Journal of Excellence in Social Sciences (2024) 3(3) 55–60
ISSN (E): 2755-3876

Efficiency of OLS and Huber M estimator in case of Outliers

Sania Shoukat[1]

Shahbaz Nawaz[2]

Muhammad Muzamil Rasheed[3]

Anam Javaid[4]

Hafiz Abdul Sami[5]

How to cite this article:

Shoukat, S., Nawaz, S., Rasheed, M. M., Javaid, A., & Sami, A. H. (2024). Efficiency of OLS and Huber M estimator in case of outliers. *Journal of Excellence in Social Sciences, 3*(3), 55–60.

Received: 19 March 2024 / Accepted: 12 May 2024 / Published online: 5 July 2024
© 2024 SMARC Publications.

Abstract

Regression analysis is a valuable tool when dealing with real-world datasets that include several types of variables. Assumption fulfillment leads to applying ordinary least square regression (OLS) on the chosen variables. When analyzing data using traditional multiple regression, the optimal technique is the ordinary least squares (OLS) estimate, provided the requirements for regression weights are satisfied. Results and estimates from samples might be misleading if the data does not match all these assumptions. Particularly problematic for least squares regression is the presence of outliers. Robust regression analysis is a typically used method. In outliers, the research aims to compare M-estimators with OLS Estimators. The effectiveness is evaluated by comparing the Huber M estimate's coefficient with the OLS estimators. To achieve the study's objective, we use Microsoft Excel to build a Monte Carlo simulation for the response and explanatory variables, which are typically distributed. The normal distribution is used to create 1,000 randomly selected integers. In order to assess the effectiveness of the estimations, outliers with varying percentages are subsequently inserted. The results demonstrate that OLS was affected by the x- and y-axis outliers. There will be no impact on the Huber M estimate from the x-axis outliers. Outliers in the Y-direction impacted the Huber M findings.

Keywords: Outliers, Robust estimators, Ordinary least square, Simulation, Efficiency

1 Introduction

[1]The Women University Multan, Pakistan
[2]Govt. of Punjab, Planning and Development Department, Pakistan
Corresponding Author: shahbazgurmani91@gmail.com
[3]School Education Department, Government of Punjab, Pakistan
[4]The Women University Multan, Pakistan
[5]Government of Punjab, Planning and Development Department, Pakistan

Copyright (c) 2024 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Regression analysis aims to identify potential independent variables by analyzing their relationships to the dependent ones (Lee, 2022; Lim et al., 2020). One of the straightforward methods used in regression analysis for component extraction is ordinary least squares (OLS) (Beattie & Esmonde-White, 2021). On the other hand, if the assumptions are not satisfied, OLS will not work well since the results will be biased (Gujrati, 2022). Depending on the sort of dependent variable or the assumptions that each form of regression analysis makes, many kinds of regression analysis, including logistic regression, ridge regression, and robust regression, may be used in these scenarios (Javaid et al., 2020). These regression methods handle a wide range of problems, such as when OLS's assumptions are not satisfied or when a dataset issue arises (Susanti et al., 2014). Outliers are a potential problem that might arise in the dataset. Outliers reduced OLS effectiveness (Begashaw et al., 2020; Javaid et al., 2019). Several types of maximum likelihood (M) estimators can be used for these problems. For example, LTS estimators, estimators of scale, and modified M estimators (MM) (Javaid et al., 2018).

Thrampoulidis et al. (2018) note that these estimators provide a range of benefits when dealing with datasets that include outliers (Sullivan et al., 2021). When compared to other types of robust estimators, M estimators are more consistent and efficient, and they do a good job when dealing with outliers (De Menezes et al., 2021; Javaid et al., 2020; Nawaz et al., 2022). Both real-world datasets and simulation studies may be used to test the efficacy of these estimators (Schmidt & Finanet, 2018). Researchers in the literature often do simulation analyses. One of the most common methods used for simulation analysis is the Monte Carlo simulation, which might be used to evaluate different techniques that have been produced (Barbu & Zhu, 2020; Mooney, 1997). Without resorting to other, more time-consuming and expensive ways, this simulation methodology allows us to assess and contrast the efficiency of the several methods that have been established (Glielmo et al., 2021; Raychaudhuri, 2008). The effectiveness of the robust estimator is also examined in this research via the use of Monte Carlo simulation analysis. To achieve this goal, the Huber M estimator is used in conjunction with weighted approaches to choose M estimators. Outliers may be produced by varying percentages in dependent and independent variables (Suhaeri et al., 2021; Sullivan et al., 2021). After that, the results are compared with the OLS to check for various outlier scenarios.

2 Methodology

The methodology used in the research can be described as follows.

2.1 Ordinary least square

Consider the standard linear regression model as in (1)

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

The basic linear regression model with one variable can be described in equation (1). The matrix form can be represented as in (2) for multiple linear regression having more than one independent variable.

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'Y \tag{2}$$

If all assumptions are satisfied, OLS are the most effective BLUE (linear unbiased) estimators (Gujrati, 2022).

2.2 M-estimators

Regarding regression models, ordinary least squares estimates are very vulnerable to (and unable to withstand) outliers (Fitrianto & Xin, 2022). Robust regression analysis, which does not need such assumptions, is useful in some instances (Berk et al., 2021). In cases when there are outliers in the data, it produces much more accurate estimates of the regression coefficients. According to

Lodhi et al. (2023), their excessive effect tends to skew the least squares coefficients. When working with a dataset that has N observations, the usual weight assigned to each observation is about 1/N. However, a weight of 15%, 20%, or even 50% may be assigned to outlying observations. The predicted coefficients become severely skewed as a result of this. Because their residuals are far less than expected, this distortion makes outliers hard to spot. Outliers are less heavily weighted in robust regression.

According to Javaid et al. (2018), the models derived as maximum likelihood models are collectively called M-estimator models. With M-estimation, first proposed by Huber (1965), the most popular and almost as efficient robust regression approach is OLS (Youssef et al., 2022). The M-estimate minimizes a function of ρ of the errors instead of minimizing the sum of squared errors, which is the original aim. Below is the M-estimate goal function.

$$min \sum_{i=0}^{n} \rho\left(\frac{e_i}{s}\right) = min \sum_{i=0}^{n} \rho\left(\frac{y_{i-X'i\,\hat{\beta}}}{s}\right) \quad (3)$$

The scale estimate can often be created by linearly combining the residuals, and it is represented by s. The objective function is given by the function of ρ, which represents the contribution of each of the residuals. These are qualities that a reasonable of ρ ought to possess:

$$\rho(e) \geq 0\,, \rho(0) = 0\,, \rho(e) = \rho(-e), and\ \rho(ei) \geq \rho(e'i)\ for\ |ei| \geq |e'i|. \quad (4)$$

The equation ρ (ei) = ei^2 applies to least squares estimate, as an example. By obtaining the partial derivatives with respect to the β and setting them as 0, the system of normal equations that may be used to solve this minimization issue can be obtained.

$$\sum_{i=0}^{n} \Psi\left(\frac{y_{i-x'\ \beta_i}}{s}\right) X_i = 0 \quad (5)$$

Where the derivative is located. How much weight to give outliers determines the function's selection. In contrast to least squares, a monotone function gives less weight to extreme outliers. As the outlying distance rises, a redescending ï function assigns more weight to an outlier until the distance reaches a certain value, reducing the weight to zero. Two approaches can be used to solve the nonlinear normal equations resulting from M-estimates: Newton-Raphson and Iteratively Reweighted Least Squares (IRLS). IRLS presents the standard equation in this way:

$$X'WX\hat{\beta} = X'Wy \quad (6)$$

Where W is $n \times n$ diagonal matrix of weights,

$$= \frac{\Psi\left(\frac{y_{i-X'i\,\hat{\beta}_0}}{s}\right)}{\left(\frac{y_{i-X'i\hat{\beta}_0}}{s}\right)} \quad (7)$$

Usually, OLS is used to get the initial vector of parameter estimations, 0. This parameter estimation is revised using IRLS using

$$\hat{\beta}_1 = (X'WX)^{-1}X'Wy \quad (8)$$

In (3), the weights are dependent on the residuals, which in turn rely on the estimated coefficients, and so on. One such method is repeatedly re-weighted least squares.

3 Result and Discussion

3.1 Data Simulation and Results

Microsoft Excel and random integers are used to build the dataset in this study. The variables are generated using a normal distribution with μ = 60 and σ= = 150. A dependent variable (Y) and

five independent variables (X1, X2, X3, X4, and X5) are generated when μ equals 150 and 250, respectively. Making outliers with a 15%, 20%, or 35% margin allows us to compare the efficiency of robust regression with ordinary least squares estimates.

Huber M estimators results are observed in all cases, and the results are observed in Table 1

Table 1: Efficiency of Huber M estimator

Outliers percentage	**X1**	**X2**	**X3**	**X4**	**X5**	**R^2**
No outliers	-8.93e-05	1.400e-02	-4.266e-02	2.456e-02	-3.76e-02	0.2
15% in X_1	-0.0217	0.0234	-0.0457	0.0345	-0.0768	0.05
20% in X_1	0.0054	0.0324	-0.0675	0.0456	-0.0567	0.08
30% in X_1	0.0034	0.0435	-0.0768	0.0278	-0.0765	0.1
15% in Y	0.0030	0.0453	-0.0564	-0.00678	-0.0675	0.09
20% in Y	-0.0019	0.0568	-0.0654	0.0257	-0.0342	0.09
30% in Y	-0.0099	0.0675	-0.0765	-0.0087	-0.0008	0.32
Total						**0.93**

Table 1 shows that adding the outliers to X's has not significantly impacted the coefficient of X's. However, by creating outliers in the Y variable at varying percentages, we may discover a change in the coefficient of X. The R2 value measures the effectiveness. In the case of X variable outliers, the R2 stays relatively unchanged. However, efficiency decreases as the number of outliers in the Y variable increases. It follows that the Huber M estimator is unaffected by X-direction outliers. However, when outliers are present, efficiency decreases in the Y direction. However, the effectiveness of the Huber M estimator decreases as the fraction of outliers increases for Y-direction outliers. In addition, if any extreme values on the X-axis are detected. As a result, when dealing with outliers, the Huber M estimator is best. A Huber M estimator will not be considered an efficient estimator if inferential observations are discovered because the efficiency of the Huber M estimator will be affected by the inferential observations. Therefore, in this scenario, the problem with outliers may be resolved using any other robust estimator, and the estimators' efficiency will not be affected.

3.2 Ordinary Least Square Result

OLS is used because it may be compared to the Huber M estimator. Findings are shown for the initial simulation with no outliers and Y and X1 variables with 15%, 20%, and 35% outliers, respectively. The goal is to see how extreme values on both the horizontal and vertical axes behave. Both robust regression analysis and the ordinary least squares estimator show that outliers reduce efficiency. See Table 2 for the OLS estimator's findings.

Table 2: Efficiency of the ordinary least square

Outliers percentage	**X1**	**X2**	**X3**	**X4**	**X5**	**R^2**
No outliers	-0.006	0.0054	-0.047	0.065	-0.087	0.009
15% in X1	-0.034	0.0076	-0.065	0.018	-0.096	0.04
20% in X1	0.0076	0.023	-0.078	0.019	-0.043	0.007
35% in X1	0.0034	0.007	-0.044	0.033	-0.063	0.009
15% in Y	0.0054	0.056	-0.098	-0.065	0.0002	0.006
20% in Y	0.0053	0.058	-0.065	-0.064	0.0003	0.009
35% in Y	-0.034	0.076	-0.098	-0.008	0.087	0.006
Total						**0.086**

Table 2 shows a difference between the datasets with no outliers and those with a lower proportion of outliers introduced. However, the regressors' coefficient values show greater fluctuation as the outlier proportion rises. Overall analysis, including the insertion of outliers, reveals a total of 0.086 variations explained by the R2 value. In all, the investigation yielded an 8.6% efficiency rate. We

may conclude the OLS estimator since it affects the outliers regarding efficiency in both the x and y directions.

4 Conclusion

OLS and Huber M estimators show the effects on results of OLS even in the situation of 15% outliers because OLS is sensitive in case of outliers, whether in the Y or X axes. Meanwhile, Huber M estimators are unaffected by outliers in the X-direction. At the same time, it is affected by Y-direction outliers. Thus, robust estimators are more efficient in situations of outliers.

5 References

Barbu, A., & Zhu, S. C. (2020). *Monte carlo methods*. Springer.

Beattie, J. R., & Esmonde-White, F. W. (2021). Exploration of principal component analysis: deriving principal component analysis visually using spectra. *Applied Spectroscopy*, *75*(4), 361–375. https://doi.org/10.1177/0003702820987847

Begashaw, G. B., & Yohannes, Y. B. (2020). Review of outlier detection and identification using a robust regression model. *International Journal of Systems Science and Applied Mathematics*, *5*(1), 4–11. https://doi.org/10.11648/j.ijssam.20200501.12

Berk, R., Buja, A., Brown, L., George, E., Kuchibhotla, A. K., Su, W., & Zhao, L. (2019). Assumption Lean Regression. *The American Statistician*, *75*(1), 76–84. https://doi.org/10.1080/00031305.2019.1592781

De Menezes, D. Q. F., Prata, D. M., Secchi, A. R., & Pinto, J. C. (2021). A review of robust M-estimators for regression analysis. *Computers & Chemical Engineering*, *147*, Article e107254. https://doi.org/10.1016/j.compchemeng.2021.107254

Fitrianto, A., & Xin, S. H. (2022). Comparisons between robust regression approaches in the presence of outliers and high leverage points. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, *16*(1), 243–252. https://doi.org/10.30598/barekengvol16iss1pp241-250

Glielmo, A., Husic, B. E., Rodriguez, A., Clementi, C., Noé, F., & Laio, A. (2021). Unsupervised learning methods for molecular simulation data. *Chemical Reviews*, *121*(16), 9722–9758. https://doi.org/10.1021/acs.chemrev.0c01195

Gujarathi, D. M. (2022). *Gujarati: Basic Econometrics*. McGraw-hill.

Huber, P. J. (1965). A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, *36*(6), 1753–1758.

Javaid, A., Akbar, A., & Nawaz, S. (2018). A review on human development index. *Pakistan Journal of Humanities and Social Sciences*, *6*(3), 357–369. https://doi.org/10.52131/pjhss.2018.0603.0052

Javaid, A., Ismail, M. T., & Ali, M. K. M. (2021). Efficient model selection for moisture ratio removal of seaweed using a hybrid of sparse and robust regression analysis. *Pakistan Journal of Statistics and Operation Research*, *17*(3), 669–681. https://doi.org/10.18187/pjsor.v17i3.3641

Javaid, A., Ismail, M., & Ali, M. K. M. (2020). Efficient model selection of collector efficiency in solar dryer using hybrid of lasso and robust regression. *Pertanika Journal of Science & Technology*, *28*(1), 193–210.

Javaid, A., Muthuvalu, M. S., Sulaiman, J., Ismail, M. T., & Ali, M. K. M. (2019, December 04). *Forecast the moisture ratio removal during seaweed drying process using solar drier* (Paper Presentation). AIP Conference Proceedings. AIP Publishing LLC. https://doi.org/10.1063/1.5136404

Lee, S. W. (2022). Regression analysis for continuous independent variables in medical research: Statistical standard and guideline of life cycle committee. *Life cycle*, *2*, Article e3. https://doi.org/10.54724/lc.2022.e3

Lim, H. Y., Fam, P. S., Javaid, A., Ali, M., & Khan, M. (2020). Ridge regression as efficient model selection and forecasting of fish drying using V-Groove hybrid solar drier. *Pertanika*

Journal of Science & Technology, *28*(4), 1179–1202

Lodhi, I., Nawaz, S., Javaid, A., Javaid, S., & Javaid, A. (2023). Geographically analysis of wheat production on annual basis. *Statistics, Computing and Interdisciplinary Research*, *5*(1), 29–36.

Mooney, C. Z. (1997). *Monte carlo simulation* (No. 116). Sage.

Nawaz, S., Shahzad, N., Fraz, T. R., Shakil, A., & Khuram, H. R. (2022). Comparison of robust estimator in case of outliers. *Webology*, *19*(2), 8994– 9001.

Raychaudhuri, S. (2008, December 07–10). *Introduction to Monte Carlo simulation* (Paper Presentation). *2008* Winter Simulation Conference. IEEE, Miami, USA. https://doi.org/10.1109/WSC.2008.4736059

Schmidt, A. F., & Finan, C. (2018). Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, *98*, 146–151. https://doi.org/10.1016/j.jclinepi.2017.12.006

Suhaeri, M. E., Alimudin, Javaid, A., Ismail, M. T., & Ali, M. K. M. (2021, November 18). *Evaluation of clustering approach with euclidean and Manhattan distance for outlier detection* (Paper Presentation). AIP Conference Proceedings. AIP Publishing LLC. https://doi.org/10.1063/5.0075570

Sullivan, J. H., Warkentin, M., & Wallace, L. (2021). So many ways for assessing outliers: What really works and does it matter? *Journal of Business Research*, *132*, 530–543. https://doi.org/10.1016/j.jbusres.2021.03.066

Susanti, Y., & Pratiwi, H. (2014). M estimation, S estimation, and MM estimation in robust regression. *International Journal of Pure and Applied Mathematics*, *91*(3), 349–360. http://doi.org/10.12732/ijpam.v91i3.7

Thrampoulidis, C., Abbasi, E., & Hassibi, B. (2018). Precise error analysis of regularized $ M $-estimators in high dimensions. *IEEE Transactions on Information Theory*, *64*(8), 5592–5628. https://doi.org/10.1109/TIT.2018.2840720

Youssef, A. H., Abonazel, M. R., & Kamel, A. R. (2022). Efficiency comparisons of robust and non-robust estimators for seemingly unrelated regressions model. *WSEAS Transactions on Mathematics*, *21*, 218–244. https://doi.org/10.37394/23206.2022.21.28