Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online): 2229-5666
Vol. 11, Issue 2 (2018): 69-80

REDESCENDING M-ESTIMATOR FOR ROBUST REGRESSION

***[1]Muhammad Noor-Ul-Amin, [2]Salah Ud Din Asghar, [3]Aamir Sanaullah and [4]Muhammad Ahmad Shehzad**
*[1,2,3]COMSATS University, Lahore, Pakistan
[4]BahauddinZakariya University, Multan, Pakistan
E Mail: *[1]nooramin@ciitlahore.edu.pk

Received February 21, 2018
Modified August 21, 2018
Accepted October 05, 2018

Abstract

In the linear regression problem, redescending M-estimators are used as an alternative method to the ordinary least square method when there are outliers in the data. Using the nonlinear transformations on the data one cannot remove the effect of outliers completely. In this paper, a redescending estimator is introduced for the robust regression to remove the effect of outliers in the data. The proposed estimator rejects the effect of outliers and provides efficient results about the parameter. The Ψ-function of the proposed objective function attains more linearity in the center before it redescends as compared to Insha (2006), Tukey (1974), Qadir (1996) and Andrews et al. (1972). The weight function of the proposed redescending M-estimator also gives improved results for the purpose it is introduced. To evaluate the prescribed results, a simulation study is conducted. A real data application is presented to demonstrate the performance of proposed estimator.

Key Words *:* Robust Regression, Redescending, M-Estimator; Outliers, Least Square Method.

1. Introduction

All-embracing work has been done by authors in the classical statistics during the last three decades. In classical statistics, ordinary least squares (OLS) method is very popular to estimate the population parameter. When there are outliers in the data it is not simple task to fit regression line as OLS estimates cannot retain their properties and do not provide efficient results to the population parameters. Due to outliers, information about the population absconds and they create a high variation in the data. To avoid this problem an alternative approach is considered to overcome this problem.

Robust regression is used when there are outliers in the data. By an improvement in the OLS method, robust regression analysis has been developed that provide efficient results in the presence of outliers. The aim of the robust regression M-estimator is to fit a model as close as to the population model. Kadilar et al. (2007) and Noor-ul-Amin et al. (2016) have used the robust regression in ratio method of estimation. The goal of the present study is to propose a robust estimator to improve the regression estimation results. A comparative study is conducted on the basis of numerical result of the proposed estimator with the other estimators i.e. Andrews et al. (1972), Ali and Qadir (2005),Insha-ullah et al. (2005), Alamgir et al. (2013) and Khalil et al. (2016).

M-estimator was introduced by Huber (1964) as a generalization of the familiar least squares criterion replacing the quadratic loss function with a symmetric function ρ (.) as,

$$Minimize_{\hat{\beta}} \sum_{i=1}^{n} \rho(r_i) \qquad (1)$$

where r_i represents the residuals. Most of the M-estimators can be solved by iteratively reweighted least square method. This estimator should satisfy the standard properties which are generally associated with an objective function of redescending M-estimator. An M-estimator is called a redescending M-estimator if it fulfills the standard properties related to it and the derivative of theρ-funtion will be a Ψ-function. However, M-estimator is not robust to the high leverage points, so it should be used in the situations where high leverage points do not occur. The weight functionw(r_i) gives less weight to the outliers and thus the estimates are less affected by outlying observations.

Differentiating equation (1) with respect to $\hat{\beta}_j$ we obtain $\Psi(r_i)$ function which gives us the following equation,

$$\sum_{i=1}^{n} \Psi(r_i) = 0 \qquad (2)$$

Dividing $\Psi(r_i)$ by r we obtained weight function given below,

$$\sum_{i=1}^{n} \mathrm{w}(r_i) X_i = \frac{\sum_{i=1}^{n} \Psi(r_i)}{r} \qquad (3)$$

It is a weighted function which assigns a weight closer to zero to the outlier very close to the zero and gives a weight very close to one if the observation lies in center of the data. It is also a weighted least square problem which requires an iterative solution called iteratively reweighted least squares.

2. Redescending M-estimator

Redescending M-estimators are non-decreasing near the origin.One of the known popular M-estimators is Huber (1973) estimator. The ρ(.) function of the Huber (1973) estimator is,

$$\begin{aligned} \rho(\mathrm{r}) &= \frac{r^2}{2} && for \quad |r| < c \\ &= c\left(|r| - \frac{c}{2}\right) && for \quad |r| \geq c \end{aligned} \qquad (4)$$

The graph of the Huber (1973) objective function is not smoothly redescending as Fig. 1 shows,

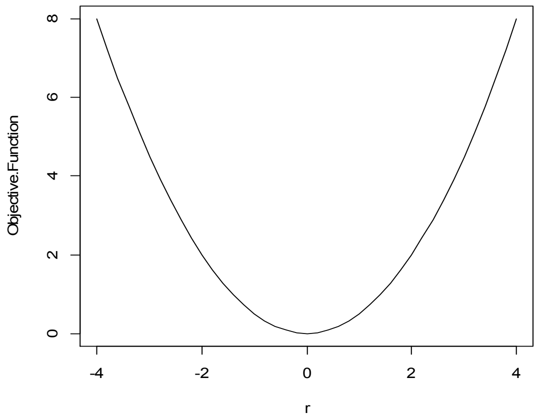


Fig.1: Huber's objective function

An objective function is redescending if its derivative is aΨ-function and fulfills the standard properties. Huber (1973)Ψ-function is,

$$\begin{aligned} \Psi(\mathrm{r}) &= \mathrm{r} \qquad && for \qquad |r| < c \\ &= \mathrm{c} && for \qquad |r| \geq c \end{aligned} \tag{5}$$

Andrews et al. (1972) proposed a three part redescending M-estimator whose Ψ-function become zero for large residuals. Andrews et al. (1972)Ψ-function is,

$$\begin{aligned} \Psi(\mathrm{r}) &= \mathrm{r} && for \qquad |r| < c \\ &= a\mathrm{sign}(\mathrm{r}) && for \qquad a < |r| \leq b \\ &= a\frac{c-|r|}{c-b} sign(r) && for \qquad b < |r| \leq c \\ &= 0 && for \qquad |r| > c \end{aligned} \tag{6}$$

where$a$,b and c are positive constant and $0 < a \leq b < c < \infty$. This function led to smoothly redescending M-estimator. After the development of this smoothly redescending M-estimator, several redescending M-estimators have been proposed.

Andrews (1974) sine function or also known and Andrews wave function is another redescending M-estimator. It has the following Ψ-function,

$$\begin{aligned} \Psi(\mathrm{r}) &= \mathrm{c}\sin\left(\frac{\mathrm{r}}{\mathrm{c}}\right) for && |r| < c \\ &= 0 && for \qquad |r| \geq c \end{aligned} \tag{7}$$

Qadir (1996) beta function has the followingΨ-function,

$$\begin{aligned} \Psi(\mathrm{r}) &= \frac{r}{16c^4}(\mathrm{c}+\mathrm{r})^2(\mathrm{c}-\mathrm{r})^2 && for \quad |r| \leq c \\ &= 0 && for \quad |r| > c \end{aligned} \tag{8}$$

Alamgir et al. (2013) proposed a modified tangent hyperbolic typeΨ-function as,

$$\begin{aligned}\Psi(\mathrm{r}) &= \frac{16re^{-2(r/c)^2}}{(1+e^{-(r/c)^2})^2} && for \quad |r| \leq c \\ &= 0 && for \quad |r| > c\end{aligned} \tag{9}$$

Khalil et al. (2007)proposed aΨ-function as,

$$\begin{aligned}\Psi(\mathrm{r}) &= \mathrm{r}\left(\frac{3}{2}\right)\left\{1-\left(\frac{r}{c}\right)^4\right\}^2 \sin\left[\left(\frac{2}{3}\right)\left\{1-\left(\frac{r}{c}\right)^4\right\}^2\right] && for \quad |r| \leq c \\ &= 0 && for \quad |r| > c\end{aligned} \tag{10}$$

3. Proposed Redescending M-estimator

We propose a new redescending M-estimator on the basis of standard properties related to it. A new objective function,

$$\rho(\mathrm{r}) = \frac{c^2}{4}\left[\frac{tan^{-1}\left(\frac{2r}{c}\right)^2}{4} + \frac{r^2c^2}{c^4+16r^4}\right] \quad for \quad |r| \geq 0 \tag{11}$$

where c[0,∞] is a tuning constant. We also discuss the shape of the objective function. The proposed function is redescending and fulfills the standard properties.The standard properties are:

- $\rho(r_i) \geq 0$
- $\rho(0) = 0$
- $\rho(r_i) = \rho(-r_i)$
- $\rho(r_i) \geq \rho(r_j)$ for $|r_i| \geq |r_j|$
- ρ is continuous (ρ is differentiable)

also Fig.2 shows the redescending nature of the proposed function.

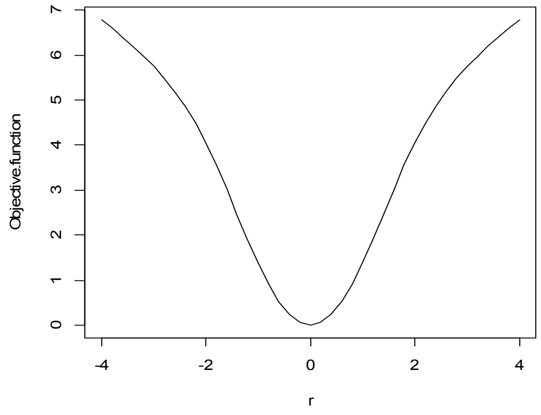


Fig. 2: Graph of Proposed objective function

Differentiating objective function w.r.t. r we obtained $\Psi(r_i)$, i.e.

$$= \frac{d}{dr}\left[\frac{c^2 tan^{-1}\left(\frac{4r^2}{c^2}\right)}{16} + \frac{c^4 r^2}{64r^4 + 4c^4}\right]$$

$$= \frac{c^2}{16}.\frac{d}{dr}\left[tan^{-1}\left(\frac{4r^2}{c^2}\right)\right] + c^4.\frac{d}{dr}\left[\frac{r^2}{64r^4 + 4c^4}\right]$$

$$= \frac{\frac{1}{\left(\frac{4r^2}{c^2}\right)^2 + 1}.\frac{d}{dr}\left[\frac{4r^2}{c^2}\right].c^2}{16} + \frac{\frac{d}{dr}[r^2].(64r^4 + 4c^4) - r^2.\frac{d}{dr}[64r^4 + 4c^4]}{(64r^4 + 4c^4)^2}c^4$$

$$= \frac{\frac{4}{c^2}.\frac{d}{dr}[r^2].c^2}{16\left(\frac{16r^4}{c^4} + 1\right)} + \frac{c^4\left(2r(64r^4 + 4c^4) - \left(64.\frac{d}{dr}[r^4] + \frac{d}{dr}[4c^4]\right)r^2\right)}{(64r^4 + 4c^4)^2}$$

$$= \frac{2r}{4\left(\frac{16r^4}{c^4} + 1\right)} + \frac{c^4(2r(64r^4 + 4c^4) - (64.4r^3 + 0)r^2)}{(64r^4 + 4c^4)^2}$$

$$= \frac{r}{2\left(\frac{16r^4}{c^4} + 1\right)} + \frac{c^4(2r(64r^4 + 4c^4) - 256r^5)}{(64r^4 + 4c^4)^2}$$

Simplifying we get,

$$= \frac{c^8 r}{(16r^4 + c^4)^2} = r\left[1 + \left(\frac{2r}{c}\right)^4\right]^{-2}$$

Hence, $\Psi(r_i)$ is,

$$\Psi(r) = r\left[1 + \left(\frac{2r}{c}\right)^4\right]^{-2} \qquad for \qquad |r| \geq 0 \qquad (12)$$

The graph of the Ψ-function of the proposed estimator is given in Fig. 3.

Dividing $\Psi(r_i)$by r we obtained weight as,

$$\mathrm{w}(r) = \left[1 + \left(\frac{2r}{c}\right)^4\right]^{-2} \qquad for \qquad |r| \geq 0 \qquad (13)$$

the proposed weight function covers the drawbacks of the preceding redescending M-estimators and provides less weights to the outliers.

The graph of the weight function of the purposed estimator is given in Fig. 4.

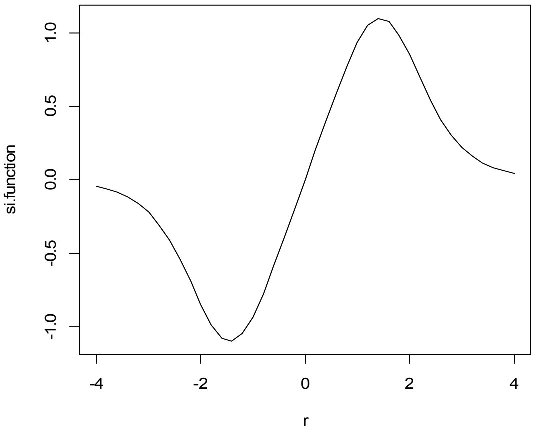


Fig. 3: Proposed Ψ-function

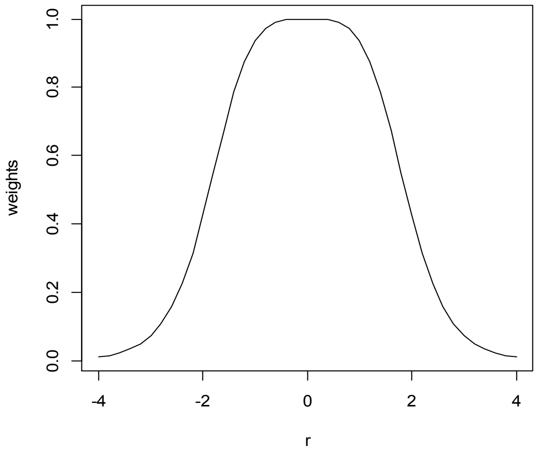
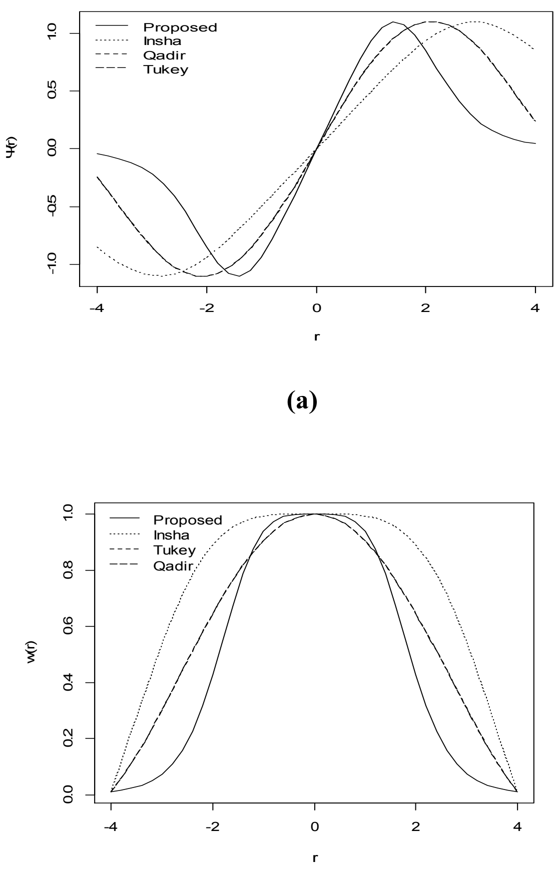


Fig. 4: Proposed weight function

4. Comparison

For the graphical comparison purpose, a multiple weight function and Ψ-function drawn on a same graph. All of the above stated redescending M-estimators work well in the presense of outliers but with some drawbacks. Andrews et al. (1972) three part estimator requires three tuning constants in the Ψ-function which is undesireable and not easy to analyze. Tukey (1974) biweight function covers some drawbacks of the Andrews et al. (1972) estimator by sacrificing some good observations.

Fig. 5 (a) represents the graph of Ψ-function and (b) represents the graph of weight function. In these graphs a comparison of the proposedweight and Ψ-function, Insha (2006), Qadir (1996) and Tukey (1974) redescending estimators has been presented. In generally Ψ-function is that better which is linear in the center so we need an estimator that treats the centeral observations linearly like ordinary least square (OLS) method and then redescends. Fig. 5 (a) and (b) shows that the proposed estimator contains more linearity in the center than any other esitmator and its Ψ-function is continuous everywhere.Grphical display of Tukey (1974) and Qadir (1996) weight and Ψ-function overcome so there are only three lines can be seen in Fig.5 (a) and (b).



(a)

(b)

Fig. 5: Comperison Graphs of Proposed, Insha's, Tukey's, Qadir's and Insha's Ψ and weight function

5. Real Data Example

To verify the effectiveness of the proposed estimator as compared to other estimators we compare the proposed redescending M-estimator with the other redescending M-estimators using the real life data example and simulation study through program R-language.

5.1 Telephone Data Example

A data of telephone calls which contains some outliers is taken from Ali and Qadir (2005) where the dependent variable is years and the independent variable is the number of phone calls made from Belgium. The scatter plot of the data is sketched in the Fig. 6. It is observed from the scatter plot that the telephone calls from 1964 to 1969 are outliers. Rousseeuw and Leory (1987) state that another recording system was used in the years where outliers occur.

In the Fig. 6 OLS line (solid line) pulled towards the outliers which is the effect of y-values towards the x-values from 1964-1969 and it is an unrepresentative fit. We also fit a line by using the proposed estimator (dotted line) and it provides the best representative fit towards the data by ignoring the effect of outliers.

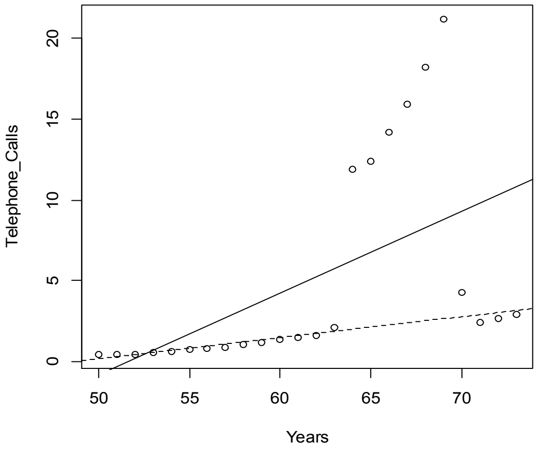


Fig. 6: Scatter plot of Telephone Calls data

The regression estimates obtained from OLS are given in Table 1 which indicates a bad fit. OLS estimates are highly influenced by outliers thus the fit represents not good estimates. From all the redescending M-estimators given in the literature Andrews (1974) gives the worst results than any other redescending M-estimator for this data. From the Table 1 it can also be observed that the proposed redescending estimator provides the most efficient results as compared to the other estimators. An estimator that has least amount of residual sum of squares (RSS) that is considered to be efficient than any other and the proposed estimator has the least RSS which represents the best fit to the data. RSS are to be used so that a real comparison can be made among difference robust methods.

Method Used	Estimates		RSS
	β_0	β_1	
OLS	-26.0059	0.5041	695.4354
Andrews	-25.6723	0.4976	218.8465
Tukey	-5.3060	0.1110	0.1362
Qadir	-5.2347	0.1099	0.1366
Asad	-5.2494	0.1102	0.1423
Insha	-5.2502	0.1103	0.3195
Uk	-5.2410	0.1101	0.3200
Proposed	-5.1981	0.1091	0.1317

Table 1: Regression estimates from different methods.

5.2 Annual Growth Rates of Price in China Example

To show the superiority of the proposed estimator another example is used. This example is taken from the Rousseeuw and Leroy (1987) for the years of 1940 to 1948, where the explanatory variable is the year and the dependent variable is the annually average growth of price. Same is scatter plot of telephone calls data in section 5.1, another scatter plot of annual growth rates of price in China is sketched in Fig. 7.In the Fig. 7 OLS line (solid line) do not provide good fit due to outliers. There are two outliers in this data. There is also a dotted line fitted through the proposed estimator which shows good fit as compared to OLS fitted line.

Numerical results of the proposed estimator as compared to other redescending M-estimators available in the litrature have been discussed in the Table 2. It can be clearly seen that the residual sum of square value of the proposed estimator is minimum which shows the better performance of the proposed estimator and gived close estimated to other robust methods. On the other hand OLS method is bad everywhere in the presence of outliers.

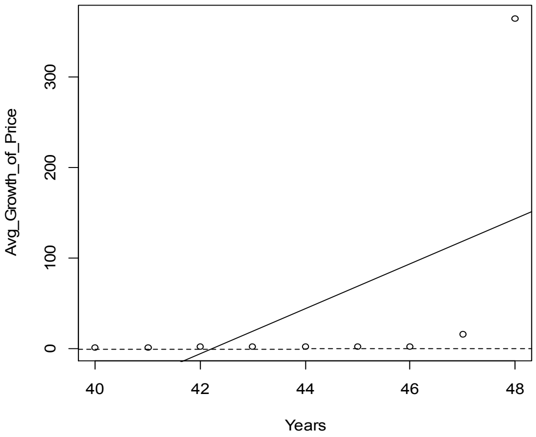


Fig. 7: Scatter plot of Telephone Calls data

Method Used	Estimates		RSS
	β_0	β_1	
OLS	-1049.47	24.845	130878
Andrews	-2.7710	0.1093	0.6175
Tukey	-2.8351	0.1108	0.6209
Qadir	-2.7535	0.1089	0.6166
Asad	-2.7294	0.1084	0.6149
Insha	-2.6514	0.1066	0.6100
Uk	-2.7711	0.1093	0.6174
Proposed	-2.6165	0.1058	0.6075

Table 2: Regression estimates from different methods.

5.3 Simulation Study

One of the common means of the comparison of different estimators is to do simulation study because the parameters of the population are unknown in the real life so for this purpose simulation study is conducted as in this situation we know the parameter of the data. The following regression equation is used,

$$y_i = \beta_0 + \beta_1 x_i + e_i \tag{13}$$

where $\beta_0 = 2, \beta_1 = 1$ and $e_i \sim N(0,1)$ and the independent variables are generated as$x_i \sim N(20,10)$ for $j = 1,2,\ldots,p$.Weight function depends onresiduals which depend upon the coefficients that are estimated and these estimated coefficientsdepend upon the weights.

An iteratively reweighted procedure is used by starting with the least square fitting. There are two cases for the simulation study. In the first case we use normal data. In the second case we took 95% observations from the first case and remaining 5% observations of the data replace by introducing outliers in the dependent variable. For this purpose we generate residuals as $e_i \sim N(50,1)$. The results are shown in the following Tables 3, 4 and 5. These results are obtained by using the average of 5000 Monto Carlo simulations where the number of samples in Table3, 4 and 5 are 100, 200 and 500 respectively. The main purpose of the simulation study is to measure the extent of the parameter estimates from the true value of the population parameter in the presence of outliers. From the following tables it is clear that the proposed estimator is providing as almost same results in the presence of outliers as the results of OLS method without outliers. The proposed estimator is also providing efficient results in the absence of outliers as OLS method.

Method Used	Case 1:Normal		Case 2: Outlier in y	
	β_0	β_1	β_0	β_1
OLS	2.0044	0.9997	4.9665	0.9762
Andrews	2.0163	0.9996	3.5441	0.9983
Tukey	1.9959	1.0002	1.9901	1.0003
Qadir	2.0023	0.9999	2.0109	0.9994
ALARM	2.0141	0.9992	2.0119	0.9994
Insha	2.0173	0.9990	1.9930	1.0002

Uk	1.9933	1.0003	2.0110	0.9993
Proposed	2.0100	0.9997	2.0044	0.9998

Table 3: Simulation results of regressionestimates from different methods, n=100

Method Used	Case 1: Normal		Case 2: Outlier in y	
	β_0	β_1	β_0	β_1
OLS	1.9962	1.0001	4.6254	0.9941
Andrews	2.0058	0.9995	3.4412	1.0027
Tukey	2.0700	0.9999	2.0038	0.998
Qadir	2.0053	0.9998	2.0060	0.9997
ALARM	2.0008	0.9999	2.0119	0.9993
Insha	2.0068	0.9997	1.9931	1.0003
Uk	1.9976	1.0000	2.0155	0.9993
Proposed	1.9962	1.0001	2.0025	0.9998

Table 4: Simulation results of regressionestimates from different methods, n=200

Method Used	Case 1: Normal		Case 2: Outlier in y	
	β_0	β_1	β_0	β_1
OLS	2.0007	0.9999	4.5271	0.9986
Andrews	1.9987	0.9999	3.3883	1.0054
Tukey	2.0026	0.9997	2.0030	0.9998
Qadir	2.0056	0.9997	1.9985	1.0000
ALARM	2.0006	0.9999	1.9968	1.0001
Insha	2.0013	0.9999	1.9931	1.0002
Uk	2.0006	0.9999	1.9918	1.0004
Proposed	1.9997	1.0000	1.9996	1.0000

Table 5: Simulation results of regression estimates from different methods, n=500

6. Conclusion

The main purpose of this study is to get information from the data more precisely in the situation where outliers exist in the data. For this purpose, a redescending M-estimator is proposed and compared with renowned estimators by conducting a simulation study. The proposed estimator is very easy to apply and it used only one tuning constant. Estimates of parameters are obtained by applying iterative least square technique. A comparative study is also presented using real life data examples for detection of outliers. The results of the real data examples showed that proposed redescending M-estimator is more efficient than other redescending M-estimators. The results of simulation study also show the superiority of the proposed estimator and the proposed estimator rejects the effect of outliers completely. It provides more precise results than any other redescending M-estimators discussed in the literature. Further, the proposed estimator is equally efficient as OLS method in the absence of outliers.

References

1. Alamgir, Ali, A., Khan, S.A., Khan D.M. and Khalil, U. (2013). A new efficient redescending M-estimator: Alamgir redescending M-estimator, Research Journal of Recent Sciences, 2(8), p. 79-91.
2. Ali, A. and Qadir, M.F. (2005). A modified M-estimator for the detection of outliers, Pakistan Journal of Statistics and Operation Research, 1(1), p. 49-64.
3. Andrews, D.F. (1974). A robust method for multiple linear regression, Technometrics, 16, p. 523-531.
4. Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. and Tukey, J.W. (1972). Robust Estimates of Location Survey and Advances, Princeton University Press, Princeton, New Jersey.
5. Beaton, A.E. and Tukey, J.W. (1974). the firing of power series meaning polynomials illustrated on band spectroscopic data, Technometrics, 16, p.147-185.
6. Huber, P.J. (1964). Robust estimation of a location parameter, The Annals of Mathematical Statistics, 35, p. 73− 101.
7. Huber, P.J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo, The Annals of Statistics, 1(5), p. 799–821.
8. Insha-Ullah, Qadir, M.F. and Ali, A. (2006). Insha's redescending M-estimator for robust regression: A comparative study, Pakistan Journal of Statistics and Operation Research, 2(2), p.135-144.
9. Kadilar, C., Candan, M. and Cingi, H. (2007). Ratio Estimation using Robust Regression, Hacettepe Journal of Mathematics and Statistics, 36, p. 181-188.
10. Khalil, U., Alamgir, Ali, A., Khan, D.M., Khan, S.A. and Qadir, F. (2016). Efficient UK's re-descending M-estimator for robust regression, Pakistan Journal of Statistics, 32(2), p. 125-138.
11. Noor-ul-Amin, M., Shahbaz, M.Q. and Kadilar, C. (2016). ratio estimators for population mean using robust regression in double sampling, Gazi University Journal of Science, 29(4), p. 793-798.
12. Qadir, M.F. (1996). Robust method for detection of single and multiple outliers, Scientific Khyber, 9(2), p. 135-144.
13. Rousseeuw, P.J. and Leroy, A.M. (1987). Robust Regression and Outlier Detection, John Wiley & Sons, New York.