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

P. I. Dalatu

Department of Mathematics, Faculty of Science, Adamawa State University, Mubi-Nigeria

Corresponding author: dalatup@gmail.com

ABSTRACT

The wild bootstrap is studied in the context of regression models with heteroscedastic disturbances. Specifically, heteroscedasticity is a systematic change in the spread of the residuals over the range of measured values. Therefore, heteroscedasticity is a problem because Ordinary Least Squares (OLS) regression assumes that all residuals are drawn from a population that has a constant variance. The violation of constancy, variance of error terms causes the problem of heteroscedasticity. Ordinary least squares estimate is no longer efficient in the presence of heteroscedasticity in a data set, because the OLS *estimates will be biased and inconsistent. To remedy this problem, we propose a Robust Wild Bootstrap for stabilizing the variance of the regression estimates where heteroscedasticity and outlier occur at the same time. However, the weakness of wild bootstrap is that, in the presence of outliers the estimates of the standard errors become large. Therefore, a robust wild bootstrap is proposed based on MM-GM6 estimator so that the problems of both heteroscedasticity and outliers can be rectified. The results show that the proposed method performs better than the existing ones such as* OLS *, Wu, and Liu.*

Keywords*: Heteroscedasticity; Outliers; Wild bootstrap; High leverage points.*

1 INTRODUCTION

Bootstrap technique was proposed by [1]. It is a statistical method that can replace theoretical formulation with extensive use of computer. This method does not depend on the distributional assumptions and is able to estimate the standard errors of parameter estimates without theoretical calculations. There are many papers which deal with bootstrap methods see [2, 3, 4]. However, according to [5], several methods can be used for linear regression analysis, including ordinary least squares (which is said to perform poorly in the case of heteroscedastic errors), weighted least squares (down-weighs data points with a high residual variance in order to address heteroscedasticity) and bootstrap (which depends on fewer assumptions on the residuals and can be used with heteroscedasticity and outliers). Within the bootstrap methods, there are variations in how to resample: the data points themselves, the residuals, or generating new residuals from a given distribution. A technique used to overcome this difficulty is the so-called wild bootstrap, developed in [6] following suggestions in [7]. Liu establishes the ability of the wild bootstrap to provide refinements for the linear regression model with heteroscedastic errors. Further evidence was provided in [8]. Both Liu and Mammen show under a variety of regularity conditions, that the wild bootstrap is asymptotically justified, in the sense that the asymptotic distribution of various statistics is the same as the asymptotic distribution of their wild bootstrap counterparts. They also show that, in some circumstances, asymptotic refinements are available that lead to agreement higher than leading order between the asymptotic distributions of the raw and bootstrap statistics [9]. According to [10], within the context of OLS regression, heteroscedasticity can be induced either through the way in which the dependent variable is being measured or through how sets predictors are being measured. Imagine if one were to analyse the amount of money spent on a family vacation as a function of the income of said family. In theory, low-income families would have limited budgets and could only afford to go to certain places or stay in certain types of hotels. High income families could choose to go on cheap or expensive vacations, depending on other factors not necessarily associated with income. Therefore, as one

progresses from lower to higher incomes, the amount of money spent on vacations would become more and more viable depending on other characteristics of family itself. [11] pointed out that the problem of classical bootstrap is the proportion of outliers in the bootstrap sample might be greater than that of the original data. However, the entire inferential procedure of bootstrap would be erroneous in the presence of outliers. Multiple regression analysis is a statistical technique used widely for modelling and analysing the relationship between one dependent variable and two or more independent variables. The standard linear regression model can be defined as:

$$Y = X\beta + \varepsilon \tag{1}$$

Where $Y = (y_1, y_2, ..., y_n)^T$, $X = (x_1, x_2, ..., x_n)^T$ and $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T$. In equation (1), β is a $k \times 1$ vector of unknown parameters, Y is an $n \times 1$ vector, X is an $n \times k$ data matrix of independent variables, and ε is an $n \times k$ vector of unobservable random errors, such that $\varepsilon \sim NID(0, \sigma^2)$. The assumption of stability of variance $Var(\varepsilon_i) = (\sigma I)$ is often violated. As a consequence, [7] proposed a wild bootstrap that can be utilized to evaluate the standard errors which are asymptotically and accurate under non stability of variance. [5], proposed another wild bootstrap method which is slightly different from Wu's method. Nevertheless, [12] stated that there is evidence that the Wild Bootstrap estimator suffer a huge set back in the presence of a few atypical observations that we often call outliers. Therefore, they incorporated the robust MM estimator. However, the MM-estimator does not have a bounded influence property. On the other hand, the GM6 estimator is robust in the X-direction. This paper, we want to investigate an alternative wild bootstrap in the wild bootstrap algorithm based on robust MM-GM6 estimator. This paper is organised as follows: Section 2 provides the materials and methods, Section 3 presents the results and discussion, Section 4 gives the conclusion of the study.

2 MATERIALS AND METHODS

2.1 Classical Wild bootstrap technique

Heteroscedasticity is a common problem in linear regression model, occurs when the variance of error terms is not stable. In this case, OLS estimator is no longer efficient. The fixed *x* bootstrapping the residual method is suggested by [1]. In section 2.2 below, we made improvement on this classical wild bootstrap which fails to identifies outliers on both X and Y directions. This bootstrapping procedure is based on the ordinary least squares residuals that can be summarized as follows:

Step 1. Fit a model $y_i = f(x_i, \beta_{ols})$ by the OLS method to the original sample of observations to get $\hat{\beta}_{ols}$ and hence the fitted model is $y_i = f(x_i, \hat{\beta}_{ols})$.

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

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

Step 4. Fit the OLS to the bootstrap value y_i^{*b} on the Fixed$-x$ to obtain $\hat{\beta}_{ols}^{*b}$.

Step 5. Repeat Steps 3 and 4 for B times to get $\hat{\beta}_{ols}^{*b1}, ..., \hat{\beta}_{ols}^{*bB}$, where B is the bootstrap replications.

This bootstrap is called $Boot_{ols}$ since it is based on the OLS method. [5] slightly modified Step 3 of the OLS bootstrapping procedure to add the weight in the residual as follows.

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

Where, $t_i's$ can be selected from a standard normal and h_{ii} is the ith leverage which represents the diagonal of $Hat\ matrix = X\left(X'X\right)^{-1}X'$.

Liu's bootstrap can be conducted by drawing random numbers t_i^* in the following way. $t_i^* = H_i D_i - E(H_i)E(D_i), i = 1,2,...,n$ and $H_1, H_2, ..., H_n$ are iid normally distributed with mean $\frac{1}{2}\left(\sqrt{\frac{17}{6}} + \sqrt{\frac{1}{6}}\right)$ and variance is $\frac{1}{2}$. As well as, $D_1, D_2, ..., D_n$ are iid normally distributed with mean $\frac{1}{2}\left(\sqrt{\frac{17}{6}} - \sqrt{\frac{1}{6}}\right)$ and variance is $\frac{1}{2}$.

2.2 Proposed Robust Wild Bootstrap Technique

The wild bootstrap is not resistant to outliers because its algorithm is based on the OLS residuals. Thus, in this paper we incorporate the MM- GM6 estimator in the wild bootstrap algorithm. The GM6-estimator aims to down weight outliers both X and Y coordinates to insure that $HLPs$ get lower weights, while the MM estimator is robust to outliers in Y coordinate. The steps of $MM - GM6$ wild bootstrap can be summarized as follows:

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

Step 2. Find the residuals of the MM estimate $\hat{\varepsilon}_i^{MM} = y_i - \hat{y}_i$. Then assign the weight of $GM6$ to each residual such that $w_i = \min\left(1, \frac{X^2_{0.95,P}}{MVE}\right)$, where MVE is the minimum-volume ellipsoid.

Step 3. The final weight residual of the MM estimate denoted by $\hat{\varepsilon}_i^{WMM}$ are formulated by multiplying the weight obtained in Step 2 with the residual of the MM estimates.

Step 4. Construct a bootstrap sample $\left(y_i^*, X\right)$, where

$$y_i^* = x_i\hat{\beta}_{MM} + t_i^* \times \min\left(1, \frac{X^2_{0.95,P}}{MVE}\right)\hat{\varepsilon}_i^{MM} \tag{3}$$

And t_i^* is a random sample following [6] procedure.

Step 5. The MM procedure is then applied to the bootstrap sample $\left(y_i^*, X\right)$ and the resultant estimate is denoted by $\hat{\beta}^{*R} = \left(X^T X\right)^{-1} X^T y^*$.

Step 6. Repeat Steps 4 and 5 for B times, where B is the bootstrap replications.

3 RESULTS AND DISSCUSION

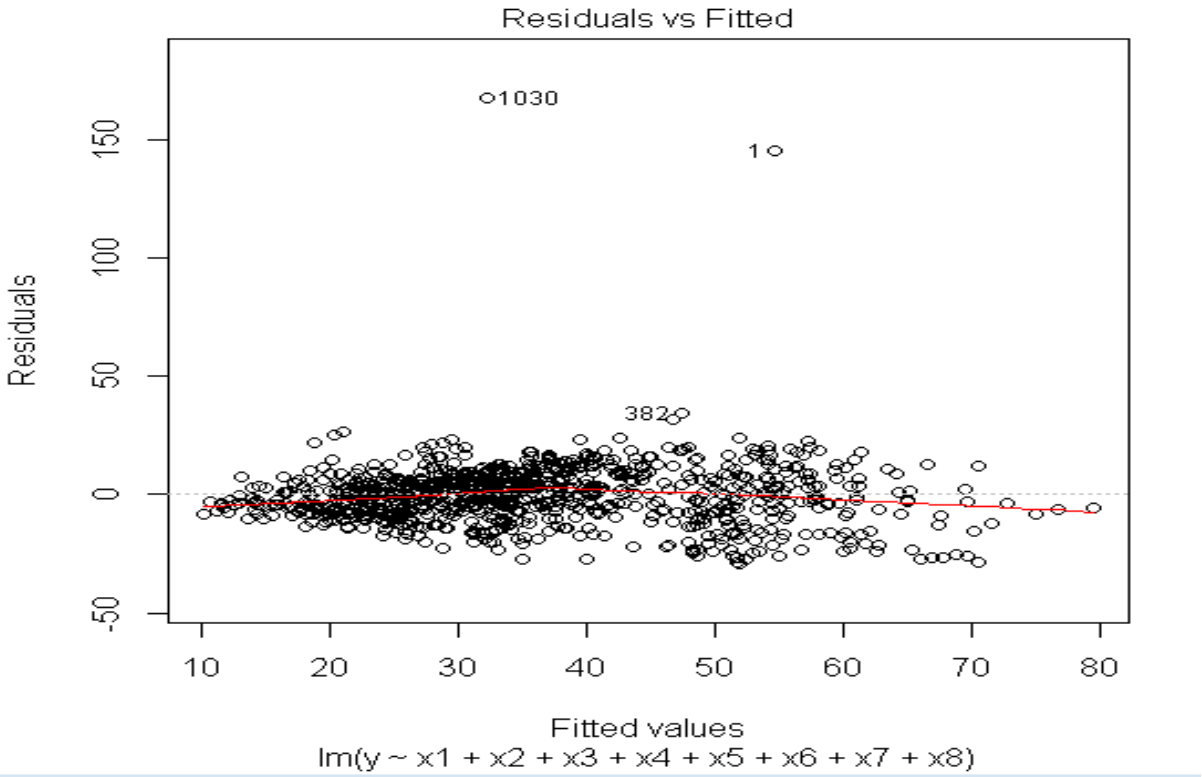

Figure 1. Residuals versus Fitted values plot of Concrete Compressive Strength data.

The residuals versus fitted values are plotted in Figure 1 that show a funnel shape suggesting a heterogeneous error variance for the data.

Table 1. Wild bootstrap standard errors of the parameters for the Concrete Compressive Strength data set.

Standard error (se)	$Boot_{ols}$	$Boot_{wu}$	$Boot_{liu}$	$Boot_{MM\text{-}GM6}$ (liu)
Intercept	31.8681	29.0002	24.5610	.8731
Cement	0.0104	0.0095	0.0075	0.0009
Blast Furnace Slag	0.0123	0.0115	0.0095	0.0011
Fly Ash	0.0153	0.0144	0.0117	0.0014
Water	0.0487	0.0447	0.0378	0.0043
Super plasticizer	0.1162	0.0986	0.0906	0.0101
Coarse Aggregate	0.0112	0.0101	0.0089	0.0010
Fine Aggregate	0.0127	0.0117	0.0098	0.0012
Age	0.0066	0.0059	0.0052	0.0006

The standard errors of the preceding methods based on 500 bootstrap samples are exhibited in Table 1. The average standard errors of the parameter estimates result are computed from data in figure1. The data which has eight predictors and has shown outliers in both X and Y directions. Based on Table 1, the classical and improved wild bootstrapping results are compared. However, at this stage by observing the results, the classical wild bootstraps are most being affected. This is because, the data has outliers in both X and Y directions, which the classical cannot cope up well to identify as improved method. However, as being observed and compared carefully, the standard errors performance under the three classical wild bootstrapping methods are most affected, especially $Boot_{ols}$, because it cannot

detect outliers in Y direction. The $Boot_{wu}$ and $Boot_{liu}$ are the two methods that have their performances are followed closely to $Boot_{MM-GM6}(Liu)$ with smaller standard errors. The two methods can detect both outliers in X and Y directions, but their performances fail short of the improved method. However, based on the data used and the results obtained shows that the wild bootstrap method has less standard errors in all the eight predictors. Therefore, we suggest that researchers may wish to use $Boot_{MM-GM6}(Liu)$ to identify outliers in both X and Y directions in their future researches.

4 CONCLUSION

This paper examines the performance of three classical wild bootstrap methods and proposed wild bootstrap method in the presence of heteroscedasticity and outliers. Standard errors results were obtained based on 500 bootstrap samples to identify heteroscedasticity and outliers. Furthermore, the numerical results show that the proposed wild bootstrap of $Boot_{MM-GM6}(Liu)$ outperforms the three classical wild bootstrap methods in all the eight predictors most especially when both outliers and heteroscedasticity are present in the data.

REFERENCES

1. Jhun, M. and Jeong, H.C. (2000). Applications of bootstrap methods for categorical data analysis. *Computational statistics & data analysis*, vol. 35, no. 1, pp.83-91.
2. Efron, B. and Tibshirani, R. (2014). Computer-Intensive Statistical Methods. *Wiley StatsRef: Statistics Reference Online*, vol. 13, no. 2, pp. 1-15.
3. Efron, B. (2012). Bayesian inference and the parametric bootstrap. *The annals of applied statistics*, vol. 6, no. 4, pp.1971-1983.
4. Candès, E. and Sabatti, C. (2020). Discussion of the Paper: Prediction, Estimation, and Attribution, by B. Efron. *Journal of the American Statistical Association*, vol. 115, no. 530, pp. 656-658.
5. Alrasheedi, M. (2013). Parametric and non-parametric bootstrap: A simulation study for a linear regression with residuals from a mixture of Laplace distributions. *European Scientific Journal*, vol. 9, no. 12, pp. 120-131.
6. Liu, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *The Annals of Statistics*, vol. 16, no. 4, pp. 1696–1708.
7. Wu, F. J. (1986). Jack-knife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, vol. 14, no. 4, pp. 1261–1350.
8. Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statics*, vol.21, pp. 225-285.
9. Flachaire, E. (2005). Bootstrapping heteroscedatic regression models: Wild bootstrap vs. pairs bootstrap. *Computational Statistics and Data Analysis*, vol. 49, no. 2, pp. 361-376.
10. Astivia, O. L. and Zumbo, B. (2019). Heteroscedasticity in multiple regression analysis: What it is, how to detect it and How to solve it with Applications in R and SPSS. *Practical Assessment, Research and Evaluation*, vol. 24, no. 1, pp. 1-17.
11. Salibian-Barrera, M. and Zamar, R. H. (2002). Bootstrapping robust estimates of regression. *The Annals of Statistics*, vol. 30, no. 2, pp. 556-582.
12. Rana, S., Midi, H. and Imon, A.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*, vol. 2012, Article ID730328, pp. 1-14.
13. Yeh, I.C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, vol. 28, no. 12, pp. 1797–1808.