Tikrit Journal of Administrative and Economic Sciences, Vol. 19, No. 64, Part (1): 712-725
Doi: www.doi.org/10.25130/tjaes.19.64.1.38

Tikrit Journal of Administrative and Economics Sciences
مجلة تكريت للعلوم الإدارية والاقتصادية
ISSN: 1813-1719 (Print)

Implementing a New Scale Technique in the M-Estimation Method to Estimate Parameters of Multiple Linear Regression: Simulation Study

Bekhal Samad Sedeeq* [A], Hogr Mohammed Qader [B], Azhy Akram Aziz [C], Dlshad Mahmood Saleh [A]

[A] Department of Statistics and Informatics/College of Administration and Economics/University of Salahaddin-Erbil Iraq

[B] Paytakht Technical Institute/private, Erbil, Iraq.

[C] Media Department/College of business and administration/Erbil polytechnic University-Erbil

Keywords:
Multiple Linear regression, robust method, Scale Estimator, OLS, RMSE.

A R T I C L E I N F O

Article history:
Received 22 Aug. 2023
Accepted 11 Sep. 2023
Available online 31 Dec. 2023

*Corresponding author:
Bekhal Samad Sedeeq
Department of Statistics and Informatics/College of Administration and Economics/University of Salahaddin-Erbil Iraq

Abstract: The goal of this study is to develop a new technique for estimating the parameters of a multiple linear regression by using M-estimation based on scale estimator to handle the influence of outlier values. In order to get new estimators, the root mean square error (RMSE) criterion is used to check the efficiency between the new technique and the classical method. The research showed that the new technique (M-estimation based on scale estimator) yields more accurate parameter estimates than the traditional approach (OLS) in all simulated cases.

Tikrit Journal of Administrative and Economic Sciences, Vol. 19, No. 64, Part (1): 712-725
Doi: www.doi.org/10.25130/tjaes.19.64.1.38

تنفيذ تقنية مقياس جديدة في طريقة تقدير- M لتقدير معلمات الانحدار الخطي المتعدد (دراسة محاكاة)

بيخال صمد صديق
قسم الاحصاء والمعلوماتية
كلية الادارة والاقتصاد
جامعة صلاح الدين- أربيل

هۆگر محمد قادر
معهد پايتخت التقني الخاص-اربيل

ئه زي أكرم عزيز
قسم الاعلام، كلية الادارة والاعمال، جامعة بوليتكنيك – اربيل،

دلشاد محمود صالح
قسم الاحصاء والمعلوماتية
كلية الادارة والاقتصاد
جامعة صلاح الدين- أربيل

المستخلص

الهدف من هذه الدراسة هو تطوير تقنية جديدة لتقدير معاملات الانحدار الخطي المتعدد باستخدام مقدرات M-الحصينة على أساس القياس S_n للتعامل مع تأثير القيم الشاذة. من أجل الحصول على مقدرات جديدة يتم استخدام معيار الجذر التربيعي لمتوسط الخطأ (RMSE) للتحقق من الكفاءة بين التقنية الجديدة والطريقة التقليدية. استنتج البحث أن التقنية الجديدة (تقدير M-الحصينة على أساس القياس) كانت لها معاملات مقدره أكثر دقة من الطريقة التقليدية (المربعات الصغرى الاعتيادية) في كل حالات المحاكاة.

الكلمات المفتاحية: الانحدار الخطي المتعدد، طريقة الحصينة، مقدر المقياس S_n، المربعات الصغرى الاعتيادية، RMSE.

1. Introduction

Multiple Regression analysis is an important statistical tool that is routinely applied in most sciences. Out of many possible regression techniques, the ordinary least squares (OLS) method has been generally adopted because of its tradition and ease of computation. However, there is presently widespread awareness of the dangers posed by the occurrence of outliers, which may be a result of keypunch errors, misplaced decimal points, recording or transmission errors, exceptional phenomena such as earthquakes or strikes, or members of a different population slipping into the sample. Outliers occur very frequently in real data, and they often go unnoticed because nowadays much data is processed by computers without careful inspection or screening. Not only the response variable can be outlying, but also the explanatory part, leading to so-called leverage points. Both types of outliers may totally spoil an ordinary LS analysis. Often, such influential points remain hidden to the user because they do not always show up in the usual OLS residual plots (Rousseeuw, and Leroy, 1987: 216). The basis of linear regression is the presumption that errors have a constant variance and are normally distributed. Outliers, which can significantly and unpredictably affect the errors, can cause this assumption to be falsified.

Outliers might result from inaccurate measurement, incorrect data entry, or unavoidable data variability. Outliers are intended to have less of an impact on the regression estimates by being given less weight or being completely disregarded in robust regression methods (Montgomery, 2012: 369).

Outliers in linear regression models can be handled with the use of robust regression techniques. Outliers are observations that drastically depart from the overall trend of the data and can skew estimations of the regression line's slope and intercept. This tutorial will teach you how to use a few popular robust regression techniques (Barnett and Lewis, 1994: 290) (Ali, T. H., & Salah, D. M., 2022: 920-939).

2. Methodology: In this part represents the theoretical aspect of multiple regression analysis will be reviewed, (outlier values), robust estimation method compared the new technique M- estimation method with the classic method (OLS) for outlier problems for estimating multiple linear regression model using the statistical criterion root mean square error (RMSE)

2-1. Multiple linear regression model: In order to represent the relationship between a scalar response or dependent variable, denoted by the letter Y, and one or more explanatory or independent variables, denoted by the letter X, we use the method of linear regression. In linear regression, unknown model parameters are inferred from the data by applying linear predictor functions to model the data (Alma, 2011: 411) (Obed, Saleh & Jamil, 2023: 1304-1324). The following is an example of a p independent variables linear regression model:

$$Y = \beta_O + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P + \varepsilon_i \quad \ldots. (1)$$

The dependent variable is Y.

Explanatory variables are X_i.

Y-intercept is called β_o.

Slope coefficients are β_p.

The model's error is ε_i.

2-2. Ordinary least squares: The multiple linear regression model and its estimation using ordinary least squares (OLS) is doubtless the most widely used tool in econometrics. It allows estimating the relation between a dependent variable and a set of explanatory variables (Ali, 2011: 331-348). The approach is based on the idea of reducing the sum of squared residuals between the actual and forecasted values. By reducing the sum of squared errors or residuals between the actual and predicted values, the OLS method

can be used to identify the best-fit line for data. Additionally, the partial derivative of the cost function with respect to the coefficients of determination must be taken into account while minimizing the sum of squares residuals in calculus (Ali, T. H. & Salah, D. M, 2021: 3388-3409). The partial derivatives must then be set to zero and the coefficients must then be solved for individually.

We estimate the parameters for a regression model using the ordinary least squares method, which minimizes the residual sum of squares, or sum of squared variances between the fitted and observed response (Almetwally& Almongy, 2018: 55-63).

$$\sum_{i=1}^{N} e_i^2 = e^T e = (Y - X\beta)^T (Y - X\beta) \quad \ldots . (2)$$

Minimization of (2) results into the least squares estimate of β which is$\hat{\beta} = (X^T X)^{-1} X^T Y$. The fitted regression model corresponding to the level of the regressor variables is $\hat{Y} = X\hat{\beta}$ The corresponding residual or error vector is $e = Y - \hat{Y} = Y - X\hat{\beta} = Y - X(X^T X)^{-1} X^T Y$. The residual sum of squares is calculated from (2) as $e^T e = \left(Y - X\hat{\beta}\right)^T (Y - X\hat{\beta})$. The residual sum of squares has degrees of freedom associated with it, since$(P + 1)$ parameters are estimated in the regression model. Thus, the mean and variance of the residual are $e_i \sim N(0, \hat{\sigma}^2)$.

Yields in the least squares estimate of, which is equal to

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad \ldots .. (3)$$

2-3. Outliers: Different scientific communities define outliers differently: An outlier is defined as an observation that does not conform to the pattern (model) suggested by the homogeneous majority of the observations in a data set. That does not conform to the linear regression line well. These observations have unusually high residual errors.

Some data sets may come from homogenous groups, while others may come from heterogeneous groups with varying features with reference to a certain variable. Outliers might result from inaccurate measurements, including data entry mistakes, or they can come from a different population from the rest of the data. Therefore, it's crucial to spot outliers for two reasons: either they point to a problem with the data that has to be corrected

or they are the first sign of a significant new trend (Rousseeuw and Leroy, 1987: 216).

Outliers are defined by (Hawkins, 1980:85) as observations that differ so significantly from other observations that it raises the possibility that they were produced by a separate mechanism. However, some definitions are considered general enough to deal with diverse types of data and methods.

- **Extreme outlier:** It refers to observation that lies at the end of the tail, for the status data upgrade or downgrade, if it is greater than (3σ)or smaller than (-3σ)in the context of figures of data under shade of standard normal distribution (Rousseeuw and Leroy, 1987: 216).
- **High-Leverage points:** High leverage points are observations that have outlying values in covariate space. In logistic regression model, the identification of high leverage points becomes essential due to their gross effects on the parameter estimates (Hawkins1980:85).

Let us consider a k-variable regression model, $Y = X\beta + \in$ the OLS residual vector can be expressed in terms of the true disturbance vector as:

$$\in = Y - \widehat{Y} = (I - H)\, Y \qquad (4)$$

Where the matrix $H = X(X^TX)^{-1}X^T$ given in Equation (4) is generally known as weight matrix or leverage matrix. Observations corresponding to excessively large $\in$ values are termed as outliers. The weight matrix H reflects joint effect of k regressors on the fitted responses. Usually the diagonal elements h_{ii} of the weight matrix H are considered as leverage values, which measure influences in the X-space. The i^{th} leverage value is defined as (Hawkins, 1980:85).

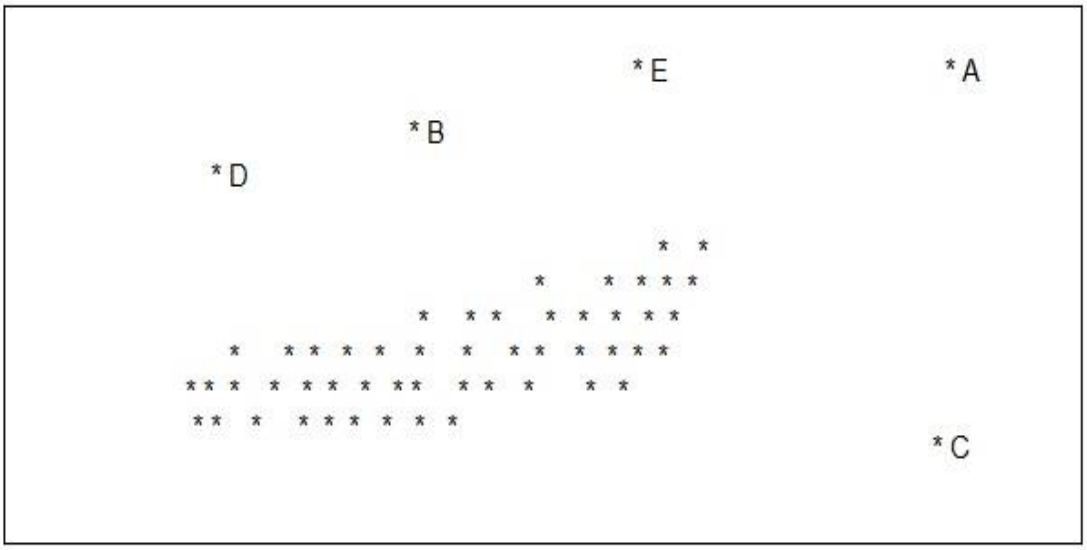
$$h_{ii} = x_i(X^TX)^{-1}x_i^T.$$

- **Influential observation:** Influential observations are those observations that, individually or collectively, excessively influence the fitted regression equation as compared to other observations in the data set (Hawkins, 1980: 85).

When presenting the following must be noted:

1. Outliers need not be influential observations.
2. Influential observations need not be outliers.
3. While observations with large residuals are undesirable, this is because least square fitting avoids large residuals.
4. Observations that have small residual do not mean necessarily, typical observation, because the gravitation of high-leverage point has small residual and influence on the success of the sample.

Doi: www.doi.org/10.25130/tjaes.19.64.1.38



The following instance illustrates the above remarks.

Adding points **A-B-C** in an insulation pattern into the typical points, the following will be produced:

In the concern point -A-: have small residual (because the value of (Y) is close to the straight through the other points), high-leverage point because it is phenomenal value of (X) and it does not have influence on the fitting of regression equation, thus the high leverage point is not influential.

In the concern point -B-: it does not possess high-leverage point (because it is located into the center (X)); whereas outlier (has great residual) and the influential point (with the entering, it does not change the slope, but it changes the straight junction point with the axis (Y)).

In the concern point -C-: it is the outlier (has great residual), high-leverage point (because of the extreme point in the space (X)) and the influential observation because it changes the fitting of regression equation (McCann, 2006: 109).

With Adding two points **E, D** with the points of the model, we can observe that:

Point -D- is: outlier but it is not influential and is not high-leverage point.

Point-E- is: influential observation because it changes the fitting of regression equation; but it does not phenomenal (it has a small residual) and it is not leverage point (Rousseeuw and Leroy, 1987:216).

2-4. New Scale Technique in M- Estimation method: If the data originate from a normal distribution, the standard deviation, the most popular scale estimate, is the most effective scale estimate. However, the standard

deviation lacks robustness in the sense that even a small change can have a significant impact on the estimated value of the standard deviation (low resistance) (CHEN, 2002: 08). Furthermore, it lacks resilience of efficiency for non-normal data. The most widely employed robust substitutes for the standard deviation are the (S_n)scale and median absolute deviation (MAD) under outliers' problem. Sigma is required to determine the robust method's parameters. In this study, we used a new technique that combines with M-Estimation. We obtain a new sigma $(\hat{\sigma}_{Sn})$, which we then apply to a robust technique to determine the parameters (Rousseeuw and Croux, 1993: 1273-1283). Huber created the M-estimation method, and it is currently the most popular robust regression methodology.

2-5. (S_n) Scale estimator alternative to the MAD: Using IRLS, this system can be solved. The likelihood works for $\beta \& \sigma$ is as follows under this circumstance:

$$L(\beta,\sigma)=\frac{1}{\sigma^n}\prod_{i=1}^{n} f\left(\frac{Y_i - X_i'\beta}{\sigma}\right) \qquad \ldots\ldots (4)$$

Where: $x_i=(1,x_{i1},x_{i2},.....,x_{ip-1})$ by replacing the ordinary least squares criterion with a robust criterion, M-estimator of β is (Hisham & Ehab, 2017: 55-63):

$$\hat{\beta}_M=\min_{\beta}\sum_{i=1}^{n}\rho\left(\frac{y_i - x_i'\beta}{\hat{\sigma}_{MAD}}\right) \qquad \ldots..(5)$$

$$\hat{\sigma}_{MAD}=C*MAD=C\left[Med\left|e_i - Med(e_i)\right|\right] \quad \ldots\ldots (6)$$

The (S_n) estimate of scale can be used as an alternative to the MAD. It shares with MAD the favored resilience traits of a restricted influence function and a 50% breakdown point. It does not depend on symmetry and also has a far greater normal efficiency (58%) than the previous method.

$$\hat{\sigma}_{Sn}=\frac{median\left[Median\left|e_i - e_j\right|\right]}{0.77519} \qquad \ldots\ldots (7)$$

In other words, we determine the median of $\left|e_i - e_j\right|$, j = 1, 2,..., n for each i. The median of these n numbers is then the estimated value of (S_n). It is found that (S_n) becomes a trustworthy estimator when the constant (1/c) is used. The chosen value is 0.77519, which is necessary for (S_n) to be a trustworthy estimator for normal data (Rousseau, P.J. and Croux, C., 1993: 1273-1283).

$$\hat{\beta}_{M-new} = \min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i' \beta}{\hat{\sigma}_{Sn}}\right) \qquad \ldots\ldots (8)$$

The (i-th) residual is indicated by the letter (e_i). The following normal equations are obtained:

$$\sum_{i=1}^{n} x_{ij} \psi\left(\frac{Y_i - X_i' \beta}{\hat{\sigma}_{Sn}}\right) = 0 \quad for \;\; j = 0 \;,\; 1 \;,\; 2,\ldots.,\; p-1 \qquad \ldots\ldots (9)$$

The iteratively reweighted least squares (IRLS) strategy was used to solve the nonlinear normal equations for M-estimates. The iterative process that follows is (Ruckstuhl, 2014:12):

1. Calculate the weights, like w_i .
2. Utilizing Eq. (7), determine a revised estimate of β .
3. Repeat steps two and three as necessary until the algorithm converges. Last but not least, the M-formula estimator's

$$\hat{\beta}_{M-NEW} = (X'wX)^{-1} X'wY \;; w = diag(w_i) \qquad \ldots. (10)$$

2-6. Evaluation Criteria: The evaluation criterion used to compare the performance of classical and robust processes in multiple linear regression models is the root mean square error (Ali, Albarwary and Ramadhan, 2023: 13):

$$RMSE = \sqrt{\frac{(Y_i - \hat{Y}_i)^2}{n-1}} \qquad \ldots\ldots (11)$$

Where:

Yi: is the actual value for the i-th observation.

Ŷi: is the predicted value for the i-th observation.

n: is the number of observations.

Doi: www.doi.org/10.25130/tjaes.19.64.1.38

3. A simulation experiment's description and analysis: This part compares the new M-estimation technique to the classic Ordinary Least Squares regression (OLS) method in a real-world setting. After studying the most crucial technique for eliminating data outliers, the comparison was made by assessing relative efficiency, which represents the root mean square of error (RMSE). The simulation experiment's implementation made use of varying degrees of the following factors: number of samples n, three sample sizes (50, 100, and 200) were used in this study. Without altering the explanatory variables, when (k) = (2, 4, and 8) and the (y) vector contain outliers (5%, 15%). A comparison of the approaches used in the estimation process represented by the new M-estimation technique with Ordinary Least Squares regression (OLS) was made for the frequency of 1000 replications. we achieve this by developing a specialized MATLAB (version 2020a) software for this work. The table below provides a summary of the algorithm used in simulation studies.

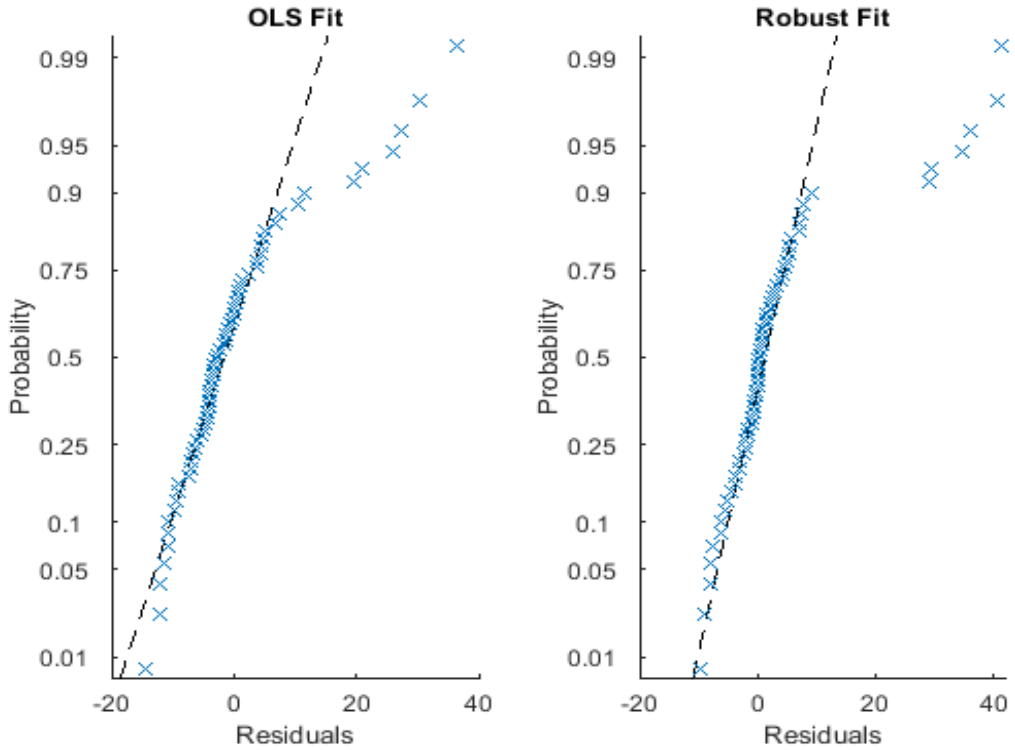


Figure (1): normal probability plot of the residuals

In figure (1) shows the residuals from the robust fit are closer to the straight line, except for the obvious outliers. This figure finds by author by using MATLAB program

Doi: www.doi.org/10.25130/tjaes.19.64.1.38

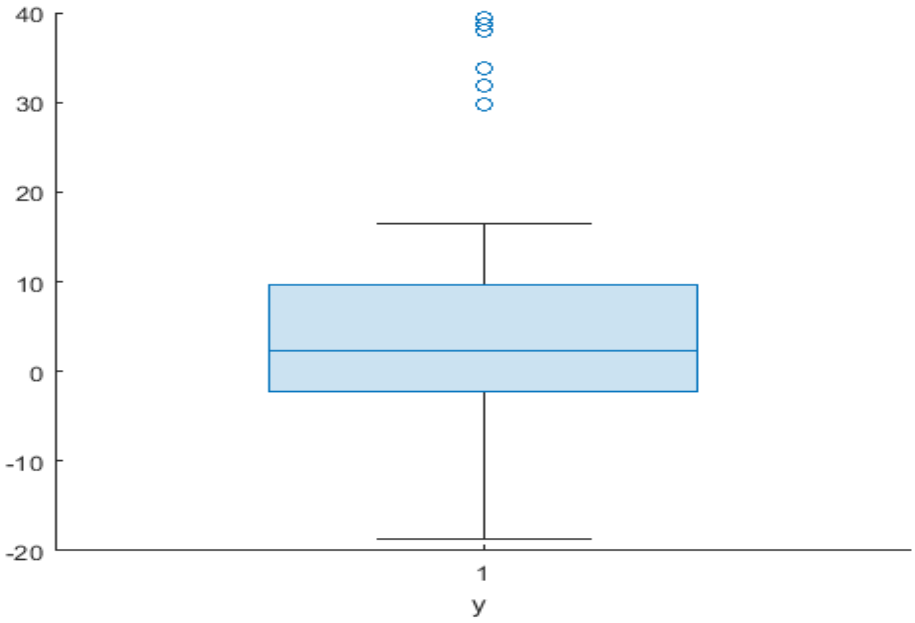


Figure (2): Box plot for (y)

By observing the figure (2), we find that the variable (Y) contains an outlier values. This figure finds by author by using MATLAB program

The simulation results of the first experiment for the classical and new technique M-estimation are summarized in the following:

a. **Linear regression model (OLS):**

y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:

	Estimate	SE	t-Stat	p-Value
(Intercept)	4.252	1.3819	3.077	0.0031688
x1	1.8632	1.4825	1.2568	0.21378
x2	-1.3023	1.7565	-0.74139	0.4614
x3	3.6857	1.4767	2.4959	0.015381
x4	1.2646	1.3422	0.94212	0.34997

Number of observations: 50

Root Mean Squared Error (RMSE): 10.8

R-squared: 0.154, Adjusted R-Squared: 0.0962

b. Linear regression model (robust fit):

y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:

Doi: www.doi.org/10.25130/tjaes.19.64.1.38

	Estimate	SE	t-Stat	p-Value
(Intercept)	1.5382	0.935	1.6452	0.10525
x1	2.2682	1.0031	2.2612	0.02744
x2	0.57771	1.1885	0.48608	0.62871
x3	3.9454	0.9991	3.9487	0.00021
x4	-0.50075	0.9085	-0.55137	0.58346

Root Mean Squared Error (RMSE): 7.3

R-squared: 0.337, Adjusted R-Squared: 0.292

From the results discussed above, we can see that the root mean square error had the lowest value of (7.3) and the highest value of R2, both of which were calculated using the new M-estimation technique. This confirms the method's superiority over the classical method (OLS) in handling outliers and obtaining a multiple linear model with high efficiency.

Table (1): shows how Y is distributed in simulation experiments.

n	The distribution of Y
50,100,200	95% $Y \sim N(0,1)$ + 5%$Y \sim N(6,30)$ 85% $Y \sim N(0,1)$ + 15%$Y \sim N(6,30)$

This table finds by authors by using MATLAB program

Table (2): the outcome of RMSE when Y is 5% contaminated and σ = 2.

Sample Size	Number of parameters	OLS	New (M-estimation)
		RMSE	RMSE
50	k=2	8.0466	3.0064
100		8.6919	2.4647
200		8.4091	1.8826
50	k=4	8.9601	5.4484
100		9.1328	4.3876
200		8.9786	3.3131
50	k=8	8.9467	7.2659
100		9.1216	6.4081
200		8.9783	5.1658

This table finds by author by using MATLAB program

Doi: www.doi.org/10.25130/tjaes.19.64.1.38

Table (3): the outcome of RMSE when Y is 5% contaminated and σ = 6.

Sample Size	Number of parameters	OLS RMSE	New (M-estimation) RMSE
50	k=2	9.3886	6.6753
100		9.9600	6.5297
200		9.7276	6.2940
50	k=4	10.1968	7.9567
100		10.3552	7.3876
200		10.2237	6.7810
50	k=8	9.8444	8.9200
100		10.3437	8.5844
200		10.2246	7.7422

This table finds by author by using MATLAB program

Table (4): the outcome of RMSE when Y is 15% contaminated and σ =2.

Sample Size	Number of parameters	OLS RMSE	New (M-estimation) RMSE
50	k=2	9.3824	3.5239
100		10.1638	3.1361
200		10.1407	3.9963
50	k=4	8.4763	5.6686
100		8.7945	4.9424
200		6.0938	4.0428
50	k=8	11.2664	5.9054
100		12.2674	5.8130
200		12.0102	5.4831

This table finds by author by using MATLAB program

Table (5): the outcome of RMSE when Y is 15% contaminated and σ =6.

Sample Size	Number of parameters	OLS RMSE	New (M-estimation) RMSE
50	k=2	11.5987	4.1289
100		8.4826	5.9850
200		11.3168	3.8699
50	k=4	9.0738	6.2586
100		6.7810	5.4918
200		7.3876	5.4227
50	k=8	9.2606	6.4739
100		9.3604	6.2827
200		10.0231	6.0798

This table finds by author by using MATLAB program

4. Conclusion

1. Based on the results of the relevant case study, it is possible to draw the conclusion that the new method in the robust method (M-estimation based on S_n scale) has successfully demonstrated its efficacy in estimating the regression model parameters with high accuracy under outlier values in the dependent variable.
2. In all simulation cases, the RMSE appears to decrease with increasing sample size.

References:

1. Ali, T. H., Estimation of Multiple Logistic Model by Using Empirical Bayes Weights and comparing it with the Classical Method with Application, Iraqi Journal of Statistical Sciences 20 (2011): 348-331.
2. Ali, T. H., Albarwari, N. H. S. and Ramadhan, D. L., Using the hybrid proposed method for Quantile Regression and Multivariate Wavelet in estimating the linear model parameters. Iraqi Journal of Statistical Sciences 20.1 (2023), 9-24.
3. Alma Ö., (2011), Comparison of robust regression methods in linear regression. International Journal of Contemporary Mathematical Sciences; 6(9): 409-21.
4. Almetwally, E. and Almongy, H., (2018), Comparison between M-estimation, S-estimation, and MM estimation methods of robust estimation with application and simulation. International Journal of Mathematical Archive; 9(11): 55-63.
5. Barnett, V. and Lewis, T., (1994), Outliers in Statistical Data. John Wiley.
6. CHEN, C., (2002), Robust Regression and Outlier Detection with the ROBUSTREG procedure [online]. SUGI Paper, SAS Institute Inc., Cary, NC.,
7. Hawkins, D. M., (1980), Identification of outliers (Vol. 11): Springer.
8. Hisham, M. A & Ehab, M. A., (2017), Comparison between Methods of Robust Estimation for Reducing the Effect of Outliers" The Egyptian Journal for Commercial Studies, Faculty of Commerce, Mansoura University, Egypt.
9. McCann, L., (2006), Robust Model Selection and Outlier Detection in Linear Regression. pdf. Unpublished, PhD Thesis, Massachusetts Institute of Technology.
10. Montgomery, D. C., Peck, E. A., & Vining, G. G., (2012), Introduction to linear regression analysis (Vol. 821): John Wiley & Sons.
11. Rousseeuw, P. J., & Leroy, A. M., (1987), Robust regression and outlier detection (Vol. 1): Wiley Online Library.
12. Rousseeuw, P.J. and Croux, C., (1993), Alternatives to the median absolute deviation, Journal of the American Stat. Assoc., 80, 1273-1283.
13. Ruckstuhl, A., (2014), Robust Fitting of Parametric Models Based on M-Estimation. Lecture notes.
14. Obed, S. A., Saleh, D. M. & Jamil, D. I., (2023), The Impact of Social Media Advertising on Customer Performance Using Logistic Regression Analysis. Qalaai Zanist Journal, 8(3), 1304–1324. https://doi.org/10.25212/lfu.qzj.8.3.54

15. Ali, T. H. & Salah, D. M., (2021), Comparison Between Wavelet Bayesian and Bayesian Estimators to Remedy Contamination in Linear Regression Model. PalArch's Journal of Archaeology of Egypt/ Egyptology, 18(10), 3388-3409.
16. Ali, T. H. & Salah, D. M., (2022), Proposed Hybrid Method for Wavelet Shrinkage with Robust Multiple Linear Regression Model with Simulation Study, Qalaai Zanistscientific JOURNAL A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq, Vol. (7), No (1), Winter 2022 ISSN 2518-6566.