Pak. J. Statist.
2021 Vol. 37(2), 121-134

ON THE PERFORMANCE OF WILD BOOTSTRAP BASED ON MM-GM6 ESTIMATOR IN THE PRESENCE OF HETEROSCEDASTIC ERRORS AND HIGH LEVERAGE POINTS

Osama, A.A. Alsattari[1] and **Paul, I. Dalatu**[2§]
[1] Al-Aqsa University, Khan Younis, Gaza strip, Palestine
Email: osama19892014@gmail.com
[2] Adamawa State University, Mubi-Nigeria
Email: dalatup@gmail.com
§ Corresponding author

ABSTRACT

The violation of constancy of variance of error terms causes the problem of heteroscedasticity. Even though the ordinary least squares (OLS) estimates are unbiased in the existence of heteroscedasticity problem in a data set, the standard errors of the parameter estimates are biased. This renders the estimator inefficient. As an alternative, a weighted residual (wild bootstrap) may be used to remedy this problem. However, the weakness of wild bootstrap is that, in the presence of outliers, the estimates of the standard errors become large. For the sake of rectifying this problem, a wild bootstrap (WB) based on MM estimates is proposed. Nevertheless, this estimator cannot handle well high leverage points (HLP_S). Thus, wild bootstrap based on MM-GM6 estimator is proposed so that the problems of both heteroscedasticity and outliers can be rectified. The performance of the proposed method denoted as WBootMM-GM6-Liu is compared with some existing techniques such as wild bootstrap of OLS (WBootOLS), wild bootstrap of Liu (WBootLiu) and wild bootstrap based on MM estimator denoted as (WBootMM-Liu). The numerical results indicate that the developed method outperformed other methods for data having both problems of heteroscedasticity and high leverage points.

KEYWORDS

Heteroscedasticity, outliers, wild bootstrap, weighted residual, MM-GM6 estimator, high leverage points

INTRODUCTION

Multiple regression analysis is a statistical technique used widely for modelling and analysing the relationship between one dependent variable and two or more predictor variables.

The standard model of linear regression can be defined as:

$$y = X\beta + \mu \tag{1}$$

where, y is an $(r \times 1)$ vector of dependent variable, X is an $(r \times k)$ data matrix of independent variables, β is a $(k \times 1)$ vector of parameters, and μ is an $(r \times 1)$ vector of

random errors with distribution of $\mu \sim NID(0, \sigma^2)$. Homoscedasticity refers to the situation when the variance of the error terms is constant. Heteroscedasticity is a common problem in a linear regression model, which occurs when the variance of the error terms are not constant (Lukman et al., 2016). In this situation, the OLS estimator is no longer efficient. There are several methods to rectify the problem of heteroscedasticity (Habshah et al., 2011). A weighted bootstrap method proposed by Wu (1986) is one of the alternative methods to rectify this problem. Liu (1988) suggested a wild bootstrap approach that, under both homoscedastic and heteroscedastic models, is slightly different from the weighted bootstrap method and worked better. Rana et al., (2012) suggested that there is evidence that the presence of outliers due to the use of ordinary least squares (OLS) in their algorithm causes such wild bootstrap estimators to suffer a huge setback. So, in the construction of the robust wild bootstrap process, they implemented the robust MM estimator. The MM estimator, however, does not have limited impact properties. Hence, in this study, we attempt to improvise the robust wild bootstrap of Rana et al., (2012) by incorporating the MM-GM6 estimator in the establishment of robust wild bootstrap.

WILD BOOTSTRAP TECHNIQUE

Efron (1979) is the first person who introduced the bootstrap technique. In this technique, the theoretical formulation could be replaced by the computer calculations. There are many authors who have used the bootstrap methods (namely Cribari-Neto and Zarkos (1999), Efron (1987), and Efron and Tibshirani (1994). In regression analysis, the most popular and widely used bootstrap technique is the fixed-*x* resampling or bootstrapping the residual suggested by Efron and Tibshirani (1986) and Rana et al., (2012). This classical bootstrap process relies on the classical OLS residuals that can be summarized as follows:

Step 1. Fit a model $y_i = x_i\beta + \varepsilon_i$ using the OLS method to the real data to obtain $\hat{\beta}_{ols}$ and hence the fitted model is $\hat{y}_i = x_i\hat{\beta}_{ols}$.

Step 2. Compute the residuals of the OLS estimate $\hat{\varepsilon}_i^{ols} = y_i - \hat{y}_i$ and each residual $\hat{\varepsilon}_i$ has equal probability, $\frac{1}{n}$.

Step 3. Draw a sample of $\varepsilon_1^*, \varepsilon_2^*, ..., \varepsilon_n^*$ randomly from $\hat{\varepsilon}_i$ with replacement and attached to $\hat{y}_i$ to obtain fixed-*x* bootstrap values y_i^{*b} where $y_i^{*b} = x_i\hat{\beta}_{ols} + \varepsilon_i^{*b}$.

Step 4. The ordinary least squares is then fitted to the bootstrap value y_i^{*b} on the fixed-*x* to obtain $\hat{\beta}_{ols}^{*b}$.

Step 5. Steps 3 and 4 were then repeated for *R* times to obtain $\hat{\beta}_{ols}^{*b1}, ..., \hat{\beta}_{ols}^{*bB}$ where *R* is the bootstrap replications.

This bootstrap is called BootOLS because it depends on the OLS method.

Liu (1988) modified Step 3 of the BootOLS method as follows:

$$y_i^{*b} = x_i\hat{\beta}_{ols} + \frac{t_i^*\hat{\varepsilon}_i}{\sqrt{1-h_{ii}}} \quad (2)$$

where $t_i's$ is a random variable from a standard normal and h_{ii} is the *ith* leverage which represents the diagonal of the projection matrix or hat matrix $\hat{H} = X(X'X)^{-1}X'$ and is denoted by "*H*". The diagonal elements of "$\hat{H}$" matrix are called the hat values denoted by h_{ii}, given by $h_{ii} = x_i^T\left(X^TX\right)^{-1}x_i, i = 1,2,...n.$ The h_{ii} values are often used as a classical diagnostic method to identify the high leverage points. However, the h_{ii} mostly fails to detect HLPs due to the fact that it suffers from the masking and swamping effects. The main reason of the proposed is to improve the masking and swamping effects (Rana et al., 2012).

Wild bootstrap based on Liu denoted as WBootLiu can be performed by selecting t_i^* in the following way.

$$t_i^* = H_iZ_i - E(H_i)E(Z_i), i = 1,2,...,n \text{ and } H_1,H_2,...,H_n \sim iid\ N\left(\frac{1}{2}\left(\sqrt{\frac{17}{6}} - \sqrt{\frac{1}{6}}\right), \frac{1}{2}\right).$$

As well as, $Z_1,Z_2,...Z_n \sim iid\ N\left(\frac{1}{2}\left(\sqrt{\frac{17}{6}} - \sqrt{\frac{1}{6}}\right), \frac{1}{2}\right)$ (Rana et al., 2012).

PROPOSED ROBUST WILD BOOTSTRAP TECHNIQUE

Wu (1986) noted that the objective of wild bootstrap is to estimate the standard errors of estimates that under heteroscedasticity are asymptotically correct. The drawback of the wild bootstrap is that the estimates of the standard errors become high in the presence of outliers. The wild bootstrap based on the MM estimator denoted as WBootMM-Liu is therefore adopted by Rana et al., (2012) further into wild bootstrap algorithm. However, this estimator cannot adequately handle high leverage points (HLP_S) because MM estimator is robust to outlier in y coordinate (Yohai, 1987). It is now evident that the GM6 is robust to high leverage points Ayinde et al., (2015). Therefore, in this paper, we incorporate the MM- GM6 estimator denoted as WBootMM-GM6-Liu in the wild bootstrap algorithm to down weight outliers in *X* and *Y* directions. The algorithm of MM-GM6 wild bootstrap can be summarized as follows:

Step 1. Fit a model $y_i = x_i\beta + \varepsilon_i$ by using the MM estimator to the real data to obtain the robust MM parameters $\hat{\beta}_{MM}$ and then the fitted model is $\hat{y}_i = x_i\hat{\beta}_{MM}$.

Step 2. The residuals of the MM estimate are obtained as $\hat{\varepsilon}_i^{MM} = y_i - \hat{y}_i$. Then, assign the weight of GM6 to each residual $\hat{\varepsilon}_i^{MM}$ to get new weighted residual $\min\left(1, \frac{x_{0.95,p}^2}{MVE}\right) \times \hat{\varepsilon}_i^{MM}$, where MVE is the minimum-volume ellipsoid.

Step 3. The MM estimate's final weighted residuals denoted as $\hat{\varepsilon}_i^{WMM}$ can be calculated by multiplying the new weight obtained in Step 2 with the value of t_i^* to get $\min\left(1, \frac{x_{0.95,p}^2}{MVE}\right) \times \hat{\varepsilon}_i^{MM} \times t_i^*$.

Step 4. A bootstrap sample $\left(y_i^*, X\right)$ is then constructed, where

$$y_i^* = x_i \hat{\beta}_{MM} + \hat{\varepsilon}_i^{WMM} \tag{3}$$

and t_i^* is randomly selected following Liu (1988) procedure.

Step 5. The MM method is then applied to the bootstrap sample $\left(y_i^*, X\right)$ and the resulting estimate can be written as $\hat{\beta}^{*R} = \left(X^T X\right)^{-1} X^T y^*$.

Step 6. Steps 3 to 5 were repeated for *R* times, where *R* is the bootstrap replications.

NUMERICAL EXAMPLE

The performance of the WBootOLS, WBootLiu, WBootMM-Liu and WBootMM-GM6-Liu is evaluated by a numerical example. A set of real data is used to test the efficiency of the preceding methods. The Education Expenditure data is taken from Chatterjee and Hadi (2015). This data set contains three predictor variables each with 50 observations. After checking the data with Diagnostic Robust Generalized Potential (DRGP) Habshah et al., (2009), it is found that observation 49 is a high leverage point and outlier in *y* direction. The WBootOLS$_S$, WBootLiu, WBootMM-Liu and WBootMM-GM6-Liu were then applied to the data set. The fitted values versus residuals are plotted in Figure 1. The two observations in Figure 1 are significant, because 7 is an outlying observation in *Y* direction and 10 is also an outlying observation in *X* direction. Therefore, the two numbered observations played an important role to test the success of are the four estimator techniques results displayed in Table 1. The proposed method is to down weight any outlying observation either in *X* direction or *Y* direction or/ both directions. It has been shown that the proposed method has the smallest standard errors. The heteroscedastic error terms are evident by the funnel shape. The standard errors of the estimates based on 500 bootstrap samples are also exhibited in Table 1. The effect of the HLP$_S$ on the estimates of the standard errors is presented in Figure 2. It is observed that the WBootOLS method perform poorly, due to the presence of outliers. It is evident that the proposed method consistently gives the best result by possessing the smallest standard errors of the parameter estimates, followed by WBootMM-Liu, WBootLiu and WBootOLS. B

Table 1
Standard Errors of the Parameter Estimates for Education Expenditure Data Set

Estimates	WBootOLS	WBootLiu	WBootMM-Liu	WBootMM-GM6-Liu
$\hat{\beta}_0$	114.9278	92.6670	76.1690	55.0694
$\hat{\beta}_1$	0.0114	0.0086	0.0069	0.0050
$\hat{\beta}_2$	0.3022	0.2334	0.1949	0.1433
$\hat{\beta}_3$	0.0496	0.0370	0.0326	0.0240

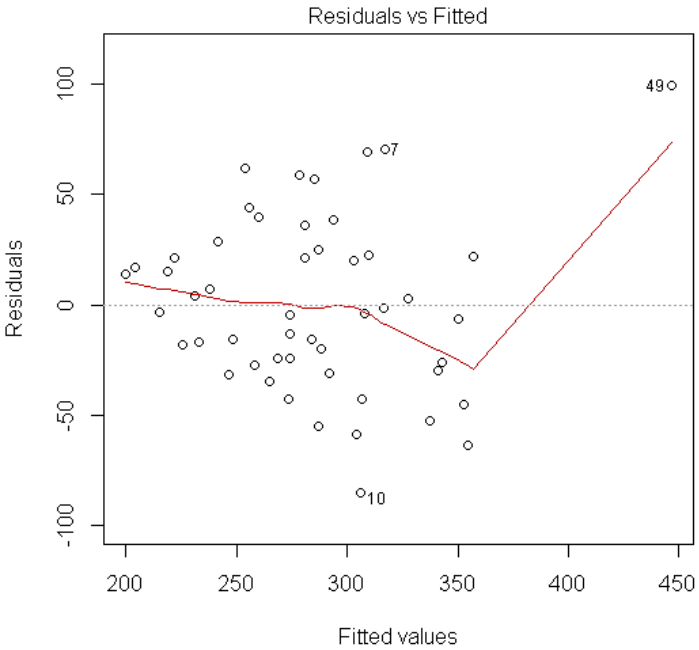

Figure 1: Fitted Values versus Residuals Plot of Education Expenditure Data

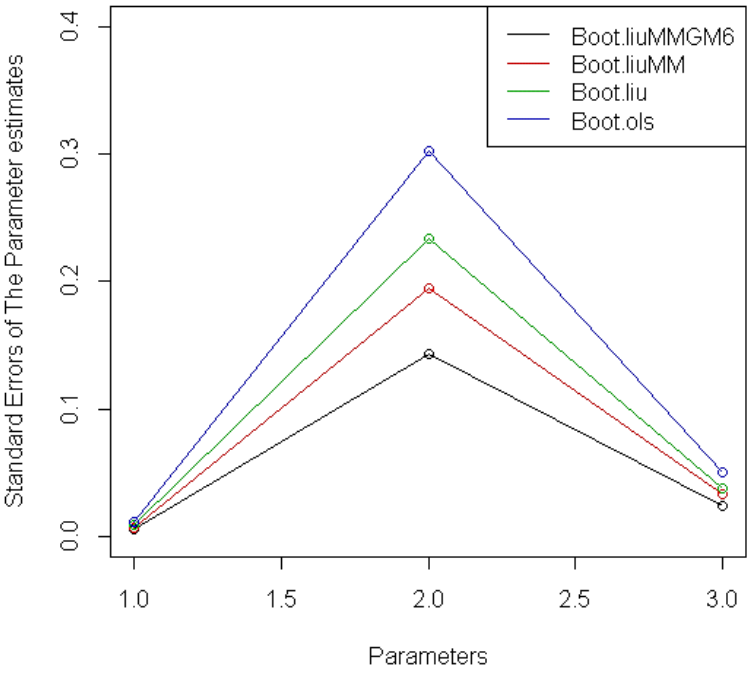

Figure 2: The standard Errors of the Parameter Estimates of WBootOLS, WBootLiu, WBootMM-Liu and WBootMM-GM6-Liu

SIMULATION STUDY

In this section, a simulation study is carried out based on the Monte Carlo procedure to investigate the performance of the proposed method denoted as WBootMM-GM6-Liu in the presence of both heteroscedasticity and high leverage points. In this paper, we consider a multiple linear regression model with two explanatory variables and different sample sizes of 20, 60, and 100. According to Liu (1988), the design of a heteroscedastic model can be written as:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \sigma_i \varepsilon_i \tag{4}$$

where x_{1i} and x_{2i} are generated from $U(0,1)$ for all the sample sizes. The parameters β_0, β_1 and β_2 are set equal to one as the true parameters of this model, and the generation function of heteroscedasticity is $\sigma_i^2 = \exp(\theta_1 x_{1i} + \theta_1 x_{2i})$, where θ_1 is to be 0.4. In this paper, the heteroscedasticity's level is $\xi = \dfrac{\max(\sigma_i^2)}{\min(\sigma_i^2)} = 4$. ε_i's where the error term

generated from $N(0,1)$ for the clean data. For 5% and 10% HLP_S, the 95% and 90% of $\varepsilon_i's$ were generated from $N(0,1)$ and the 5% and 10% were generated from $N(0,20)$. The simulation for each sample size involves a total of 500000 replications with 1000 replications and 500 bootstrap samples each. This simulation was performed based on the procedure of Cribari-Neto and Zarkos (1999) and Furno (1997). The four estimation methods such as WBootOLS, WBootLiu, WBootMM-Liu and WBootMM-GM6-Liu were then applied to the simulated data. The outcomes of simulation study are summarized in Tables (2-4). The standard errors of WBooOLS, WBootLiu, WBootMM-Liu and WBootMM-GM6-Liu are presented in Table 2. When the problem of heteroscedasticity is presented in the simulated data without outliers, the performance of all three methods is close to each other, but WBootMM-GM6-Liu is slightly better than the other classical and robust methods. It can be observed that with the increase in the percentage of HLP_S, the standard errors of the parameter estimates of the classical wild bootstrap increase for various sample sizes. However, our proposed method is less affected by the presence of HLP_S. Table 3 shows the bias of different methods. It can be observed that the bias of WBootOLS, WBootLiu increases with the increase in the percentage of HLP_S. Furthermore, our proposed BootMM-GM6-Liu is slightly biased with the increase in the level of HLP_S. It can be seen from Table 4 that the value of RMSE increases with the increase in the percentage of HLP_S, while it is decreases with the increase in the sample size.

Table 2
Standard Errors of the WBooOLS, WBootLiu, WBootMM-Liu and WBootMM-GM6-Liu Estimates

%outliers	Coeff	WBooOLS	WBootLiu	WBootMM-Liu	WBootMM-GM6-Liu
Sample Size $n = 20$					
0%	β_0	1.9860	2.0207	2.2416	1.6056
	β_1	2.1496	2.1872	2.4263	1.7378
	β_2	2.2862	2.3261	2.5804	1.8482
5%	β_0	12.3313	6.9925	2.7469	1.3614
	β_1	13.3472	7.5685	2.9733	1.4736
	β_2	14.1948	8.0492	3.1621	1.5672
10%	β_0	16.9743	10.4052	3.5968	0.9521
	β_1	18.3727	11.2624	3.8932	1.0868
	β_2	19.5395	11.9777	4.1404	1.0088
Sample Size $n = 60$					
0%	β_0	0.8734	0.8439	0.68964	0.61325
	β_1	1.2106	1.1696	0.95585	0.84998
	β_2	1.0389	1.0038	0.82032	0.72945
5%	β_0	5.28304	3.4899	0.7751	0.6687
	β_1	7.32243	4.8371	1.2897	1.1246
	β_2	6.28411	4.1512	1.1308	0.9962
10%	β_0	6.90025	4.7175	1.0377	0.9151
	β_1	9.56393	6.5386	1.7840	1.5775
	β_2	8.20776	5.6114	1.4519	1.2896
Sample Size $n = 100$					
0%	β_0	0.6001	0.5978	0.4037	0.3578
	β_1	0.9439	0.9404	0.6349	0.5628
	β_2	0.7989	0.7959	0.5374	0.4764
5%	β_0	3.10377	1.9934	0.4960	0.4577
	β_1	4.88213	3.1355	0.8938	0.7744
	β_2	4.13234	2.6539	0.7331	0.6389
10%	β_0	5.0275	3.4509	0.6692	0.5955
	β_1	7.9081	5.4282	1.1742	0.9781
	β_2	6.6935	4.5945	0.9405	0.7912

Table 3
Bias of the WBootOLS, WBootLiu, WBootMM-Liu and WBootMM-GM6-Liu Estimates

%outliers	Coeff	WBootOLS	WBootLiu	WBootMM-Liu	WBootMM-GM6-Liu
Sample Size $n = 20$					
0%	β_0	0.2341	0.1445	0.0136	0.0305
	β_1	0.3425	0.1590	0.0238	0.0080
	β_2	0.4129	0.1877	0.0121	0.0358
5%	β_0	1.4978	0.0927	0.0504	0.0227
	β_1	0.4243	0.1502	0.1599	0.0300
	β_2	0.2191	0.1250	0.0013	0.0479
10%	β_0	3.5791	0.0839	0.1523	0.0479
	β_1	0.3255	0.2403	0.0822	0.0868
	β_2	0.3989	0.0387	0.1051	0.0087
Sample Size $n = 60$					
0%	β_0	0.0932	0.0140	0.0073	0.0176
	β_1	0.0789	0.0172	0.0204	0.0424
	β_2	0.0645	0.0316	0.0050	0.0038
5%	β_0	1.7955	0.0769	0.0163	0.0346
	β_1	0.6089	0.1241	0.0530	0.0596
	β_2	0.2337	0.0468	0.0074	0.0318
10%	β_0	3.6490	0.2816	0.0316	0.0773
	β_1	0.3126	0.0753	0.1086	0.0415
	β_2	0.4250	0.2709	0.0103	0.0026
Sample Size $n = 100$					
0%	β_0	0.0543	0.0251	0.0002	0.0071
	β_1	0.0761	0.0651	0.0008	0.0040
	β_2	0.0346	0.0095	0.0137	0.0183
5%	β_0	1.5640	0.0743	0.0246	0.0152
	β_1	0.8576	0.0331	0.0269	0.0180
	β_2	0.9013	0.0292	0.0076	0.0106
10%	β_0	3.5980	0.1555	0.0574	0.0020
	β_1	0.5160	0.2651	0.0199	0.0704
	β_2	0.8585	0.1263	0.0237	0.0174

Table 4
***RMSE* of the WBootOLS, WBootLiu, WBootMM-Liu and WBootMM-GM6-Liu Estimates**

%outliers	Coeff	WBootOLS	WBootLiu	WBootMM-Liu	WBootMMGM6-Liu
Sample Size $n = 20$					
0%	β_0	1.8155	2.1407	2.7537	1.6805
	β_1	2.2349	2.6078	3.2519	2.0856
	β_2	2.1599	2.4831	3.1675	1.9397
5%	β_0	9.4067	3.8727	3.1629	1.3031
	β_1	13.4186	10.2546	3.6883	1.6157
	β_2	10.4096	3.1602	3.8059	1.5250
10%	β_0	12.9611	4.0081	3.8915	1.3139
	β_1	15.9545	9.5572	4.7758	1.6134
	β_2	15.0627	4.8645	4.7253	1.5661
Sample Size $n = 60$					
0%	β_0	0.7165	0.7188	0.6000	0.4992
	β_1	1.2094	1.1729	0.9929	0.9133
	β_2	1.0419	1.0644	0.8949	0.7422
5%	β_0	4.7503	3.1186	0.7525	0.6276
	β_1	8.2838	7.1643	1.2228	1.0559
	β_2	6.3874	5.2139	1.0859	0.9349
10%	β_0	7.2108	4.7696	0.9540	0.8875
	β_1	11.7072	10.5537	1.6308	1.4928
	β_2	8.3711	6.7133	1.3163	1.2957
Sample Size $n = 100$					
0%	β_0	0.5512	0.5755	0.3653	0.3508
	β_1	0.9691	0.9937	0.6666	0.6119
	β_2	0.8047	0.8106	0.5618	0.4818
5%	β_0	3.4920	2.4610	0.4738	0.4360
	β_1	4.7654	3.5670	0.8388	0.7418
	β_2	3.9779	2.5507	0.6932	0.6156
10%	β_0	6.4221	4.1885	0.6320	0.5618
	β_1	8.3462	6.5994	1.0978	0.9012
	β_2	7.1185	5.5802	0.8852	0.7471

The effect of HLP_S on the standard errors of the parameter estimates is displayed in Figure 3-8. It can be observed from the plots that the standard errors of the parameter estimates of the proposed WBootMM-GM6-Liu outperforms other methods at both percentages of 5% and 10% HLP_S. This is evident by having the smallest standard errors.

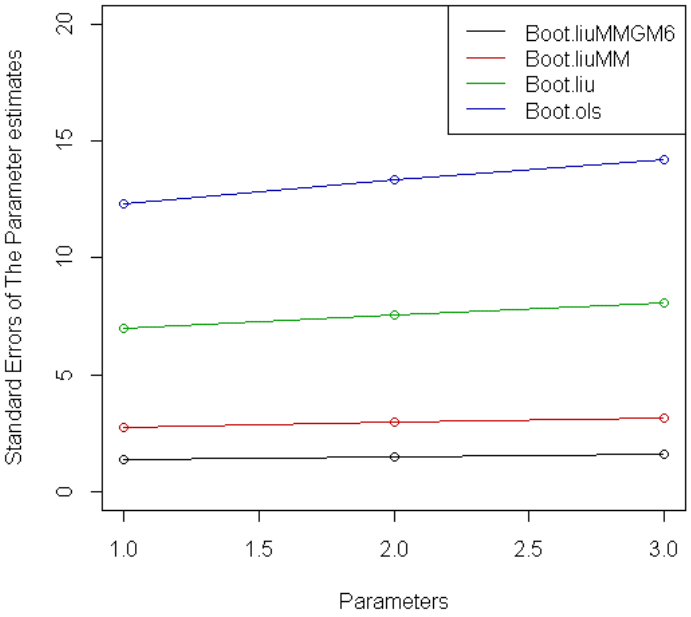

Figure 3: The Effect of 5% HLPs on the Standard Errors of the Parameter Estimates when Sample Size $n = 20$

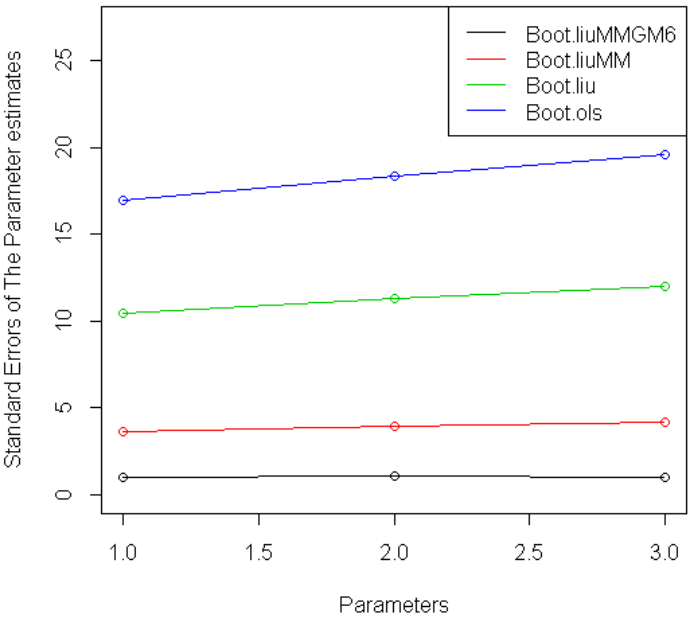

Figure 4: The Effect of 10% HLPs on the Standard Errors of the Parameter Estimates when Sample Size $n = 20$

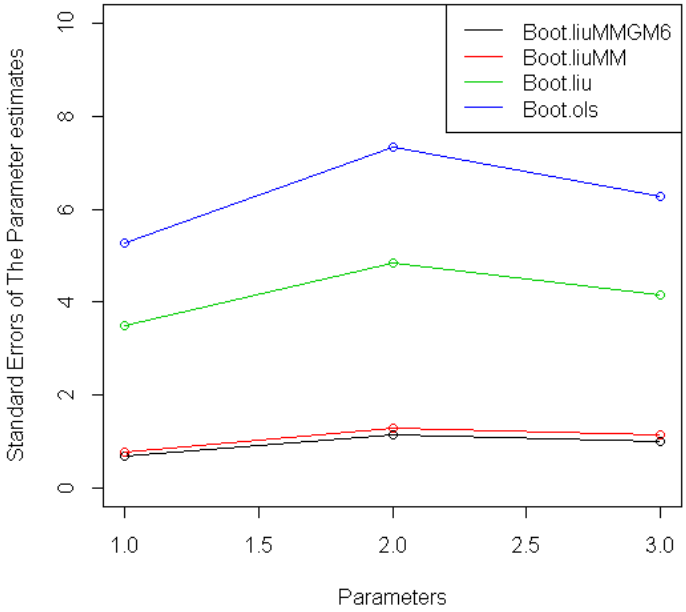

Figure 5: The Effect of 5% HLPs on the Standard Errors of the Parameter Estimates when Sample Size $n = 60$

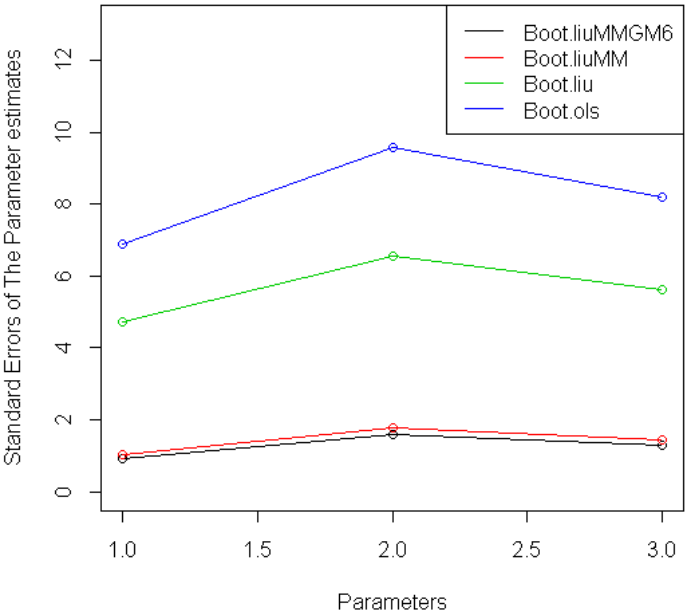

Figure 6: The Effect of 10% HLPs on the Standard Errors of the Parameter Estimates when Sample Size $n = 60$

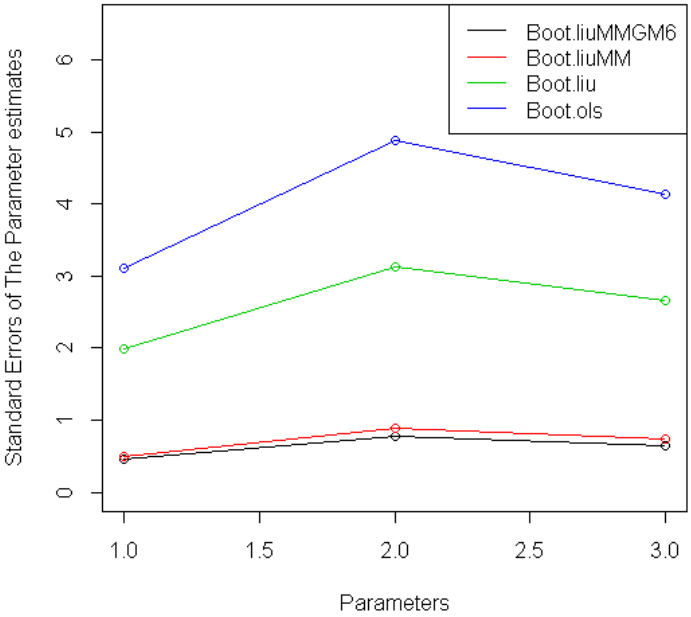

Figure 7: The Effect of 5% HLPs on the Standard Errors of the Parameter Estimates when Sample Size $n = 100$

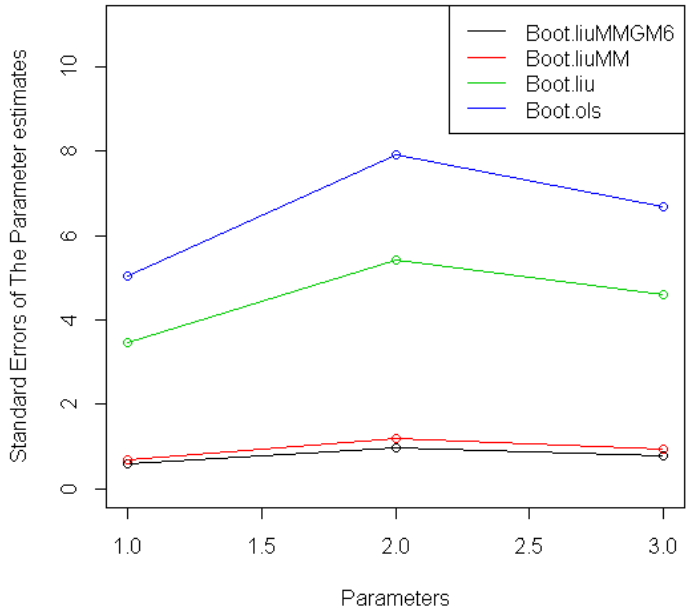

Figure 8: The Effect of 10% HLPs on the Standard Errors of the Parameter Estimates when Sample Size $n = 100$

CONCLUSION

The main objective of this paper is to develop a robust wild bootstrap method for multiple regression model in the presence of heteroscedasticity and high leverage points. In this regard, we proposed robust wild bootstrap method, namely, WBootMM-GM6-Liu based on the MM-GM6 estimator. It can be observed from the simulation study and the real data set that the suggested method has a good performance compared with other existing methods in the existence of heteroscedasticity and high leverage points.

REFERENCES

1. Ayinde, K., Lukman, A.F. and Arowolo, O.T. (2015). Robust regression diagnostics of influential observations in linear regression model. *Open Journal of Statistics*, 5, 1-11.
2. Chatterjee, S. and Hadi, A.S. (2015). *Regression analysis by example*. John Wiley & Sons.
3. Cribari-Neto, F. and Zarkos, S.G. (1999). Bootstrap methods for heteroskedastic regression models: evidence on estimation and testing. *Econometric Reviews*, 18(2), 211-228.
4. Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in Statistics* (pp. 569-593). Springer, New York, NY.
5. Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1), 54-75.
6. Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397), 171-185.
7. Efron, B. and Tibshirani, R.J. (1994). *An Introduction to the Bootstrap*. CRC press.
8. Furno, M. (1997). A robust heteroskedasticity consistent covariance matrix estimator. *Statistics: A Journal of Theoretical and Applied Statistics*, 30(3), 201-219.
9. Habshah, M., Norazan, M.R. and Rahmatullah Imon, A.H.M. (2009). The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics*, 36(5), 507-520.
10. Lukman, A.F., Oranye, E., Okegbade, I. and Arowolo, O.T. (2016). Weighted Cochrane Two Stage Estimator for Handling Autocorrelation and Heteroscedasticity. *Journal of the Nigeria Association of Mathematical Physics*, 34(1), 209-212.
11. Midi, H., Rana, M.S. and Imon, A.R. (2009). The performance of robust weighted least squares in the presence of outliers and heteroscedastic errors. *WSEAS Transactions on Mathematics*, 8(7), 351-361.
12. Rana, S., Midi, H., and Imon, A.H.M.R. (2012). Robust wild bootstrap for stabilizing the variance of parameter estimates in heteroscedastic regression models in the presence of outliers. *Mathematical Problems in Engineering*, Volume 2012, Article ID 730328, 1-14.
13. Liu, R.Y. (1988). Bootstrap procedures under some non-iid models. *The Annals of Statistics*, 16(4), 1696-1708.
14. Wu, C.F.J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4), 1261-1295.
15. Yohai, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15(2), 642-656.