A New Robust Method for Estimating Linear Regression Model in the Presence of Outliers

Taoheedat Alanamu, M.Sc.[*1] **and Gafar Matanmi Oyeyemi Ph.D.**[2]

Department of Statistics, University of Ilorin, PMB 1515, Ilorin, Nigeria.

E-mail: taoheedatalanamu@yahoo.com[*]

ABSTRACT

Ordinary Least-Squares (OLS) estimators for a linear model are very sensitive to unusual values in the design space or outliers among response values. Even single atypical value may have a large effect on the parameter estimates. In this paper, we propose a new class of robust regression method for the classical linear regression model. The proposed method was developed using regularization methods that allow one to handle a variety of inferential problems where there are more covariates than cases. Specifically, each outlying point in the data is estimated using case-specific parameter. Penalized estimators are often suggested when the number of parameters in the model is more than the number of observed data points. In light of this, we propose the use of Ridge regression method for estimating the case-specific parameters.

The proposed robust regression method was validated using Monte-Carlo datasets of varying proportion of outliers. Also, performance comparison was done for the proposed method with some existing robust methods. Assessment criteria results using breakdown point and efficiency revealed the supremacy of the proposed method over the existing methods considered.

(Keywords: robust regression, case indicator, Ridge regression, outlier)

INTRODUCTION

The most widely used technique for fitting models to data is Regression analysis. The multiple linear regression model in terms of the observations can be written in matrix notation as:

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

where $y$is an $n \times 1$ vector of observed response values, **X** is the $n \times p$ matrix of the predictor variables, $\boldsymbol{\beta}$ is the $p \times 1$, and **ε** is the $n \times 1$ vector of random error terms. The aim of regression analysis is to find the estimates of unknown parameters. When a regression model is fitted using ordinary least squares, we get a few statistics to describe a large set of data. These statistics can be highly influenced by a small set of data that is different from the bulk of the data (McCann, 2006). OLS is not robust (even a single outlier can totally offset the OLS estimator). Some statistical techniques have been developed that are not so easily affected by outliers. They are robust methods, such as Least Median of Squares (LMS), Least Trimmed Squares (LTS), Huber M Estimation, MM Estimation, Least Absolute Value Method (LAV) and S Estimation (Yu et al., 2014).

M-estimates (Huber, 1981) are solutions of the normal equation with appropriate weight functions. They are resistant to unusual y observations, but sensitive to high leverage points on x, hence the breakdown point of an M-estimate is $1/n$. The quartile based estimators, Least Median of Squares (LMS) (Siegel 1982) which minimize the median of squared residuals, Least Trimmed Squares (LTS) (Rousseeuw, 1983) which minimize the trimmed sum of squared residuals, and S-estimates (Rousseeuw and Yohai, 1984) which minimize the variance of the residuals all have high breakdown point but with low efficiency. Generalized S-estimates (GS-estimates) (Croux et al., 1994) maintain high breakdown point as S-estimates and have slightly higher efficiency.

MM-estimates proposed by Yohai (1987) can simultaneously attain high breakdown point and efficiencies. Mallows Generalized M-estimates (Mallows, 1975) and Schweppe Generalized M-estimates (Handschin et al., 1975) down weight the high leverage points on x but cannot distinguish “good” and “bad” leverage points, thus resulting in a loss of efficiencies. In addition, these two estimators have low breakdown points when p, the number of explanatory variables, is large. Schweppe one-step (S1S) and Generalized M-estimates (Coakley and Hettmansperger, 1993) overcome the problems of Schweppe Generalized M-estimates and are calculated in one step. They both have high breakdown points and high efficiencies. Recently, Lee et al. (2011) and She and Owen (2011) proposed a new class of robust methods based on the regularization of case specific parameters for each response. The case specific parameter stands as the bedrock of the current study. The approach treats an outlier as missing observation and try to estimates it using regularized as approach because the observation is used as a covariate on its own. This induces collinearity triggered by few observations.

In this study, we review and describe some available robust methods. In addition, a simulation study and a real-life data application are used to compare different existing robust methods. The efficiency and breakdown point (Yu et al. 2014) are two traditionally used important criteria to compare different robust methods. The efficiency is used to measure the relative accuracy of the robust estimate compared to the OLS estimate when the error distribution is exactly normal and there are no outliers. Breakdown point is used to measure the proportion of outliers an estimate can tolerate before it goes to infinity. In this paper, finite sample breakdown point (Yu et al., 2014) is used and defined as follows: Let $z_i = (x_i, y_i)$. Given any sample $z = (z_i, \dots, z_n)$, denote $T(z)$ the estimate of the parameter β. Let $z^{'}$ be the corrupted sample where any m of the original points of z are replaced by arbitrary bad data. Then the finite sample breakdown point δ^* is defined as:

$$\delta^*(z, T) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{z^{'}} \| T(z^{'}) - T(z) \| = \infty \right\},$$

where $\| . \|$ is Euclidean norm.

ROBUST REGRESSION BASED ON REGULARIZATION OF CASE-SPECIFIC PARAMETERS

She and Owen (2011) and Lee et al. (2011) proposed a new class of robust regression methods using the case-specific indicators in a mean shift model with regularization method. A mean shift model for the linear regression is:

$$y = \boldsymbol{X\beta} + \boldsymbol{\gamma} + \varepsilon, \varepsilon \sim N(0, \sigma^2 I) \qquad \textbf{(2)}$$

Where $y = (y_1, y_2, \dots, y_n)^T$, $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)^T$, and the mean shift parameter γ_i is non-zero when the *i*th observation is an outlier and zero otherwise. Due to the sparsity of γ_is, She and Owen (2011) and Lee et al. (2011) proposed to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ by minimizing the penalized least squares using L_1 penalty:

$$L(\beta, \gamma) = \frac{1}{2}\{y - (X\beta + \gamma)\}^T\{y - (X\beta + \gamma)\} + \lambda \sum_{i=1}^{n} |\gamma_i| \qquad \textbf{(3)}$$

Where λ are fixed regularization parameters for $\boldsymbol{\gamma}$. Given the estimate of $\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\beta}}$ is the OLS estimate with y replaced by y − γ. She and Owen (2011) and Lee et al. (2011) proved that the above estimate is in fact equivalent to the M-estimate if Huber’s ψ function is used. However, their proposed robust estimates are based on different perspective and can be extended to many other likelihoods based models.

Proposed Robust Regression Method

The robust regression method proposed here is based on regularization of case specific parameter method originally developed by She and Owen (2011) and Lee et al. (2011). We intend to modify the method using the following steps:

1. Identification of influential (outlying) observations,

2. Estimating the mean shift parameter $\boldsymbol{\gamma}$ using Ridge Regression method,

3. Estimating $\boldsymbol{\beta}$using OLS given $\hat{\boldsymbol{\gamma}}$ from step 2.

Ridge Regression

The Ordinary Least Square (OLS) methods tend to produce estimates that are imprecise and unstable, leading to poor prediction. In order to prevent the difficulties of the OLS method, Hoerl and Kennard (1970) suggested the ridge regression as an alternative procedure to the least square method in regression analysis. The ridge regression which is the linear transformation of the least square method is based on adding a biasing constant λ to the diagonal of X'X matrix before computing β's. Therefore, ridge regression is given by:

β_{ridge} = (X'X+ λI)$^{-1}$X'Y. $\lambda \geq 0$ **(4)**

where λ is the ridge parameter and I is the identity matrix. The value of λ is appropriate between the interval of (0, 1). Note that if λ = 0 the ridge estimator becomes the ordinary least square. The values of λ will be selected by the analyst. The corresponding values of the ridge parameter produces different regression coefficient. As the value of λ increases from zero the smaller the variance, the greater the biased introduced. It is always difficult to select the optimal value of λ that produces the stable regression coefficients. Some various methods are used in selecting the appropriate λ.

Derivation of the Proposed Procedure

Consider the mean shift model:

$$y = X\beta + X_0\gamma + \varepsilon \qquad \textbf{(5)}$$

Where X_0 is the design matrix for the mean shift parameter γ, whose column is either 1 or 0 and sum of each column is 1.

Merging the columns of X and X_0 to obtain X_* and likewise merging the parameters β and γ to obtain θ, thus (15) becomes:

$$y = X_*\theta + \varepsilon \qquad \textbf{(6)}$$

If we define the constrained $\varepsilon'\varepsilon$ as $[\varepsilon'\varepsilon]_\lambda$ given as:

$$[\varepsilon'\varepsilon]_\lambda = \min_\theta (y - X_*\theta)'(y - X_*\theta) + \lambda\theta'\theta$$

Our interest is to estimate the regularization parameter γ using ridge regression method thus; minimizing $[\varepsilon'\varepsilon]_\lambda$. Since $\gamma\epsilon\theta$, estimating θ leads to estimating γ.

Now, the partial derivative of $[\varepsilon'\varepsilon]_\lambda$ with respect to γ is:

$$\frac{\partial(y - X_*\theta)'(y - X_*\theta)}{\partial\theta} = 2X_*'(y - X_*\theta)$$

and

$$\frac{\partial\lambda\theta'\theta}{\partial\theta} = 2\lambda\theta$$

Gathering together gives the first order condition:

$$X_*'y = X_*'X_*\theta + \lambda\theta$$

$$X_*'y = [X_*'X_* + \lambda I]\theta$$

$$\hat{\theta} = [X_*'X_* + \lambda I]^{-1}X_*'y$$

Given the estimate of $\hat{\gamma}$, $\hat{\beta}$ is the OLS estimate with y replaced by y − $\hat{\gamma}$, specifically:

$$\hat{\beta} = (X'X)^{-1}X'(y - \hat{\gamma}) \qquad \textbf{(7)}$$

Simulation Study

In this section, we compare different robust methods and report the mean squared errors (MSE) and relative efficiency of the parameter estimates for each estimation method. We compare the OLS estimate with seven other commonly used robust regression estimates: the M estimate using Huber's ψ function (M-H), the M estimate using Tukey's bi-square function (M-T), the S estimate, the LTS estimate, the LMS estimate, the MM estimate (using bi-square weights and k_1 = 4.68), the LAD and the proposed method (PM). The data generation processes that follow were adapted from Yu et al. (2014).

We generated n samples from the model:

$$Y = X\boldsymbol{\beta} + \varepsilon,$$

where $\boldsymbol{X} = [X_1, X_2, X_3] \sim N_3(\boldsymbol{0}, I)$, $\boldsymbol{\beta} = [1,\ 1,\ 1]$. In order to compare the performance of different methods, we consider the following three cases for the error density of ε and independent variables X.

Case I (With y direction Outlier):

$\varepsilon \sim (1-\pi)N(0,1) + \pi N(0,10^2)$ - contaminated normal mixture with $\pi = 0.1, 0.2, 0.3, 0.4, 0.5$ proportions of contamination.

Case II (With x direction Outlier):

$\boldsymbol{X} \sim (1-\pi)N_3(\boldsymbol{0}, I) + \pi N_3(\boldsymbol{0}, 10^2 I)$ - contaminated multivariate normal mixture with $\pi = 0.1, 0.2, 0.3, 0.4, 0.5$ proportions of contamination.

Case III (With x, y direction Outlier):

$\boldsymbol{X} \sim (1-\pi)N_3(\boldsymbol{0}, I) + \pi N_3(\boldsymbol{0}, 10^2 I)$ – contaminated multivariate normal mixture and $\varepsilon \sim (1-\pi)N(0,1) + \pi N(0, 10^2)$ with $\pi = 0.1, 0.2, 0.3, 0.4, 0.5$ proportions of contamination, and overall cases. The replication size was fixed at 1000.

CRITERIA FOR ASSESSING THE ESTIMATORS PERFORMANCE

Mean Square Error of Parameter

$$MSE_i = \frac{\sum_{j=1}^{p} (\hat{\beta} - \beta)^2}{p}$$

$$AMSE = \frac{\sum_{i=1}^{I} MSE_i}{I}$$

Relative Efficiency of Robust Estimators

$$RE = \frac{AMSE_{OLS}}{AMSE_{RE}}$$

Where I is the number of iteration and p is the number of parameters in the model.

RESULTS

In this section, we present the results of average MSE and RE over iterations used. The proportion of outliers were varied from 0.1 - 0.5. The proposed method denoted by (PM) is compared to eight other methods to examine their robustness and overall efficiency.

Table 1: Average MSE of Various Methods over all Sample Sizes for Case I at $\pi = 0.1, 0.2, 0.3, 0.4, 0.5$.

π	OLS	LA	H-M	B-M	MM	LMS	LTS	S	PM
0.1	0.3504	0.0904	0.0982	0.0902	0.0677	1.9829	0.2115	0.1105	0.03720
0.2	0.6947	0.1918	0.2403	0.2092	0.1182	7.6956	0.2907	0.1365	0.0450
0.3	1.0874	0.4212	0.5151	0.4752	0.4158	11.284	0.5122	0.3925	0.0597
0.4	1.4611	0.7532	0.845	0.8359	0.9152	13.392	1.4485	0.9707	0.1316
0.5	1.8004	1.1028	1.2276	1.2752	1.4132	17.4091	2.6971	1.7331	0.2434

Table 2: Average RE of Various Methods over all Sample Sizes for Case I at $\pi = 0.1, 0.2, 0.3, 0.4, 0.5$.

π	OLS	LA	H-M	B-M	MM	LMS	LTS	S	PM
0.1	7.6400	46.7900	55.6300	69.5900	72.3400	11.1400	13.9100	22.9500	89.2700
0.2	3.6800	32.8800	31.1800	51.9800	53.2700	10.7200	12.9800	20.9400	76.7500
0.3	2.2800	22.6200	15.5800	34.0400	32.7400	10.7600	12.6300	16.6400	65.1500
0.4	1.6200	14.0700	6.7700	16.4800	14.5900	9.4800	11.7400	12.9000	49.5500
0.5	1.2700	8.4100	2.9900	5.5800	4.7300	7.1400	9.7100	7.8200	39.0800

Table 3: Average MSE of Various Methods over all Sample Sizes for Case II at $\pi = 0.1, 0.2, 0.3, 0.4, 0.5$.

π	OLS	LA	H-M	B-M	MM	LMS	LTS	S	PM
0.1	0.5335	0.5013	0.5125	0.4515	0.0714	0.6750	0.1911	0.1153	0.0399
0.2	0.6551	0.6913	0.6610	0.6675	0.1176	0.9939	0.2165	0.1446	0.0489
0.3	0.7118	0.7494	0.7199	0.7290	0.4781	0.7520	0.2431	0.1993	0.0661
0.4	0.7368	0.7665	0.7448	0.7535	0.6983	0.8745	0.3689	0.6320	0.1073
0.5	0.7499	0.7760	0.7552	0.7611	0.7517	1.8071	0.7786	0.8003	0.2266

Table 4: Average RE of Various Methods over all Sample Sizes for Case II at $\pi = 0.1, 0.2, 0.3, 0.4, 0.5$.

π	OLS	LA	H-M	B-M	MM	LMS	LTS	S	PM
0.1	8.04	6.65	7.77	7.48	40.60	11.35	14.03	21.49	86.72
0.2	5.44	4.50	5.22	4.99	19.70	10.55	13.00	18.55	74.66
0.3	4.52	3.92	4.36	4.18	7.41	9.42	11.82	12.85	63.14
0.4	4.16	3.76	4.03	3.89	4.61	3.52	6.49	4.44	49.02
0.5	4.06	3.73	3.98	3.89	4.02	1.72	3.48	3.57	39.14

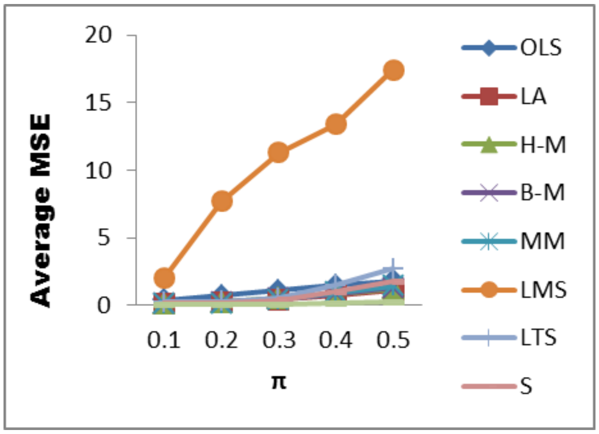

Figure 1: Plot of Average MSE against Outlier Percentages for Case I.

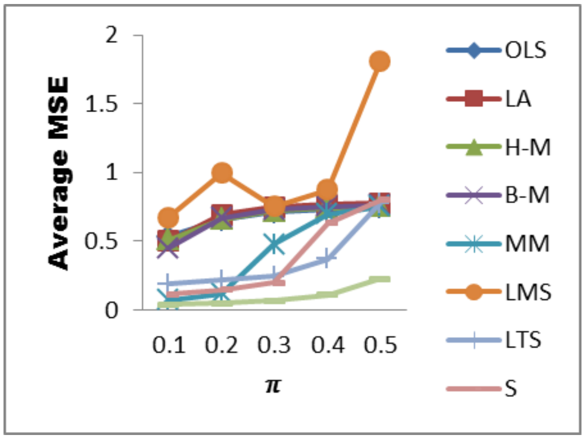

Figure 3: Plot of Average MSE against Outlier Percentages for Case II.

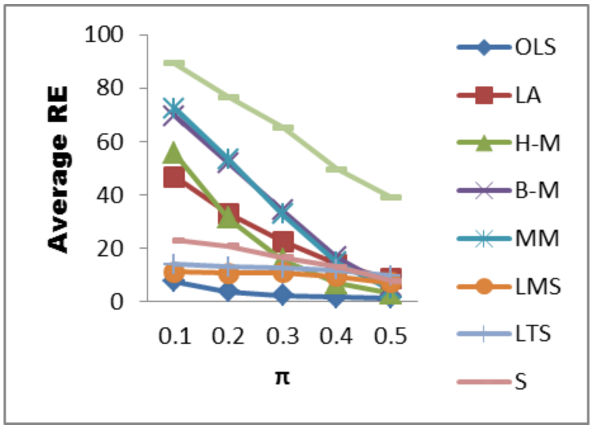

Figure 2: Plot of Average RE against Outlier Percentages for Case I.

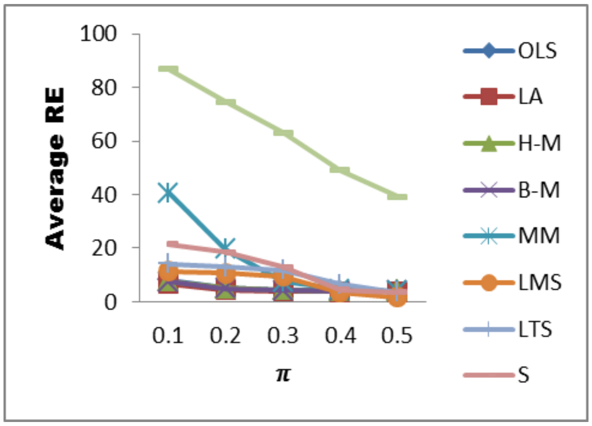

Figure 4: Plot of Average RE against Outlier Percentages for Case II.

Table 5: Average MSE of Various Methods over all Sample Sizes for Case III at $\pi = 0.1, 0.2, 0.3, 0.4, 0.5$.

π	OLS	LA	H-M	B-M	MM	LMS	LTS	S	PM
0.1	0.6607	0.3936	0.4458	0.2482	0.0772	0.5076	0.1676	0.1175	0.0387
0.2	0.8868	0.7285	0.7370	0.5461	0.1402	0.9364	0.2704	0.1750	0.0552
0.3	1.0247	0.8763	0.8760	0.7968	0.2185	0.7385	0.2502	0.2226	0.0748
0.4	1.2362	1.0559	1.0492	1.0050	0.4893	1.1614	0.3922	0.3783	0.1426
0.5	1.2439	0.9327	1.0392	1.0076	0.7203	0.9175	0.2918	0.3927	0.2459

Table 6: Average RE of Various Methods over all Sample Sizes for Case III at $\pi = 0.1, 0.2, 0.3, 0.4, 0.5$.

π	OLS	LA	H-M	B-M	MM	LMS	LTS	S	PM
0.1	5.67	5.88	6.35	15.72	39.86	11.84	15.28	21.55	87.35
0.2	2.98	3.22	3.40	4.00	21.77	10.85	12.60	17.91	74.30
0.3	2.46	2.74	2.86	2.92	11.60	10.60	12.97	14.51	63.96
0.4	1.89	2.24	2.26	2.33	5.28	9.64	11.59	9.98	49.09
0.5	1.93	2.66	2.35	2.40	3.82	6.86	10.69	5.78	39.73

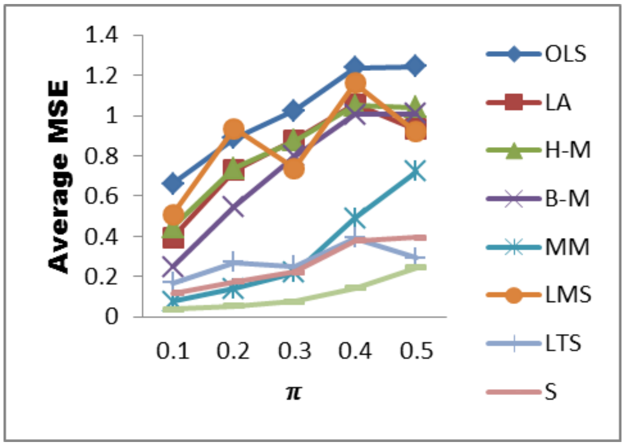

Figure 5: Plot of Average MSE against Outlier Percentages for Case III.

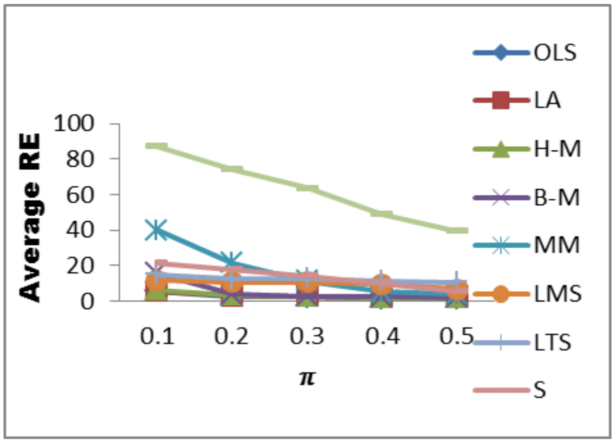

Figure 6: Plot of Average RE against Outlier Percentages for Case III.

DISCUSSION OF RESULTS AND CONCLUSION

The results presented cover mean squared errors (MSE) and relative efficiency (RE) of the parameter estimates for each estimation methods in 1000 replications. Table 1 gives the result for the simulated linear regression model with normal distributed error and 0.1 – 0.5 proportion of outlier (Case I), the proposed method (PM) has the smallest MSE over all proportion of outlier, H-M, B-M and MM have similar lower MSE but they are not as efficient as PM. LS, LMS, LTS, and S has relative larger MSE due to outlier effect.

Table 2 presents the corresponding relative efficiency of the methods in terms of comparing their MSE to the MSE of LS fitting for model without outlier. The result shows that proposed PM method was still able to achieve about 90% efficiency at 0.1 proportion of outlier. The next estimator is MM which has about 72% relative efficiency at 0.1 proportion of outlier. This result indicates that at 0.1 proportion of outlier in Y direction, the proposed PM method is far better than all existing methods considered. Similar results were observed at 0.2, 0.3, 0.4 and 0.5 proportion of outlier in y direction.

In general, for Y direction outlier, the proposed PM method was found to be relatively far better than any of the existing method in terms of efficiency. Table 1 and 2 were also used to assess the breakdown of the methods, it was observed that the breakdown of the proposed PM method is 0.5 compared to other methods considered.

Moving to leverage outlier (X- direction), similar results as observed under y direction outlier were observed. It was also observed that leverage outlier has more effect than the y direction outlier earlier discussed. Further decline in efficiency were observed for the (X, Y) direction outlier condition. Specifically, none of the existing could reasonable withstand 0.1 proportion of outlier.

Whereas, the PM method still yielded about 90% efficiency at 0.1 proportion of outlier. Thus, based on the simulation results for Y, X, and X,Y directions outlier, the best estimator in terms of efficiency and breakdown point of at least 0.5 is PM.

REFERENCES

1. Coakley, C.W. and T.P. Hettmansperger. 1993. "A Bounded Influence, High Breakdown, Efficient Regression Estimator". *Journal of American Statistical Association*. 88:872-880.

2. Croux, C., P.J. Rousseeuw, and O. H¨ossjer. 1994. "Generalized S-estimators". *Journal of American Statistical Association*. 89:1271-1281.

3. Donoho, D.L. and P.J. Huber. 1983. "The Notation of Break-down Point". In: A Festschrift for E. L. Lehmann, Wadsworth

4. Fan, J. and R. Li. 2001. "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties". *Journal of the American Statistical Association*. 96: 1348-1360.

5. Freedman, W.L., C.D. Wilson, and B.F. Madore. 1991. "New Cepheid Distances to Nearby Galaxies Based on BVRI CCD Photometry". *Astrophysical Journal, 372,455-470*.

6. Gervini, D. and V.J. Yohai. 2002. "A Class of Robust and Fully Efficient Regression Estimators". The Annals of Statistics. 30: 583-616.

7. Handschin, E., J. Kohlas, A. Fiechter, and F. Schweppe. 1975, "Bad Data Analysis for Power System State Estimation". *IEEE Transactions on Power Apparatus and Systems*. 2:329-337.

8. Huber, P.J. 1981. *Robust Statistics*. John Wiley and Sons: New York, NY.

9. Jackel, L.A. 1972. "Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals". *Annals of Mathematical Statistics*. 5: 1449-1458.

10. Lee, Y., S.N. MacEachern, and Y. Jung. 2011. "Regularization of Case-Specific Parameters for Robustness and Efficiency". Submitted to the Statistical Science.

11. Mallows, C.L. 1975. "On Some Topics in Robustness". Unpublished memorandum, Bell Telephone Laboratories: Murray Hill, NJ.

12. Maronna, R.A., R.D. Martin, and V.J. Yohai. 2006. *Robust Statistics*. John Wiley: New York, NY.

13. Naranjo, J.D. and T.P. Hettmansperger. 1994. "Bounded Influence Rank Regression". *Journal of the Royal Statistical Society B*. 56:209-220.

14. Rousseeuw, P.J. 1982. "Least Median of Squares regression". Research Report No. 178, Centre for Statistics and Operations Research, VUB: Brussels, Belgium.

15. Rousseeuw, P.J. 1983. "Multivariate Estimation with High Breakdown Point". Research Report No. 192, Center for Statistics and Operations Research. VUB Brussels: Belgium,.

16. Rousseeuw, P.J. and C. Croux. 1993. "Alternatives to the Median Absolute Deviation". *Journal of American Statistical Association*. 94: 388-402.

17. Rousseeuw, P.J. and V.J. Yohai. 1984. *Robust Regression by Means of S-Estimators*. Springer: Berlin, Germany.

18. She, Y. 2009. "Thresholding-Based Iterative Selection Procedures for Model Selection and Shrinkage". *Electronic Journal of Statistics*. 3, 384C415.

19. She, Y. and A.B. Owen. 2011. "Outlier Detection Using Nonconvex Penalized Regression". *Journal of American Statistical Association*. 106:626-639.

20. Siegel, A.F. 1982. "Robust Regression Using Repeated Medians". *Biometrika*. 69:242-244.

21. Stromberg, A.J., D.M. Hawkins, and O. H¨ossjer. 2000. "The Least Trimmed Differences Regression Estimator and Alternatives". *Journal of American Statistical Association*. 95:853-864.

22. Yohai, V. J. 1987. "High Breakdown-Point and High Efficiency Robust Estimates for Regression". \, 15, 642-656.

23. Yu, C. and W. Yao. 2017. “Robust Linear Regression: A Review and Comparison”. *Communications in Statistics-Simulation and Computation*. 1-22.

ABOUT THE AUTHORS

Taoheedat Alanamu, is a postgraduate student at the University of Ilorin. She is a registered Statistician and is a member of the Nigerian Statistical Association. She holds a Master of Science (M.Sc.) in Statistics from the University of Ilorin. Her research interests are in multivariate and regression analysis.

Dr. Gafar Matanmi Oyeyemi, is a Reader in the Department of Statistics, University of Ilorin. He holds a Ph. D. degree in Multivariate Analysis and Application. His research interests are in the area of multivariate analysis and application, statistical quality control, and total quality management.

SUGGESTED CITATION

Alanamu, T. and G.M. Oyeyemi. 2018. “A New Robust Method for Estimating Linear Regression Model in the Presence of Outliers”. *Pacific Journal of Science and Technology.* 19(1):125-132.

Pacific Journal of Science and Technology