HASIL CEK_document_14

by Math Document_14

Submission date: 06-May-2023 09:32AM (UTC+0700)
Submission ID: 2085610239
File name: document_14.pdf (609.95K)
Word count: 4832
Character count: 25519

EKSAKTA
Journal of Sciences and Data Analysis

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI: 10.20885/EKSAKTA.vol4.iss1.art2

Research Article

Comparasion of The M, MM and S Estimator in Robust Regression Analysis on Indonesian Literacy Index Data 2018

Ditya Anggraheni Rahayu[1,*], Ummu Fitrotin Nursholihah[2], Galang Suryaputra[3], Sugiyarto Surono[4]

[1,2,3,4] Mathematics Department Ahmad Dahlan University, Jl. Ringroad Selatan, Kragilan, Tamanan, Kec. Banguntapan, Kabupaten Bantul, Daerah Istimewa Yogyakarta 55191

[*] *Corresponding author: ditya1900015067@webmail.uad.ac.id*

Received: 12 January 2023; Accepted: 13 January 2023; Published: 16 January 2023

Abstract: Regression analysis is a method used to determine the relationship between one dependent variable and one or more independent variables. However, the existence of outliers in the 2018 Community Literacy Development Index data led to the application of statistical methods not sensitive to pencils for analysis. This was the reason for the adoption of robust regression methods which include the M, S, and MM estimations. The three estimation methods are estimators with high damage points. This study aims to compare whether the three estimation methods are better in estimating the regression coefficient in terms of the residual standard error and adjusted r-square values. The smaller the residual standard error and the greater the adjusted r-square, the better the estimation method. Descriptive and inferential analysis with robust regression was used due to the existence of several outlier data and to provide good regression model results with unbiased values. It was discovered that the S-estimator and MM-estimator are the best methods because they have the smallest Residual Standard Error (RSE) of 1.856 and $R2$ of 0.9778.

Keywords: robust, regression analysis, outliers, estimators, literacy index

Introduction

Regression analysis is a method normally used to determine the relationship between one dependent and one or more independent variables [1]. Moreover, the estimation method is usually applied to estimate the value of the response variable influenced by the independent variable, and this is often conducted using the Ordinary Least Squares (OLS) or least squares method.

The classical approach in linear regression models is the Ordinary Least Square (OLS) technique where the sum of square of errors is minimized. Minimization of this sum means that the deviation of observation Yfrom the fitted regression line is minimized. Three underlying assumptions must be fulfilled by the errors term in linear regression analysis. These assumptions are normality, constant variance, and independence assumptions, [2]. In general, the errors are assumed to be independently and identically distributed random variables from normal distribution with mean zero and constant variances. The violation of these assumptions will cause a misleading analysis or even disturb the validity of the linear regression model fitted to the variables [3][4]. The estimations computed by linear regression will become unreliable and cannot provide useful information about the data if assumptions violate, [5][6]. In real-life data, unusual observations are a widespread issue in data analysis, [7]. One of the possible unusual observations is observed in the response variable's direction. In that case, it is called an outlier while theextreme value contained in the predictor variables is called a high leverage point. In regression, an outlier is defined as observation that does not follow the general pattern of the whole data

set [8]. Meanwhile, a high leverage point represents a x-value that lies far away from the rest of the data, [9]. Not all outliers and high leverage points are influential points, [10]. An observation is categorized as influential if removing that particular point singly or in combinationwill cause changes to the fitted model and hence the parameters of estimation. An influential point will affect the analysis's precision based on the OLS regression method. The situation worsens when such a point's presence causes the violation in linear regression assumptions,[11]. In regression analysis, in order to deal with the effect of an outlier or high leverage point which is influential, robust regression is introduced, ([12]-[13]).

The S and MM estimation methods were also compared in [14] and it was reported that S estimation is effective in the case of reading ability data for a group of children. LTS and MM have also been compared by [15] and it was discovered that MM is not more efficient than the LTS method which as has a smoother objective function, thereby, leading to its high sensitivity to local effects and the existence of a high breakdown point value.

Therefore, this research aims to determine a robust regression model with M, MM, and S estimations for the literacy index in the provinces throughout Indonesia and compare their effectiveness.

Data and Methods

Data

This research simulated the 2018 data on Literacy Index, Collection Sufficiency, Sufficiency of Library Staff, and Number of Libraries with NES (National Education Standards) from every province in Indonesia. The data were obtained from the Center for Library Development and Reading Interest Correction (P3MB) of the National Library of the Republic of Indonesia (Perpusnas RI). Moreover, the research variables include the Literacy Index per province as the dependent variable (Y) while Collection Sufficiency (X_1), Library Power Sufficiency (X_2), and Number of Libraries with NES(X_3) were used as the independent variables (X).

Methods

The analysis process applied is described using the following flow chart:

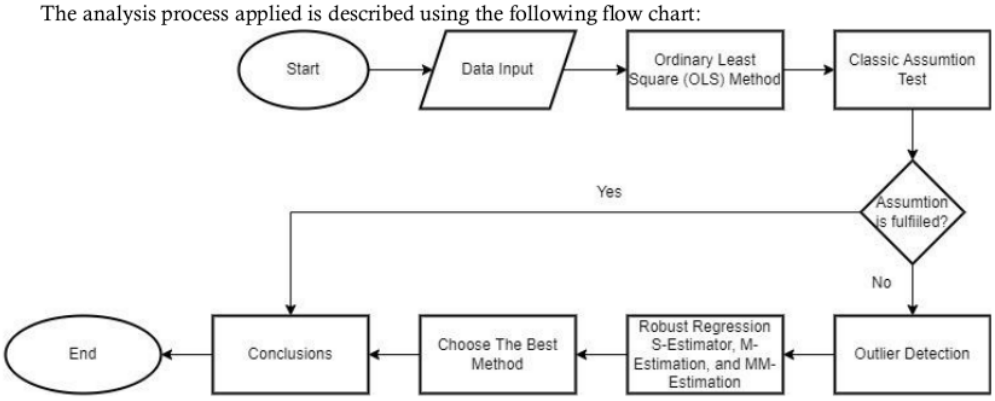

Figure 1. flowchart research

The stages in Figure 1. Are as follows:

1. Input Data
   The research process begins with inputting data on the Literacy Index in Indonesia in 2018 using R-3.6.1 software.
2. Ordinary Least Square (OLS) Method

Perform analysis of Ordinary Least Square (OLS) parameter estimation using R-3.6.1 software.

3. Classic Assumption Test
   Carrying out classic assumption tests, namely the normality test, linearity test, heteroscedasticity test, autocorrelation test, and multicorrelation test. If the classical assumption test is not met, it can be suspected that there are outliers, then the analysis process is continued with outlier detection using R-3.6.1 software.
4. Outlier Detection
   Detecting outliers using R-3.6.1 software using the R_student, DfFITS, and plots methods.
5. Robust Regression S-Estimator, M-Estimator, and MM-Estimator
   Performing analysis of estimation of robust regression parameter S Estimation, M Estimation, and MM Estimation using R-3.6.1 software.
6. Choose The Best Method
   Comparing the results of the three estimates and selecting the best estimation method in terms of the MSE and R^2 values using software R-3.6.1.
7. Conclusion

Multiple Linear Regression Model

Regression is a statistical analysis to model the relationship between response and predictor variables [16]. The models can also be used to determine the significance of the dependent variable on the independent variable. A multiple regression model usually has more than one independent variable and can be denoted as follows:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \quad (1)$$

Where,

Y_i = dependent variable

$x_{i1}, x_{i2}, \ldots, x_{ik}$ = the value of the independent variable on the $i-$ th observation

$\beta_0, \beta_1, \ldots, \beta_k$ = regression parameters

ε_i = remainder in the normally distributed $i-$ th observation

$i = 1,2, \ldots, n$ and n states the number of observations while $j = 1,2, \ldots, k$ and k declares the predictor variable.

Ordinary Least Square

Ordinary Least Square or OLS is often used to estimate regression model parameters but it can provide inefficient results due to the presence of outliers in the data [16]. The principle associated with this method is to minimize the sum of the squares of the remainder to obtain the estimated value of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ as follows:

$$JKS = S(\beta_j) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})^2 \quad (2)$$

The partial derivative of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ is subsequently determined and equated to zero in order to obtain the estimator value of the linear regression model.

Regression Analysis Assumption Test

The regression model obtained from OLS has a regression coefficient that meets the characteristics of an unbiased and best linear estimator commonly known as the Best Linear Unbiased Estimator (BLUE) [17]. Moreover, classical multiple regression assumption tests can be implemented on the residual data by determining the difference between the observed and estimated data in the multiple regression model. Those discussed in this research include the Normality, Heteroscedasticity, and Autocorrelation which were all conducted on residual data from literacy index data to determine the influence of internal libraries with the focus on the adequacy of library collections, adequacy of library staff, and NES libraries in all provinces in Indonesia.

The normality test was used to determine whether the residual data processed were normally distributed or not and it was conducted using Kolmogorov-Smirnov and Shapiro-Wilk tests. The basic

concept of the Kolmogorov-Smirnov test is to compare the distribution of the data to be tested with the standard normal distribution based on the criterion that the data is normally distributed when it shows similar distribution without significant difference from the standard normal distribution [3]. The heteroscedasticity test was also applied to ensure the regression model had an inequality or similarity in residual variance from one observation to another [18]. It was conducted using the Breusch-Pagan test such that heteroscedasticity is believed to have occurred when the $\boldsymbol{\rho - value <}$ and $\mathbf{H_0}$ is rejected but the assumption is met when the $\mathbf{H_0}$ is accepted. Moreover, the autocorrelation test was used to determine the correlation between residuals in the regression model at irregular intervals due to the fact that autocorrelation often occurs in data containing an element of time (time series). It was detected using the Durbin-Watson test [19], thereby it is believed to exist in the residuals when the $\boldsymbol{\rho - value <}$ and $\mathbf{H_0}$ is rejected but the assumption test can be fulfilled when the $\mathbf{H_0}$ is accepted. Furthermore, a multicollinearity test was used to determine whether the independent variables had a significant relationship or not, and this was achieved using the Variance Inflation Factor (VIF) value [20]. The criterion is that there is no multicollinearity when the $\mathbf{VIF < 10}$ and $\mathbf{H_0}$ is accepted and this indicates the fulfillment of the assumption.

2.4 Outlier Detection

Outliers are data that do not follow the overall data pattern or the general pattern for the regression model produced (Seheult, A. H., et all, 2005). An outlier can be identified using Cook's Distance method with the test statistic determined using the following Equation (9).

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' x' x (\hat{\beta}_{(i)} - \hat{\beta})}{kMSE} = \frac{(y_i - \hat{y})^2}{kMSE}\left[\frac{h_{ii}}{(1 - h_{ii})^2}\right] \quad (3)$$

Robust

The term "robust" was introduced into the statistical literature by Box in 1953 [22] even though it has been used as a pruned method sporadically for more than a century as indicated in Anonymous (1821) [23]. However, Tukey 1960 was the first to recognize the extreme sensitivity of some conventional statistical procedures to small deviations from assumptions [24]. The realization that statistical methods optimized for conventional Gaussian models are unstable under small perturbations was also observed to be essential for further theoretical developments initiated by Huber (1964) [25] and Hampel (1968) [26].

According to [27], the estimation methods in robust regression include:

a. M-estimation (Maximum likelihood type) is a simple estimation method both in the calculation and theory introduced by Huber (1973). It analyzes the data by assuming that most of the outliers are detected in the dependent variable.
b. LTS (Least Trimmed Squares) estimation is a method with a high breakdown point introduced by Rousseeuw (1984). The breakdown point is a measure of the minimum proportion of data contaminated with outliers compared to all observational data.
c. S (Scale) estimation is a method with a high breakdown point introduced by Rousseeuw and Yohai (1984). It has a higher efficiency than LTS at the same breakdown value.
d. MM estimation (Method of Moment) is a combination of high breakdown point and M estimation by Yohai (1987) and is observed to have a higher efficiency than the S estimation.

M estimation

M estimation is an extension of the maximum probability and robust estimation methods [28].

E-ISSN: 2720-9326
P-ISSN: 2716-0459

Table 1. M Estimation Algorithm

Algorithm 1 Estimation of M

1. Estimation of the regression coefficient on the data using OLS (Ordinary Least Square).
2. Classical assumption test of the regression model
3. Detection of the outliers in the data.
4. Calculation of the parameter estimates $\widehat{\boldsymbol{\beta}}^{\mathbf{0}}$ with OLS.
5. Calculation of the residual value $\mathbf{e_i} = \mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}$.
6. Calculation of the $\hat{\boldsymbol{\sigma}}_\mathbf{i} = \mathbf{1.4826}$ MAD.
7. Calculation of the $\mathbf{u_i} = \frac{\mathbf{e_i}}{\hat{\boldsymbol{\sigma}}_\mathbf{i}}$ value
8. Calculation of the weighting value using

$$\mathbf{w_i} = \begin{cases} \left[1 - \left(\frac{\mathbf{u_i}}{\mathbf{c}}\right)^2\right]^2, & |\mathbf{u_i}| \le \mathbf{4.685} \\ \mathbf{0}, & |\mathbf{u_i}| > \mathbf{4.685} \end{cases}$$

9. Calculation of $\widehat{\boldsymbol{\beta}}_\mathbf{M}$ using the weighted least squares (WLS) method with a weight of $\mathbf{w_i}$
10. Repeat steps 5-8 to obtain the convergent value of $\widehat{\boldsymbol{\beta}}_\mathbf{M}$.
11. Conduct a test to determine whether the independent variable has a significant effect on the dependent variable.

S Estimation

According to Rousseeuw and Yohai [29], the S estimator is an estimate with a high breakdown point but low efficiency. It is normally obtained from the minimization of the M estimator based on the residual scale. The weakness of the M estimation is that it pays less attention to the distribution of the data and is not a function of the overall data because it only uses the median as a weighted value. Therefore, the S estimator uses the residual standard deviation to overcome this weakness.

Table 2. Estimation Algorithm S

Algorithm 2 S estimation

1. Estimate the regression coefficient on the data using OLS.
2. Classical assumption test of regression model
3. Detect the presence of outliers in the data
4. Calculate $\widehat{\boldsymbol{\beta}}^{\mathbf{0}}$ with OLS.
5. Calculate the residual value $\mathbf{e_i} = \mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}$
6. Calculate

$$\hat{\boldsymbol{\sigma}}_\mathbf{i} = \begin{cases} \frac{\mathbf{median}|\mathbf{e_i} - \mathbf{median}\ \mathbf{e_i}|}{\mathbf{0.6745}}, & \mathbf{iterasi} = \mathbf{1}; \\ \mathbf{0}, & \mathbf{iterasi} > \mathbf{1} \end{cases}$$

7. Calculate the value of $\mathbf{u_i} = \frac{\mathbf{e_i}}{\hat{\boldsymbol{\sigma}}}$
8. Calculate weighted value

$$\mathbf{w_i} = \begin{cases} \begin{cases} \left[1 - \left(\frac{\mathbf{u_i}}{\mathbf{1.547}}\right)^2\right]^2, & |\mathbf{u_i}| \le \mathbf{1.547} \\ \mathbf{0}, & |\mathbf{u_i}| > \mathbf{1.547} \end{cases}, & \mathbf{iterasi} = \mathbf{1} \\ \frac{\boldsymbol{\rho}(\mathbf{u})}{\mathbf{u^2}}, & \mathbf{iterasi} > \mathbf{1} \end{cases}$$

9. Calculate $\widehat{\boldsymbol{\beta}}_\mathbf{S}$ using the WLS method with weights $\mathbf{w_i}$
10. Repeat steps 5-8 to obtain the value of $\widehat{\boldsymbol{\beta}}_\mathbf{S}$ convergent.
11. Test to determine whether the independent variable has a significant effect on the dependent variable.

E-ISSN: 2720-9326
P-ISSN: 2716-0459

MM estimation

The P was continued with the M estimator. This is the reason the MM estimator procedure involves estimating the regression parameters using the S estimator in order to minimize the residual scale. The aim is to have a high breakdown point and more efficiency. It is important to note that the breakdown value is a general measure of the proportion of outliers that can be overcome before they affect the model [27]. The MM estimator method is in the following form:

$$w_i(u_i) = \frac{\psi(u_i)}{u_i} = \begin{cases} \left(1 - \left(\frac{u_i}{c}\right)^2\right)^2, & |u_i| < c \\ 0, & |u_i| \geq c \end{cases} \quad (4)$$

Where, $\hat{\sigma}$ is the standard deviation obtained from the estimated residual $\rho(u_i)$ and used as the objective function of Tukey Bisquare

$$\rho(u_i) = \begin{cases} \frac{u_i^2}{2} - \frac{u_i^4}{2c^2} + \frac{u_i^6}{6c^2}, & -c \leq u_i < -c; \\ \frac{c^2}{6}, & u_i < -c \text{ or } u_i > c \end{cases} \quad (5)$$

Where, is the value $u_i = \frac{e_i}{\hat{\sigma}}$ and $\hat{\sigma}$ is the estimated scale and this means Equation (5) is changed to

$$S(\beta_j) = \sum_{i=1}^{n} \rho\left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta_j}{\hat{\sigma}}\right) \quad (6)$$

According to [30], the choice of population estimation for $\hat{\sigma}$ is $\hat{\sigma}_{sn}$ which is fixed and indicates the $\hat{\sigma}$ scale of the S estimator in the nth iteration.

Table 3. MM Estimation Algorithm.

Algorithm 3 MM estimation

1. Estimate the regression coefficients on the data using OLS (Ordinary Least Square)
2. Classical assumption test of regression model
3. Detect the presence of outliers in the data
4. Robust regression coefficient estimation using MM estimator.
   a. Calculate the initial estimator of coefficient $\boldsymbol{\beta}$ using the robust regression method of the S estimator.
      1. Calculate the parameters $\hat{\boldsymbol{\beta}}^{\mathbf{0}}$ using OLS
      2. Calculate the residual value $\mathbf{e_i} = \mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}$ from the S estimator
      3. Calculate the value of $\hat{\boldsymbol{\sigma}}_\mathbf{i} = \hat{\boldsymbol{\sigma}}_{\mathbf{sn}}$

$$\hat{\boldsymbol{\sigma}}_\mathbf{i} = \begin{cases} \dfrac{\mathbf{median}|\mathbf{median}(\mathbf{e_i})}{\mathbf{0.6745}}, & \mathbf{iterasi = 1} \\ \sqrt{\dfrac{\mathbf{1}}{\mathbf{nK}}\displaystyle\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{w_i e_i^2}}, & \mathbf{iterasi > 1} \end{cases}$$

      4. Calculate the value of $\mathbf{u_i} = \frac{\mathbf{e_i}}{\hat{\boldsymbol{\sigma}}_\mathbf{i}}$
      5. Calculate the weighting value

$$\mathbf{w_i} = \begin{cases} \left[\mathbf{1} - \left(\dfrac{\mathbf{u_i}}{\mathbf{4.685}}\right)^2\right]^2, & |\mathbf{u_i}| \leq \mathbf{1.547}, \quad \mathbf{iterasi = 1} \\ \mathbf{0}, & |\mathbf{u_i}| > \mathbf{1.547} \\ \dfrac{\boldsymbol{\rho}(\mathbf{u})}{\mathbf{u^2}}, & \mathbf{iterasi > 1} \end{cases}$$

      6. Calculate the parameters $\hat{\boldsymbol{\beta}}_\mathbf{S}$ using the WLS method with weighting $\mathbf{w_i^0}$
      7. Repeat steps $\mathbf{2 - 5}$ until a convergent $\hat{\boldsymbol{\beta}}_\mathbf{S}$ is obtained
   b. Calculate residual value $\mathbf{e_i} = \mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}$
   c. Calculate $\hat{\boldsymbol{\sigma}}_\mathbf{i} = \hat{\boldsymbol{\sigma}}_{\mathbf{sn}}$

E-ISSN: 2720-9326
P-ISSN: 2716-0459

d. Calculate $\mathbf{u_i} = \frac{\mathbf{e_i}}{\hat{\sigma}_i}$

e. Calculate weight

$$\mathbf{w_i} = \begin{cases} \left[1 - \left(\frac{\mathbf{u_i}}{4.685}\right)^2\right]^2, & |\mathbf{u_i}| \leq 4.685 \\ 0, & |\mathbf{u_i}| > 4.685 \end{cases}$$

f. Calculate parameters $\hat{\boldsymbol{\beta}}_S$ with the WLS method with weighting $\mathbf{w_i^0}$

g. Repeat steps $\mathbf{b} - \mathbf{e}$ until the value is obtained $\hat{\boldsymbol{\beta}}_{MM}$ the convergent

8. Calculate $\hat{\boldsymbol{\beta}}_M\mathbf{M}$ using the WLS method with weights $\mathbf{w_i}$
9. Repeat steps 5-8 to get the convergent value of $\hat{\boldsymbol{\beta}}_M\mathbf{M}$.
10. Test to determine whether the independent variable has a significant effect on the dependent variable.

Results and Discussions

Ordinary Least Square (OLS) Method

The relationship between literacy index and internal influences such as the adequacy of library collections, adequacy of the library, and library staff with NES per province in Indonesia for the 2018 period was analyzed using multiple regression. Meanwhile, a linear test was applied to determine the existence of linear relationships between two or more variables tested, and the $p - value = 0{,}6794$ which is more than $\alpha = 0{,}05$. This means the model is linear and feasible to use. Furthermore, the parameter estimation results obtained through the OLS method are presented in the following Table 4.

Table 4. OLS parameter estimation

Parameter	Estimated value	R^2
β_0 (Intercept)	−0.1998	0.6702
β_1	146.5686	
β_2	−5695.8156	
β_3	2721.8545	

Table 4 shows that the initial regression model using the OLS method can be defined using the following Equation 1

$$Y = -0.1998 + 146.5686X_1 - 5695.8156X_2 + 2721.8545X_3 \quad (7)$$

The equation model (7) does not fully explain the dependent variable due to errors. Based on Table 4. The R^2 value of 0.6702 is obtained, which means that the dependent variable (Y) can be explained by the variables X_1, X_2, X_3 of 67.02% while the rest is explained by other variables.

This method has the ability to produce the best approach when the classical assumptions have been fulfilled in order to avoid biased values and ensure valid interpretation in the acquisition of regression coefficients.

E-ISSN: 2720-9326
P-ISSN: 2716-0459

Classic Assumption Test

Several assumptions are required to be fulfilled in using regression analysis and these include the normality, homoscedasticity, autocorrelation, and multicollinearity tests. The normality tests conducted using the Shapiro-Wilk test produced $p-value = 9{,}043 \times 10^{-7}$ which is smaller than $\alpha = 0{,}05$ and this means the data is normally distributed and the assumption is fulfilled. Moreover, the homoscedasticity was determined using the Breusch-Pagan test, and the $p-value < 2{,}2 \times 10^{-16}$ obtained was found to be smaller than $\alpha = 0{,}05$, thereby, indicating the assumption of residual homoscedasticity is not satisfied. The autocorrelation test was performed using the Durbin-Watson test and the $p-value = 0{,}2455$ obtained is greater than $\alpha = 0{,}05$, thereby, indicating the assumption has been satisfied.

Outlier Detection

The inability to satisfy the homoscedasticity assumption led to the detection of the outliers using the Cook Distance (Cook's D) method and the results are presented in Figure 2.

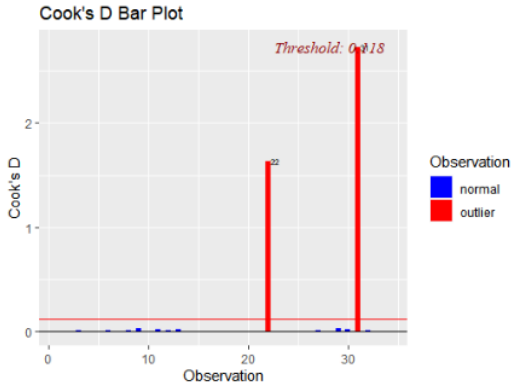

Figure 2. outlier detect

It was discovered from the figure that there are two outliers in the data are the 22nd data and 31st data.

Robust Regression

The Robust Regression parameters estimated using the M-estimator are presented in the following Table 5.

Table 5. Estimation using M estimator

Parameter	Estimated value	RSE
β_0 (Intercept)	2,5018	2,551
β_1	105,2857	
β_2	2324,9940	
β_3	4083,6711	

E-ISSN: 2720-9326
P-ISSN: 2716-0459

It was discovered that the RSE value is 2,551 and the regression model using an M-estimate is indicated in the following Equation 1.

$$Y = 2{,}5018 + 105{,}2857X_1 + 2324{,}9940X_2 + 4083{,}6711X_3 \quad (8)$$

This implies an increase of one unit in X_1, X_2, and X_3 is expected to make Y increase by 105,2857 units, 2324,9940 units, and 4083,6711 units respectively. Meanwhile, the estimation of Robust Regression parameters using the S-estimator is presented in Table 6.

Table 6. estimation using S estimator

Parameter	Estimated value	RSE	R^2
β_0 (Intercept)	1,457	1,856	0,9778
β_1	89,150		
β_2	17258,317		
β_3	5398,865		

The table shows that the RSE value is 1,856 and $R^2 = 0{,}9778$ while the regression model using S estimation is defined in Equation 1.

$$Y = 1{,}457 + 89{,}150X_1 + 17258{,}317X_2 + 5398{,}865X_3 \quad (9)$$

This shows an increase of one unit in X_1, X_2, and X_3 is expected to increase Y by 89,150 units, 17258,317 units, and 5398,865 units respectively. Furthermore, the estimation of Robust Regression parameters using the MM-estimator is presented in Table 7.

Table 7. estimation using MM estimator

Parameter	Estimated value	RSE	R^2
β_0 (Intercept)	2,663	1,856	0,8646
β_1	96,328		
β_2	5855,951		
β_3	4721,923		

The table shows that RSE = 1,856 and $R^2 = 0{,}8646$ while the Regression model using the MM-estimation is presented in the following Equation 1.

$$Y = 2{,}663 + 96{,}328X_1 + 5855{,}951X_2 + 4721{,}923X_3 \quad (10)$$

This indicates that an increase of one unit in X_1, X_2, and X_3 is expected to increase Y by 96,328 units, 5855,951 units, and 721,923 units respectively.

E-ISSN: 2720-9326
P-ISSN: 2716-0459

The Best Method

The best estimate is selected based on the smallest RSE and the most significant R^2 from the values presented in the following Table 8.

Table 8. Comparison of RSE and R^2 values for M, S, and MM estimator

Estimate	Estimation-M	Estimation-S	Estimation-MM
β_0 **(Intercept)**	2,5018	1,457	2,663
β_1	105,2857	89,150	96,328
β_2	2324,9940	17258,317	5855,951
β_3	4083,6711	5398,865	4721,923
RSE	2,551	1,856	1,856
R^2		0,9778	0,8646

Table 8 shows that the best Robust Regression method is the S-estimate using the model presented in Equation 1.

$$Y = 1{,}457 + 89{,}150X_1 + 17258{,}317X_2 + 5398{,}865X_3 \quad (11)$$

The Robust Regression Model with S-estimation also has an R^2 value of 0,9778 and this means the dependent variable Y can be influenced by variable X at 97,78% while the rest is explained or influenced by other variables outside the model.

The regression equation (11) can be described as follows:

1. The regression coefficient of the Collection Sufficiency variable (X_1) is 89,150, meaning that assuming the other independent variables are constant, for every change of 1 Collection Sufficiency unit, the literacy index per province will change by 89.150 units.
2. (X_1), (X_2), and Number of Libraries with NES(X_3) were used as the independent variables (X).
3. The regression coefficient of the Library Power Sufficiency variable (X_2) is 17258,31, meaning that by assuming the other independent variables are constant, for every change of 1 unit of Library Power Sufficiency, the literacy Index per province will change by 17258.31 units
4. The regression coefficient of the variable Number of Libraries with NES (X_2) is 5398.865, meaning that by assuming the other independent variables are constant, for every change of 1 unit of Library Power Sufficiency, the literacy index per province will change by 5398.865 units.

Conclusions

Regression analysis is a method normally applied to determine the relationship between one dependent variable and one or more independent variables while the Ordinary Least Squares (OLS) method is usually applied for estimation. However, the existence of outliers in the 2018 Community Literacy Development Index data requires the use of a statistical analysis method that is not sensitive to outliers.

This research was conducted to overcome the problem of regression analysis when the existing data assumptions are not met due to different reasons such as the presence of outliers. Robust regression

methods including M, S, and MM estimations were, therefore, used and the findings showed that the S estimation was the best model to determine the factors mostly influencing the literacy index in each province of Indonesia in 2018. The R^2 value was also recorded to be 0.9778 and this implies 97.78% of the dependent variable was explained by the independent variables while the remaining is associated with other variables outside the model.

It is recommended that other robust methods with higher accuracy are used in future studies to reduce the overall effect of outlier interference while increasing the prediction accuracy. Attention should also be placed on more complex problems such as the use of multivariable robustness.

Acknowledgement

The author thanks to Mr. Sugiyarto, Ahmad Dahlan University and all parties who have helped so that this research can be completed. This research was funded by the LPPM UAD the fiscal year 2022.

References

[1] Z. Aflakhah, J. Jajang, and A. T. Br. Sb., Kajian Metode Ordinary Least Square Dan Robust Estimasi M Pada Model Regresi Linier Sederhana Yang Memuat Outlier, J. Ilm. Mat. dan Pendidik. Mat., 11(1) (2020).
[2] D. Alita, A. D. Putra, and D. Darwis, Analysis of classic assumption test and multiple linear regression coefficient test for employee structural office recommendation, Indonesian J. Comput. Cybern. Syst., 15(3) (2021) 295.
[3] Y. Bee Wah and N. Mohd Razali, Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests, J. Stat. Model. Anal., 2(11) (2011) 21–33.
[4] A. F. Schmidt and C. Finan, Linear regression and the normality assumption, J. Clin. Epidemiol., 98 (2018) 146–151.
[5] S. Hadi, A. S., and Chatterjee, Regression analysis by example. John Wiley & Sons., 2015.
[6] M. Williams, C. A. Gomez Grajales, and D. Kurkiewicz, Assumptions of Multiple Regression: Correcting Two Misconceptions - Practical Assessment, Research & Evaluation, Evaluación práctica, Investig. y evaluación, 18(11) (2013) 1–16.
[7] O. Ayinde, K., Lukman, A. F., and Arowolo, Robust regression diagnostics of influential observations in linear regression model, Open J. Stat., 5(4) (2015) 273.
[8] C. L. Su, X., and Tsai, Outlier detection, Wiley Interdiscip. Rev. Data Min. Knowl. Discov., 1(3) (2011) 261–268.
[9] C. Yu and W. Yao, Robust linear regression: A review and comparison, Commun. Stat. Simul. Comput., 46(8) (2017) 6261–6282.
[10] R. P. Flora, D. B., LaBrish, C., & Chalmers, Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. , Front. Psychol., 2(55) (2012).
[11] Ö. G. Alma, Comparison of Robust Regression Methods in Linear Regression, Int. J. Contemp. Math. Sci., 6(9) (2011) 409–421.
[12] D. Huang, R. Cabral, and F. Dela Torre, Robust Regression, IEEE Trans. Pattern Anal. Mach. Intell., 38(2) (2016) 363–375.
[13] G. G. Montgomery, D. C., Peck, E. A., and Vining, Introduction to linear regression analysis. John Wiley & Sons., 2021.
[14] A. Semar, F. Virgantari, and H. Wijayanti, Perbandingan Estimasi S (Scale) Dan Estimasi Mm (Method of Moment) Pada Model Regresi Robust Dengan Data Pencilan, Statmat J. Stat. Dan Mat., 2(1) (2020) 21.
[15] A. Shodiqin, A. N. Aini, and M. R. Rubowo, Perbanding Dua Metode Regresi Robust yakni Metode Least Trimmed Squares (LTS) dengan metode Estimator-MM (Estmasi-MM) (Studi Kasus Data Ujian Tulis Masuk Terhadap Hasil IPK Mahasiswa UPGRIS)," J. Ilm. Teknosains, 4(1) (2018) 35–42.
[16] R. V Hogg and A. T. Craig, Introduction to Mathematical Statistics, Fourth. 1970.
[17] Algifari, Analisis Regresi, Teori, Kasus dan Solusi. Yogyakarta: BPFE UGM, 2000.
[18] S. Andriani, Uji Park Dan Uji Breusch Pagan Godfrey Dalam Pendeteksian Heteroskedastisitas Pada Analisis Regresi, Al-Jabar J. Pendidik. Mat., 8(1) (2017) 63–72.
[19] J. Durbin and G. S. Watson, Testing for Serial Correlation in Least Squares Regression II,

Biometrika, 38 (1951) 159–177.
[20] M. Sriningsih, D. Hatidja, and J. D. Prang, Penanganan Multikolinearitas Dengan Menggunakan Analisis Regresi Komponen Utama Pada Kasus Impor Beras Di Provinsi Sulut, J. Ilm. Sains, 18(1) 92018) 18.
[21] A. M. Seheult, A. H., Green, P. J., Rousseeuw, P. J., & Leroy, Robust Regression and Outlier Detection. John wiley & sons., 2005.
[22] G. E. P. Box, Non-Normality and Tests on Variances, Biometrika, 40, (1953).
[23] P. J. Huber, Robust Statistics. Berlin Heidelberg: Springer, 2011.
[24] J. W. Tukey, Conclusions vs Decisions, Technometrics, 2 (2012) 423–433.
[25] D. H. Huber Jr, E. E., and Ridgley, Magnetic Properties of a Single Crystal of Manganese Phosphide, Phys. Rev., 1964.
[26] F. R. Hampel, Contributions to the Theory of Robust Estimation., 1968.
[27] C. Chen, Robust Regression and Outlier Detection with the ROBUSTREG Procedure, SAS Inst. Inc., 9 (2002) 25–27.
[28] Y. dan Susanti, Estimasi-M dan Sifat-sifatnya pada Regresi Linear Robust., Math-Info, 2008.
[29] P. R. & V. Yohai, Robust Regression by Means of S-Estimators. New York: Springer, 1984.
[30] M. S.-B. RA Maronna, RD Martin, VJ Yohai, Robust Statistics: Theory and Methods (with R). 2019.

HASIL CEK_document_14

ORIGINALITY REPORT

4%	4%	0%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Islam Indonesia Student Paper	4%

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	On		