Robust multi-stage method (MM) and least median square (LMS) evaluation on handling outlier for multiple regression

Khusnul Khotimah, Kusman Sadik*, and Anang Kurnia
Department of Statistics, IPB University, Indonesia

*E-mail : kusmans@apps.ipb.ac.id

Abstract. Outliers are observation with extreme values that are very different from most other data. The least squares method (OLS) is a method that is very sensitive to outliers because it can cause the modeling accuracy decrease. Throwing out outliers with the aim of improving the suitability of the regression equation cannot be done carelessly because it will provide imprecise estimation precision. This study examines the performance of robust method of the least median squares (LMS) and the multi-stage method (MM) compared to OLS in a regression analysis of data which contains outliers. Data analysis was performed on simulation data and oil palm production data. Based on the average parameter estimate bias value, MM method has the best performance in each scenario condition of data size and outlier percentage, while based on the average root mean squares error (RMSE), LMS has better performance than MM when the data size is 25. Analysis of Indonesian oil palm production data in 2018 which data size 25 and contains 44% outliers resulted the conclusion that LMS method produced smallest RMSE and highest $R^2$, namely 38.81 and 99.78%, respectively. MM method is in the second best position, while OLS produces largest RMSE and highest $R^2$.

1. Introduction

1.1 Research Background

Linear regression method is a statistical method used to evaluate the linear relationship between quantitative response variables and one or more quantitative explanatory variables. The least squares method (OLS) is the most popular approach used in estimating linear regression model parameters [1]. However, this least squares method has limitations, which is strongly influenced by outlier data [2]. Outliers are an observation with extreme values that are very different from most other data [3]. The existence of these outlier observations can interfere with the normal assumptions of residuals and can reduce the power of the prediction method [4].

It is only possible to remove outliers from the data set if the cause of this value is known, such as measurement or analysis errors, errors in recording data, 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 influence on the model that is formed. Therefore, discarding outliers with the aim of improving the suitability of the regression equation cannot be done carelessly because it will provide imprecise estimation precision [5].

The solution to the problem of outlier in the regression analysis process can be done using the robust regression method. Robust regression is a regression method used when the data contain outliers observed values [5]. The examination of the least median square (LMS) and multi-stage method (MM) robust methods in simple regression analysis through simulation and their application
in 2017 rice production data produced the conclusion that the LMS method was the best method based on low RMSE values and the high coefficient of determination [6]. The robust MM method is a combination of the robust M method and the robust S method which has a high breakdown point [7]. LMS also has a high high breakdown point because the basic principle of this method is based on a median value that is more robust to outliers than the average value [8].

The two methods of handling outliers will be reused in this study to be applied to multiple regression, namely regression with more than one explanatory variable. The selection of the best method will be based on the estimated bias values of the regression parameters and the root mean square error (RMSE). Therefore, the researcher raised the topic of "Robust multi-stage method (MM) and least median square (LMS) evaluation on handling outlier for multiple regression". Comparisons are made through simulating data with various data sizes, types of outliers, and percentage of outliers, then applying them to agricultural data, namely oil palm production data (in thousand tons) in 2018.

1.2. Research Objective
The objectives of this study are as follows:
1. Compare and determine the best regression method for various types of outliers based on the parameter estimation bias values and RMSE through a simulation process on various data sizes, types of outliers, and outlier contamination percentage.
2. Determine the best method for estimating the regression model for oil palm production data (in thousand tons) in 2018 which contains outliers.

2. Research Methods
2.1. Data Description
2.1.1. Simulation Data. This research involves generating data that contains outliers for the purpose of robust multiple linear regression simulation. The simulation process is carried out using data sizes \((n)\) of 25 data for small sizes, 50 data for medium sizes, and 300 data for large sizes. There are 3 types of outliers that are generated, namely vertical outliers, good leverage points, and bad leverage points. The percentage of outliers \((m)\) used is 0%, 10%, 20%, and 30% for each type of outlier. The number of explanatory variables used in this study amounted to four explanatory variables. The data generated from the combination of scenarios will then be used in the parameter estimation process using the ordinary least squares (OLS), LMS, and MM methods. The parameters \(\beta_0, \beta_1, \beta_2, \beta_3, \text{ and } \beta_4\) used in this study are equal to 3 for all parameters.

2.1.2. Actual Data. This study also uses actual data as an application to compare the OLS, MM and LMS methods. The actual data used is oil palm production data (in thousand tonnes) as a response variable. The explanatory variables used are data on the area of state-owned oil palm plantations in thousand hectares \((X_1)\), data on the area of private plantations in thousand hectares \((X_2)\), data on the area of smallholder plantations in thousand hectares \((X_3)\), and data on the area of damage in thousand hectares \((X_4)\). The actual data used were obtained from the Indonesian Palm Oil Statistics Catalog by the Central Statistics Agency in 2018 [9]. The data will be analyzed using multiple linear regression with OLS, LMS, and MM. The data consists of 25 observations which are oil palm producing provinces in Indonesia.

2.2. Procedures of Analysis
2.2.1. 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 \(\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 3\)
2. Generating vectors explanatory variable (\(x\)) with normal distribution as much as \(n\) data with average and variety as follows:
   i. \(X = [x_1 \ x_2 \ x_3] \sim MVN \left( \begin{bmatrix} 5 \\ 5 \\ 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right)\) for general data
   
   ii. \(X^* = [x_1^* \ x_2^* \ x_3^*] \sim MVN \left( \begin{bmatrix} 30 \\ 30 \\ 30 \\ 30 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right)\) for outlier data

3. Generates a vector of error values (\(\epsilon\)) as much as \(n\) data with the following details:
   i. \(\epsilon \sim N(0,1)\) for general data
   
   ii. \(\epsilon^* \sim N(30,1)\) and \(\epsilon^{**} \sim N(-20,1)\) for outlier data

4. Determine vector of \(y\) based on below steps:
   i. **Vertical Outlier**
      
      \[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + r, \]

      with \(r\) being a vector of error components obtained from a combination of sampling data \(\epsilon \sim N(0,1)\) as much \((1 - m) \times n\) and data \(\epsilon^* \sim N(30,1)\) as much as \(m \times n\).

   ii. **Good leverage point**

      \[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 s + \epsilon, \]

      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 \(m \times n\).

   iii. **Bad leverage point**

      \[ y = \beta_0 + \beta_1 x + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + t \]

      with \(t\) is a vector of error components where the first \(\frac{1}{2}m \times n\) data and the last \(\frac{1}{2}m \times n\) data sequentially obtained from sampling data \(\epsilon^* \sim N(30,1)\) and \(\epsilon^{**} \sim N(-20,1)\), the rest data is taked from sampling data \(\epsilon \sim N(0,1)\) as much \((1 - m) \times n\).

      Then, as much \(\frac{1}{2}m \times n\) first data and \(\frac{1}{2}m \times n\) last data of last explanation variable add by 20.

      Illustration of the appearance of the three types of outliers that will be generated can be seen in Figure 1 below [10]:

      ![Figure 1. Outliers types](image)

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 \(\beta\) parameter

\[ \hat{\beta} = (X'X)^{-1}(X'y) \]
(2) LMS method regression algorithm. The parameter estimation steps with the LMS method summarized in a PROGRESS algorithm by Rousseeuw and Hubert [10] on the MASS R software package are as follows:

a. Determine the value of \( g \)

\[
g = \left\lfloor \frac{n + k + 1}{2} \right\rfloor
\]

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 \( v = 1, 2, 3, ..., g \) and index \( d \) is the number of subsets formed, \( d = 1, 2, 3, ..., C_g^n \).

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 \( (\hat{\sigma}_{LMS}) \)

\[
\hat{\sigma}_{LMS} = 1.4826 \left( 1 + \frac{5}{(n - g)} \right) \sqrt{\frac{n}{M}}
\]

g. Calculate LMS estimator weights \( (w_{i,LMS}) \):

\[
w_{i,LMS} = \begin{cases} 1, & |e_i| \leq 2.5 \\ 0, \text{ lainnya} \end{cases}
\]

\( w_{i,LMS} \) is a diagonal element of the \( W^{LMS} \) matrix that has a size of \( n \times n \) and other elements worth 0.

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

h. Estimating regression parameters using a weighted regression method

\[
\hat{\beta}_{LMS} = \left( X'W^{LMS}X \right)^{-1} \left( X'W^{LMS}y \right)
\]

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.

\( \hat{\beta}_{LMS} \): 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 \( \hat{\sigma}_s \)

\[
\hat{\sigma}_s = \begin{cases} \text{median}|e_i - \text{median}(e)| & ; q = 1 \\ \frac{1}{nK} \sum_{i=1}^{n} w_{i,s} e_i^2 & ; q > 1.
\end{cases}
\]

with \( K = 0.199 \) and \( q \) is an iteration.
d. Calculate the estimator weighting \((w_{is})\) as follows:

\[
\begin{align*}
   w_{is} &= \begin{cases} 
   \left[1 - \left(\frac{u_i}{c}\right)^2\right]^2, & |u_i| \leq c; q = 1 \\ 
   0, & |u_i| > c; \\
   \frac{\rho(u_i)}{u_i^2}, & q > 1
   \end{cases}
\end{align*}
\]

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

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

with \(u_i = \frac{e_i}{\hat{\sigma}_s}\) and \(c = 4.685\).

e. Calculating \(\hat{\beta}_s\) using weighted OLS with weight \(w_{is}\).

f. Repeating stage b-e until the convergent \(\hat{\beta}_s\) value is obtained.

g. After obtaining the value of \(\hat{\beta}_s\) 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 \((w_{i,MM})\):

\[
\begin{align*}
   w_{i,MM} &= \begin{cases} 
   \left[1 - \left(\frac{u_i}{4.685}\right)^2\right]^2, & |u_i| \leq 4.685; \\
   0, & |u_i| > 4.685.
   \end{cases}
\end{align*}
\]

d. Calculating \(\hat{\beta}_{MM}\) uses the least weighted square with weight \(w_{i,MM}\)

\[
\hat{\beta}_{MM} = (X'W_{MM}X)^{-1}(X'W_{MM}y)
\]

e. Repeating stage a-c until a convergent \(\hat{\beta}_{MM}\) value is obtained.

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

6. Repeat steps 1 to 6 as many as \(r = 100\) replications.

7. Comparing the average value of parameter estimation bias and RMSE.

\[
\text{average of bias} (\hat{\beta}_j) = \frac{1}{r} \sum_{l=1}^{r} \left| \beta_j - \hat{\beta}_{ij} \right|, j = 1, 2, 3, 4
\]

\[
\text{average of RMSE} = \frac{1}{r} \sum_{l=1}^{r} \sqrt{\frac{\sum_{i=1}^{n} (y_{il} - \hat{y}_{il})^2}{n-p}}
\]

with,

- \(n\) : number of observations
- \(y_{il}\) : responses of the \(i\)-data and \(l\)-replication
- \(\hat{y}_{il}\) : predicted response of the \(i\)-data on the \(l\)-replication
- \(p\) : number of parameters
- \(r\) : number of repetitions
- \(\hat{\beta}_{ij}\) : an estimation of the \(j\)-parameter in the \(l\)-replication, \(l = 1,2,\ldots,r\)

8. Summarizes the results of evaluation values on each combination of data size, type of outlier, outlier percentage, and estimation method used.
2.2.2. *Actual Data*. The steps to analyze the actual data of oil palm production in 2018 are as follows:

1. Plot data collection.
2. Estimating the regression model with the OLS method and calculating the residuals value.
3. Identify the vertical outliers, good leverage points, and bad leverage points by calculating the value of the generalized hat matrix and generalized RStudent.
4. Estimating the regression model with the OLS, LMS, and MM methods for data containing these outliers.
5. Compare the RMSE and $R^2$ values of the OLS, LMS, and MM methods in stage 6.
6. Determine the best guess of the model, namely the alleged model with the smallest RMSE value and highest $R^2$.

3. *Results and Discussion*

3.1 *Simulation Result Evaluation*

The value of the criteria for the goodness of the method used to evaluate the performance of the method is the absolute relative bias of the parameter estimates of the explanatory variables and RMSE. These values are presented in the form of an average of 100 replications performed. A good method is a method that produces an average absolute relative bias in the estimated parameters and low RMSE.

1) Mean relative absolute bias of parameter estimates
   a. Data without outliers

| Table 1. Mean relative absolute bias values of estimated parameters of each method in each explanatory variable in the data without outliers |
|---------------------------------------------------------------|
| **Bias relatif $\hat{\beta}_1$** | **Bias relatif $\hat{\beta}_2$** | **Bias relatif $\hat{\beta}_3$** | **Bias relatif $\hat{\beta}_4$** |
| OLS | 0.0371 | 0.0392 | 0.0408 | 0.0375 |
| LMS | 0.0668 | 0.0636 | 0.0694 | 0.0639 |
| MM | 0.0383 | 0.0415 | 0.0425 | 0.0384 |

The mean of the lowest absolute relative bias values for the data without outliers as shown in Table 1 is obtained at the values generated by the OLS method. This condition applies to all estimated parameters on the four explanatory variables, namely 0.0371, 0.392, 0.0408, and 0.0375, respectively. This shows that the OLS method is able to produce a good parameter estimate value in the data conditions without outliers. On the other hand, an inappropriate method used in estimating multiple regression parameters in the type 1 data condition is the LMS method. This statement is supported by the average absolute relative bias value of the parameter estimates generated by the LMS method which is the highest among other methods, namely, respectively, reaching values of 0.0668, 0.0636, 0.0694, and 0.0639. The difference in the size of the data used gives a difference in the resulting relative bias values. Table 2 shows that increasing the size of the data used decreases the resulting mean relative absolute bias. This applies to all methods used.

| Table 2. Mean relative absolute bias values of estimated parameters of each method in each data size used in the data without outliers |
|---------------------------------------------------------------|
| **Method** | **Data size** | 25 | 50 | 300 |
|------------------|--------------|---|---|---|
| OLS | 0.0603 | 0.0401 | 0.0161 |
| LMS | 0.1201 | 0.0589 | 0.0186 |
| MM | 0.0612 | 0.0423 | 0.0169 |
b. Data Contains Vertical Outlier

The mean absolute relative bias values of the estimated parameters $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$, and $\beta_4$ generated by each of the methods used have similar characteristics for each outlier percentage. The relative bias values are presented in Table 3. The method that produces the lowest mean absolute relative bias values is the MM method, followed by the LMS method with not much different values. On the other hand, the method that produces the highest average bias value is found in the OLS method. The average relative bias value of the estimated parameters by OLS increased as the percentage of outliers that contaminated the data increased. This condition applies to the four estimated parameters $\beta$ of the explanatory variable. Thus, based on the average absolute relative bias values presented in Table 3, it can be concluded that the LMS and MM methods are proven to be more appropriate to use than OLS for data analysis containing vertical outliers.

### Table 3. Mean relative absolute bias values of the estimated parameters of each explanatory variable on data contaminated with vertical outliers

| Method | $\beta_1$ 0.1 | $\beta_1$ 0.2 | $\beta_1$ 0.3 | $\beta_2$ 0.1 | $\beta_2$ 0.2 | $\beta_2$ 0.3 |
|--------|---------------|---------------|---------------|---------------|---------------|---------------|
| OLS    | 0.5405        | 0.7842        | 0.8747        | 0.5671        | 0.7853        | 0.8355        |
| LMS    | 0.0598        | 0.0540        | 0.0578        | 0.0600        | 0.0579        | 0.0543        |
| MM     | 0.0409        | 0.0443        | 0.0514        | 0.0393        | 0.0470        | 0.0465        |

Table 4 presents the mean relative absolute bias values for the estimated parameters for each percentage of good leverage point outliers. The resulting value characteristics differ from those when the data is contaminated with vertical outliers. OLS produces the highest average value when the data is contaminated with vertical outliers, but not when the data is contaminated with outliers of good leverage points. The average value produced by OLS is actually the lowest compared to LMS and MM. Therefore, OLS is still recommended for data that is contaminated with outliers of good leverage points.

### Table 4. Mean relative absolute bias values of parameter estimates for each explanatory variable on data contaminated with good leverage point outliers

| Method | $\beta_1$ 0.1 | $\beta_1$ 0.2 | $\beta_1$ 0.3 | $\beta_2$ 0.1 | $\beta_2$ 0.2 | $\beta_2$ 0.3 |
|--------|---------------|---------------|---------------|---------------|---------------|---------------|
| OLS    | 0.0369        | 0.0368        | 0.0386        | 0.0397        | 0.0392        | 0.0389        |
| LMS    | 0.0620        | 0.0712        | 0.0632        | 0.0644        | 0.0661        | 0.0605        |
| MM     | 0.0400        | 0.0399        | 0.0398        | 0.0417        | 0.0411        | 0.0405        |

Table 5 presents the mean relative absolute bias values of the estimated parameters in the contaminated data with bad leverage point outliers. Similarly, when the data is contaminated with vertical outliers, MM and LMS produce an average absolute relative bias
value of low parameter estimates when the data is contaminated with bad leverage point outliers. The mean relative bias of the estimated parameters was found in the MM method. Meanwhile, OLS produced the highest average bias value. This condition applies to each estimated parameter $\beta_0, \beta_1, \beta_2, \beta_3,$ and $\beta_4$.

**Table 5.** Average absolute relative bias values of the estimated parameters of each explanatory variable on data contaminated with bad leverage point outliers

| Method | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ |
|--------|-----------------|-----------------|-----------------|-----------------|
| OLS    | 1.1485 | 1.7927 | 2.2262 | 1.2769 | 1.7230 | 2.2287 |
| LMS    | 0.0594 | 0.0551 | 0.0496 | 0.0547 | 0.0550 | 0.0603 |
| MM     | 0.0408 | 0.0446 | 0.0452 | 0.0378 | 0.0458 | 0.0551 |

The mean absolute relative bias of the overall parameter estimates for each data sizes.

**Table 6.** Mean relative absolute bias values of parameter estimates in each data size and type of outlier

| Method | Vertical outlier | Good leverage point | Bad leverage point |
|--------|-----------------|---------------------|-------------------|
| Data size | 25 | 50 | 300 | 25 | 50 | 300 | 25 | 50 | 300 |
| OLS     | 1.1592 | 0.7588 | 0.3025 | 0.0474 | 0.0305 | 0.0124 | 2.4680 | 1.6639 | 0.7594 |
| LMS     | 0.1032 | 0.0515 | 0.0179 | 0.0950 | 0.0461 | 0.0142 | 0.0101 | 0.0480 | 0.0179 |
| MM      | 0.0718 | 0.0467 | 0.0176 | 0.0507 | 0.0322 | 0.0129 | 0.0714 | 0.0442 | 0.0176 |

The effect of data size on the average relative bias value for each type of outlier is presented in Table 6. An increase in the size of the data used will reduce the average absolute relative bias value of the parameter estimates generated by the overall method in each type of outlier. Thus, it can be said that a large data size will increase the accuracy of parameter estimation.

2) The average of RMSE

a. Data without outliers

Table 7 presents the average RMSE value by each method for each data size used obtained from 100 replications. The three methods used resulted in a relatively low RMSE average with the lowest value obtained in the MM method, followed by LMS, and OLS. However, for data without outliers the most appropriate method is OLS.

**Table 7.** Average RMSE value of each method in each data size used in the data without outliers

| Method | Data size | 25 | 50 | 300 |
|--------|-----------|----|----|-----|
| OLS    |           | 0.9930 | 0.9877 | 0.9945 |
| LMS    |           | 0.7600 | 0.8821 | 0.9557 |
| MM     |           | 0.7855 | 0.8209 | 0.8553 |

b. Data Contains outlier vertical outliers

Table 8 presents the average RMSE value of all methods for each data size and the percentage of vertical outliers. LMS and MM proved to be appropriate for data analysis containing vertical outliers. This is indicated by the low RMSE value of the two methods. LMS produces the
lowest RMSE in each percentage of outliers at data size 25, while MM produces the lowest RMSE when data sizes are 50 and 300. OLS produces high average RMSE values, reaching more than 23 when the percentage of vertical outliers in the data is 0.3. Overall, an increase in the percentage of outliers and data size will increase the average RMSE value generated by each method.

Table 8. Average RMSE for each method in each data size and the percentage of vertical outliers

| Method | 25     | 50     | 300    |
|--------|--------|--------|--------|
|        | 0.1    | 0.2    | 0.3    | 0.1    | 0.2    | 0.3    | 0.1    | 0.2    | 0.3    |
| OLS    | 13.8329| 20.3072| 23.1510| 15.1615| 23.1361| 15.0247| 20.0535| 22.9852|
| LMS    | 0.5632 | 0.7628 | 0.8303 | 0.9542 | 0.9689 | 0.9230 | 0.9580 | 0.9256 | 0.8823 |
| MM     | 0.7887 | 0.7652 | 0.7186 | 0.8212 | 0.7987 | 0.7502 | 0.8497 | 0.8221 | 0.7920 |

c. Outliers of good leverage points

The outliers of good leverage points in the dataset used did not adversely affect the performance of each method. This is indicated by the average RMSE value for each method shown in Table 9, which remains at a relatively small value. The average RMSE value generated by each method will decrease when the percentage of good leverage points is 0.2 and go back up when the percentage is 0.3. This applies to each data size used.

Table 9. Average RMSE value for each method in each data size and the percentage of outliers of good leverage point

| Method | 25     | 50     | 300    |
|--------|--------|--------|--------|
|        | 0.1    | 0.2    | 0.3    | 0.1    | 0.2    | 0.3    | 0.1    | 0.2    | 0.3    |
| OLS    | 0.9896 | 0.9808 | 1.0019 | 0.9975 | 0.9864 | 1.0077 | 0.9997 | 0.9970 | 0.9998 |
| LMS    | 0.5200 | 0.4757 | 0.4946 | 0.8700 | 0.8675 | 0.9118 | 0.9676 | 0.9550 | 0.9619 |
| MM     | 0.7870 | 0.7753 | 0.7832 | 0.8193 | 0.8080 | 0.8356 | 0.8564 | 0.8522 | 0.8593 |

d. Bad leverage point

The characteristics of the effect of increasing the percentage of bad leverage point outliers on the average RMSE value tend to be similar to vertical outliers. Table 10 shows that as the percentage of bad leverage points increased, the average RMSE value increased. This applies to all methods used except the MM method. Increasing the percentage of bad leverage point outliers actually decreases the average RMSE value generated by the MM method. The LMS and MM methods produce a low mean RMSE value compared to OLS. LMS is superior to MM only when the data size is 25 with the outlier percentage of bad leverage points 0.1, besides MM is superior to LMS. The OLS method produces an average high RMSE value reaching 54 when the percentage of bad leverage points is 30%. Increasing the size of the data only decreases the average RMSE value generated by the MM method, while other methods do not have a down or up pattern of differences in data sizes.

Table 10. Average RMSE value of each method in each data size and the percentage of bad leverage points

| Method | 25     | 50     | 300    |
|--------|--------|--------|--------|
|        | 0.1    | 0.2    | 0.3    | 0.1    | 0.2    | 0.3    | 0.1    | 0.2    | 0.3    |
| OLS    | 35.0476| 45.7083| 54.4763| 32.0408| 47.9738| 55.5752| 32.3992| 45.3720| 55.3515|
| LMS    | 0.6268 | 0.8184 | 0.8862 | 0.9338 | 1.0122 | 0.9360 | 0.9507 | 0.9384 | 0.8736 |
| MM     | 0.7642 | 0.7638 | 0.7151 | 0.8286 | 0.8127 | 0.7557 | 0.8464 | 0.8299 | 0.7924 |

e. Overall average RMSE for each type of outlier

The effect of different types of outliers that contaminated the data set on the average RMSE value generated by each method is presented in Table 11. The outliers of vertical outliers and
bad leverage points in the data resulted in a high average RMSE value, while the good leverage point type was not. Each method still produces a small average RMSE value when the data is contaminated with good leverage point outliers. The average RMSE value generated by the outliers of bad leverage points, which contaminated the data, was greater than that of vertical outliers. However, both of them adversely affect the performance of methods other than LMS and MM. The LMS and MM methods appear to be able to overcome contamination by all types of outliers as indicated by the stable mean RMSE values at small numbers.

| Method | Vertical outlier | Good leverage point | Bad leverage point |
|--------|-----------------|---------------------|-------------------|
| OLS    | 19.3347         | 0.9956              | 44.8827           |
| LMS    | 0.8631          | 0.7805              | 0.8862            |
| MM     | 0.7896          | 0.8196              | 0.7899            |

### Table 11. Average RMSE values in data on data type 2 for each type of outlier

#### 3.2. Actual Data Analysis

1) Data

The data distribution plots of oil palm production with each of the explanatory variables used are presented in Figure 2. Exploration shows that some data are located far from other data sets. These data have the potential to become outlier. Therefore, it is necessary to look at the standardized residual value to identify vertical outliers, while good leverage point outliers are identified by using the plot of the diagonal value of the hat matrix.

![Diagram of actual data plot](image)

**Figure 2.** Actual data plot

2) Multicollinearity detection

The existence of a high and real correlation indicates a multicollinearity problem. The correlation values between the explanatory variables used along with the p-values for each correlation are presented in Table 12. The real correlation value is indicated by the p-value which is less than the 5% real level. P-values less than 0.05 were found in the relationship between the explanatory variables $X_1X_3$, $X_2X_3$, and $X_3X_4$ with correlations of 0.4590, 0.5257, and 0.7322, respectively. This means that the correlation that occurs between $X_1X_3$, $X_2X_3$, and $X_3X_4$ is real. Thus, there is a multicollinearity problem in the data set used.
Table 12. Pearson correlation values and p-values among explanatory variables

| Explanatory variable | $X_1$ | $X_2$ | $X_3$ |
|----------------------|-------|-------|-------|
| Pearson Correlation  | $X_2$ | 0.2861 |       |
| Nilai-p              |       | 0.1655 |       |
| Pearson Correlation  | $X_3$ | 0.4590 | 0.5257|
| Nilai-p              |       | 0.0210 | 0.0070|
| Pearson Correlation  | $X_4$ | 0.2878 | 0.3060| 0.7322|
| Nilai-p              |       | 0.1629 | 0.1368| 0.0000|

3) Outliers detection

Figure 3 shows that 11 observations were detected as outliers. There is one observation detected in the upper left room, that observation is an outlier type of vertical outlier (3rd observation). There are as many as four observations as outliers of good leverage points, namely observations 1, 6, 13 and 16. These outliers of good leverage points are detected in the right center space. Furthermore, there are six bad leverage point outliers, namely observations 2, 4, 6, 9, 14, and 15. These observations are detected in the upper right room. The observations of vertical outliers, good leverage points, and bad leverage points in Figure 3 are presented with red, green, and purple dots, respectively.

![Figure 3](image)

**Figure 3.** Plot of generalized hat matrix vs generalized RStudent

4) Regression Analysis

The results of the regression coefficients or parameter estimates obtained from each of the methods used in this study are presented in Table 13. The characteristics of the estimated parameters produced by LMS and MM tend not to be much different. On the other hand, the regression coefficient by OLS has different characteristics. This difference can be seen clearly in the estimated parameter $\beta_1$ where LMS and MM show a negative value while OLS shows a positive number. This means that the estimation of parameter $\beta_1$ by LMS and MM is not suitable because they interpret that any increase in the country's land area by one thousand hectares will reduce the estimation of oil palm production. This discrepancy is most likely due to the multicollinearity in the data which in this study this problem is not a major concern.
Table 13. Regression coefficients of each methods

| Method | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ |
|--------|-----------------|-----------------|-----------------|-----------------|
| OLS    | 6.5900          | 2.8630          | 2.8390          | -1.9090         |
| LMS    | -3.6059         | 1.4774          | 3.0393          | -0.9996         |
| MM     | -3.9790         | 1.4820          | 3.1400          | -2.3900         |

5) Evaluation model goodness of fit

The criteria for the goodness of the model used in this study to compare the performance between methods in actual data regression analysis were the values of RMSE and $R^2$. The best method is the method that produces the smallest RMSE value and the high $R^2$. The RMSE and $R^2$ values of each method in estimating the oil palm data regression model are presented in Table 14. The smallest RMSE and $R^2$ values were generated by the LMS method, namely 38.81 and 0.9978, respectively. Meanwhile, OLS produced the largest RMSE value and the lowest $R^2$. Therefore, the LMS method is the best method in multiple regression analysis to determine the effect of several explanatory variables used on total oil palm production in Indonesian provinces in 2018.

Table 14. RMSE and $R^2$ value of actual data regression

| Method | RMSE     | $R^2$  |
|--------|----------|--------|
| OLS    | 647.8000 | 0.8940 |
| LMS    | 38.8100  | 0.9978 |
| MM     | 80.7100  | 0.9891 |

4. Conclusion

4.1. Conclusion

The simulation study shows that OLS has a good performance on the data without outliers and bad leverage point outliers. LMS and MM are robust against all types of outliers, this is indicated by a stable parameter estimation bias average and RMSE value at low values. LMS is more appropriate to use for data analysis with a small size, in this study it is 25. MM is superior to LMS when the data size is medium and large, in this case 50 and 300, respectively. The application of OLS, LMS, and MM to the actual data on total oil palm production in Indonesia in 2018 which has 25 data size resulted the conclusion that the LMS method was the best method. This is indicated by the lowest RMSE value and the highest $R^2$. Three types of outliers detected in the actual data used, these are vertical outliers, good leverage points, and bad leverage point with total 44% outliers in the data. This means that the conclusions obtained from estimating parameters of the actual data on oil palm production in 2018 are in line with the results of estimating multiple regression parameters through simulation that for small data size (in this case is 25) LMS has best performance.

4.2. Suggestion

Multiple regression involves more than one independent variable. This allows for a correlation between independent variables which is often known as multicollinearity. The multicollinearity can affect the parameter estimation results. Therefore, it is necessary to conduct further research to address data that contains outliers as well as multicollinearity problems. In addition, this research did not study the comparison of method performance in data analysis that contained more than one type of outlier in one data set. Future research is expected to examine that case problem.

References

[1] D’Urso P and Massari R 2013 Weighted least squares and least median squares estimation for the fuzzy linear regression analysis Metron vol 71 pp 279–306
[2] Atilgan YK and Gunay S 2011 Least median of squares solution of multiple linear regression models through the origin Communication in Statistics—Theory and Methods vol 40 pp 4125–37
[3] Montgomery DC, Peck EA and Vining GG 2012 *Introduction to Linear Regression Analysis* ed 5 Wiley

[4] Cankaya S and baci SH 2015 A comparative study of some estimation method in simple linear regression model for different sample size in presence outliers *Turkish Journal of Agriculture-Food Science and Technology* vol 3(6) pp 380–6

[5] Montgomery DC, Peck EA, Vining GG. 2012. *Introduction to Linear Regression Analysis*. Ed ke-5. Wiley

[6] Khotimah K, Sadik K and Rizki A 2020 Kajian regresi kekar menggunakan Method penduga-MM and kuadrat median terkecil *Indonesian Journal of Statistics and Its Applications* vol 4(1) pp 97–115

[7] Almetwally EM and 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* vol 9(11) pp 1-9

[8] De Nijs R and Klausen TL 2013 On the expected difference between mean and median *Electronic Journal of Applied Statistical Analysis* vol 6(1) pp 110–7

[9] [BPS] Badan Pusat Statistik Jakarta Pusat 2019 *Statistik Kelapa Sawit Indonesia 2018* Jakarta Pusat Badan Pusat Statistik

[10] Rousseeuw PJ and Hubert M 1997 Recent developments in PROGRESS *L1-Statistical Procedure and Related Topics* vol 31 pp 201–214.