Earthline Journal of Mathematical Sciences
ISSN (Online): 2581-8147
Volume 8, Number 2, 2022, Pages 295-304
https://doi.org/10.34198/ejms.8222.295304

The Performance of Redescending M-Estimators when Outliers are in Two Dimensional Space

Anekwe Stella Ebele[1,*] and Onyeagu Sidney Iheanyi[2]

[1] Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria
e-mail: stellaanekwe@gmail.com

[2] Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria
e-mail: si.onyeagu@unizik.edu.ng

Abstract

M-estimators are robust estimators that give less weight to the observations that are outliers while redescending M-estimators are those estimators that are built such that extreme outliers are completely rejected. In this paper, redescending M-estimators are compared using both the Monte Carlo simulation method and the real life data to ascertain the method that is more efficient and robust when outliers are in both x and y directions. The results from the simulation study and the real life data indicate that Anekwe redescending M-estimator is more efficient and robust when outliers are in both x and y directions.

1. Introduction

1.1. Redescending M-estimators

Redescending M-estimators are estimators with their influence functions (ψ-functions) redescending to zero. They are those M-estimators that reject extreme outliers. Some of these estimators are:

1.1.1. Hampel redescending M-estimator

Hampel's three-part redescending M-estimator was proposed by [7] in the Princeton Robustness study. Its ψ-function is given as

Received: November 18, 2021; Revised: December 9, 2021; Accepted: December 14, 2021

2020 Mathematics Subject Classification: 49-XX, 49J15, 49K15, 49Mxx.

Keywords and phrases: outliers, robustness, M-estimators, redescending M-estimators, efficiency.

*Corresponding author

$$\psi(r) = \begin{cases} r & ; \text{ if } |r| \leq a \\ a \text{ sign } (r) & ; \text{ if } a < |r| \leq b \\ \frac{(c-|r|)}{(c-b)} a \text{ sign } (r) & ; \text{ if } b < |r| \leq c \\ 0 & ; \text{ if } |r| > c \end{cases} \quad (1)$$

where a, b, c are positive constants and $0 < a \leq b < c < \infty$ and r are the residuals scaledover Median Absolute Deviation MAD.

1.1.2. Tukey's biweight redescending M-estimator

[4] proposed Tukey's biweight M-estimator and its ψ-function is given as

$$\psi(r) = \begin{cases} r\{1 - (\frac{r}{c})^2\}^2 & ; |r| \leq c \\ 0 & ; \text{otherwise} \end{cases} \quad (2)$$

where c is arbitrary value known as tuning constant and r are the residuals scaled over MAD. For Tukey's biweight, c = 4.685 gives 95% efficiency on normal distribution.

1.1.3. Alarm redescending M-estimator

[1] proposed the Alarm's Redescending M-estimator for robust regression and outlier detection. Its ψ -function is given as

$$\psi(r) = \begin{cases} \frac{16r\,(e^{-(r/c)^2}}{(1+e^{-(r/c)^4}} & ; |r| \leq c \\ 0 & ; |r| > c \end{cases} \quad (3)$$

where c is the tuning constant and r are the residuals scaled over MAD.

1.1.4. Anekwe redescending M-estimator

[3] proposed the Anekwe's redescending M-estimator for robust regression and outlier detection. Its ψ -function is given as

$$\psi(r) = \begin{cases} r\left(1 - \left(\frac{r}{c}\right)^2\right)^2 \left(1 + \left(\frac{r}{c}\right)^2\right)^2 & ; |r| < c \\ 0 & ; |r| \geq c \end{cases} \quad (4)$$

where c is the tuning constant for the ith observation and the variable r are the residuals scaled over MAD.

2. Simulation Design

Monte Carlo simulation method is used to generate random data from different

probability distributions. We took the true parameters to be 1, 2, and 5 for β_0, β_1, and β_2 respectively. Each simulation case was replicated $M = 1000$ times. The estimates of each estimator were calculated in each of the iteration and the Mean of the M replicated estimates given by

$$\hat{\beta}_j = \frac{\sum_{i=1}^{M} \hat{\beta}_{ji}}{M} \qquad \text{for } j = 0, 1, 2, \cdots, p \tag{5}$$

was recorded for each estimator.

For comparison, the parameters estimates of the Mean Square Error (MSE) and the absolute bias (BIAS) of the [7], [4] and [1] and the [3] redescending M-estimators are computed.

Robustness of an estimator is measured using absolute bias given as

$$AbsBias(\hat{\beta}_j) = |\beta_j - \hat{\beta}_j| \qquad \text{for } j = 0, 1, 2, \cdots, p \tag{6}$$

Efficiency of an estimator is measured using the MSE (mean square error) defined as

$$MSE(\hat{\beta}_j) = \frac{\sum_{i=1}^{M} (\beta_j - \hat{\beta}_{ji})^2}{M} \qquad \text{for } j = 0, 1, 2, \cdots, p \tag{7}$$

and the variance of the estimator is defined as

$$Var(\hat{\beta}_j) = MSE(\hat{\beta}_j) - \left(Bias(\hat{\beta}_j)\right)^2 \qquad \text{for } j = 0, 1, 2, \cdots, p. \tag{8}$$

The estimator with lowest MSE is the most efficient; the smaller the MSE the more efficient is the estimator.

Simulated data were generated (including percentage mixtures of contaminated and uncontaminated data) in simple and multiple regressions, using two sample sizes, $n = 20$ and 200. The choices of the distributions used and the range choices for each distribution were chosen to use the idea of [13].

Results and Discussions from the Simulation Study

The Simulated results for the Hampel, Bisquare (Biweight), Alarm and Anekwe's redescending M-estimators are discussed as follows:

Discussion of stimulated results for data with outliers in both x and y directions.

At 10% outliers for both x and y axes in a simple regression, the result from Table 1

indicates that the Anekwe's estimator takes the lead as the most efficient and robust method but followed closely by the Alarm estimator. The Hampel and Bisquare estimators performed badly in this category.

Since M-estimators cannot perform very well when outliers are in the x-direction, the results of Hampeland Bisquare estimators got worse by the increase of outliers at both axes, that is, 15% outliers for both x and $y$axes in a simple regression as shown in Table 2. The Anekwe estimator is still the best with respect to efficiency and robustness but followed closely by the Alarm estimator.

At 20% outliers in both axes in a simple regression as shown in Table 3, the Anekwe and Alarm estimators are more efficient and robust compared to other estimators. All the estimators do not perform very well in this category with Bisquare estimator on the lead.

Table 4 presents the result for 10% outliers in both axes in a multiple regression model. The Anekwe estimator is more efficient and robust compared to other estimators. Alarm and Hampel estimator follow closely as the second and third best estimators respectively. The Bisquare estimator has high parameters estimates for β_1.

Furthermore, the Anekwe estimator takes the lead as the most efficient and robust estimator as shown in Table 5 for 15% outliers for both x and y axes in a multiple regression. Alarm estimator came second while Hampel's estimator was the third most efficient and robust estimator. The Bisquare estimator is also the least efficient and robust estimator in this category.

Lastly, Table 6 presents the result for 20% outliers in both axes in a multiple regression model. The Anekwe estimator outperformed other estimators as the most efficient and robust estimator. The second most efficient and robust estimator is the Alarm estimator which performs better than Hampel and Bisquare estimators.

Table 1: Simulated MSE and BIAS on simple regression for 10% outliers in x and y-axes

Sample Size	**Beta**	**Criteria**	**Bisquare**	**Hampel**	**Alarm**	**Anekwe**
20	$\boldsymbol{\beta_0}$	**BIAS**	0.0084	0.0913	0.0297	0.0130
	$\boldsymbol{\beta_0}$	**MSE**	0.1651	0.1832	0.0903	0.0870
20	$\boldsymbol{\beta_1}$	**BIAS**	1.9782	1.9831	0.3536	0.3026
	$\boldsymbol{\beta_1}$	**MSE**	3.9194	3.9387	0.9525	0.8435
200	$\boldsymbol{\beta_0}$	**BIAS**	0.0138	0.0733	0.0204	0.0092
	$\boldsymbol{\beta_0}$	**MSE**	0.0161	0.0224	0.0079	0.0072
200	$\boldsymbol{\beta_1}$	**BIAS**	1.9747	1.9812	0.0072	0.0064
	$\boldsymbol{\beta_1}$	**MSE**	3.9000	3.9256	0.0252	0.0242

Table 2: Simulated MSE and BIAS on simple regression for 15% outliers in x and y-axes

Sample Size	**Beta**	**Criteria**	**Bisquare**	**Hampel**	**Alarm**	**Anekwe**
20	$\boldsymbol{\beta_0}$	**BIAS**	0.0344	0.1965	0.0539	0.0203
	$\boldsymbol{\beta_0}$	**MSE**	0.1939	0.2917	0.1469	0.1302
20	$\boldsymbol{\beta_1}$	**BIAS**	1.9850	1.9965	0.5903	0.5300
	$\boldsymbol{\beta_1}$	**MSE**	3.9452	3.9912	1.5426	1.3790
200	$\boldsymbol{\beta_0}$	**BIAS**	0.0310	0.1949	0.0634	0.0290
	$\boldsymbol{\beta_0}$	**MSE**	0.0216	0.0636	0.0138	0.0100
200	$\boldsymbol{\beta_1}$	**BIAS**	1.9875	1.9983	0.0163	0.0270
	$\boldsymbol{\beta_1}$	**MSE**	3.9504	3.9936	0.0649	0.0816

Table 3: Simulated MSE and BIAS on simple regression for 20% outliers in x and y-axes

Sample Size	Beta	Criteria	Bisquare	Hampel	Alarm	Anekwe
20	β_0	**BIAS**	0.0809	0.5218	0.1471	0.0571
	β_0	**MSE**	0.2606	0.9700	0.2870	0.2245
20	β_1	**BIAS**	1.9939	2.0231	1.0304	0.9843
	β_1	**MSE**	3.9797	4.0984	2.5685	2.3292
200	β_0	**BIAS**	0.0916	0.4125	0.1772	0.0743
	β_0	**MSE**	0.0354	0.2180	0.0547	0.0247
200	β_1	**BIAS**	1.9970	2.0180	0.3748	0.6731
	β_1	**MSE**	3.9885	4.0727	0.8216	1.4156

Table 4: Simulated MSE and BIAS on multiple regression for 10% outliers in x and y-axes

Sample Size	Beta	Criteria	Bisquare	Hampel	Alarm	Anekwe
20	β_0	**BIAS**	0.0007	0.0384	0.0131	0.0051
	β_0	**MSE**	0.1750	0.1096	0.1054	0.1072
20	β_1	**BIAS**	1.7762	0.3924	0.3628	0.3328
	β_1	**MSE**	3.5352	1.0484	0.9481	0.9278
20	β_2	**BIAS**	0.0162	0.0138	0.0021	0.0044
	β_2	**MSE**	0.1574	0.1004	0.0893	0.0100
200	β_0	**BIAS**	0.0154	0.0286	0.0107	0.0049
	β_0	**MSE**	0.0146	0.0109	0.0069	0.0064
200	β_1	**BIAS**	1.7170	0.5577	0.0656	0.0343
	β_1	**MSE**	3.0553	0.9027	0.0718	0.0473
200	β_2	**BIAS**	0.0246	0.0274	0.0036	0.0043
	β_2	**MSE**	0.0131	0.0120	0.0052	0.0049

Table 5: Simulated MSE and BIAS on multiple regression for 15% outliers in x and y-axes

Sample size	**Beta**	**Criteria**	**Bisquare**	**Hampel**	**Alarm**	**Anekwe**
20	$\boldsymbol{\beta_0}$	**BIAS**	0.0680	0.1109	0.0407	0.0182
	$\boldsymbol{\beta_0}$	**MSE**	0.5347	0.2455	0.1569	0.1462
20	$\boldsymbol{\beta_1}$	**BIAS**	1.9679	0.5967	0.4631	0.4363
	$\boldsymbol{\beta_1}$	**MSE**	4.0595	1.6517	1.2286	1.1852
20	$\boldsymbol{\beta_2}$	**BIAS**	0.1498	0.1103	0.0369	0.0132
	$\boldsymbol{\beta_2}$	**MSE**	0.4868	0.3000	0.1460	0.1378
200	$\boldsymbol{\beta_0}$	**BIAS**	0.0122	0.1560	0.0297	0.0117
	$\boldsymbol{\beta_0}$	**MSE**	0.0208	0.0553	0.0116	0.0091
200	$\boldsymbol{\beta_1}$	**BIAS**	1.8639	1.3730	0.2419	0.0741
	$\boldsymbol{\beta_1}$	**MSE**	3.5381	2.7039	0.3560	0.0787
200	$\boldsymbol{\beta_2}$	**BIAS**	0.0999	0.2670	0.0148	0.0020
	$\boldsymbol{\beta_2}$	**MSE**	0.0294	0.1325	0.0104	0.0063

Table 6: Simulated MSE and BIAS on multiple regression for 20% outliers in x and y-axes

Sample size	**Beta**	**Criteria**	**Bisquare**	**Hampel**	**Alarm**	**Anekwe**
20	$\boldsymbol{\beta_0}$	**BIAS**	0.3308	0.6816	0.1465	0.0603
	$\boldsymbol{\beta_0}$	**MSE**	1.0812	1.9598	0.4146	0.2479
20	$\boldsymbol{\beta_1}$	**BIAS**	1.9990	1.0853	0.8032	0.6188
	$\boldsymbol{\beta_1}$	**MSE**	4.1447	3.3236	1.9788	1.6481
20	$\boldsymbol{\beta_2}$	**BIAS**	0.5267	0.5560	0.1276	0.0440

	β_2	**MSE**	1.5443	1.6220	0.4391	0.2376
200	β_0	**BIAS**	0.1171	0.8758	0.1856	0.0510
	β_0	**MSE**	0.0476	1.0360	0.0847	0.0192
200	β_1	**BIAS**	1.9319	1.9476	0.8974	0.2640
	β_1	**MSE**	3.7705	4.0105	1.8030	0.3753
200	β_2	**BIAS**	0.2784	1.0724	0.2496	0.0205
	β_2	**MSE**	0.1216	1.4194	0.1884	0.0161

3. Real-life Data

For comparison, we applied the redescending M-estimators to real-life data and the dataset had been extensively used by other researchers in the area of robust regression.

The Hawkins, Bradu, and Kass data

[8] generated artificial data for testing the performance of robust estimators. The data contains **75** observations in four dimensions (one response and three explanatory variables). The first 10 observations are bad leverage points, and the next four points are good leverage points (i.e., their x_i are outlying, but the corresponding y_i fit the model quite well).

Table 7: Estimates of the model parameters for Hawkins, Bradu and Kass data

Parameter	**Hampel**	**Biweight**	**Alarm**	**Proposed**
β_0	-0.181	-0.946	-0.181	-0.181
β_1	0.081	0.145	0.082	0.081
β_2	0.040	0.197	0.040	0.040
β_3	-0.052	0.180	-0.052	-0.052
Data points used	65	71	65	65
Residual standard error	0.77	0.63	0.56	0.56

The summary of the results for estimates of the model parameters for Hawkins, Bradu and Kass data for the estimators are presented in Table 7. Alarm, Hampel and Anekwe's method detected 10 outliers in the robust fit while Biweight estimator detected 4 outliers in the analysis.

Conclusion

Simulation studies were done to ascertain the effectiveness of the redescending M-estimators when outliers are in both x and y directions. Mean square error (MSE) and BIAS were used for comparison under two different sample sizes.

From the stimulated results, it was obvious that when outliers are both in the leverage points and the response, the Anekwe estimator is the most efficient and robust estimator compared to the Hampel, Alarm and Bisquare estimators.

References

[1] A. A. Alamgir, S. A. Khan, D. M. Khan and U. Khalil, A new efficient redescending M-estimator: Alamgir redescending M-estimator, *Research Journal of Recent Sciences* 2(8) (2013), 79-91.

[2] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers and J. W. Tukey, *Robust Estimates of Location: Survey and Advances*, Princeton, NJ: Princeton University Press, 1972.

[3] S. Anekwe and S. Onyeagu, The redescending M-estimator for detection and deletion of outliers in regression analysis, *Pakistan Journal of Statistics and Operations Research* 17(4) (2021), 997-1014. https://doi.org/10.18187/pjsor.v17i4.3546

[4] A. E. Beaton and J. W. Tukey, The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data, *Tecnometrics* 16(2) (1974), 147-185. https://doi.org/10.1080/00401706.1974.10489171

[5] N. R. Draper and H. Smith, *Applied Regression Analysis*, 3rd ed., New York: John Wiley and Sons, 1998. https://doi.org/10.1002/9781118625590

[6] F. R. Hampel, The influence curve and its role in robust estimation, *Journal of the American Statistical Association* 69(346) (1974), 383-393. https://doi.org/10.1080/01621459.1974.10482962

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

[8] D. M. Hawkins, D. Bradu and G. V. Kass, Location of several outliers in multiple regression data using elemental sets, *Technometrics* 26 (1984), 197-208. https://doi.org/10.1080/00401706.1984.10487956

[9] P. J. Huber, Robust estimation of location parameter, *The Annals of Mathematical Statistics* 35 (1964), 73-101. https://doi.org/10.1214/aoms/1177703732

[10] T. D. Nguyena and R. Welch, Outlier detection and least trimmed squares approximation using semi-definite programming, *Comput. Stat. Data* 54 (2010), 3212-3226. https://doi.org/10.1016/j.csda.2009.09.037

[11] P. J. Rousseeuw, *Least Median of Squares Regression*, Research Report No. 178, Centre for Statistics and Operations Research, VUB Brussels, 1982.

[12] P. J. Rousseeuw, *Multivariate Estimation with High Breakdown Point*, Research Report No. 192, Centre for Statistics and Operations Research, VUB Brussels, 1983.

[13] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, New York: Wiley-Interscience, 1987. https://doi.org/10.1002/0471725382

[14] J. W. Tukey, *Exploratory Data Analysis*, Reading, MA: Addison-Wesley, 1977.