	0.039125	0.003394	0.058258
		β_1	0.501369	0.000797	0.001369	2.51E-06	0.001584
n=40	5%	β_0	0.152038	0.040838	0.047962	0.003968	0.062993
		β_1	0.498571	0.001085	0.001429	3.22E-06	0.001794
	10%	β_0	0.162953	0.04049	0.037047	0.003012	0.054881
		β_1	0.497957	0.001014	0.002043	5.2E-06	0.002281
	20%	β_0	0.177539	0.041913	0.022461	0.002261	0.047552
		β_1	0.498262	0.000775	0.001738	3.62E-06	0.001903
n=80	5%	β_0	0.170064	0.026584	0.029936	0.001603	0.040036
		β_1	0.500064	0.000944	6.36E-05	8.94E-07	0.000946
	10%	β_0	0.173411	0.027136	0.026589	0.001443	0.037991
		β_1	0.499728	0.0007	0.000272	5.64E-07	0.000751
	20%	β_0	0.179892	0.028379	0.020108	0.00121	0.034781
		β_1	0.499531	0.000515	0.000469	4.85E-07	0.000697
n=100	5%	β_0	0.160896	0.023674	0.039104	0.00209	0.045712
		β_1	0.501346	0.000891	0.001346	2.61E-06	0.001614
	10%	β_0	0.168682	0.024053	0.031318	0.001559	0.039489
		β_1	0.499911	0.000701	8.87E-05	4.99E-07	0.000706
	20%	β_0	0.15866	0.024896	0.04134	0.002329	0.048257
		β_1	0.500544	0.000524	0.000544	5.7E-07	0.000755
			Boot = 5000				
n=20	5%	β_0	0.182848	0.042315	0.017152	0.002085	0.045659
		β_1	0.501472	0.00188	0.001472	5.7E-06	0.002388
	10%	β_0	0.176173	0.042488	0.023827	0.002373	0.048713
		β_1	0.500956	0.000845	0.000956	1.63E-06	0.001276
	20%	β_0	0.160875	0.042412	0.039125	0.00333	0.057702
		β_1	0.501401	0.000784	0.001401	2.58E-06	0.001605
n=40	5%	β_0	0.150642	0.041122	0.049358	0.004127	0.064243
		β_1	0.498549	0.001148	0.001451	3.42E-06	0.00185
	10%	β_0	0.161715	0.040754	0.038285	0.003127	0.055916
		β_1	0.497948	0.001052	0.002052	5.32E-06	0.002306
	20%	β_0	0.176369	0.041673	0.023631	0.002295	0.047907
		β_1	0.498275	0.000759	0.001725	3.55E-06	0.001884
n=80	5%	β_0	0.170486	0.026599	0.029514	0.001579	0.039732
		β_1	0.500049	0.000987	4.9E-05	9.76E-07	0.000988
	10%	β_0	0.173829	0.027183	0.026171	0.001424	0.037734
		β_1	0.499722	0.000716	0.000278	5.9E-07	0.000768
	20%	β_0	0.181834	0.028373	0.018166	0.001135	0.033691
		β_1	0.499504	0.000531	0.000496	5.28E-07	0.000727
n=100	5%	β_0	0.161172	0.024058	0.038828	0.002086	0.045677
		β_1	0.501362	0.00092	0.001362	0.09082	0.001363
	10%	β_0	0.170131	0.024543	0.022986	0.109416	0.33078
		β_1	0.499883	0.000704	0.001117	0.089931	0.299884
	20%	β_0	0.158904	0.025652	0.041096	0.117004	0.042059
		β_1	0.50055	0.000522	0.00055	0.090331	0.000551

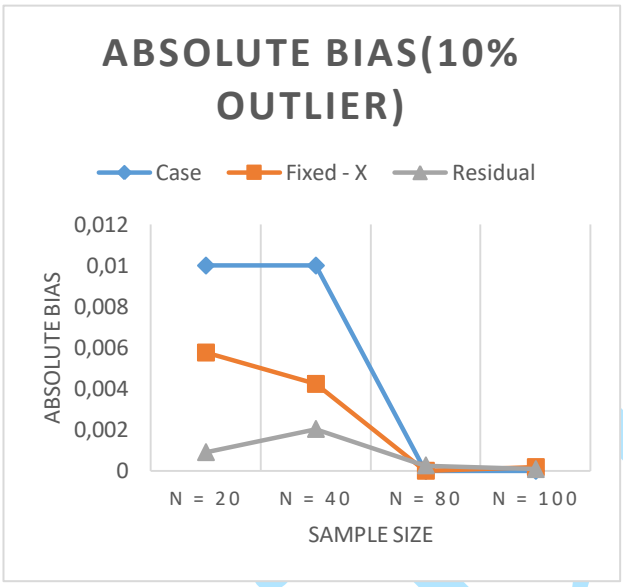
Table 4: Absolute Bias and Standard Error (in bracket) based on bootstrapping of 1000 and 5000 runs for $\hat{\beta}_1$

Outliers	Method	Runs = 1000				Runs = 5000			
		N = 20	N = 40	N = 80	N = 100	N = 20	N = 40	N = 80	N = 100
5%	Case	0.057119	0.009989	0.010985	0.000216	0.060171	0.009316	0.009816	9.32E-05
		[0.119506]	[0.125977]	[0.11678]	[0.006531]	[0.123811]	[0.124013]	[0.121364]	[0.004198]
	Fixed - X	0.000358	0.008301	0.003791	0.000305	0.006536	0.004998	0.000252	0.000763
		[0.104939]	[0.224644]	[0.05023]	[0.017752]	[0.492664]	[0.21193]	[0.057293]	[0.025261]
	Residual	0.00141	0.001429	6.36E-05	0.001346	0.001472	0.001451	4.9E-05	0.001362
		[0.001884]	[0.001085]	[0.000944]	[0.000891]	[0.00188]	[0.001148]	[0.000987]	[0.00092]
10%	Case	0.009989	0.009989	0.0000	6E-06	0.009316	0.010724	0.00000	1.15E-05
		[0.125977]	[0.125977]	[8.89E-02]	[0.000769]	[0.124013]	[0.120946]	[8.92E-02]	[0.000759]
	Fixed - X	0.005759	0.004233	8.1E-06	0.000187	0.008421	0.000844	4.1E-06	4E-06
		[0.290573]	[0.076575]	[0.003555]	[0.00382]	[0.301482]	[0.080481]	[0.003649]	[0.003842]
	Residual	0.000914	0.002043	0.000272	8.87E-05	0.000956	0.002052	0.000278	0.001117
		[0.000855]	[0.001014]	[0.0007]	[0.000701]	[0.000845]	[0.001052]	[0.000716]	[0.000704]
20%	Case	0.009989	0.010985	4.19E-05	8.29E-05	0.009316	0.009816	3.65E-05	4.18E-05
		[0.125977]	[0.11678]	[0.000456]	[0.000552]	[0.124013]	[0.121364]	[0.000464]	[0.00053]
	Fixed - X	0.001231	1.51E-05	0.000136	0.0001	0.000791	6.27E-05	3.43E-05	9.5E-06
		[0.079335]	[0.003829]	[0.002352]	[0.002397]	[0.086807]	[0.003802]	[0.002405]	[0.002379]
	Residual	0.001369	0.001738	0.000469	0.000544	0.001401	0.001725	0.000496	0.00055
		[0.000797]	[0.000775]	[0.000515]	[0.000524]	[0.000784]	[0.000759]	[0.000531]	[0.000522]

Table 5: Comparison between bootstrapping approaches using MSE

Sample size	Coefficient		Case Bootstrapping		Residual Resampling Bootstrapping		Fixed-X Bootstrapping	
			Boot		Boot		Boot	
			1000	5000	1000	5000	1000	5000
n=20	5%	β_0	0.005657	0.005433	0.001923	0.002085	0.102694	0.098809
		β_1	0.017544	0.01895	5.54E-06	5.7E-06	0.254964	0.24276
	10%	β_0	0.004004	0.003938	0.002414	0.002373	0.081176	0.074964
		β_1	0.01597	0.015466	1.57E-06	1.63E-06	0.084466	0.090962
	20%	β_0	0.004004	0.003938	0.003394	0.00333	0.061197	0.063183
		β_1	0.01597	0.015466	2.51E-06	2.58E-06	0.006296	0.007536
n=40	5%	β_0	0.004004	0.003938	0.003968	0.004127	0.039241	0.036911
		β_1	0.01597	0.015466	3.22E-06	3.42E-06	0.050534	0.044939
	10%	β_0	0.004004	0.004041	0.003012	0.003127	0.030003	0.02893
		β_1	0.01597	0.014743	5.2E-06	5.32E-06	0.005882	0.006478
	20%	β_0	0.003984	0.003973	0.002261	0.002295	0.031809	0.030221
		β_1	0.013758	0.014826	3.62E-06	3.55E-06	1.47E-05	1.45E-05
n=80	5%	β_0	0.003984	0.003973	0.001603	0.001579	0.015066	0.250252
		β_1	0.013758	0.014826	8.94E-07	9.76E-07	0.002537	0.003283
	10%	β_0	0.000449	0.000449	0.001443	0.001424	0.01466	0.013858
		β_1	0.007906	0.007948	5.64E-07	5.9E-07	1.26E-05	1.33E-05
	20%	β_0	0.000852	0.000865	0.00121	0.001135	0.015394	0.014777
		β_1	2.1E-07	2.16E-07	4.85E-07	5.28E-07	5.55E-06	5.78E-06
n=100	5%	β_0	0.00055	0.000578	0.00209	0.002086	0.010362	0.010602
		β_1	4.27E-05	1.76E-05	2.61E-06	0.09082	0.000315	0.000639
	10%	β_0	0.000571	0.000616	0.001559	0.109416	0.010875	0.010875
		β_1	5.91E-07	5.77E-07	4.99E-07	0.089931	1.46E-05	1.48E-05
	20%	β_0	0.000622	0.000635	0.002329	0.117004	0.011762	0.011699
		β_1	3.11E-07	2.83E-07	5.7E-07	0.090331	5.76E-06	5.66E-06





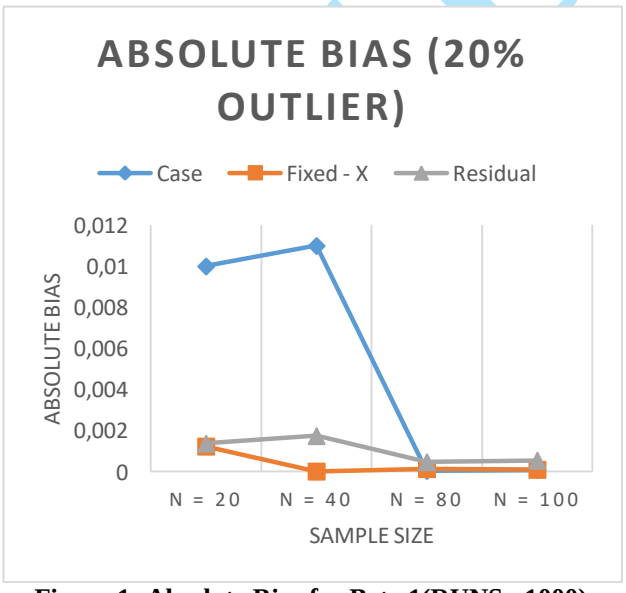
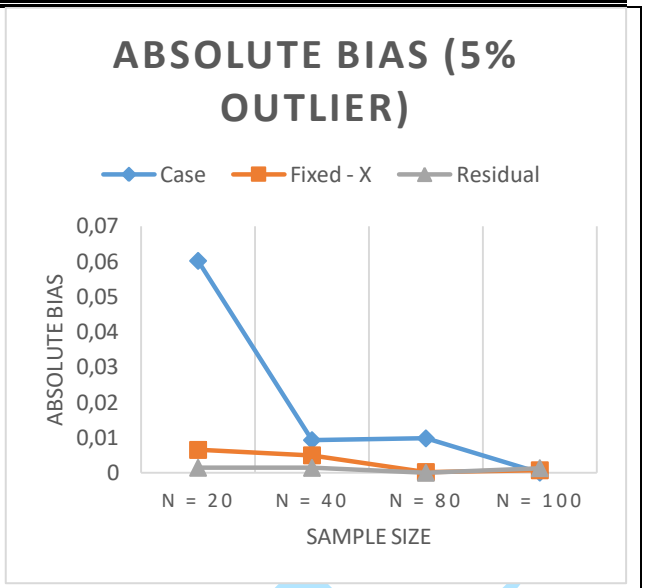
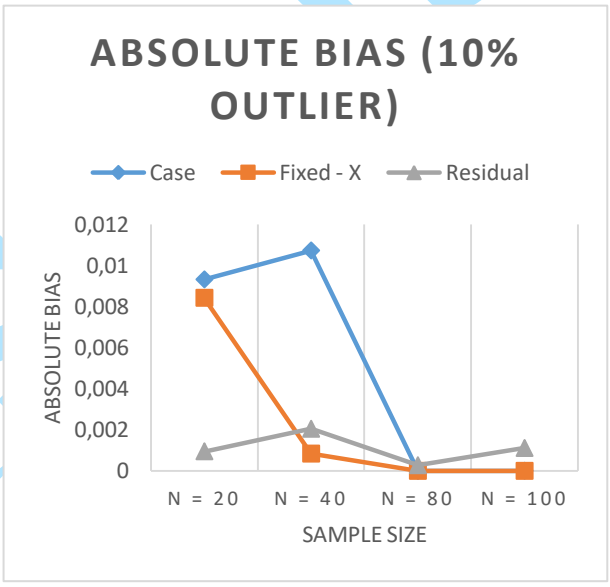


Figure 1: Absolute Bias for Beta 1(RUNS =1000)





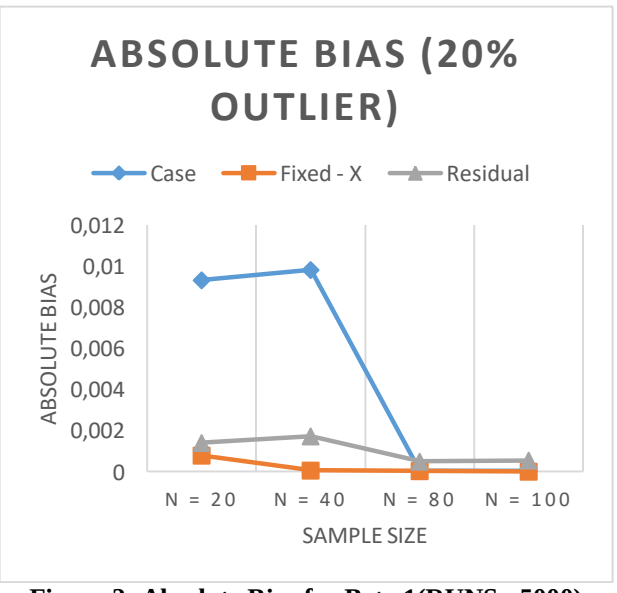
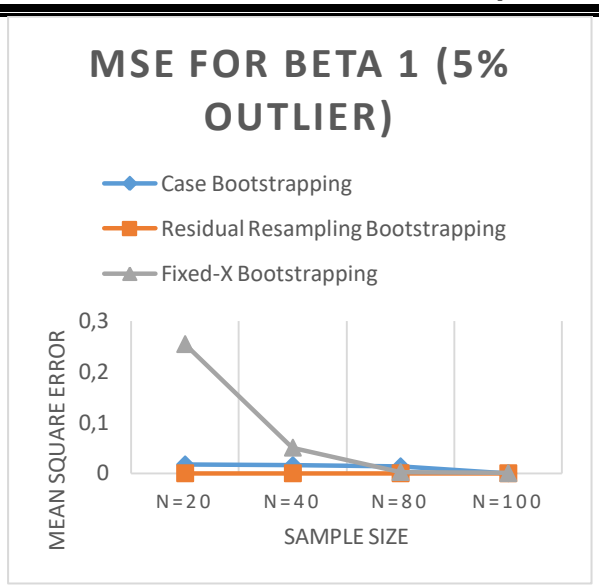
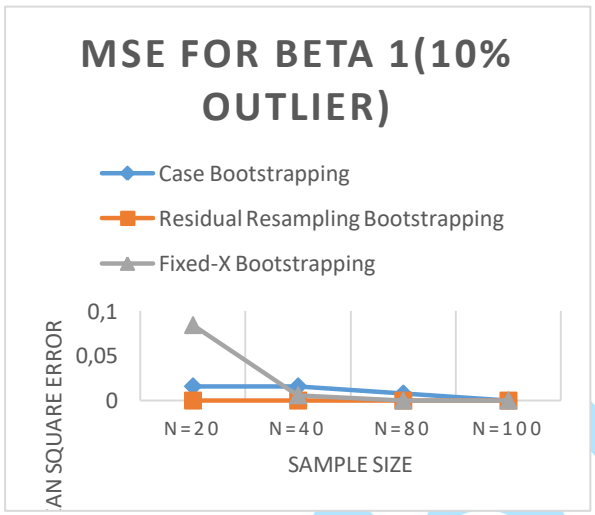


Figure 2: Absolute Bias for Beta 1(RUNS =5000)





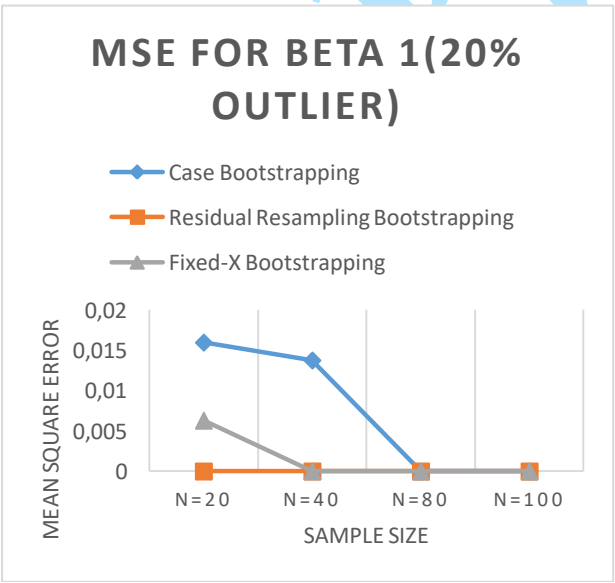
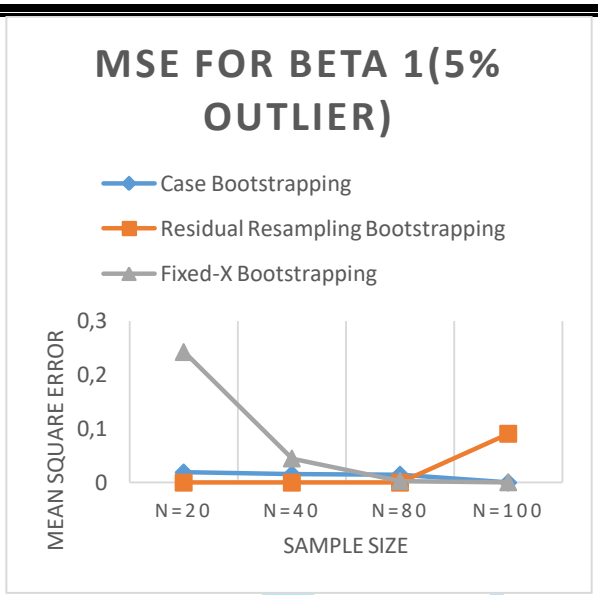
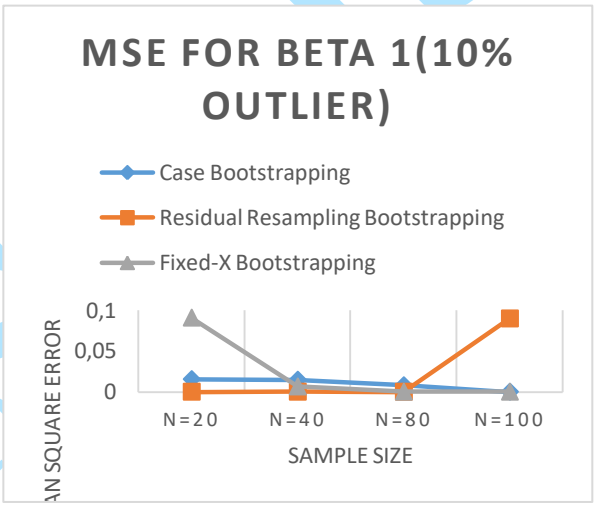


Figure 3: Mean square error for beta 1(RUNS = 1000)





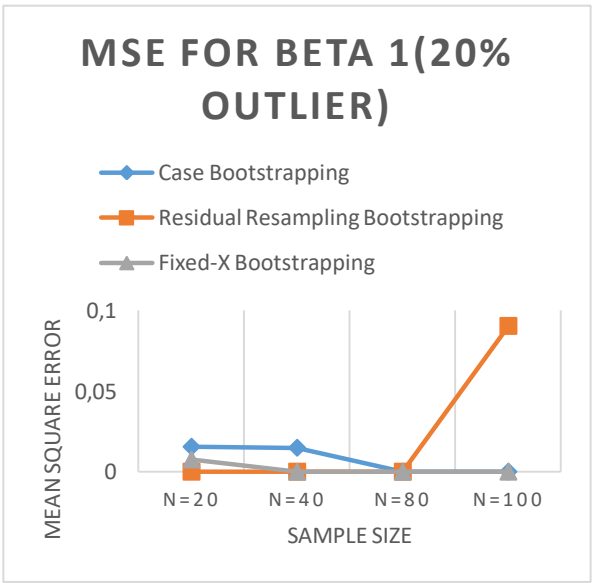


Figure 4: Mean square error for beta 1(RUNS = 5000)

Table 4 shows the summary of the results of the three bootstrap estimators using absolute bias and standard error by percentage outlier, sample size and number of runs respectively. Absolute bias explains how far the estimates are from the true values, the estimator with the least absolute bias is therefore the best.

At 5% outliers, the residual resampling method produces the least absolute bias estimates and standard errors at all the sample sizes considered followed by the Fixed-X method.

At 10% and 20% contamination levels, the residual resampling method performs best with least absolute bias estimates and standard errors obtained only at N = 20 and N = 40, while the case bootstrap method is best at sample sizes N = 80 and N = 100. However, no clear pattern is seen in the performances of the estimators as the number of runs increase from 1000 to 5000 but in most cases, the estimates produced by the methods are smaller at 5000 runs.

4.1 Comparison between Case Bootstrapping, Residual Resampling Method, and Fixed-X Bootstrapping using MSE

Table 5 shows the comparison of the three techniques under consideration in terms of the mean square error. Mean square error shows the degree of squared deviations of the errors produced by the estimators. Residual Resampling bootstrapping technique produces smaller MSEs for 5%, 10% and 20% outlier. The bootstrap method exhibits an asymptotic property, that is, the mean square errors of the estimators decrease consistently as the sample size increases across all number of boots considered.

The Figures 1-4 show the graphical display of the behaviours of the slope, $\hat{\beta}_1$ estimators using absolute bias and mean square error.

CONCLUSION

In this paper, we evaluated different bootstrapping estimators for estimating the parameters of a linear model contaminated different percentages of outliers. The bootstrap techniques used are; residual resampling, Fixed-X Bootstrap and Case Bootstrap methods.

The simulations showed that Fixed-X bootstrapping performs better than the Case Bootstrapping and Residual Resampling when considering 5% 10% and 20% levels of outliers. It was also deduced that the performance of these bootstrapping approaches varied when considered under different Booting level that is, 1000 and 5000. The higher the booting, the better or closer the estimates to the true value.

Another interesting result of this study is the good performance of bootstrap method. The bootstrap method displayed an asymptotic property with the mean square errors of the estimators decreasing as the sample size increases at boot 1000 to 5000.

In conclusion, when considering a linear model with inclusion of outliers in the dataset and the Bootstrapping approach is to be used, Case Bootstrapping is recommended. Theoretically, the residual bootstraps always generate datasets with the same design as the original data, it is therefore expected to perform better in situations where the design is not similar for every individual.

REFERENCES

[AA13] **S. Alamgir, A. Amjad** - *Split Sample Bootstrap Method.* World Applied Sciences Journal 21 (7): 983-993, 2013.

[U+09] **H. S. Uraibi, H. Midi, B. A. Talib, J. H. Yousif** - *Linear Regression Model Selection Based on Robust Bootstrapping Technique*. American Journal of Applied Sciences 6 (6): 1191-1198, 2009.

[ZRM14] **S. M. Zahari, N. M. Ramli, B. Mokhtar** - *Bootstrapped Parameter Estimation in Ridge Regression with Multicollinearity and Multiple Outliers*. Journal of Applied Environmental and Biological Sciences, J. Appl. Environ. Biol. Sci., 4(7S)150-156, 2014.