Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6, 8506-8513
2024
Publisher: Learning Gate
DOI: 10.55214/25768484.v8i6.3825

Comparison between classical and robust methods for linear model parameters in the presence of outlier's values

Hayder Raaid Talib[1*], Sarah Adel Madhloom[2], Kareem Khalaf Aazer[3]
[1,3]University of Sumer / College of Administration & Economics; haider.raid@uos.edu.iq (H.R.T.) Kareemalataby28@gmail.com (K.K.A.).
[2]Middle Technical University Suwaira Technical Institute; Sarah_adel@mtu.edu.iq (S.A.M.).

Abstract: The subject of regression is one of the subjects that has become widely applied by those interested in various social and economic sciences, because it describes the relationship between variables in the form of an equation. The linear equation that includes explanatory variables is called the linear regression equation. To describe any phenomenon using regression models, the model assumptions must be met, as well as the importance of the accuracy of the data and its analysis on the accuracy of the results to be achieved in any scientific research. Hence, the interest in studying the integrity of the data, which is considered a necessary issue for the integrity of the results. It is obvious in any scientific research that we purify the data from outliers, if any, and then the presence of outliers in the selected sample helps to shade the results of the statistical analysis, and certainly the shading is greater the greater the number of those values or the greater the deviation of these values from the rest of the observations of the selected sample. The importance of this research came in estimating the parameters of the regression model using the usual methods (least squares) and robust methods (M method and MM method). The methods were compared using the root mean square error criterion using the simulation method, where different sample sizes were generated with four different ratios of outlier's values. It was noted that the superiority of robust methods was observed for all ratios of outlier's values and sample sizes.

Keywords: *M and MM methods, Outliers values, Robust methods.*

1. Introduction

The main steps of statistics were divided into several stages, and each stage became a major field of its fields, and the most important stages on which the accuracy of the results expected from the statistical model adopted in the study is based is the stage of estimating the parameters of the model, because the capabilities of the parameters of the model are what give the character of the phenomenon to be studied, and it has been suggested There are several statistical methods or methods to estimate the unknown parameters of the statistical model in the event that a number of assumptions or basic conditions are met. Among the most famous of these methods is the Ordinary Least Squares (OLS) method and the maximum likelihood method that is used to estimate the parameters of the linear regression model due to the good characteristics of the estimators of these methods. It is characterized as the best unbiased linear estimate that has the least variance among other unbiased linear estimates. However, the development taking place in the field of scientific research has imposed itself in this field, and the ideas and methods proposed for estimation methods have become very numerous, as some of them were designed to treat cases related to some phenomena, and another section was developed when some values of a phenomenon behave in a special way that differs from their peers, which is what is called outliers. The first idea or article published about the disturbances caused by outliers of the least square's method was by researcher Legendre in 1815. [7][9]

* Correspondence: haider.raid@uos.edu.iq

1.1. Research Objective

The research objective is to compare the classical methods represented by the least squares method with the robust methods represented by the M and MM methods using the simulation method using the statistical indicator mean square error (MSE).

1.2. Outlier Observation

It is that observation that is distant from the rest of the group's data for any of the variables of a phenomenon or a group of phenomena in terms of that value being large or small and located at one end of the group of observations arranged in ascending or descending order.

It is statistically defined as the observation that comes from a community different from the study population, but mathematically it is the observation whose Studentized Residual (RI) value is large compared to other observations in the data set.[6][12][14]

Anomalous observations can appear in more than one variable, as in the case of multivariate. The anomalous observation takes the form of a vector or part of a sub-vector.

1.3. Linear Regression Models

The study of linear regression models came to determine two points: the first is to define the relationship between the two variables from the graph of the shape of the regression line, and the second is to specify the linear parameters in the model. Most statistical studies and research are concerned with the second point.

2. Estimation Methods

2.1. Classic Methods of Estimation

There are several methods for estimating the parameters of the regression model, but these methods are considered vulnerable, meaning that the presence of anomalies in the data used affects the properties of their outputs. Among these methods are the least squares method, which are among the most famous and most used methods in estimating the parameters of the regression model. In order to understand the method.[3]

2.2. Least Square Method

As it is known that the least squares method is based on the idea of minimizing the sum of squares of errors or residuals, and the same idea has been adopted in the case of linear models, so we can express the sum of squares of residuals, also called classical least squares, and this method is one of the widely used methods in statistical applications and depends on the existence of a relationship between two or more variables. The principle of this method is based on finding the straight line that passes through the points of the diffusion shape in a way that makes the sum of the squares of the points' dimensions as little as possible, i.e., determining the value (β) that makes this sum as little as possible. The final formulas for the estimated parameters of the regression model are:[1][7]

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \bar{Y}\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2 - \bar{X}\sum_{i=1}^{n} X_i} \qquad \text{or} \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{1}$$

$$\therefore \hat{\beta}_0 = \frac{\sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} X_i Y_i)}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} \qquad \text{or} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{2}$$

This means that the value ($\hat{\beta}_1$) of what is nothing but the quotient of the common variance between (y, x) divided by the variance of (X).

The values of ($\hat{\beta}_1$) determine the best straight line that cuts through the diffusive shape, expressing the relationship between (Y, X). it is noted that both $\hat{\beta}_0$ and $\hat{\beta}_1$ are a function of the measurements of the sample items (β). Estimators can be obtained by the least square method, which will be as follows:

$$\hat{\beta} = (X'X)^{-1}(X'Y) \tag{3}$$

DOI: 10.55214/25768484.v8i6.3825

2.3. Robust Estimation

Following classical methods in estimating model parameters is not safer than applying robust estimation methods because the conditions required to apply classical methods are not easy, such as the absence of outliers or following a random error distribution other than the distribution that suits the method adopted in estimation, and the presence of a single outlier observation may destroy the good advantages of least squares estimators.[12][17]

2.4. M-method

This method is considered one of the most famous robust methods and is based on replacing the square of the residuals $\left(r_i^2\right)$ used in the least square's method with another function for the residuals while maintaining the main objective of the method, which is to make the amount $\sum_{i=1}^{n} \rho(r_i)$ The least possible. Note that $\rho(\cdot)$is a symmetric function with a single minimum endpoint at zero. Deriving this function relative to the regression parameters produces:[2][5]

$$\sum_{i=1}^{n} \psi(r_i)x_i = 1 \qquad (4)$$

Where $\psi(\cdot)$ is the derivative of $\rho(\cdot)$ and is called the influence function, while xi is the set of values of the explanatory variables for observation i.

In the end, the M estimators are obtained by solving nonlinear equations, which are numbered by the number of model parameters. This method has advantages, the most important of which are:

1- M-estimators are identical to OLS estimators when $\psi(t) = t$.

2- The M method is identical to the L1regression method when $\psi(t) = \sin(t)$

3- The M estimators do not have the characteristic of (Scale Invariant). In order for these estimators to have this characteristic, the random errors must be standardized by subtracting the arithmetic mean of these errors from them and dividing them by the standard deviations, which is calculated in a robust manner. One of the most reliable methods for calculating $\hat{\sigma}$is the MAD (median absolute deviation) method, according to which the formula for $\hat{\sigma}$ is as follows: $\hat{\sigma} = c \operatorname{med}(\,|\, ri - \operatorname{med} ri \,|\,) \quad i = 1,2,\dots,n$

Where c is a constant quantity and equals 1.4826.

4- Since the resulting robust estimators depend on the function $\rho(\cdot)$, several formulas for this function have been proposed. The following are the derivatives of some of these proposed formulas:[8][11]

- **Huber Minimax function**

$$\psi(t) = \begin{cases} t & \text{if } |\,t\,| < b \\ b \ \text{sign}\ (t) & \text{if } |\,t\,| \geq b \end{cases} \qquad (5)$$

Where b is a constant.

- **Descending Minimax function**

$$\psi(t) = \begin{cases} t & \text{if } |\,t\,| < a \\ b \ \text{sign}\ (t) \tan h \left[\frac{1}{2} b\ (c - [\ t\ ]\ )\right] & \text{if } a \leq |\,t\,| < c \\ c & \text{o.w.} \end{cases} \qquad (6)$$

Where a, b, and c are constants.

- **Hampel function**

$$\psi(t) = \begin{cases} t & \text{if } |\,t\,| < a \\ a \ \text{sign}\ (t) & \text{if } a \leq |\,t\,| < b \\ \{\ (\ c - |\,t\,|\ )\ /\ (\ c - d\ )\}\ a \ \text{sign}\ (t) & \text{if } b \leq |\,t\,| < c \\ o & \text{o.w.} \end{cases} \qquad (7)$$

DOI: 10.55214/25768484.v8i6.3825

Where a, b, and c are constants.

- **4) Andrew function**

$$\psi(t) = \begin{cases} \text{Sin}\,(t) & \text{if } -\pi \le t < \pi \\ 0 & \text{o.w.} \end{cases} \qquad (8)$$

- **Tukey function**

$$\psi(t) = \begin{cases} t(1 - (t / c)^2)^2 & \text{if } |t| < c \\ 0 & \text{o.w.} \end{cases} \qquad (9)$$

2.5. MM Estimator of Regression

Before we begin to define the estimator of MM, we will make some remarks about the estimation of (MM) in the framework of standard regression. We assume that (X) is a data matrix of size nxp containing the explanatory variables in columns, and we assume that the data of (Y) of size nx1 contains the dependent variables, and the ith row of Y, X contains information about the ith sample, which is denoted by ((Yi,Xi respectively). Let us assume a linear regression model, and the least squares estimator of β is defined as follows:[4]

$$\hat{\beta}_{Ls} = \text{argmin} \sum_{i=1}^{n} (y_i - x_i\beta)^2 \qquad (10)$$

It is known as the optimal estimator because it has the least variance and is an unbiased estimator if the error term is distributed normally. In the event that the error comes from other distributions such as heavy-tailed distributions, the least squares estimators lose their optimality in reaching the best estimators. The robust estimators known as (MM) estimators are obtained by replacing the square of the residuals in the above equation used in the least square method with the outlier function ρ.[10]

$$\hat{\beta}_{M} = \text{argmin} \sum_{i=1}^{n} \rho\,(y_i - x_i\beta)^2 \qquad (11)$$

The outlier function must be symmetric and non-decreasing (increasing) and be

$\rho(u) = u^2 \qquad (12)$

Let $r_i = y_i - x_i\beta$

It refers to the residuals in the objective function. The weight for observation i is defined as follows:

$W_i^r = \rho(r_i) / r_i^2 \qquad (13)$

$$\hat{\beta}_{M} = \text{argmin} \sum_{i=1}^{n} W_i^r\,(y_i - x_i\beta)^2 \qquad (14)$$

The estimator (MM) represents the weighted least squares (LS) and the weights depend on the estimator β. the MM regression estimator provides immunity against only vertical outliers which are within the error range. The other types of outliers are the leverage points which come from observations xi which are within the prediction range. The MM robust estimator provides protection against both vertical outliers and leverage points. Therefore, the MM robust estimator is as follows:[6][18]

$$\hat{\beta}_{RM} = \text{argmin}_{\beta} \sum_{i=1}^{n} wi^r\, wi^x (y_i - x_i\beta)^2 \qquad (15)$$

2.6. Data Generation

There are several methods for generating random data that follow a normal distribution, including the Box-Muller method and the approximate asymptotical method, and the Box-Muller method is more popular for its ease of use and accuracy of its results, and for generating data distributed according to a multivariate normal distribution with a mean vector μ and a variance matrix ∑, X ~ N (μ, ∑), by following the following algorithm:

1. Generating zi values using the (Box Muller) method
   Zi ~ N (0, 1), i= 1, 2, … P
   Z ~ MNp (0, 1) Since Z is an nxp matrix.
2. Partition of the variance matrix ∑ using the (Cholesky decomposition) method to obtain C. $C'C = \sum$ where C is a lower triangular matrix of PxP order and is the square root of the matrix ∑. The segmentation process takes place as follows:
   a) The di vector is calculated by the following formula:
   $$d_i = a_{ij} - \sum_{k=1}^{j-1} \ell_{ik}^2 d_k \; ; \; \ell_{ij} = \left[\left(a_{ij} - \sum_{k=1}^{j-1} \ell_{ik}\ell_{ik} d_k\right)\right] d_j^{-1}, \text{j, k= 1,2, …, P}$$
   b) The C matrix is extracted according to the following formula:
   $$C = \left[LD^{\frac{1}{2}}\right] \text{ where D = diag (di) and L = } \left[\ell_{ij}\right]$$

3. To calculate the matrix X whose data is distributed in a multivariate normal distribution with a mean vector μ and a covariance matrix and covariances is ∑ the following formula is used: $X' = \mu + CZ'$ and X ~ MNp (μ, ∑)

2.7. Generating Regression Model Data

In addition to generating the X matrix, which represents the explanatory variables. The random error that follows a normal distribution is generated with a mean (0) and a variance (σ_e^2). Then the values of the response variable are calculated from the following linear relationship: Y = X β + e

The values of β are imposed according to the model to be studied and in a manner consistent with the nature of the studied phenomenon, depending on the theoretical background of the phenomenon.

The sample: Four sample sizes were used, small, medium, two sizes, then large, as follows (60, 100, 150), and the error variance was in four values (1, 1.5), and we will address the problem of data outlier and the case of the MAR outlier mechanism with outlier rates of 2%, 6%, 10% and 15% and by repeating the experiment 500 times in order to determine which algorithms are more efficient.

2.8. Analyzing the Results of Simulation Experiments

Table 1.
MSE for the model to the estimation methods and outlier ratios when (e ~ N (1,1)) and sample size 60.

Outlier ratio **Methods**	**%2**	**%6**	**%10**	**%15**
Huber minimax estimators	1.563	1.508	1.598	1.583
Descending minimax estimators	1.593	1.608	1.519	1.599
Hampel estimators	1.569	1.527	1.582	1.518
Andrew estimators	1.611	1.500	1.567	1591
Tukey estimators	1.572	1.549	1.578	1.611
MM estimators	1.583	1.609	1.658	1.681
OLS estimators	1.892	1.916	1.698	1.627
Best	Huber minimax	Andrew	Descending minimax	Hampel

Table 2.
MSE for the model to the estimation methods and outlier ratios when (e ~ N (1,1)) and sample size 100.

Outlier ratio Methods	%2	%6	%10	%15
Huber minimax estimators	1.537	1.588	1.512	1.514
Descending minimax estimators	1.558	1.569	1.583	1.513
Hampel estimators	1.521	1.534	1.502	1.593
Andrew estimators	1.602	1.514	1.594	1.528
Tukey estimators	1.511	1.631	1.528	1.575
Mm estimators	1.796	1.761	1.541	1.566
Ols estimators	1.899	1.927	1.863	1.983
Best	Tukey	Andrew	Hampel	Descending minimax

Table 3.
MSE for the model to the estimation methods and outlier ratios when (e ~ N (1,1)) and sample size 150.

Outlier ratio Methods	%2	%6	%10	%15
Huber minimax estimators	1.293	1.298	1.295	1.237
Descending minimax estimators	1.236	1.269	1.263	1.284
Hampel estimators	1.248	1.246	1.312	1.213
Andrew estimators	1.313	1.254	1.297	1.211
Tukey estimators	1.263	1.284	1.225	1.243
Mm estimators	1.285	1.214	1.247	1.277
Ols estimators	1.585	1.692	1.758	1.581
Best	Descending minimax	MM	Tukey	Andrew

Table 4.
MSE for the model to the estimation methods and outlier ratios when (e ~ N (1,1.5)) and sample size 60.

Outlier ratio Methods	%2	%6	%10	%15
Huber minimax estimators	1.626	1.694	1.658	1.677
Descending minimax estimators	1.641	1.683	1.673	1.688
Hampel estimators	1.708	1.708	1.638	1.676
Andrew estimators	1.703	1.708	1.593	1.612
Tukey estimators	1.703	1.708	1.695	1.607
Mm estimators	1.636	1.669	1.663	1.684
Ols estimators	1.948	1.946	1.912	1.913
Best	Huber minimax	Descending minimax	Andrew	Tukey

Table 5.
MSE for the model to the estimation methods and outlier ratios when (e ~ N (1, 1.5)) and sample size 100.

Outlier ratio methods	%2	%6	%10	%15
Huber minimax estimators	1.437	1.438	1.462	1.414
Descending minimax estimators	1.458	1.469	1.483	1.413
Hampel estimators	1.421	1.434	1.422	1.493
Andrew estimators	1.412	1.484	1.494	1.448
Tukey estimators	1.512	1.431	1.428	1.475
Mm estimators	1.526	1.431	1.441	1.366
Ols estimators	1.899	1.727	1.863	1.883
Best	Andrew	Tukey	Hampel	MM

Table 6.
MSE for the model to the estimation methods and outlier ratios when (e ~ N (1, 1.5)) and sample size 150.

Outlier ratio methods	%2	%6	%10	%15
Huber minimax estimators	1.098	1.095	1.037	1.039
Descending minimax estimators	1.069	1.063	1.084	1.036
Hampel estimators	1.046	1.052	1.013	1.099
Andrew estimators	1.054	1.097	1.011	1.084
Tukey estimators	1.084	1.025	1.043	1.078
MM estimators	1.114	1.047	1.077	1.089
OLS estimators	1.292	1.258	1.361	1.438
Best	Hampel	Tukey	Andrew	Descending minimax

2.9. Analysis of the Results of Simulation Experiments of Outlier Patterns

First: In the case of the error variance, it is 1, From Tables (1,2 and 3) a, we note the following:

In the case of a sample size of 60 we notice that the Huber Minimax method gave the best results for all outlier ratios %2, and the Andrew method gave the best results for all outlier ratios %6, and the Descending Minimax method gave the best results for all outlier ratios %10, and the Hampel method gave the best results for all outlier ratios %15.

In the case of a sample size of 100 we notice that the Tuky method gave the best results for all outlier ratios %2, and the Andrew method gave the best results for all outlier ratios %6, and the Hampel method gave the best results for all outlier ratios %10, and the Descending Minimax method gave the best results for all outlier ratios %15.

In the case of a sample size of 150 we notice that the Descending Minimax method gave the best results for all outlier ratios %2, and the MM method gave the best results for all outlier ratios %6, and the Tuky method gave the best results for all outlier ratios %10, and the Andrew method gave the best results for all outlier ratios %15.

Second: In the case of the error variance, it is 1.5, From Tables (4,5 and 6) a , we note the following:

In the case of a sample size of 60 we notice that the Huber Minimax method gave the best results for all outlier ratios %2, and the Descending Minimax method gave the best results for all outlier ratios %6, and the Andrew method gave the best results for all outlier ratios %10, and the Tuky method gave the best results for all outlier ratios %15.

In the case of a sample size of 100 we notice that the Andrew method gave the best results for all outlier ratios %2, and the Tuky method gave the best results for all outlier ratios %6, and the Hampel method gave the best results for all outlier ratios %10, and the MM method gave the best results for all outlier ratios %15.

In the case of a sample size of 150 we notice that the Hampel method gave the best results for all outlier ratios %2, and the Tuky method gave the best results for all outlier ratios %6, and the Andrew method gave the best results for all outlier ratios %10, and the Descending Minimax method gave the best results for all outlier ratios %15.

Given the difficulty of providing all the assumptions of the linear model in experiments, and to reduce the effect of outliers if they are present in the data, the research recommends using the studied methods according to the sample size and benefiting from the scientific programs prepared by the researcher in applying these methods.

Copyright:

References

[1] Basu, S. & Heitjan, D. F. (2016) "Distinguishing 'Missing at Random' & 'Missing completely at Random' The JASA, vol. 23, No. 4, p. 123- 132.

[2] Beal, M.J., & Ghahramani, Z. (2016) "The variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model structures", www.stat.cmu.edu/fienberg/stat36-756-16.html .

[3] Bickel,D.R.,(2012). "Robust and efficient estimation of the mode of continuous data "http://www.mcg.edu/research/biostat/bickel.html.

[4] Campbell, N. A (1999) "Robust procedures in Multivariate Analysis I: Robust covariance Estimation" Applied statistics, vol. 23, P. 213- 224.

[5] Daniel P. Normolle .(2013)."An Algorithm for Robust nonlinear Analysis of Radioimmunoassay and Othre Bioassays "Ann Arbor,MI 48119-2129 , 313-936-1113 .

[6] David, T. O (2014) "Applied Robust statistics" Chapman & Hall, New York.

[7] Dempster, A. P. & Larid, N. M. & Rubin, D. B. (1977) "Maximum Likelihood from Incomplete data via the EM Algorithm" J. Royal Statist. Soc. Ser. B, vol. 11, P. 1-22.

[8] ftp://ftp.cea.fr/pub/dsv/madic/pubils/roche-em .

[9] Honaker, J.; Joseph, A.; King, G. & Scheve, K. (2011) "Analyzing Incomplete political science data: An Alternative Algroithm for Multiple Imputation" American political science review, vol. 68, No. 2, P. 123-137.

[10] Jank, W. (2016) "The EM Algorithm, it's stochastic Implementation and Global Optimization: some challenges and opportunities for OR", www.rhsmith.umd.edu/faculty/wjank/.../page1118.htm .

[11] Jurckova, J. & Sen, P. K. (1996) "Robust statistical procedures; Asymptotic and Interrelations" John Wile & Sons, New York.

[12] Lin, C. & Rubin, D. B. (2012) "Maximum Likelihood Estimation of Factor Analysis using the ECME Algorithm with complete and Incomplete Data", statistica sinica 8, p. 437-446.

[13] Little, R. J. A. & Rubin, D. B (2013)" statistical analysis with missing data", 2nd ed., John Wile & sons, New York.

[14] Little, R. J. A. & Rubin, D. B. (2012) "Statistical Analysis with Missing Data", New York: John. Wiley.

[15] Maronna, R. A. (1976) "Robust M- Estimators of Multivariate location and scatter" The Annals of statistics, vol. 3, P. 33- 48.

[16] Morrison, D. F. (1971) "Expectation and variances of Maximum likelihood Estimates of the Multivariate Normal Distribution parameters with missing data", JASA, vol. 53, P. 326- 339.

[17] Peter, J. H (1981) "Robust statistic" John Wile & Sons, New York.

[18] Roche. A. (2016) "EM Algorithm and variants: an informal tutorial",

[19] Rubin, D. B. (1974) "Characterizing the Estimation of parameters in Incomplete- Data problems" Journal of the American statistical association.