Webology (ISSN: 1735-188X)
Volume 19, Number 2, 2022

Comparison Of Robust Estimator In Case Of Outliers

Shahbaz Nawaz[1] , Naeem Shahzad[2*] , Tayyab Raza Fraz[3] , Anas Shakil [4] , Hafiza Rukhsana Khuram [5]

[1]Bureau of Statistics Govt. of Punjab Planning and Development Department.

[2]College of Statistical and Actuarial Sciences, University of the Punjab, Lahore, Pakistan.

[3]Faculty, Department of Statistics University of Karachi.

[4]Department of Statistics University of Karachi.

[5]Govt. Post Graduate College for Women Samundri, GCU Faisalabad.

ABSTRACT

Outliers is a big problem in real life data analysis. In case of outliers, simple linear regression cannot perform well. For this problem, robust type of estimators are present. In this study, a simulation study is done from normal distribution having a sample size of 2500. Outliers with different percentages are generated to observe the efficiency of the robust type estimators. Three types of maximum likelihood (M) and modified maximum likelihood (MM) are used for the purpose of analysis. The efficiency is observed for each estimator and the coefficients are noted. The comparison is made with ordinary least square (OLS) in case of no outliers and for different percentages of outliers in the dataset. The results are observed in each case. Overall the Huber M showed the better efficiency than other estimators in the generated scenarios.

Keywords: robust regression, outliers, least square, simulation. M estimators

INTRODUCTION

Ordinary least square (OLS) is considered as a best technique in model selection only under some assumptions are met (Zuur et al., 2009). The problem in the dataset occurs in case of outlier are present in the dataset. The linear regression estimates got effected in presence of outliers. The

efficiency of OLS get reduced in such kind of dataset. So for such kind of problems, robust methods are available for hadleing the issue as (Gad & Qura, 2016) reviewed in their study the different types of robust methods for handeling the outliers. Many kind of robust estimators are available as maximum likelihood type estimators (M estimators), modified M estimators (MM) and estimators of scale (S) estimators (Susanti et al.,2014). But mostly researchers preferred M estimators (Sinova & Van Aelst, 2018). The main purpose of the robust regression is to provide efficient estimates even in case of outliers Draper and Smith (1998). In robust M estimators, the weighted functions are reduced at the tails in comparison of the least squae estimators in which weight one is given to all observations (Stuart, 2011). Robust type of estimators are used by (Dupuis & Victoria, 2013) for developing the variance inflation factor (VIF) regression for dealing with outliers. later on (Amini & Roozbeh, 2016) introduced robust ridge regression with the help of some robust estimtors for the problems of outliers in the dataset. One of the work was of Lukman et al.(2017) , a comparison was made from them for M, MM, LTS, LAD, OLS consisted of six economic variables from 1947 to 1962. Later on, the shrinkage robust estimators are developed by (Norouzirad et al., 2017) for combined problem of multicollinearity and outliers.
In previous research, there are many types of robust estimators were developed but the most common type is the M estimators due to its advantages and properties (Sinova & Van Aelst, 2018). In this research, M and MM estimators are used. The comparison is made with OLS in case of outlier problems. The results indicated that overall the M estimator shows better efficiency than the MM and OLS estimators.

METHODOLOGY

In this study, OLS, M and MM estimators are used. Three kind of weighted functions named as Huber, Hampel and Tukey's Bisquare are used for the purposed of analysis.

Ordinary least square

According to Stauart (2011), Let the design matrix X is to be defined as with the vector Y and ε, then the estimates can be calculated as

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ & \vdots & \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}, \quad Y = \begin{pmatrix} y1 \\ y2 \\ \vdots \\ yn \end{pmatrix} \text{ and } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

So the classic linear model is considered as $Y = X\beta + \varepsilon$ and the aim for least square estimate in to minimization of

$$\sum_{i=1}^{n}\varepsilon_i^2 = \varepsilon^T\varepsilon = \left(Y - X\beta\right)^T\left(y - X\beta\right)$$

$$= Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta$$

at minimumm of

$$\frac{\partial}{\partial\beta}\left(\sum_{i=1}^{n}\varepsilon_i^2\right) = \frac{\partial}{\partial\beta}\left(Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta\right)$$

$$= 0 - X^TY - X^TY + 2\left(X^TX\right)\beta$$

Thus the estimate of least square $\hat{\beta}$ is solution to $X^TX\hat{\beta} = X^TY$. As it minimizes $\hat{\varepsilon}^T\hat{\varepsilon} = \sum_{i=1}^{n} r_i^2$ so in case of X^TX as non singular then the estimates of least square can be directly estimated from data $\hat{\beta} = (X^TX)^{-1}X^TY$.

Maximum likelihood type Estimators

Huber's function

Kumar (2009) defined penalty function and influence function for Huber M estimator as follows

$$\rho_{HUBER(x)} = \frac{x^2}{2v^2}, \qquad \text{for} \qquad |x| \le kv^2$$

(3.17)

$$= \frac{k^2v2}{2} - k|x| \qquad \text{for} \quad |x| > kv$$

and

$$\Psi_{HUBER(x)} = \frac{\boldsymbol{x}}{\boldsymbol{v}^2} \qquad \text{for} \quad |x| \le kv^2$$

Hampel's function

Zaman and Bulut (2019) defined the Penalty fumction and influence functions for this estimator can be given as

$$P_{HAMPLE(x)} = \frac{x^2}{2} \text{ for } |x| \le a$$

$$= \frac{a^2}{2} - a|x| \text{ for } a < |x| \le b$$

$$a(\frac{c-|x|}{c-b}|x|) \quad \text{for} \quad b <|x|\leq c$$

= 0 for $|x| > c$

and

$$\psi_{HMAPLE(x)} = x \quad \text{for} \quad |x|\leq a$$

$$= asgn\left(x\right) \quad \text{for } a <|x|\leq b$$

$$= a\,\text{sgn}(x) = \frac{c-|x|}{c-b} \quad \text{for } b <|x|\leq c$$

= 0 for $|x| > c$

The choice for constant a is $a = kv^2$ and $b = 2kv^2$ that depends on robustness measures which is derived from influence function**.**

Tukey's bisquare function
it was suggested by Tukey (1977) can be defined as

$$\rho\left(y\right) = \frac{1}{6}(1 - (1 - (\frac{y}{k})^2)^3) \quad \text{for} \quad |y| \leq k$$

$$= \frac{1}{6} \quad \text{for} \quad |y| > k$$

When $k = 5$ or $k = 6$.

These three weighted functions are applied in this research. The results are noted in case of no outliers and for different percentages of outliers in dataset.

RESULT AND DISCUSSION

Data Simulation and Results

In this research, OLS, Huber M, Hampel M, Bisquare M, Huber MM, Hampel MM and Bisquare MM methods are analyzed. OLS is compared with all the robust estimators in term of efficiency based on the effecrives performance of coefficient of determination (R^2). R^2 in this study is found as

$R^2 = 1 - \frac{SS_{error}}{SS_{total}}$. On the basis of the value obtained from the formula, all mehods are compared in the analaysis.

The dataset in this study is generated from normal distribution with mean 100 and standard deviation 500 i.e. ~ (100,500). Six independent random variables are generated with one dependent variable. The sample size for all the variables are kept as 2500.
To generate outliers in the dataset, lower quartile (Q1), upper quartile (Q3) and inter quartile range (IQR) are calculated for each variable. Lower boundries and upper boundries for the IQR is calculated. The observations outside these limits is considered as now outliers. So, for the purpose of introducing outliers, 5%, 10% and 20% dataset is replaced at random with the data observations in each variable. The overall simulation is performed one time. The results of each method is analysed in case of no outliers, 5% outliers, 10% and 20% outliers respectively. The efficiency of each method used in this study is observed on the basis of R^2. R software is used for the analysis purpose.

The results are analyzed for each method in Table 1.

Table 1: Comparison of robust regression

	R2 values				
Techniques	**No Outliers**	**5% Outliers**	**10% Outliers**	**20% Outliers**	**Sum**
OLS	0.0010	0.0014	0.0009	0.0038	0.00713
Huber M	**0.0012**	**0.0024**	**0.0013**	**0.0043**	**0.009**
Hampel M	0.0011	0.0018	0.0011	0.0042	0.0082
BisquareM	0.0012	0.0023	0.0013	0.0043	0.009
Huber MM	0.0012	0.0022	0.0013	0.0043	0.009
Hampel MM	0.0012	0.0022	0.0013	0.0043	0.009
Bisquare MM	0.0012	0.0022	0.0013	0.0043	0.009

Table 1 provide the results of the simulation study used in the analysis. The performance of each method is observed in case of no outliers and with different percentages of outliers. Overall the sum is calculated for observing the total efficiency of the proposed method. From the overall efficiency, clearly the M and MM estimators are better than OLS. While in case of individual

analysis, when there is even no outlier in the dataset, OLS is less efficient than the robust estimators. In case of 5% outliers in the dataset, Huber M is showing the more efficiency as compare to other methods. For 10% outliers, OLS is again less efficient in the simulation study. Here Hampel M is less efficient as compare to other robust methods. In case of there is 20% outliers in the dataset. Robust estimators are still better than OLS. If the comparison is made with M and MM estimators only, then the MM estimators are showing the consistent performance with all the weighted functions used in the analysis as compare to M. The weighted functions used in M estimators are showing the different efficiency in each case. Over all, the Huber M method can be preferred based on the consistently better efficiency in all cases than the other methods.

The coefficients for each method is also observed and the behavior of the techniques in each case is analyzed. Table 2 present the coefficients obtained from each technique used in the analysis

Table 2: Coefficients of independent factors

Technique	**Coefficients of Variables**					
OLS	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**
No outliers	0.021	-0.011	0.006	0.005	0.013	-0.016
5% Outliers	-0.010	0.015	-0.020	0.023	0.003	-0.001
10% Outliers	-0.001	0.001	-0.026	0.004	-0.0001	-0.016
20% Outliers	0.029	0.007	-0.011	-0.005	0.033	0.036
Huber M	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**
No outliers	0.02	-0.02	0.01	-0.01	0.011	-0.014
5% Outliers	-0.02	0.013	-0.03	0.03	0.012	-0.014
10% Outliers	-0.01	0.004	-0.03	0.01	-0.015	-0.012
20% Outliers	0.04	0.01	-0.01	-0.003	0.039	0.03
Hampel M	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**
No outliers	0.022	-0.016	0.007	-0.001	0.011	-0.016
5% Outliers	-0.015	0.016	-0.022	0.026	0.006	-0.011
10% Outliers	-0.008	0.003	-0.029	0.005	-0.006	-0.014
20% Outliers	0.033	0.007	-0.001	-0.005	0.036	0.036
Bisquare M	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**
No outliers	0.020	-0.022	0.009	-0.007	0.010	-0.017
5% Outliers	-0.021	0.013	-0.024	0.031	0.011	-0.014
10% Outliers	-0.007	0.002	-0.032	0.008	-0.013	-0.012
20% Outliers	0.037	0.008	-0.008	-0.003	0.039	0.035
Huber MM	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**
No outliers	0.020	-0.022	0.009	-0.007	0.010	-0.017
5% Outliers	-0.021	0.013	-0.024	0.031	0.010	-0.014

10% Outliers	-0.007	0.002	-0.032	0.008	-0.013	-0.012
20% Outliers	0.036	0.008	-0.008	-0.008	0.039	0.035
Hampel MM	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**
No outliers	0.02	-0.02	0.01	-0.01	0.01	-0.02
5% Outliers	-0.021	0.013	-0.024	0.031	0.010	-0.014
10% Outliers	-0.007	0.002	-0.032	0.008	-0.013	-0.012
20% Outliers	0.04	0.08	-0.01	-0.03	0.04	0.04
Bisquare MM	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**
No outliers	0.020	-0.022	0.009	-0.007	0.010	-0.017
5% Outliers	-0.021	0.013	-0.024	0.031	0.010	-0.014
10% Outliers	-0.007	0.002	-0.032	0.008	-0.013	-0.012
20% Outliers	0.036	0.008	-0.008	-0.003	0.039	0.035

From Table 2, it is observed that the coefficients in OLS effected a lot in case of outliers in the dataset. There is a great difference in value of each independent factors. When there are no outliers in the dataset and when the outliers are present in the analaysis. In the M estimators and MM estimators, the coefficient values are not so much effected in presence of outliers in the dataset.
So, in case of outliers, OLS cannot be considered as a good choice. The robust estimators can provide the better estimates as compared OLS in outlier situation in dataset. In the present simulation study, Huber M can be preferred than all others in term of getting more efficient results. Also the coefficients for Huber M got not so effected in case of outliers. So, when there is need to deal with the problem of outliers, Huber M can be used in this kind of situation for getting efficient results.

CONCLUSION

The results indicates that the robust estimators showed the better performance in case of outliers. The coefficients for OLS changed a lot in every generated scenarios. Therefore, OLS is considered as very sensitive in problem of outliers. On the other hand, robust estimators got less effected. Among the robust estimators, Huber M showed the overall better efficiency than the other weighted functions used in the analaysis. Hampel M is considered the less efficient estimators than other robust estimators in this research. The MM estimators with each kind of weighted function showed almost the consistent behavior in all cases. So, in case of outliers, Huber M estimator can be used to get the more efficient results.

REFERENCES

1 Amini, M., & Roozbeh, M. (2016). Least trimmed squares ridge estimation in partially linear regression models. Journal of Statistical Computation and Simulation, 86(14), 2766-2780.
2 Draper, N. R., & Smith, H. (1998). Applied regression analysis (Vol. 326). John Wiley &

Sons.

3 Dupuis, D. J., & Victoria-Feser, M. P. (2013). Robust VIF regression with application to variable selection in large data sets. The Annals of Applied Statistics, 7(1), 319-341.

4 Gad, A. M., & Qura, M. E. (2016). Regression Estimation in the Presence of Outliers : A Comparative Study. International Journal of Probability and Statistics, 5(3), 65–72.

5 Kumar, T. A. (2009). A robust multiuser detection based scheme for crosstalk mitigation in DMT VDSL with non-Gaussian noise. 2009 International Conference on Signal Acquisition and Processing, ICSAP 2009, (i), 234–238.

6 Lukman, A. F., Ayinde, K., Adegoke, A. S., Tosin, D., & State, O. (2017). Some Robust Liu Estimators 1. Zimbabwe Journal of Science & Technology, 12(2409–0360), 8–14.

7 Norouzirad, M., Arashi, M., & Ahmed, S. E. (2017). Improved robust ridge M-estimation. Journal of Statistical Computation and Simulation, 87(18), 3469–3490.

8 Sinova, B., & Van Aelst, S. (2018). Advantages of M-estimators of location for fuzzy numbers based on Tukey's biweight loss function. International Journal of Approximate Reasoning, 93, 219-237.

9 Stauart, C. (2011). Robust regression. Guide to Statistics.

10 Susanti, Y., Pratiwi, H., Sulistijowati H., S., & Liana, T. (2014). M Estimation, S Estimation, and Mm Estimation in Robust Regression. International Journal of Pure and Apllied Mathematics, 91(3).

11 Tukey, J. W. 1977. Exploratory data analysis. Boston, MA: Addison Wesley

12 Zaman, T., & Bulut, H. (2019). Modified regression estimators using robust regression methods and covariance matrices in stratified random sampling. Communications in Statistics-Theory and Methods, 1-14.

13 Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). Limitations of Linear Regression Applid on Ecological Data. Springer science plus Business media