JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 10(4), DECEMBER, 2022
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com

Performance Evaluation of Some Robust Estimators in the Presence of Outliers using a Default and Adjusted Turning Constant Values

S. Abdullahi[1*], B. A. Rasheed[2], A. Ahmed[3]
[1]*Bauchi State College of Agriculture, Bauchi State, Nigeria*
[2]*Federal Polytechnic Bauchi, Bauchi State, Nigeria*
[3]*Abubakar Tafawa Balewa University Bauchi, State, Nigeria*

ABSTRACT
This research work was conducted to evaluate the performance of S, M and MM robust estimators in the presence of outliers in regression model using a default and adjusted turning constant values of the Bisquare redescent weights, so as to see the impact of increasing or reducing the turning constant values on the resistibility of the robust estimators. Since it has been observed from the literature, that the previous researchers stick to the used of the default turning constant values despite the fact that the turning constant can be adjusted. A simulation study was used to compare the performance of the robust methods with five independent variables on the sample sizes 50, 150 and 300 respectively. The result of the study showed that increasing or decreasing the turning constant values for S and M estimators have a significant impact on the resistibility of the estimators on sample sizes 50, 150, and 300 respectively, while adjusting the turning constant value for MM-estimator has no impact on the robust regression result. The study also recommends that analyst should use smaller value than the default turning constant for M-estimation to have more resistibility against outlying observations in the regression models, because the smaller the turning constant value the more resistant the M-estimator. Secondly the study recommended that analyst should use a value greater than the default K value for the S-estimation, because the higher the turning constant the more resistant the S-estimator against outlying observations.

ARTICLE INFO
Article History
Received: June, 2022
Received in revised form: July, 2022
Accepted: September, 2022
Published online: December, 2022

KEYWORDS
Outlier, Turning Constant, S-Estimation, M-estimation, MM-estimation.

INTRODUCTION

Regression analysis is conceptually method for determining functional relationships between the dependent and a set of independent variables. The relationship between yand $x_1, x_2, \dots, x_n$can be approximated by the regression model:

$y = \mathrm{f}(x_1, x_2, \dots, x_n)\beta + e_{ij}$... 1.1

Where e_{ij} is assumed to be a random error representing the discrepancy in the approximation, it accounts for the failure of the model to fit the data exactly. To judge how well the estimated regression model fits the data, we can look at the size of the residuals, which is:

$e_{ij} = \mathrm{y_i} - (\mathrm{f}(x_1, x_2, \dots, x_n)\beta)$....1.2

A point which lies far from the line and often has a large residual value is called an outlier. However, it is well known that the OLS estimate is extremely sensitive to the outliers. However, regression modeling is guided by a number of assumptions. When the analyst estimates regression models and tests its assumptions, the frequently finds that the assumptions are substantially violated then, and parameter estimated poorly fits the data. Often, however, a transformation will not eliminate the leverage of influential outliers that bias the prediction and distort the significance of parameter estimates. When such occur, robust regression that is resistant to the influence of outliers may be the reasonable recourse (Yuliana

Corresponding author: Abdullahi, S. ✉ sagirabdullahignscoa@gmail.com ✉ *Bauchi State College of Agriculture, Bauchi.*

et al., 2014). The robust methods have been defined to deal with the influential points in regression analysis, where the value of the estimation by using this method is not much affected with outliers. The characterization of robust methods is that they fit the bulk of the data well if the data contain outliers, while if a small proportion of outliers are present the robust method gives approximately the same results as the classical (non-robust) method. As a consequence of fitting the bulk of the data well, robust methods provide a very reliable method of detecting outliers, even in high dimensional multivariate situation.

Based on the literature review, we have noticed that most of the previous researchers on this topic stick to the used of default turning constant value C or K of the redescent weights for the respective robust regression methods despite the fact that the C or K value can be adjusted. Justo and Calandra (2021) evaluated the performance of some robust estimators against OLS for the Measurement of a Topographic Altimetric Network, where they adjusted the constant value for M-estimation method from the default value 4.685 to 1.345 which led to increases in the break down point of the M estimator. Therefore, this study intends to examine what will happen when there is a change on the default values of the tuning constant? This is where I based my research since the values of the tuning constant for the redescent weight can be set (apart from the pre-set default), this study aimed to verify whether there is an improvement in the high breakdown point modeling when the values are increase or decrease. In this research, we want to study a situation when all the estimators are operating under the same redescent weight and adjusting the tuning constants and measure the performance of such operators and verify which one outperform others. The aim of this study is to compare the performance of M, S and MM robust estimators in the presence of outliers using a default and adjusted turning constant value of the redecent weights and also to compare the performance of the robust estimators with default and adjusted turning constant values respectively. The findings of this study will serve as an alternative for researchers to choose between the default and adjusted turning constant value for the robust estimators at different sample sizes.

LITERATURE REVIEW

Robust Regression

Robust regression is a regression method that is used when the distribution of residual is not normal or there are some outliers that affect the model (Yuliana et al., 2014). This method is an important tool for analyzing the data which is affected by outliers so that the resulting models are stout against outliers (Drapper & Smith, 1998). A robust regression is an iterative procedure that is designed to overcome the problem of outliers and influential observations in the data and minimize their impact over the regression coefficients (Zaman *et al*., 2001). Most. The main objective of robust estimation is to obtain reliable estimates/inferences for unknown parameters in the presence of outliers. The robust procedure replaces the sum of squared residuals of the OLS with some other function that is being less influenced by the unusual observations. These procedures first fit a regression to the data and then identify the outliers as those observations having large residuals. Robust techniques have three desiring properties, namely, efficiency, breakdown point, and bounded influence. The breakdown point is the smallest fraction of the unusual observations that an estimator can tolerate before giving an incorrect result. It is always a value between zero and 0.5. It measures the degree of robustness, the robustness of an estimator increases as the breakdown point increases. For example, OLS has a breakdown point of 0% which represents that even a single outlier is sufficient to distort the OLS estimators. The robust techniques have 50% of breakdown point which is considered as the highest breakdown point. The property of bounded influence measures the resistance of the estimator against bad observations. It encounters the tendency of the least squares to allow the leverage points to exhibit greater influence.

Outliers

Outliers are unusual data values that occur almost in all research projects involving

Corresponding author: Abdullahi, S. ✉ sagirabdullahignscoa@gmail.com ✉ *Bauchi State College of Agriculture, Bauchi.*

data collection. Outliers are observations that have extreme value relations. The term outlier is defined as the data which are far away from the bulk of the data, or more generally, from the pattern set by the majority of the data (Hampel et al 1986). According to Huber (1981). An outlier is a data point that is far outside the norm for a variable or population. An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism (Ranjit, 2014)). Beside this it is better to point out some definition about terms which is going hand by hand with outlier as follows. Leverage is a measure of how an independent variable deviates from its mean. An observation with an extreme value on a predictor variable is a point with high leverage. Influence an observation is said to be in influential if removing that observation substantially changes the estimation of the coefficients. A datum that is "influential" is one for which the regression estimate changes considerably if it is removed. Rejection Point is the point beyond which the influence function becomes zero. That is the contribution of the points beyond the rejection point to the final estimate is comparatively legible. According to Begashaw and Yonnes, (2020). The presence of outliers can lead to inflated error rates and substantial distortions of parameter and statistic estimates when using either parametric or nonparametric tests. Casual observation of the literature suggests that researchers rarely report checking for outliers of any sort. Outliers can have deleterious effects on statistical analyses. First, they generally serve to increase error variance and reduce the power of statistical tests. Second, if non-randomly distributed they can decrease normality (and in multivariate analyses, violate as assumptions of sphericity and multivariate normality), altering the odds of making both Type I and Type II errors. Third, they can seriously bias or influence estimates that may be of substantive interest.

Causes of Outliers

Outliers from incorrect recording data: Outliers are often caused by human error, such as errors in data collection, recording or entry. Data from an interview can be recorded incorrectly, or mistaken upon data entry. Errors of this nature can often be corrected by returning to the original documents or even the subjects if necessary and possible and entering the correct value. Outliers from intentional or motivated mis-reporting: There are times when participants purposefully report incorrect data to experimenters or surveyors. A participant may make conscious effort to sabotage the research (Yohai *et al*., 2001) or may be acting from other motives. Social desirability and self-presentation motives can be powerful. This can also happen for obvious reasons when data are sensitive. Outliers from sampling error: Another cause of outliers is sampling. It is possible that a few members of a sample were inadvertently drawn from a different population than the rest of the sample. For-example: in education, in advert entry sampling academically gifted or mentally retorted students is a possibility and (depending on the goal of the study) might provide undesirable outliers. These cases should be removed as they do not reflect the target population. Outliers as legitimate cases sampled from the correct population: It is possible that an outlier can come from the population being sampled legitimately through random chance, it is important to note that sample size plays a role in the probability of outlying values. Within a normally distributed population, it is more probable that a given data point will be drawn from the most densely concentrated area of the distribution, rather than one of the tails. Yua and Sullivian (2004) have said as a researcher casts a wider net and the data set becomes larger, the more the sample resembles the population from which it was drawn and thus the likelihood of outlying values become greater. In other words, there is only about one percentage chance you will get an outlying data point from a normally distributed population, this means that, on the average, about one percentage of your subjects should be three standard deviations from the mean.

Outliers Management.

Outliers being inconsistent observations and largely deviated from the majority of the observations in data need proper handling as they pose serious threat to the

Corresponding author: Abdullahi, S. ✉ sagirabdullahignscoa@gmail.com ✉ *Bauchi State College of Agriculture, Bauchi.*

regression model and its estimated coefficients and, as a result, give misleading outcomes. Two types of outliers can happen in the regression dataset. One with extremely large values in the response is referred to as vertical outliers, whereas observations with extremely large values in the explanatory variable are called leverage points. Outliers being influential to the classical regression require methods that are insensitive to it. Two approaches are commonly used to cope with this problem, popularly known as diagnostic approach and robust procedures (Rousseeuw and Leroy, 1984). The diagnostic approaches try to identify the unusual observations through diagnostic statistics and remove it from the data and then classical procedures, for example, least square estimation procedure is then applied to the remaining clean dataset. This approach is suitable for simple data or when there are one or two outliers, but it becomes inefficient to detect outliers in case of multiple outliers present in a multidimensional dataset. Therefore, an appropriate procedure to deal with outliers is robust procedures that not only detect multiple outliers in complex data but also give efficient results.
(Khan *et al*., 2021)

METHODOLOGY

M Estimator: One of the robust regression estimation methods is the M estimation. The letter M indicates that M estimation is an estimation of the maximum likelihood type. If estimator at M estimation is $\hat{\beta}_M = \beta_n(x_1, x_2, \dots, x_n)$ then:

$$E\left[\hat{\beta} = \beta_n(x_1, x_2, \dots, x_n)\right] = \beta \qquad 3.1$$

This implies that equation (1) is an unbiased estimator with its corresponding variance given as:

$$Var(\hat{\beta}) = \frac{[\tilde{\beta}]^2}{n.E\left[\frac{d}{d\beta} In\, f(x_i;\beta)\right]^2} \qquad 3.2$$

Where $\tilde{\beta}$is the linear unbiased estimator for β. M estimation is an extension of the maximum likelihood estimate method and a robust estimation (Yuliana & Susanti, 2008).

Algorithm 1: M - Estimator

1. Estimate regression coefficients on the data using OLS.
2. Test assumptions of the regression model
3. Detect the presence of outliers in the data.
4. Calculate estimated parameter β_0 with OLS.
5. Calculate residual value $e_i = y_i - \hat{y}_i$
6. Calculate value $\hat{\sigma}_i = 1.4826\, MAD$
7. Calculate value $u_i = e_i/\hat{\sigma}_i$
8. Calculate the weighted value using:

$$w_i = \begin{cases} \left[1 - \left(\frac{u_i}{4.685}\right)^2\right]^2 ; & |u_i| \le 4.685 \\ 0 ; & |u_i| \le 4.685 \end{cases}$$

9. Calculate $\hat{\beta}_M$using weighted least squares (WLS) method with weighted w_i
10. Repeat steps 5-8 to obtain a convergent value of $\hat{\beta}_M$ and test to determine whether independent variables have significant effect on the dependent variable.

S Estimator: The weakness of M-estimation is the lack of consideration on the data distribution and not a function of the overall data because only using the median as the weighted value. This method uses the residual standard deviation to overcome the weaknesses of M- estimator. According to Salibian and Yohai (2006), the S-estimator is defined by $\beta_S = Min_\beta \hat{\sigma}_s(e_1, e_2, \dots, e_n)$ with determining minimum robust scale estimator.

Algorithm 2: S - Estimator

1. Estimate regression coefficients on the data using OLS.
2. Test assumptions of the regression model
3. Detect the presence of outliers in the data.
4. Calculate estimated parameter β_0 with OLS.
5. Calculate residual value $e_i = y_i - \hat{y}_i$
6. Calculate value $\hat{\sigma}_i$;

$$\hat{\sigma}_i = \begin{cases} \frac{Median.e_i.(Median)(e_i)}{0.6745}, & Iteration = 1 \\ \frac{1}{nK}\sum w_i e_i^2, & Iteration > 1 \end{cases}$$

7. Calculate the value $u_i = \frac{e_i}{\hat{\sigma}_i}$
8. Calculate the weighted value, w_i

Corresponding author: Abdullahi, S. ✉ sagirabdullahignscoa@gmail.com ✉ *Bauchi State College of Agriculture, Bauchi.*

$$w_i = \begin{cases} \left\{\left[1-\left(\frac{u_i}{1.547}\right)^2\right]^2, & |u_i| \le 1.547 \quad Iteration = 1 \\ 0, & |u_i| > 1.547 \end{cases} \\ \frac{\rho(u)}{u^2} \qquad\qquad Iteration > 1$$

9. Calculate $\hat{\beta}_S$ with WLS method with weighted w_i
10. Repeat steps 5 – 8 to obtain a convergent value of $\hat{\beta}_S$ and test to determine whether independent variables have significant effect on the dependent variable.

MM Estimator: MM estimation provide a high breakdown value and more efficient than M and S estimators. Breakdown value is a common measure of the proportion of outliers that can be addressed before these observations affect the model [3]. MM-estimator is the solution of;

$$\sum_{i=1}^{n} \rho_1(u_i)X_{ij} = 0 \; or \sum_{i=1}^{n} \rho_1\left(\frac{Y_i - \sum_{j=0}^{k} X_{ij}\beta_j}{S_{MM}}\right)X_{ij} = 0 \qquad 3.3$$

Where S_{MM} is the standard deviation of S estimation and ρ is the Turkey's biweight function defined by:

$$\rho(u_i) = \begin{cases} \frac{u_i^2}{2} - \frac{u_i^4}{2c^2} + \frac{u_i^6}{6c^2}, & -c \le u_i \le c \\ \frac{c^2}{6}, & u_i < -c \; or \; u_i > c \end{cases} \qquad 3.4$$

Algorithm 3: MM Estimator

1. Estimate regression coefficients on the data using OLS.
2. Test assumptions of the regression model
3. Detect the presence of outliers in the data.
4. Calculate residual value $e_i = y_i - \hat{y}_i$ of S estimate
5. Calculate $\hat{\sigma}_i = \hat{\sigma}_{sn}$
6. Calculate $u_i = \frac{e_i}{\hat{\sigma}_i}$
7. Calculate the weighted value w_i:

$$w_i = \begin{cases} \left[1-\left(\frac{u_i}{4.685}\right)^2\right]^2, & |u_i| \le 4.685 \\ 0, & |u_i| > 4.685 \end{cases}$$

8. Calculate $\hat{\beta}_{MM}$ using weighted least square method with weighted w_i
9. Repeat step 5 – 8 to obtain a convergent value of $\hat{\beta}_{MM}$ and test whether the independent variables have a significant effect on the dependent variable.

Simulation Study: Generating Data Set

Simulation methodology was adopted are as follows: Let n be sample size, and p the number of predictor variables; then n cases will be generated randomly in order to test the performances of the different estimators considered.

Table 1: Factors and levels for the simulated data sets

Number of Independent Variables 'p'	Sample Size 'n'		
5	50	150	300

NB: The calculations were done using statistical programing language R, the R language provides a wide variety of statistical and graphical techniques and strongly functional language in computer modelling.

RESULTS AND DISCUSSIONS

Table 1: Result of M-estimation, S-estimation and MM-estimation on simulated data with five levels on sample size 50.

	RSE	RSE	RSE
	C=4.685	C=3.685	C=6.685
M-estimation	0.3291	0.3224	0.3296
MM-estimation	0.2994	0.2994	0.2994
S-estimation	k=1.547	k=0.547	k=2.547
	0.2996	0.8474	0.182

Source: Authors' computation aided by R package v 4.1.1.

Corresponding author: Abdullahi, S. ✉ sagirabdullahignscoa@gmail.com ✉ *Bauchi State College of Agriculture, Bauchi.*

Table 2: Result of M-estimation, S-estimation and MM-estimation on simulated data with five levels on sample size 150.

	RSE	RSE	RSE
	C=4.685	C=2.685	C=5.685
M-estimation	0.3402	0.3368	0.3423
MM-estimation	0.3208	0.3208	0.3208
S-estimation	C=1.547	C=0.547	C=2.547
	0.321	0.9078	0.195

Source: Authors' computation aided by R package v 4.1.1.

Table 3: Result of M-estimation, S-estimation and MM-estimation on simulated data with five levels on sample size 300.

	RSE	RSE	RSE
	C=4.685	C=2.685	C=5.685
M-estimation	0.3282	0.3103	0.3258
MM-estimation	0.2993	0.2993	0.2993
S-estimation	C=1.547	C=0.547	C=2.547
	0.2995	0.8471	0.1819

Source: Authors' computation aided by R package v 4.1.1.

Table: 4 Comparing the performance of the robust methods with 5 levels when n=50, 150 and 300.

n=50					**n=150**				**n=300**
Methods	**RSE**				**RSE**				**RSE**
M	0.3291	0.3224	0.3296	0.3402	0.3368	0.3423	0.3282	0.3103	0.3258
MM	0.2994	0.2994	0.2994	0.3208	0.3208	0.3208	0.2993	0.2993	0.2993
S	0.2996	0.8474	0.182	0.321	0.9078	0.195	0.2995	0.8471	0.1819

Source: Authors' computation aided by R package v 4.1.1.

DISCUSSION OF RESULTS

Table1 to 3 showed the result of M, MM and S estimators' methods of robust regression with five (5) independent variables on sample sizes 50, 150 and 300 respectively. Table 4 compared the performance of the three methods with 5 independent variables on sample sizes 50, 150 and 300 respectively. As we can see from the table 4 above, for n=50 as we decrease the default value (4.685) to (3.685) the Residual standard error (RSE) also decreases from (0.3291) to (0.3224) meaning that, the resistibility of the M-estimator increases by reducing the default value. Also we have seen that by increasing the default value to (6.685) the RSE also increases from (0.3291) to (0.3296), meaning that the resistibility of the estimator against outliers' decreases. But for the MM-estimator under the sample size 50, the RSE remain unchanged meaning that decreasing or increasing the default turning constant has no any impact on the estimator. While for the S-estimator, we can see that, as we decrease the turning constant K the RSE increases from (0.2996) to (0.8474) meaning that increasing the default value seriously affect the resistibility of the estimator S. But if the value K increases from the default to (2.547) the RSE decreases seriously from (0.2712) to (0.182) meaning that for the S-estimator increasing the default value gives a better result. Likewise, for n=150 case is same as that of n=50 for the M, MM and S estimators respectively. Also for n=300 case is same as that of n=50 and 150 except that for M-estimator, where increasing the default value **c** gives better result than the default.

CONCLUSION

Based on the findings of this study, we can say that adjusting the turning constant value

Corresponding author: Abdullahi, S. ✉ sagirabdullahignscoa@gmail.com ✉ *Bauchi State College of Agriculture, Bauchi.*

for M, MM and S estimators have a significant impact on the resistibility of the estimators against outlying observations. From the result of the study, we can conclude that decreasing the default turning constant value for M-estimation from (4.685) to (2.685) will increase the resistibility of the estimator against outlying observations on all sample sizes. The study also concludes that maintaining the default **c** value for MM-estimator gives a better result. The findings of the study also conclude that increasing the K constant value of S-estimator from (1.547) to (2.547) increases the resistibility of the estimator by having smaller RSE on all the sample sizes 50, 150 and 300 respectively. This study recommends that analyst should use smaller value than the default turning constant for M-estimation to have more resistibility against outlying observations in the regression models. Because the smaller the turning constant value the more resistant the M-estimator. Secondly the study recommended that analyst should use a value greater than the default K value for the S-estimation, because the higher the turning constant the more resistant the S-estimator against outlying observations.

REFERENCE

Begashaw, A. B. and Yohannes, Y. B. (2020). Review of outlier detection and identifying using robust regression. *International Journal of Systems Science and applied mathematics.* 5(1): 4-11.

Draper, N. R. & Smith, H. (1998). Applied Regression Analysis, Third Edition, Wiley Interscience Publication, United States, 1998.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). Robust Statistics, the Approach Based on Influence Functions, *John Wiley & Sons*, New York.

Huber, P. J. (1981), Robust Statistics. John Wiley & Sons, New York.

Justo, C. E. and Calandra, M. V. (2021). Evaluation of Robust Linear Regression Methods for the Measurement of a Topographic Altimetric Network. A *Journal of Scientific and Engineering Research, 8(4):30-39*

Khan *et al*. (2021). Applications of robust regression techniques: An econometric approach. *Journal of Hindawi Mathematical Problems in engineering*, volume 2021, 9 pages.

Lianng Yuh and Vill A. Sullivan (2004). Robust Estimation Using Sas·Softvare, *Department of Biostatistics Merrell Dow Research Institute Cincinnati*, Ohio 45215 Mathsoft, Inc.Seattle, WA, 255-298.

Rousseeuw, P. J. and A. M. Leroy (1984), *Robust Regression and Outlier Detection*, John Wiley & Sons, New York, NY, USA.

Ranjit, K. P. (2014), Some Methods of Detection of Outliers in Linear Regression Model, Iasri, *Library Avenue*, New Delhi-110012.

Salibian, M. and Yohai, V. J. (2006). A Fast Algorithm for S-Regression Estimates. *Journal of Computational and Graphical Statistics*, 15, No. 2 (2006), 414-427.

Yuliana, S., Hasih, P. Sri, S. H., and Twenty, L. (2014) M-estimation, S-estimation, and MM-estimation in robust regression. *International Journal of Pure and Applied Mathematics* Volume 91, 349-360.

Yuliana, A and Susanti, A. (2008). Estimasi M dan sifat-sifatnya pada Regresi Linear Robust, *Journal of Math*-Info, 1, No. 11 8-16.

Zaman, A, Rousseeuw, P. J, and Orhan. M, (2001). Econometric applications of high breakdown robust regression techniques, *Economics Letters*, vol. 71, no. 1, pp. 1–8.

Corresponding author: Abdullahi, S. ✉ *sagirabdullahignscoa@gmail.com* ✉ *Bauchi State College of Agriculture, Bauchi.*