Study of Robust Regression Modeling Using MM-Estimator and Least Median Squares

Khusnul Khotimah [1], Kusman Sadik[2], Akbar Rizki[3]
{khusnulrfa@gmail.com[1], kusmansadik@gmail.com[2], akbar.ritzki@gmail.com[3]}

Statistics, IPB University, Bogor, 16680,Indonesia[1], *Statistics, IPB University, Bogor, 16680,Indonesia*[2], *Statistics, IPB University, Bogor, 16680,Indonesia*[3]

Abstract. Ordinary least squares (OLS) is a method commonly used to estimate regression equations. One solution handle OLS limitation to outlier problem is to use the robust regression method. This study used least-median squares (LMS) and multi-stage method (MM) robust regression. Simulation results of regression analysis in various scenarios are concluded that LMS and MM methods have better performance compared to OLS on data containing vertical and bad leverage point outliers. MM method has lowest average parameter estimation bias, followed by LMS, then OLS. LMS has smallest average root mean squares error (RMSE) and highest average R^2 is followed by MM then OLS. The results of the regression analysis comparison of the three methods on Indonesian rice production data in 2017 which contains 10% outliers were concluded that the LMS is the best method. The LMS produces the smallest RMSE of 4.44 and the highest R^2 that is 98%.

Keywords: least median squares, multi-stage method, outliers, robust regression, root mean squares error.

1 Introduction

The development of Statistics and Mathematics is supported by technology that produces various methods that have been applied in various fields of life. One method of analysis that is widely applied in various scientific fields is the linear regression method. Linear regression method is a statistical method used to evaluate the linear relationship between the quantitative response variables with one or more quantitative explanatory variables. The ordinary least squares method is the most popular approach used in estimating the linear regression estimation parameters [1]. However, this least squares method has limitations, which are strongly influenced by outlier data [2]. Outliers are an extreme value observation that is very much different from most other data [3].

Disposal of outliers from the data set can only be done if the value is known to cause, such as measurement or analysis errors, data recording errors, or failure of measuring instruments. Outliers sometimes contain information that is more important than the value of other data observations so that it will have a big effect on the model formed. Therefore, removing the outlier value to improve the compatibility of the regression equation cannot be done carelessly because it will provide precision estimation that is not right [3].

Completion of outlier problems in the regression analysis process can be done using the robust regression method. Robust regression is a regression method that is used when data

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290533

contains outlier observation values [4]. Ehab Mohamed A and Hisham Mohamed A [5] and Arista Oktarinanda [6] examined the comparison of several robust methods. Almetwally and Almongy [5] compared the M, S, and MM methods and concluded that the MM method was the best method based on the smallest criteria of bias and mean square error (MSE). Oktarinanda's study [6] compared the LMS and LTS methods and concluded that the LMS method was more accurate in predicting models based on smaller RMSE values.

The results of this study lead this research in comparing the MM regression estimators and LMS estimators and determine the best method. The selection of the best method will be based on the value of the regression parameter bias, root mean square error (RMSE), and R-Square (R^2). Therefore, the researcher raised the topic of "Study of Robust Regression Modeling Using the MM-Estimation and Least Median Squares". Comparison is done through simulating data with various data sizes, outliers, and outliers, then applied to agricultural data, namely rice production data (in million tons) in 2017.

2 Materials

2.1 Simulation Data

This study involves generating data that contains outlier observations for robust linear regression simulations. The simulation process carried out using data size (n) as many as 50 data for small size, 200 data for medium size, and 1000 data for large size. There are 3 types of outliers that are generated, namely vertical outliers, good leverage points, and bad leverage points. Outliers percentage (m) used is 0%, 5%, 10%, 15%, 20%, and 30% for each type of outlier. The generation data from the combination of scenarios will then be used in the parameter estimation process using the ordinary least squares method (OLS), LMS, and MM. The parameters β_0 and $\beta_1$used in this study are 5 and 2 respectively so that the linear regression model is formed as esquation (1) follows:

$$\boldsymbol{y} = 5 + 2\boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{1}$$

with,
$\boldsymbol{y}$: response data vector size $n \times 1$.
$\boldsymbol{x}$: explanatory variable data vector $n \times 1$
$\boldsymbol{\varepsilon}$: error vector size $n \times 1$

All simulation results are based on 1000 repetitions carried out with the help of R software. The process of comparing the goodness of the method is done by looking at the parameter estimation bias ($bias\ (\hat{\beta}_0)$ and $bias\ (\hat{\beta}_1)$), RMSE, and R^2.

2.2 Actual Data

This study also uses actual data as a comparison application of OLS, MM and LMS methods. The actual data used is data on rice production (in million tons) and data on the amount of use of organic fertilizer (in thousand tons) in 2017 obtained from the Ministry of Agriculture's Data and Information System Center of the Republic of Indonesia. The data will be analyzed using simple linear regression to determine the linear relationship between explanatory variables, namely data on the amount of organic fertilizer with the response variable, namely rice production data in 2017. The data consists of 34 observations which are provinces in Indonesia.

3 Methods

3.1 Simulation Data

The process of data analysis in this study uses R software with the help of "MASS" and "robustbase" packages. The simulation steps carried out in this study are as follows:

1. Set $\beta_0 = 5$ and $\beta_1 = 2$
2. Generating vectors explanatory variable $(\boldsymbol{x})$ with normal distribution as much as n data with average and variety as follows:
    i. $x \sim N\ (5, 1)$ for data not outliers.
    ii. $x^* \sim N\ (30, 1)$ for outlier data.
3. Generates a vector of error values (ε) as much as n data with the following details:
    i. $\varepsilon \sim N\ (0, 1)$ for data not outliers.
    ii. $\varepsilon^* \sim N\ (30, 1)$ for outlier data.
4. Determine y value vector based on the following scenario:
    i. Vertical Outlier
    Determine y vector based on model in equation (2)

$$\boldsymbol{y} = 5 + 2\boldsymbol{x} + \boldsymbol{r}\ , \qquad (2)$$

with r being a vector of error components obtained from a combination of sampling data $\varepsilon \sim N\ (0, 1)$ as much $(1 - m)\ \times\ n$ and data $\varepsilon^* \sim N\ (30, 1)$ as much as $m\ \times\ n$.The vector y will then be regressed with vector x at the next stage.

    ii. Good leverage point
    Determine the vector y based on the modelin equation (3)

$$\boldsymbol{y} = 5 + 2\boldsymbol{s} + \boldsymbol{\varepsilon}, \qquad (3)$$

with s is a vector of explanatory variables obtained from a combination of data sampling $x \sim N\ (5, 1)$ as much $(1 - m)\ \times\ n$ and data $x^* \sim N\ (30, 1)$ as much as $m\ \times\ n$.The vector y will then be regressed with vector s in the next stage.

    iii. Bad leverage point
    Determine y vector based on model in equation (4)

$$\boldsymbol{y} = 5 + 2\boldsymbol{x} + \boldsymbol{t}, \qquad (4)$$

with t is a vector of error components obtained from a combination of sampling data $\varepsilon^* \sim N\ (30, 1)$ as much as $\frac{1}{2}m\ \times\ n$ with data $\varepsilon \sim N\ (0, 1)$ as much $(1 - \frac{1}{2}m)\ \times\ n$. Then as much as $m \times n$ the first data in the explanatory vector (x) is added by 2.The vector y will then be regressed with the new x vector at the next stage.

Illustration of the appearance of the three types of outliers that will be generated can be seen in Figure 1 below [7]:

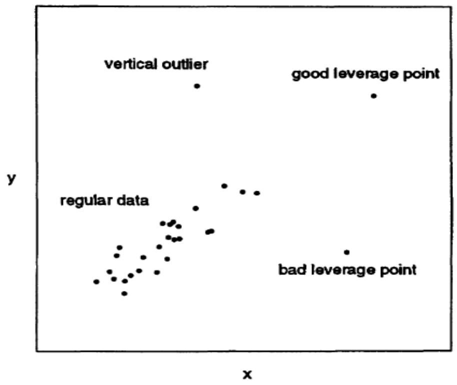

Fig. 1. Outliers types

5. Regress all datasets using OLS, LMS, and MM with details of the following steps:

(1) OLS method regression algorithm

a. Arrange data vectors y and X matrix. X matrix is a data matrix of explanatory variables measuring $n \times (k + 1)$ where k is the number of explanatory variables and the first column contains vector 1.

b. Calculates the estimation coefficient of the β parameter using equation (5)

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{y}) \ . \tag{5}$$

(2) LMS method regression algorithm.

The parameter estimation steps with the LMS method summarized in a PROGRESS algorithm by Rousseeuw and Hubert [7] on the MASS R software package are as follows:

a. Determine the value of g using equation (6) below:

$$g = \left[\frac{n+k+1}{2}\right] . \tag{6}$$

b. Taking a random set of g-sized data sets from a data set measuring n observations, so that there will be $f = C_g^n$ the subdata set.

c. Estimating regression parameters from each set of data using OLS

d. Calculates the value of M_d or the median of the residuals square (e_{vd}^2) in each set of data. Index v is an index for the number of observations in each set of data $v = 1,2,3, \dots, g$ and index d is the number of subsets formed, $d = 1,2,3, \dots, C_h^n$.

e. Determine the value of M, namely the minimum median e_{vd}^2 based on the results of stage (d).

f. Make an initial estimate of the standard deviation of the LMS ($\hat{\sigma}_{LMS}$) using equation (7) below:

$$\hat{\sigma}_{LMS} = 1.4826\left(1+\frac{5}{(n-g)}\right)\sqrt{M}\ . \tag{7}$$

g. Calculate LMS estimator weights ($w_{i_{LMS}}$) using equation (8);

$$w_{i_{LMS}} = \begin{cases} 1\,, \left|\frac{e_i}{\hat{\sigma}_{LMS}}\right| \leq 2.5 \\ 0\,, lainnya \end{cases}, \tag{8}$$

$w_{i_{LMS}}$ is a diagonal element of the $\boldsymbol{W^{LMS}}$ matrix that has a size of $n \times n$ and other elements worth 0.

$$\boldsymbol{W^{LMS}} = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{pmatrix}$$

h. Estimating regression parameters using a weighted regression method as equation (9)

$$\boldsymbol{\hat{\beta}_{LMS}} = (\boldsymbol{X}'\boldsymbol{W^{LMS}X})^{-1}(\boldsymbol{X}'\boldsymbol{W^{LMS}y}) \tag{9}$$

with,

y : vector data response size $n \times 1$.

X : data matrix explanatory variable size $n \times (k + 1)$ with the first column containing vector 1 and k is the number of explanatory variables.

$\boldsymbol{\hat{\beta}_{LMS}}$: vector regression coefficient size $(k + 1) \times 1$.

(3) MM estimator algorithm

The MM estimator method is obtained through two stages. First, calculate scale estimate or standard deviation ($\hat{\sigma}_s$) using the estimator method based on the following steps:

a. Estimating regression coefficients on data using the OLS method
b. Calculates residuals value $e_i = y_i - \hat{y}_i$
c. Calculates $\hat{\sigma}_s$ using equation (10) below:

$$\hat{\sigma}_s = \begin{cases} \frac{median\ |e_i - median\ (e)|}{0.6745} & ; q = 1 \\ \sqrt{\frac{1}{nK}\sum_{i=1}^{n} w_{i_S} e_i^2} & ; q > 1. \end{cases} \tag{10}$$

with $K = 0.199$ and q is an iteration.

d. Calculate the estimator weighting (w_{i_S})using equation (11) as follows:

$$w_{i_S} = \begin{cases} \begin{cases} \left[1-\left(\frac{u_i}{c}\right)^2\right]^2 & ,|u_i| \leq c; q = 1 \\ 0 & ,|u_i| > c; \end{cases} \\ \frac{\rho(u_i)}{u_i^2} \qquad\qquad q > 1 \end{cases} \tag{11}$$

The function $\rho(u_i)$ if using the Tukey weighting function is:

$$\rho(u_i)_{Tukey} = \begin{cases} \frac{c^2}{6}\left(1-\left(\frac{u_i}{c}\right)^2\right)^3 & ,|u_i| \leq c \\ \frac{1}{6}c^2 & ,|u_i| > c \end{cases} \tag{12}$$

with $u_i = \frac{e_i}{\hat{\sigma}_S}$ and c = 4.685.

e. Calculating $\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}}$ using weighted OLS with weight w_{i_S}.

f. Repeating stage b-e until the convergent $\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}}$value is obtained.

g. After obtaining the value of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}}$which converges then calculates $\hat{\sigma}_S$ in the last iteration to then be used as a scale estimate in the calculation of the next MM estimator.

Second, predict the regression parameters by doing iteratively reweighted least squares (IRLS). The calculation steps are as follows:

a. Calculating the value of $u_i = \frac{e_i}{\hat{\sigma}_S}$, the value of $\hat{\sigma}_S$ is obtained from the stage (g).

b. Calculates the weighting value of the MM method ($w_{i_{MM}}$) using equation (13):

$$w_{i_{MM}} = \begin{cases} \left[1-\left(\frac{u_i}{4.685}\right)^2\right]^2 & ,|u_i| \leq 4{,}685; \\ 0 & ,|u_i| > 4{,}685. \end{cases} \tag{13}$$

c. Calculating $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$uses the least weighted square with weight $w_{i_{MM}}$ with equation (14)

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}} = (\boldsymbol{X}'\boldsymbol{W}^{\boldsymbol{MM}}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{W}^{\boldsymbol{MM}}\boldsymbol{y}) \tag{14}$$

d. Repeating stage a-c until a convergent $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$value is obtained.

The MM estimator algorithm process is carried out using the help of the robustbase R software package.

6. Check the diagnostic model of each method in step 5 using the following test:
   (1) The normal assumption of the side is tested using the Kolmogorov-Smirnov test,
   (2) Assume homogeneity of residual using Breusch-Pagan.
   (3) The assumption of free (non-autocorrelated) freedom was tested using the Runs-test.
7. Repeat steps 1 to 6 as many as $r = 1000$ replications.
8. Comparing the average value of parameter estimation bias, RMSE, and R^2. See equation (15), (16), (17) respectively.

$$average\ of\ bias(\hat{\beta}_j) = \frac{1}{r}\sum_{l=1}^{r}\left|\beta_j - \hat{\beta}_{lj}\right|, j = 0,1 \qquad (15)$$

$$average\ of\ RMSE = \frac{1}{r}\sum_{l=1}^{r}\sqrt{\left(\sum_{i=1}^{n}\frac{(y_{il}-\hat{y}_{il})^2}{n-p}\right)} \qquad (16)$$

$$average\ of\ R^2 = \frac{1}{r}\sum_{l=1}^{r}\left(1 - \frac{\sum_{i=1}^{n}(y_{il}-\hat{y}_{il})^2}{\sum_{i=1}^{n}(y_{il}-\overline{y_l})^2}\right) \qquad (17)$$

with,

n : number of observations
y_{il} : respons of the i-data and l-replication
$\overline{y_l}$: the average of the response data in l-replication
$\hat{y}_{il}$: predicted response of the i-data on the l-replication
p : number of parameters
r : number of repetitions
$\hat{\beta}_{lj}$: an estimation of the j-parameter in the l-replication, $l = 1,2,\dots,r$

9. Summarizes the results of evaluation values on each combination of data size, type of outlier, outlier percentage, and estimation method used.

3.2 Actual Data

The steps to analyze the actual data of rice production in 2017 are as follows:

1. Plot data collection on the use of organic fertilizer (thousand tons) and data on the amount of rice production (million tons).
2. Estimating the regression model with the OLS method and calculating the residuals value.
3. Identify vertical outliers through plots of standardized residuals. Vertical outlier outliers are indicated by the standardized values of residuals of more than 2 or 2.5 relatives to the conditions of residuals scattering.
4. Identify outliers of good leverage points through plots of diagonal matrix hat values. Data that has a matrix hat diagonal value greater than twice the average overall diagonal matrix hat element are outliers of good leverage points.
5. Identify outliers of bad leverage points. Data identified as vertical outliers and also outliers for good leverage points are data outliers for bad leverage points.
6. Estimating the regression model with the OLS, LMS, and MM methods for data containing these outliers.
7. Compare the RMSE and R^2 values of the OLS, LMS, and MM methods.
8. Determine the best guess of the model, namely the alleged model with the smallest RMSE value and highest R^2.

4 Results and Discussion

4.1 Simulation Study

Regression analysis simulation using OLS, LMS, and MM methods on data containing outliers obtained the following results:

1. The average estimation bias value of the parameter β_0
   a. Vertical outlier

Table 1 shows the average value of the $\hat{\beta}_0$ bias in each percentage of vertical outliers and for each data size. The average value of bias $\hat{\beta}_0$ when the data does not contain outlier produces the smallest average value of $\hat{\beta}_0$ bias lies in the OLS method, followed by the MM method, then the LMS method. This applies to the overall size of the data generated, both sizes 50, 200, and 1000. However, the average value of the bias $\hat{\beta}_0$ OLS method has significantly increased every percentage of vertical outliers increase in the data. On the other hand, the LMS and MM methods tend to maintain the average value of $\hat{\beta}_0$ bias at a small value. When viewed based on the size of the data used, every size of the data increases, the average value of the $\hat{\beta}_0$ bias of the three methods has decreased. Overall, the MM method has the smallest value of the $\hat{\beta}_0$ bias. The existence of vertical outliers greatly affects the estimation of the OLS method $\hat{\beta}_0$ parameter.

Table 1 The average value of the $\hat{\beta}_0$ bias on various data sizes and percentage of vertical outlier

Data sizes	Method	Percentage of outliers (%)					
		0	5	10	15	20	30
50	OLS	0.30	7.16	11.58	13.82	17.1	22.5
	LMS	0.72	0.68	0.67	0.65	0.65	0.62
	MM	0.31	0.32	0.33	0.33	0.34	0.36
200	OLS	0.15	4.45	7.11	9.96	12.7	17.9
	LMS	0.39	0.40	0.38	0.37	0.37	0.36
	MM	0.15	0.15	0.16	0.16	0.17	0.17
1000	OLS	0.06	3.16	6.10	9.09	12.1	18.2
	LMS	0.22	0.21	0.21	0.21	0.21	0.20
	MM	0.07	0.07	0.07	0.07	0.07	0.08

b. Good leverage point

The average value of $\hat{\beta}_0$ bias from the data containing outliers of good leverage points is presented in Table 2. Table 2 shows that the MM and OLS methods have an average value of $\hat{\beta}_0$ bias which tends to be the same and close to 0. The LMS method has the biggest an average $\hat{\beta}_0$ bias value compared to OLS and MM. However, every the data size and outlier percentage are increasing, the average $\hat{\beta}_0$ bias of the LMS method decreases. Overall, the three methods have good performance in the case of data containing outliers of good leverage points. This is indicated by the small value of $\hat{\beta}_0$ bias for the three methods.

Table 2 The average value of the $\hat{\beta}_0$ bias on various data sizes and percentage good leverage points outliers

at a siz	M et ho d	Percentage of outliers (%)

		0	5	10	15	20	30
	OLS	0.30	0.09	0.08	0.08	0.08	0.08
50	LMS	0.71	0.52	0.30	0.23	0.21	0.20
	MM	0.31	0.14	0.09	0.08	0.08	0.09
	OLS	0.14	0.04	0.04	0.04	0.04	0.04
200	LMS	0.40	0.15	0.12	0.11	0.11	0.12
	MM	0.15	0.04	0.04	0.04	0.04	0.04
	OLS	0.06	0.02	0.02	0.02	0.02	0.02
1000	LMS	0.22	0.07	0.06	0.06	0.07	0.07
	MM	0.06	0.02	0.02	0.02	0.02	0.02

c. Bad leverage point

The average value of $\hat{\beta}_0$ bias from the data containing outliers of bad leverage points for each data size is presented in Table 3. Table 3 shows that when the percentage of bad leverage is 0%, the average value of the smallest $\hat{\beta}_0$ bias lies in the OLS method, followed by the MM method and finally is the LMS method. However, when the data contains bad leverage points, the OLS method produces the highest average value of the $\hat{\beta}_0$ bias compared to the other two methods. Furthermore, every increase in the percentage of outliers in the data, the average value of the $\hat{\beta}_0$ OLS and LMS bias decreases, while the average value of the $\hat{\beta}_0$ MM bias method increases. The average value of the $\hat{\beta}_0$ bias method MM is greater than that of LMS when the data size is 200 with an outlier percentage of 30% and when the data size is 1000 with an outlier percentage of 15%, 20%, 30%. Overall, the increasing size of the data causes the average value of the $\hat{\beta}_0$ bias of the three methods to decrease. The increase or decrease in the average value of the $\hat{\beta}_0$ bias is not significant. This shows that the observation of outlier bad leverage points in the data does not have a significant effect on the average value of the $\hat{\beta}_0$ bias of the three methods.

Table 3 The average value of the $\hat{\beta}_0$ bias on various data sizes and percentage bad leverage points outliers

Data sizes	Method	Percentage of outliers (%)					
		0	5	10	15	20	30
	OLS	0.12	0.54	0.55	0.55	0.55	0.54
50	LMS	0.30	0.29	0.29	0.29	0.29	0.30
	MM	0.13	0.16	0.16	0.17	0.18	0.22
	OLS	0.06	0.53	0.54	0.54	0.54	0.53
200	LMS	0.18	0.18	0.17	0.18	0.18	0.19
	MM	0.06	0.08	0.10	0.13	0.16	0.21
1000	OLS	0.03	0.53	0.54	0.54	0.54	0.53
	LMS	0.10	0.10	0.10	0.10	0.10	0.12

	MM	0.03	0.05	0.10	0.16	0.19	0.21

2. The average value of the estimation bias parameter β_1

a. Vertical outlier

The average value of the bias parameter estimation β_1 presented in Table 4 shows that the OLS method has the smallest average bias value only when there is no outlier in the data or when the outlier percentage is 0%. Furthermore, when the data contains outliers, the average estimation bias value β_1 by OLS always increases and it is at the highest average bias value compared to the other two methods. The average estimation β_1 bias value LMS and MM method for all data sizes remain consistent at a relatively small value so that the LMS and MM method is well used for estimating the parameter β_1 on data containing vertical outliers.

Table 4 The average value of the β_1 bias on various data sizes and percentage of vertical outlier

Data sizes	Method	Percentage of outliers (%)					
		0	5	10	15	20	30
	OLS	0.06	1.36	2.06	2.36	2.78	3.21
50	LMS	0.14	0.13	0.13	0.13	0.13	0.12
	MM	0.06	0.06	0.06	0.06	0.07	0.07
	OLS	0.03	0.72	1.00	1.21	1.37	1.59
200	LMS	0.08	0.08	0.07	0.07	0.07	0.07
	MM	0.03	0.03	0.03	0.03	0.03	0.03
	OLS	0.01	0.33	0.45	0.54	0.59	0.68
1000	LMS	0.04	0.04	0.04	0.04	0.04	0.04
	MM	0.01	0.01	0.01	0.01	0.01	0.01

b. Good leverage point

Observation of the outlier type of good leverage points contained in the data set does not affect the performance of the OLS method in estimating parameters β_1. This is indicated by the average value of the estimation bias parameter β_1 OLS method which is always smaller than the other two methods, even as the outlier percentage increase the estimated bias value β_1 OLS decreases closer to the value of 0. A decrease in the bias value of the estimated β_1 also occurs in LMS and MM along with the increase in outliers percentage with the average estimated bias value of β_1 by MM is always smaller than the LMS method. Table 5 shows that the three methods have an estimated bias value of β_1 close to 0. The presence of outlier good leverage points does not have a negative influence on the estimation of β_1 by the three methods.

Table 5 The average value of the $\hat{\beta}_1$ bias on various data sizes and percentage good leverage points outliers

Data sizes	Method	Percentage of outliers (%)					
		0	5	10	15	20	30
	OLS	0.06	0.01	0.01	0.01	0.01	0.00
50	LMS	0.14	0.09	0.04	0.02	0.01	0.01
	MM	0.06	0.02	0.01	0.01	0.01	0.01
	OLS	0.03	0.00	0.00	0.00	0.00	0.00
200	LMS	0.08	0.02	0.01	0.01	0.01	0.01
	MM	0.03	0.01	0.00	0.00	0.00	0.00
	OLS	0.01	0.00	0.00	0.00	0.00	0.00
1000	LMS	0.04	0.01	0.00	0.00	0.00	0.00
	MM	0.01	0.00	0.00	0.00	0.00	0.00

c. Bad leverage point

The average value of $\hat{\beta}_1$ bias from data containing outliers bad leverage points in Table 6 shows that when the percentage of outliers is 0%, for all data sizes the average value of the smallest $\hat{\beta}_1$ bias lies in the OLS method, followed by the MM method, then the method LMS. Furthermore, every increase in the percentage of outliers in the data, the average value of the $\hat{\beta}_1$ bias of the three methods is increasing, with the highest average bias value owned by the OLS method for all data sizes. The increase in the average value of the $\hat{\beta}_1$ bias MM method is greater than the increase in the average value of the $\hat{\beta}_1$ bias the LMS method. This is indicated by the average value of $\hat{\beta}_1$ bias MM which is higher than the average bias $\hat{\beta}_1$ LMS on certain outliers percentage and data sizes. Both the LMS and MM method can maintain the average bias $\hat{\beta}_1$ which is consistent with values close to 0 for all data sizes and various outliers. Thus, the LMS and MM method is more robust to the presence of outliers of bad leverage points compared to the OLS method.

Table 6 The average value of the $\hat{\beta}_1$ bias on various data sizes and percentage bad leverage points outliers

Data sizes	Method	Percentage of outliers (%)					
		0	5	10	15	20	30
	OLS	0.12	1.21	1.22	1.23	1.23	1.24
50	LMS	0.27	0.27	0.28	0.28	0.29	0.32
	MM	0.12	0.21	0.23	0.26	0.30	0.41
	OLS	0.05	1.19	1.21	1.23	1.23	1.23
200	LMS	0.16	0.16	0.17	0.18	0.19	0.23
	MM	0.06	0.11	0.17	0.25	0.33	0.45

	OLS	0.03	1.19	1.22	1.23	1.23	1.24
1000	LMS	0.09	0.09	0.09	0.10	0.12	0.18
	MM	0.03	0.09	0.22	0.35	0.42	0.47

3. Average Determination Coefficient Value

a. Vertical outlier

The average value of R^2 by the three methods namely OLS, LMS and MM are presented in Table 7. OLS has a high average value of R^2 only when there is no outlier in the data. Then when the data contains outliers, the average value of R^2 by the OLS has decreased. This shows that even the slightest outlier vertical outlier has a bad effect on the coefficient of determination or R^2. The average value of the LMS method for all data sizes is at the highest value compared to OLS and MM, which is close to 1. The average value of the MM method R^2 is also consistently at a value close to 1 for all data sizes.

Table 7 The average value of the R^2 on various data sizes and percentage of vertical outlier

Data sizes	Method	Percentage of outliers (%)					
		0	5	10	15	20	30
	OLS	0.94	0.04	0.03	0.03	0.03	0.02
50	LMS	1.00	1.00	1.00	0.99	0.99	0.98
	MM	0.95	0.95	0.95	0.95	0.95	0.94
	OLS	0.94	0.03	0.02	0.01	0.01	0.01
200	LMS	1.00	0.99	0.99	0.99	0.99	0.98
	MM	0.95	0.95	0.95	0.95	0.95	0.94
	OLS	0.94	0.02	0.01	0.01	0.01	0.01
1000	LMS	0.99	0.99	0.99	0.98	0.98	0.97
	MM	0.95	0.95	0.95	0.95	0.95	0.94

b. Good leverage point

Table 8 shows that the OLS, LMS, and MM methods on data containing outliers of good leverage points have an average value R^2 which is very high, which is above 90% for all data sizes. Having a good leverage point does not worsen the average value of R^2 produced by each method. The higher the outlier percentage will increase the average value of R^2 for the three methods. The average value of the R^2 by LMS method is at the highest position, while the MM method tends to have the same R^2 value as the OLS method.

Table 8 The average value of the R^2 on various data sizes and percentage good leverage points outliers

at a siz	M et ho d	Percentage of outliers (%)

		0	5	10	15	20	30
	OLS	0.94	1.00	1.00	1.00	1.00	1.00
50	LMS	1.00	1.00	1.00	1.00	1.00	1.00
	MM	0.94	0.99	1.00	1.00	1.00	1.00
	OLS	0.94	1.00	1.00	1.00	1.00	1.00
200	LMS	1.00	1.00	1.00	1.00	1.00	1.00
	MM	0.94	1.00	1.00	1.00	1.00	1.00
	OLS	0.94	1.00	1.00	1.00	1.00	1.00
1000	LMS	0.99	1.00	1.00	1.00	1.00	1.00
	MM	0.94	1.00	1.00	1.00	1.00	1.00

c. Bad leverage point

Table 9 presents the average value of R^2 on various data sizes containing outliers of bad leverage points. When data containing 0% bad leverage points, the average R^2 value for all three methods for all data sizes is in the range of values greater than 75%. However, when the data begins to contain bad leverage points outliers, the average value of R^2 by OLS method decreases to a range of 40%. Conversely, the average value of the R^2 methods of LMS and MM has increased to reach a value close to 100%. The average value of the LMS R^2 method is at the highest value compared to the other two methods for the overall data size. The presence of outliers of bad leverage points in the data greatly affects the performance of the OLS method but not the LMS and MM methods.

Table 9 The average value of the R^2 on various data sizes and percentage bad leverage points outliers

		Percentage of outliers (%)					
Data sizes	Method	0	5	10	15	20	30
	OLS	0.80	0.47	0.46	0.45	0.44	0.40
50	LMS	0.85	0.98	0.98	0.99	0.99	0.99
	MM	0.83	0.91	0.93	0.94	0.96	0.98
	OLS	0.80	0.48	0.47	0.45	0.44	0.40
200	LMS	0.82	0.97	0.98	0.99	0.99	0.99
	MM	0.83	0.90	0.94	0.96	0.98	0.98
	OLS	0.80	0.48	0.46	0.45	0.44	0.40
1000	LMS	0.82	0.97	0.98	0.99	0.99	0.99
	MM	0.83	0.91	0.96	0.97	0.98	0.98

4. Average RMSE Value

a. Vertical outlier

The average RMSE value of the OLS, LMS, and MM methods in Table 10 shows that the OLS method has a small RMSE value approaching the value of 0 only when there is no outlier in the data. Furthermore, when the data is contaminated with vertical outlier outliers, the average RMSE value by OLS has increased very high, while the average RMSE value by LMS and MM is at a relatively small value.

Table 10 The average value of RMSE on various data sizes and percentage of vertical outlier

Data sizes	Method	Percentage of outliers (%)					
		0	5	10	15	20	30
	OLS	0.50	11.89	18.19	21.05	24.26	27.78
50	LMS	0.42	0.44	0.46	0.48	0.49	0.51
	MM	0.49	0.52	0.58	0.62	0.70	0.89
	OLS	0.50	13.12	18.05	21.48	24.07	27.57
200	LMS	0.47	0.48	0.49	0.50	0.50	0.51
	MM	0.50	0.53	0.58	0.63	0.69	0.87
	OLS	0.50	13.09	18.02	21.44	24.02	27.51
1000	LMS	0.48	0.48	0.49	0.50	0.50	0.50
	MM	0.50	0.53	0.57	0.62	0.68	0.86

b. Good leverage point

Table 11 shows that the existence outliers of the type of good leverage point is not too bad for the performance of each method. The average RMSE value of the three methods remains at a value close to 0. There is no increase in the average RMSE value which is very high for each method. The average RMSE value for the MM method is almost the same as the average RMSE value of the OLS method, while the average RMSE value of the LMS method is at the lowest value compared to the OLS and MM methods.

Table 11 The average value of RMSE on various data sizes and percentage good leverage points outliers

Data sizes	Method	Percentage of outliers (%)					
		0	5	10	15	20	30
	OLS	0.50	0.50	0.50	0.50	0.50	0.50
50	LMS	0.43	0.43	0.43	0.43	0.43	0.43
	MM	0.49	0.49	0.49	0.49	0.49	0.49
	OLS	0.50	0.50	0.50	0.50	0.50	0.50
200	LMS	0.47	0.47	0.47	0.47	0.47	0.47
	MM	0.50	0.50	0.50	0.50	0.50	0.50
1000	OLS	0.50	0.50	0.50	0.50	0.50	0.50

	LMS	0.48	0.48	0.48	0.48	0.48	0.48
	MM	0.50	0.50	0.50	0.50	0.50	0.50

c. Bad leverage point

The average RMSE value of the OLS, LMS, and MM methods for each data size and the percentage of bad leverage points is presented in Table 12. When the outlier percentage is 0%, the average RMSE value of the three methods is quite small for all data sizes. Furthermore, as the percentage of bad leverage points increases in the data, the average RMSE value of each method increases. The highest increase was experienced by the OLS method. The average increase in the RMSE value of the LMS and MM method is not significant, with the smallest average RMSE value being the average RMSE value generated by the LMS method.

Table 12 The average value of RMSE on various data sizes and percentage bad leverage points outliers

Data sizes	Method	Percentage of outliers (%)					
		0	5	10	15	20	30
50	OLS	1.00	4.66	5.64	6.45	7.17	9.01
	LMS	0.56	0.67	0.71	0.76	0.81	0.93
	MM	0.86	0.87	0.87	0.88	0.90	1.07
200	OLS	1.00	3.80	5.14	6.20	7.11	8.63
	LMS	0.81	0.86	0.89	0.92	0.94	1.03
	MM	0.87	0.87	0.88	0.91	0.96	1.12
1000	OLS	1.00	3.78	5.12	6.17	7.07	8.59
	LMS	0.89	0.91	0.93	0.94	0.94	1.10
	MM	0.87	0.87	0.90	0.95	1.00	1.14

5. Interaction Effect of Factor Used

Scenarios of variation of outliers percentage affect the evaluation value. The increasing percentage of vertical outliers and bad leverage points will cause the parameters and RMSE bias values of OLS and MM methods to increase with the highest increase experienced by OLS. Increasing the percentage of vertical outliers and bad leverage points does not have a major effect on the results of parameter estimation bias by the LMS method. This is indicated by the bias value of the parameter estimation LMS method decreases every increase in the percentage of vertical outliers and tends to be stable with increasing outliers of bad leverage points. However, for the type of outreach to good leverage points, the increase in the percentage of outliers in the refractive value of parameter estimation by OLS, LMS, and MM methods is decreasing. Increasing the size of the data will cause the parameter estimation bias to decrease. This applies to all types of outlier and all levels of outlier.

Likewise, the scenario of variation of outliers percentage affects the coefficient of determination. The increasing percentage of vertical outliers and bad leverage points will cause the coefficient of determination of the OLS method to decrease very high. Increasing the percentage of vertical outliers and bad leveerage points did not significantly influence the results of the coefficient of determination by the LMS and MM methods. This is shown by the coefficient of determination LMS and MM methods which tend to be stable at high values. However, the increase in the percentage of good leverage point cause coefficient of determination by OLS, LMS, and MM methods is increasing. The increase in data size does not have a big effect on the resulting coefficient of determination.

The increasing percentage of outlier vertical outliers and bad leverage points will cause the RMSE of OLS, LMS and MM methods to increase with the highest increase experienced by OLS. However, for the type of outlier the good leverage point, the increase in the percentage of outliers does not affect the RMSE value by the OLS, LMS, or MM methods. Increasing the data size will cause the RMSE values of all methods to increase. However, the increase that occurred was not significant.

4.2 Actual Data Study

Data. The data scatter plot from data used is shown in Figure 2. The data distribution plot in Figure 2 shows exploratively that there are several data located far from other data sets. These data have the potential to become outliers. Therefore, it is necessary to look at the standardized residual value to identify vertical outliers, while the good leverage points are identified by using a diagonal matrix hat value plot.

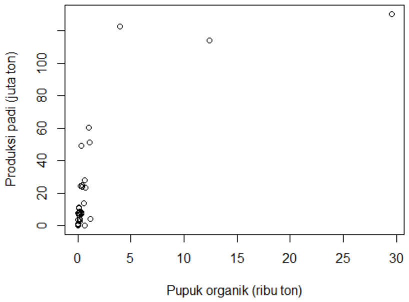

Fig. 2. Data distribution plot

Outlier Check. The residual standardized value based on the plot in Figure 3 is at the interval of the values of 0 to 4. The majority of the data has standardized residuals that are spread out in the value 0-2. However, there is 1 observation that has a value above 2.5, which is the 12th data so that this data is a vertical outlier data. On the other hand, based on the plot of diagonal matrix diagonal values in Figure 4, there are 2 data which have a diagonal matrix hat value greater than 0.12 (twice the average diagonal matrix hat), which is the 13th and 15th data. Both data are data outlier good leverage point. There is no data identified as vertical outlier outliers as well as good leverage points, so there is no bad leverage point outlier in the actual data used.

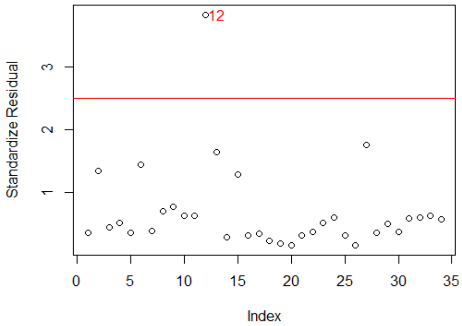

Fig. 3. Standaridized residuals plot of actual data

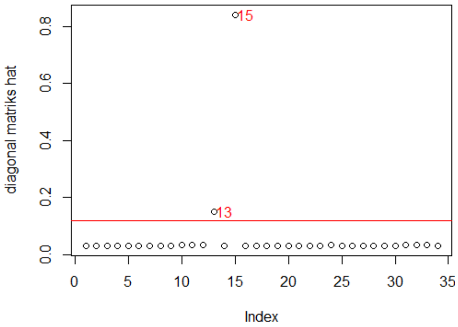

Fig. 4. Hat diagonal matrix plot of actual data

Regression Analysis. The plot of the regression equation for the OLS, LMS, and MM methods is presented in Figure 5. The black, blue, and red lines in Figure 5 are respectively the regression equation lines using the OLS, LMS, and MM methods. Figure 5 shows that the LMS and MM regression lines tend to be almost the same, while the OLS regression line is much different from LMS and MM. The OLS regression line is attracted by the existence of outlier observations. This is supported by the estimation value of the third parameter method. Table 13 presents the regression coefficient values or parameter estimates generated by the three methods. The regression coefficients produced by the LMS and MM methods tend to have values that are not much different, while the regression coefficients by OLS have different characteristics. Estimation value of the parameter β_0 of the LMS and MM method is smaller than the estimated value of the parameter β_1. Conversely, the estimation value of the parameter β_0 by OLS is greater than the estimated value of the parameter β_1.

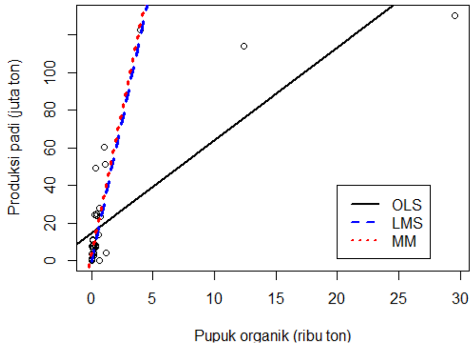

Fig. 5. Regression line of OLS, LMS, and MM method

Table 13 Regression coefficient

Method	$\hat{\beta}_0$	$\hat{\beta}_1$
OLS	14.62	4.93
LMS	-1.01	30.42
MM	2.53	30.87

Regression Assumptions Check. The regression assumptions p-values are presented in Table 14. The p-values of KS-test and Breusch-Pagan by OLS are 0.0124 and 0.0195 respectively, less than the 0.05 significance level so that the assumptions of normality and homogeneity in the range of predictions using OLS are not met. The assumption of randomness of residuals of estimation using the OLS method is fulfilled as indicated by the Runs-test p-value of more than 0.05, which is equal to 0.1278. The p-value of the KS-test, BP-test, and Runs-test for the LMS and MM robust method were at a value of more than 0.05. This means that the assumptions of normality, homogeneity, and randomness of residuals of LMS and MM methods are fulfilled.

Table 14 The actual data regression assumptions p-values

Method	KS-*test*	BP-*test*	*Runs-test*
OLS	**0.0124**	**0.0195**	0.1278
LMS	0.2791	0.3959	0.4860
MM	0.4421	0.4017	0.1635

Criteria for the goodness of the model. The RMSE and R^2 OLS, LMS, and MM values in estimating the regression model of rice production containing outliers are presented in Table 15. The method that produces the largest to smallest RMSE value in a row is the OLS method of 23.15, the MM method is equal to 6.78, and the LMS method is 4.44. The highest R^2 value is generated by the LMS method, which is equal to 98%, then followed by the MM method which produces a value of R^2 of 96%, while OLS produces the lowest R^2 value of 58%.

A good method used in regression analysis is a method that produces a small RMSE value and conversely produces the largest R^2value. Table 15 shows that the LMS method in this study produced the smallest RMSE value of 4.44 and the highest R^2 value of 98%. Therefore, the LMS method is the best method in simple regression analysis to determine the effect of data on the amount of use of organic fertilizer on total rice production in the provinces of Indonesia in 2017. The linear regression simulation results show that in small size data, size 50 observations which contains outliers of 5% to 10%, the LMS method produces the smallest RMSE and the largest R^2 value. This means that the conclusions obtained from estimating parameters for the actual data of rice production in 2017 are in line with the results of estimating regression parameters through simulation. The best regression line equations formed by the LMS method are as follows:

$$\hat{y}_i = -1.0100 + 30.4200x_i$$

The estimation of β_0 ($\hat{\beta}_0$) is negative, which is equal to -1.0100. The interpretation of the $\hat{\beta}_0$ value is that when no organic fertilizer is used, the estimated average production of rice produced will decrease by 1.0100 million tons. On the other hand, the value of $\hat{\beta}_1$ is positive, which is equal to 30.4200. This value illustrates that every addition of one thousand tons of use of the amount of organic fertilizer, rice production will increase by 30.4200 million tons.

Tabel 15 RMSE dan R^2 value of actual data regression model

Method	RMSE	R^2
OLS	23.1500	0.5800
LMS	4.4400	0.9800
MM	6.7800	0.9600

5 Conclusion and Suggestion

5.1 Conclusion

The simulation results of estimating data regression parameters containing outliers show that the LMS and MM method is a good method used when data contains vertical outlier outliers, good leverage points, and bad leverage points. This is based on the average value of the parameter estimation bias and the average RMSE value is quite small and the average value of R^2 is high. OLS method is only good to use when the type of outliers contained in the data is a good leverage point. The presence of outlier vertical outliers and bad leverage points affect the estimation results and normal alignment assumptions by the OLS method. The normal assumption of the OLS method is not fulfilled when there are vertical outliers and bad leverage points in the data.

The application of OLS and the LMS and MM robust regression method on actual data on total rice production in Indonesia in 2017 resulted in the conclusion that the robust LMS method was the best method. This is indicated by the lowest RMSE value, which is equal to 4.44 and the highest R^2 value, which reaches 98%. The actual data used was detected to contain vertical outlier outliers and good leverage points. Therefore, the conclusions of the best methods produced in applying the actual data of rice production in 2017 are in line with the results obtained in the simulation process. The linear regression simulation results show that in small size data, that is 50 observations with a 5% to 10% outlier content, the LMS method produces the smallest RMSE average and the largest R^2 value average.

5.2 Suggestion

Robust regression methods for outliers have been developed. This study uses only two types of robust methods in estimating data regression parameters containing outliers, namely the LMS and MM methods. Subsequent research is expected to be able to use a comparison of other robust regression so that it can compare more robust regression methods. In addition, the simulation process in this study did not notice the effect of the distribution of data and the variety of data on the estimation results. Therefore, it is expected that the next study also evaluates the effect of distribution and a variety of data on the performance of the regression model estimation method.

References

[1] D'Urso P, Massari R. 2013. Weighted least squares and least median squares estimation for the fuzzy linear regression analysis. *Metron.* 71:279-306

[2] Atilgan YK, Gunay S. 2011. Least median of squares solution of multiple linear regression models through the origin. *Communication in Statistics—Theory and Methods.* 40:4125-4137

[3] Montgomery DC, Peck EA, Vining GG. 2012. *Introduction to Linear Regession Analysis*. 5th Ed. Wiley

[4] Chen C, SAS Institute Inc, Cary NC. 2002. Robust regression and outlier detection with the robustreg procedure. *SUGI Proceedings.* 265-270

[5] Almetwally EM, Almongy HM. 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): 1-9

[6] Oktarinanda A. 2014. Perbandingan efisiensi metode least trimmed square (LTS) dan metode least median square (LMS) dalam penduga parameter regresi robust. *Jurnal Statistik*. 2(3):177-180

[7] Rousseeuw PJ dan Hubert M. 1997. Recent developments in PROGRESS. *L_1-Statistical Procedure and Related Topics.* 31:201-214