Mathematics and Statistics 8(4): 373-381, 2020
DOI: 10.13189/ms.2020.080401 http://www.hrpub.org

Robust Regression Analysis Study for Data with Outliers at Some Significance Levels

Waego Hadi Nugroho*, Ni Wayan Surya Wardhani, Adji Achmad Rinaldo Fernandes, Solimun

Department of Statistics, Faculty of Mathematics and Natural Science, Brawijaya University, Veteran Street Malang 65145, Indonesia

Received March 22, 2020; Revised April 23, 2020; Accepted May 20, 2020

Abstract Robust regression analysis is an analysis that is used if there is an outlier in a regression model. Outliers cause data to be abnormal. The most commonly used parameter estimation method is Ordinary Least Squares (OLS). However, outliers in models cause the estimator of the least-squares in the model to be biased, so handling of outliers is required. One of the regressions used for outliers is robust regression. Robust regression method that can be used is M-Estimation. By using Tukey's Bisquare weighted function, a robust M-estimation method can estimate parameters in a model, for example in malnutrition data in East Java Province 2017 to 2012. This study aims to compare the robust method of M-estimation and OLS method on data with several different levels of significance, which is 1%, 5%, and 10%. The predictor variables used in this study were the percentage of poor society, population density, and some health facilities. R^2 is used to compare the OLS method and the robust method of M-estimation. The results obtained that robust regression is the best method to handle the model if there are outliers in the data. It was supported by almost all results of the value of R^2 on each data that M-estimation has a higher value than the OLS method.

Keywords Malnutrition, M-Estimation, Tuckey Bisquare

1. Introduction

Regression analysis is a statistical analysis that is used to determine the relationship between response variables and predictor variables, both and more. According to (Draper & Smith, 1992), regression analysis is used to determine conclusions from data that has a related relationship between response and predictors variables. If there is more than one predictor variable, then the analysis can be said as multiple linear regression analysis (Fernandes et al., 2018).

There are several methods to predict the parameters in regression, one of the methods is Ordinary Least Square (OLS) (Hidayat et al., 2019). Estimating parameters with OLS must fulfill some prescribed assumptions, with errors mutually independent and normal with a middle value 0 and variance σ^2 (Fernandes et al, 2014). Multiple regression analysis must fulfill several other assumptions namely the assumption of normality, multicollinearity, heteroscedasticity, and autocorrelation (Gujarati, 2003).

Various problems can occur in estimating model parameters, one of which is if the assumption of normality isn't fulfilled, one of which is due to outliers in the data (Fernandes et al., 2019). According to (Draper & Smith, 1992), data coming out of the regression line is an outlier. Outliers are data that have different characteristics from other non-outliers data. The existence of outliers can affect the process of data analysis and errors in making conclusions, but the existence of outliers can be overcome by one method, namely robust regression (Kutner et al., 2004).

The robust method is used when a data contains an outlier and has an abnormal distribution that will affect the parameter estimator (Risnawati et al., 2019). The purpose of the robust method is to detect outliers in the data and give strong results to the effect of outliers (Rousseeuw & Leroy, 1987). There are several kinds of estimators available in robust regression, such as RM-Estimator, LMS-Estimator, LTS-Estimator, S-Estimator, LWS-Estimator, M-Estimator, LAD-Estimator, MM-Estimator, and REWLSE-Estimator. In this study, the analysis used is a robust M-estimator analysis. This estimator is one of the simplest predictors in both computational and theoretical first introduced by Huber in 1973.

Before conducting a robust M-estimator regression analysis, outliers are estimated on the data. There are several ways to detect the existence of outliers by considering the determination of the value of α, which is by Stundetized Deleted Residual (TRES) and Cook Distance.

However, the detection research used was (TRES) with an significant levels value of 10%, 5%, and 1%.

This study will examine how different the significance levels vary between 5% and 10% with a robust M-estimator analysis on the data about the effect of the response variable, namely malnutrition. There are 3 predictor variables, namely the poor society (%), population density (/km) and the number of health facilities in cities in the province of East Java.

2. Literature Review

2.1. Linear Regression

Multiple linear regression analysis is a statistical analysis used only if there is more than one predictor variable (x) which can affect the response variable (y). The model for multiple linear regression according to (Gujarati, 2003) is as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i \quad (1)$$

where $\beta_1,...,\beta_p$ are regression coefficients of the dependent variables-*i* and $x_{1i},...,x_{pi}$ the independent variables-*i* with their parameters. Some of the assumptions underlying the multiple regression model are error normality, multicollinearity, heteroscedasticity, and autocorrelation.

2.1.1. Normality Test

Residual normality testing is carried out using the Kolmogorov-Smirnov test. The test statistics used are:

$$D_n = \text{maximum}\left|F_n(x) - F_0(x)\right| \quad (2)$$

Where the maximum upright D_n between the empirical distribution functions $F_n(x)$ with the normal distribution function $F_0(x)$. If $D_n > D_n(\alpha)$ with $D_n(\alpha)$ is the critical point of the Kolmogorov-Smirnov test, the decision taken against H_0 with an error rate of α or can be said to be residual does not spread normally (Kutner et al., 2004).

2.1.2. Multicollinearity

Multicollinearity occurs because of the linear correlation between two or more predictor variables. The consequences of multicollinearity are very difficult to separate the influence of each independent variable on the non-independent variables (Kutner et al., 2004). One method that can be used to detect whether there is multicollinearity or not is by looking at the value of the VIF (Variance Inflation Factor) of each predictor variable. VIF values can be defined as follows (Gujarati, 2003):

$$VIF = \frac{1}{1 - R_j^2} \quad (3)$$

When R_j^2 is coefficient of determination between the predictor x_j variables with other predictor variables. If the value of the VIF obtained is high, the greater the correlation between the independent variables. If there is a VIF $\geq$ 10, the correlation between predictor variables is very high and vice versa (Bowerman & O'Connel, 1990).

2.1.3. Heteroscedasticity

According to (Gujarati, 2003), the result of heteroscedasticity in regression is that the estimates obtained are inefficient, both in small and large samples. One of the ways to overcome heteroscedasticity in regression is by testing the Spearman ranking correlation. Spearman's rank correlation (r_s) can be calculated using the formula:

$$r_s = 1 - 6\left(\frac{\sum d_i^2}{N\left(N^2 - 1\right)}\right) \quad (4)$$

where :

d_i : difference in standard deviation (S) ranking and absolute value error, $e_i = y_i - \bar{y}$

N: number of samples

2.1.4. Autocorrelation

If the regression model has autocorrelation, it means that the existing sample variant cannot describe the population variant. To see whether there is autocorrelation, one method that can be used is by testing through the Durbin-Watson test. With hypothesis testing, namely:

H_0: there is no positive correlation

H_1: there is a positive correlation

Durbin-Watson's statistics are as follows (Kutner et al., 2004):

$$d = \frac{\sum_{i=2}^{n}\left(e_i - e_{i-1}\right)^2}{\sum_{i=1}^{n} e_i^2} \quad (5)$$

Durbin and Watson have created tables with statistics d Durbin-Watson with a real level of 5% and 1%. In the table, there are upper bound d_u and lower bound d_l for various values of n (= number of samples) and k (number of independent variables). After calculating the value from d in the formula [5], then the results obtained, compared with d_l in the table. If the correlation is positive (ie 0 <p <1), d will tend to have a small value. In other words, if the value of d is smaller than the value of d_l, the decision obtained is to reject H_0. It can be concluded that there is autocorrelation.

2.2. Parameter Estimation

If the regression analysis has fulfilled all the assumptions for the residual, then parameter estimation is done in multiple regression. The best method for estimating regression parameters is by the principle of minimizing the number of square error, this method is called the ordinary least squares method (OLS) (Draper & Smith, 1992). So for each i[th] location, the parameter estimation of the OLS model is done by the matrix operation as follows:

$$\hat{\underset{\sim}{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underset{\sim}{y} \tag{6}$$

One problem that often occurs in the multiple regression model is the presence of outliers. Outliers are observations that may have a large effect on regression coefficients caused by the location of data that does not follow the pattern in general (Kutner et al., 2004). Outliers cause the least-squares estimator to be biased. Outlier problems can also affect confidence intervals to be of greater value and estimates become inefficient and inconsistent (Barnett & Lewis, 1994).

2.3. Outliers

According to (Cohen et al., 2003) in his book, he explained that three characteristics could potentially be used to detect outliers namely Leverage, Discrepancy, and Influence. There are two types of measures of the effect that can be used, first is a measure of global effect, namely DFFITS (difference in standardized) and Cook's Distance. One way that can be used to detect and identify an outright according to (Bowerman & O'Connel, 1990) is TRES. TRES is a test statistic used to check for outliers in the response variable. The hypothesis underlying this test is (Kutner et al., 2004):

H_0: The i[th] observation is not an outlier

H_1: The i[th] observation is an outlier

$$TRES_i = e_i\left[\frac{n-p-2}{SSE(1-h_i)-e^2}\right]^{\frac{1}{2}} \tag{7}$$

where:

e_i: $y_i - \hat{y}_i$

h_i: leverage value for i[th] observation

SSE : sum square error

p : number of predictor variants

n : number of observations (i = 1, 2, ..., n)

with the testing criteria underlying the decision are:

$$\left|TRES_i\right| = \begin{cases} \leq t_{\frac{\alpha}{2}n-p-2} & , H_0 accepted \\ > t_{\frac{\alpha}{2}n-p-2} & , H_1 accepted \end{cases} \tag{8}$$

Where:

$t_{\frac{\alpha}{2}n-p-2}$: the distribution of t with the degree of freedom $n-p-2$

All possible values of $\left|TRES_i\right|$ follow the distribution of t with the degree of freedom $n-p-2$. The DFITS method is a measure to find out how big the deviation of the Y value is for a data without the first observation, and is used to detect whether a particular value influences the y value. In theory, the DFITS method can be defined according to (Cohen et al., 2003) as follows:

$$DFITS_i = \frac{y_i - y_{i(i)}}{\sqrt{MS_{residual(i)}\,h_i}} \tag{9}$$

where:

y_i : predictive value when the i[th] case is not deleted

$y_{i(i)}$: predictive value when the i[th] case is written off

$MS_{residual\ (i)}$: the variance value of an error when the i[th] case is deleted

h_i : leverage value for i[th] observation

Test criteria that underlie decisions namely (Kutner et al., 2004):

$$\left|\left(DFITS\right)_i\right| = \begin{cases} \leq 2\sqrt{\frac{p+1}{n}} & , H_0 accepted \\ \geq 2\sqrt{\frac{p+1}{n}} & , H_1 accepted \end{cases} \tag{10}$$

The hypothesis underlying testing is:

H_0 : The observation *i* has no effect

H_1 : The observation *i* has effect

Cook's Distance (Di) is used to detect outliers affecting all regression coefficients. In theory, it can be defined as follows:

$$D_i = \frac{r_i^2}{p}\left[\frac{h_i}{(1-h_i)}\right] \tag{11}$$

Where r_i^2 has the following formula:

$$r_i = \frac{e_i}{MS_{Residual}\sqrt{1-h_i}} \tag{12}$$

After knowing the formula of r_i^2, equation (12) can be substituted in equation (11) so that:

$$D_i = \frac{(e_i)^2}{p.\mathrm{MS}_{\mathrm{Residual}}}\left[\frac{h_i}{(1-h_i)^2}\right] \tag{13}$$

Where:
e_i : i[th] residual
r_i^2 : studentized residual
p : number of parameters including β_0
h_i : leverage value, for observation

$$\mathrm{MS}_{\mathrm{Residual}} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-\left(k+1\right)}$$

The hypothesis underlying testing is:
H_0: The observation *i* has no effect
H_1: The observation *i* has effect
The criteria used to test the hypothesis are as follows:

$$D_i = \begin{cases} \leq F_{\alpha,p,n-p} & ,H_0 accepted \\ > F_{\alpha,p,n-p} & ,H_1 accepted \end{cases} \quad (14)$$

According to (Gujarati, 2003), it is necessary to select the best model between the regression models obtained from parameter estimation using robust M-estimator with a regression model obtained from estimating parameters using the ordinary least squares method OLS). One of the best model selection criteria can use the value R^2.

The value of R^2 in linear regression cannot be used to compare two models with many different predictor variables x. When the number of variable predictors is added, the proportion of the variable response y explained by the predictor x variable will always increase. R^2 can be used to choose which model is best based on the number of predictor variables x used. The value of R^2 can be calculated using a formula:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \quad (15)$$

where:
y_i : the response variable for the i[th] observation
$\hat{y}_i$: estimator of the y_i response variable for the i[th] observation
$\bar{y}$: the average response variable y_i for i[th] observation
k : number of predictor variables
n : the number of observations

2.4. Robust Regression

Rousseeuw & Leroy (1987) explain that estimating parameters with the least squares method has weaknesses, one of which is if there are influential outliers. Not only in the response variable, but it can affect the predictor variables. If there is an outlier, it will lead to a deviation from one of the assumptions, namely the assumption of a normality, so that the resulting prediction becomes invalid. (Draper & Smith, 1992), explain one of the deviations that occur if there are outliers, which can cause data to be distributed abnormally, so that assumptions are not met. The right method to overcome the deviation from normal assumptions is by using robust regression. Kutner et al. (2004) explained that to overcome the existence of outliers, robust regression can be used to reduce the effect of outliers when compared to using OLS so that strong predictors are generated and not affected by outliers. Robust analysis can be used to analyze and match regression models and overcome outlier points that have large side values without removing data but find models that match most of the data as robust solutions.

2.5. Robust Regression M-Estimator

Robust M-estimator regression was first introduced by Hubber in 1973. Robust M-estimator is one of the most frequently used methods and is considered good for estimating parameters caused by outliers. According to (Draper & Smith, 1992), robust regression is the M-estimator of the maximum likelihood type.

In the OLS equation according to (Yuliana et al., 2014), the purpose of this method is to minimize the number of equal squares. Following are the results of the OLS method:

$$\sum_{i=1}^{n}\left(y_i - x_i^{'}\beta\right)^2 = \sum_{i=1}^{n} e_i^2 \quad (16)$$

In the robust M-estimator, replace e_i^2 with $\rho\left(u_i\right)$ in equation (16), where the value is $u_i = \frac{e_i}{s}$. So that the robust M-estimator minimizes the objective function:

$$\sum_{i=1}^{n}\rho\left(u_i\right) = \sum_{i=1}^{n}\rho\left(\frac{e_i}{\sigma}\right) = \sum_{i=1}^{n}\rho\left(\frac{y_i - \boldsymbol{x_i^{'}\beta}}{\sigma}\right) \quad (17)$$

The function ρ contributes to each side with the following conditions:

$$\rho(u_i) \geq 0$$
$$\rho(0) = 0$$
$$\rho(u_i) = \rho(-u_i)$$
$$\rho(u_i) \geq \rho(u_i) \text{for} |e_i| \geq |u_i|$$

The equation for the robust M-estimator is as follows (Montgomery & Peck, 1992):

$$\min\sum_{i=1}^{n}\rho\left(u_i\right) = \min\sum_{i=1}^{n}\rho\left(\frac{e_i}{\sigma}\right) = \min\sum_{i=1}^{n}\rho\left(\frac{y_i - \boldsymbol{x_i^{'}\beta}}{\sigma}\right) \quad (18)$$

M-Estimator as a solution to equation (16), needs to set a scale to produce equation (17). The robust estimator scale is *s*, with the following formula:

$$s = \frac{median\left|e_i - median\left(e_i\right)\right|}{0.6745} = \frac{MAD}{0.6745} \quad (19)$$

For the function ρ, you can use the weighted function Tukey's Bisquare as follows:

$$\rho\left(u_i\right) = \begin{cases} \frac{u_i^2}{2} - \frac{u_i^4}{2c^2} + \frac{u_i^6}{6c^4} & \left|u_i\right| \le c \\ \frac{c^2}{6} & \left|u_i\right| > c \end{cases} \quad (20)$$

In equation (19) the median is resistant to outliers, therefore the median is used in the calculation of the robust estimator for σ. If *n* is large, the constant value of 0.6745 makes the value *s* as an estimator that approaches the bias of σ.

Next, a solution is needed for equation (17) by using the first partial derivative of ρ with β_j (j = 0, 1, ... , p)and then with zero. If it is known that $\psi = \rho'$ and X_{ij} is the first observation at the j-point, then:

$$\sum_{i=1}^{n} X_{ij} \psi\left(\frac{y_i - \boldsymbol{x_i'\beta}}{s}\right) = 0 \quad (21)$$

To provide a solution to equation (20), a weighted function is defined as follows:

$$\omega\left(e_i\right) = \frac{\psi\left(\frac{y_i - \boldsymbol{x_i'\beta}}{s}\right)}{\left(\frac{y_i - \boldsymbol{x_i'\beta}}{s}\right)} \quad (22)$$

Note that $u_i = \frac{e_i}{s}$, then equation (22) can be written as follows:

$$\omega_i = \begin{cases} \left[1 - \left(\frac{u_i}{c}\right)^2\right]^2, & \left|u_i\right| \le c \\ 0 & \left|u_i\right| > c \end{cases} \quad (23)$$

Value c = 4.685 for the weighting function of Tukey's Bisquare. So equation (21) becomes:

$$\sum_{i=1}^{n} x_{ij} \omega_i \left(y_i - x_i'\beta\right) = 0 \quad (24)$$

Equation (24) can be written in the form of a matrix to:

$$\hat{\beta}_{Robust} = \left(X'W_0X\right)^{-1} X'W_0y \quad (25)$$

Where W_0 is the main diagonal matrix of the weighting with the i-diagonal element is w_{i0}. Equation [25] is known as the Weighted Smallest Squares (WLS) equation (Yuliana et al., 2014). To calculate the M-estimator, you can use an iteration from a weighted or IRLS (Iteratively Reweighted Least Square) OLS. The steps to do the calculation are as follows:

1. Calculates the value of estimator $\hat{\beta}_{(0)}$ from the regression model with the OLS method. Then calculate the errors (e_i).
2. Calculates the initial s value (s_0) and the initial weighting W_{i0} from the initial edge.
3. Get the first new predictive value ($\hat{\beta}_1$) which is robust with IRLS as follows

$$\hat{\beta}_1 = \left(X'W_0X\right)^{-1} X'W_0y$$

4. Change the parameter estimator in step 3 as $\hat{\beta}_0$ in step 1 and get a new errors, new value, and new weighted value.
5. Then do the 3rd step again and again until it gets convergent, namely $\hat{\beta}_k$-$\hat{\beta}_{k_{k-1}} < 1 \times 10^{-6}$, where k is the iteration index.

Weighted of Tukey's Bisquare has a tuning constant (c) of 4.685 that has each weight (Kutner et al., 2004).

3. Research Method

This study uses secondary data sourced from the Central Statistics Agency (BPS) [16] in East Java, including data on the percentage of poor society (%), life expectancy (years), school expectations (years) and per capita expenditure for food (percent) in districts / cities East Java Province in 2016. The stages of analysis carried out are: (1) Testing of residual normality of classical linear regression models; (2) Testing of spatial heterogeneity; (3) Outlier detection; (4) Establishment of the GWR model; (5) Estimating parameters of the GWR model; (6) IRLS Process; (7) Testing estimators of RGWR parameters; (8) Calculating the coefficient of local determination; (9) Test for residual normality of the RGWR model; (10) Making a map of the 2016 East Java poverty rate.

4. Result and Discussion

4.1. Normality Test

Residual normality testing was carried out using the Kolmogorov Smirnov test. The value of the $\boldsymbol{D_n}$ test statistic is as follows:

Table 1. Residual Normality Test Results in Malnutrition Data

Year	(D_n)	*P-value*
2017	0.60526	7.16E-14
2016	0.62162	2.53E-14
2015	0.71053	2.22E-16
2014	0.57895	1.30E-12
2013	0.57895	1.30E-12
2012	0.75952	2.22E-16

Therefore H_0 is rejected so that it can be concluded that the residuals of classical linear regression models are not normally distributed. Residuals that are not normally distributed can be caused by an outlier in the data.

4.2. Robust Regression

Outlier detection is done using TRES detection. Then the results of TRES are compared with $t_{\frac{\alpha}{2}n-p-2}$. Outlier detection is done using TRES detection. Then the results of TRES are compared with $t_{\frac{\alpha}{2}n-p-2}$. Results can be seen in Table 2.

Table 2. Results of Detecting Outliers with TRES

Year	Alpha 1%	Alpha 5%	Alpha 10%
2017	1(10)	2(10,20)	3(10,11,20)
2016	2 (8,10)	2 (8,10)	2 (8,10)
2015	2 (8,10)	3 (8,10,18)	3 (8,10,18)
2014	2 (8,10)	2 (8,10)	2 (8,10)
2013	2 (6,10)	2 (6,10)	3 (6,8,10)
2012	2 (12,21)	3 (10,12,21)	3 (10,12,21)

Based on the results from Table 2, it is known that all malnutrition data know 2012-2017 has outliers. Testing the best model at the level of 10% in Table 3 shows that in 2017, the value of OLS R^2 is higher than the Robust estimator R^2 value. It can be concluded that the OLS model is able to explain the response variable better than the Robust M-estimator model. But in 2013, the R^2 robust M-estimator is higher than the R^2 OLS value. Then it can be concluded that the Robust M-estimator model is able to explain the response variable better.

Table 3. R^2 value results in outlier data with a significance level of 10%

Year	R^2 - OLS	R^2 - Robust *M*
2017	0,2480	0,1698
2013	0,3464	0,5977

In the data of malnutrition in 2017 with a significance level of 10%, parameter estimation uses OLS. By using software R, the estimation results for malnutrition data in 2017 with a 10% significance level are as follows:

Table 4. OLS estimation results on outlier data with a significance level of 10%

Year	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
2017	-28,1890	4,3769	0,0030	0,0537

In the 2013 malnutrition data with a significance level of 10%, parameter estimation uses Robust M-estimator. By using software R, the estimation results for malnutrition data in 2013 with a significance level of 10% are as follows on Table 5.

Table 5. Robust Estimation Results on 2013 data with a significance level of 10%

Year	Maximum iteration	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\left\|\hat{\beta}_k - \hat{\beta}_{k-1}\right\|$
2013	11	-36,654	4,3879	0,0051	0,0506	0,0000
	12	-36,655	4,3878	0,0051	0,0506	0,0000

Test results for OLS model parameters on 2017 data use R software, and the results are as follows (α = 10%) in Table 6.

Table 6. Test results of OLS model parameters on 2017 data with a significance level of 10%

Year	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	F_{hit}
2017	-28,1890	4,3769*	0,0030	0,0537*	3,5185*

In 2017, the F_{hit} value is greater than F_{Table} at 2,247, showing that F_{hit} is significant. So there is an influence between predictor variables and overall response variables. Viewed in the partial test, $\hat{\beta}_1$ and $\hat{\beta}_3$ have a significant influence on the response variable.

Test results of the Robust model parameters on the 2013 data use R software, and the results are as follows (α = 10%) on Table 7.

Table 7. Test results of the Robust model parameters on 2013 data with a significance level of 10%

Year	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	F_{hit}
2013	-36,6553	4,3878*	0,0051	0,0506*	3,9908*

Note: The * sign is significant at the level of 10%

In 2013, there was an influence between the response variable and the overall predictor variable, because Fhit showed significant results with F_{Table} values of 2.247. The results obtained in the partial test show that $\hat{\beta}_1$ and $\hat{\beta}_3$ have a significant effect on the response variable.

The selection of the Best Model with R^2 in outlier data with a 5% significance level using Software R, results in a value of R^2 in the regression model with OLS and robust M-estimator for each malnutrition data.

Table 8. R^2 value on outlier data with a significance level of 5%

Year	R^2 for OLS	R^2 for M-Robust
2017	0,1909	0,2207
2015	0,1739	0,4716
2012	0,2694	0,3362

It can be seen that all values of R^2 in Robust M-estimator are higher than the value of R^2 in OLS. It can be concluded that the Robust model is able to explain the response variable better than the OLS model when data contains outliers.

In the data of malnutrition in 2017, 2015, and 2012 with a significance level of 5%, parameter estimation using Robust is conducted. By using software R, the estimation results for malnutrition data in 2017, 2015 and 2012 with a significance level of 5% are as follows on Table 9.

Table 9. Maximum Iteration of M Robust Regression on outlier data with a significant of 5%

Year	Maximum Iteration	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\|\hat{\beta}_k - \hat{\beta}_{k-1}\|$
2017	7	-18,5708	3,9055	0,0021	0,0522	0,0000
	8	-18,5701	3,9055	0,0021	0,0522	0,0000
2015	17	-63,0099	6,2573	0,0094	0,0504	0,0000
	18	-63,0113	6,2567	0,0094	0,0504	0,0000
2012	21	-144,9612	10,7267	0,0291	0,1030	0,0000
	22	-144,959	10,7266	0,0291	0,1030	0,0000

Based on table 9, we have obtained robust parameter estimation with the difference in the value of $\hat{\beta}_k$ with $\hat{\beta}_{k-1}$ less than 1×10^{-6} at the end of the iteration. So that it can be concluded, in table 9 it is a new model to explain the amount of malnutrition in each Regency / City in East Java with a significance level of 5%. After obtaining robust data, the parameter testing of robust data is repeated. With the same test with the previous test, the following results are obtained, with α = 5% on Table 10.

Table 10. Robust Regression Parameter Test Results on outlier data with a significance level of 5%

Year	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	F_{hit}
2017	-18,5701	3,9055*	0,0021	0,0522*	1,9894
2015	-63,0113	6,2567*	0,0094*	0,0504*	3,4184*
2012	-144,959	10,7266*	0,0291*	0,1030*	2,6839

By using R software, it can be concluded that in 2017, the F_{hit} value is smaller than F_{Table} at 2,866, which shows that F_{hit} t is not significant. So there is no influence between predictor variables and overall response variables. Seen in the partial test, there are $\hat{\beta}_1$ and $\hat{\beta}_3$ which have a significant influence on the response variable.

In 2015, there was an influence between the response variable and the overall predictor variable, because F_{hit} showed significant results with F_{Table} values of 2.874. The results obtained in the partial test indicate that the overall regression coefficient has a significant effect on the response variable.

In 2012, where F_{Table} = 2,874, F_{hit} was declared insignificant. Then it can be concluded that overall there is no influence between predictor variables and response variables. Judging by a partial test, all regression coefficients significantly influence the response variable.

The selection of the Best Model with R^2 in outlier data with a significance level of 1% using Software R, results in a value of R^2 in the regression model with OLS and robust M-estimator for each malnutrition data. Results can be seen as follows:

Table 11. R^2 value in outlier data with a significance level of 1%

Year	R^2 for OLS	R^2 for M- Robust
2016	0.2063	0.3489
2014	0.2270	0.2961

Estimation M is higher than the value of in R^2 OLS. It can be concluded that the Robust model is able to explain the response variable better than the OLS model when data contains outliers. Just like before, parameter estimation needs to be done for each existing variable. According to table 4.14 on data on malnutrition in 2017, and 2012 with a significance level of 1%, parameter estimation using Robust is done. By using software R, the estimation results for malnutrition data in 2016 and 2014 with a significance level of 1% are as follows on Table 12.

Based on table 12, we have obtained robust parameter estimation with the difference in the value of $\hat{\beta}_k$ with $\hat{\beta}_{k-1}$ less than 1×10^{-6} at the end of the iteration.

So that it can be concluded, in Table 12 it is a new model to explain the amount of malnutrition in each district / city in East Java with a significance level of 1%.

Table 12. Maximum Iteration of M Robust Regression on outlier data with 1% significance

Year	Maximum Iteration	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\|\hat{\beta}_k - \hat{\beta}_{k-1}\|$
2016	9	-17.9282	3.9108	0.0020	0.0520	0.0000
	10	-17.9276	3.9108	0.0020	0.0520	0.0000
2014	8	-61.0303	5.4852	0.0088	0.0822	0.0000
	9	-61.0314	5.4853	0.0088	0.0821	0.0000

As in the previous data, it is necessary to test parameters on data with a significance level of 1%. By using software R, the results are as follows, with α = 1% on Table 13.

Table 13. Test results model parameters on outlier data with a significance level of 1%

Year	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	F_{hit}
2016	-17.9276	3.9108*	0.0020	0.0520*	2.9314
2014	-61.0314	5.4853*	0.0088*	0.0821*	2.5897

By using R software, it can be concluded that in 2016, the F_{hit} value was greater than F_{Table} at 4,343, which shows that Fhit is significant. So there is an influence between predictor variables and overall response variables. Viewed in the partial test, $\hat{\beta}_1$ and $\hat{\beta}_3$ have a significant influence on the response variable.

In 2013, there was an influence between the response variable and the overall predictor variable, because F_{hit} showed significant results with F_{Table} values of 4,343. The results obtained in the partial test indicate that the overall regression coefficient has a significant effect on the response variable.

5. Conclusions

Based on the results of the analysis it can be concluded that, the model with parameter estimators obtained from the M-Estimator method can be effectively used to predict Malnutrition in East Java Province. This is supported by the value of R^2 in the Robust M-estimator method which produces a significant value for the value of R^2 in the OLS estimator.

The results of the parameter estimation equation with the OLS method do not differ significantly from the results of the parameter estimation equation using the robust M-estimator regression method both at the 1%, 5% or 10% significance level. This is supported by the results of parameter testing which shows that the regression coefficients that have a significant effect are almost entirely located in $\hat{\beta}_i$ which is the same both in the OLS method and in the robust M-estimator.

REFERENCES

[1] Barnett, V. and Lewis, T. Outliers in Statistical Data (3rd ed.). New York: John Wiley and Son Inc. 1994.

[2] Bowerman, B. L. and O'Connel, R. T. Linear Statistical Models, an Applied Approach. Boston: PWS-Kent Publishing Company. 1990.

[3] Cohen, J., Cohen, P., West, S. G. and Aiken, L. S. Applied Multiple Regression/Correlation Analysis for The Behavioral Sciences. New Jersey: Lawrence Erlbaum Associate. 2003.

[4] Draper, N.R and Smith, H. Applied Regression Analysis (2nd ed.) New York: John Wiley and sons, Inc. 1992.

[5] Fernandes, A.A.R.,Nyoman Budiantara, I.,Otok, B.W., & Suhartono. Reproducing Kernel Hilbert space for penalized regression multi-predictors: Case in longitudinal data. International Journal of Mathematical Analysis, 8(40), 1951-1961, 2014.

[6] Fernandes, A.A.R, Jansen, P., Sa'adah, U, Solimun, Nurdjannah, Amaliana, L., & Efendi, A. Comparison of Spline Estimator at Various Levels of Autocorrelation in Smoothing Spline Nonparametric Regression For Longitudinal Data. Communications in Statistics – Theory and Methods, 46 (24), 12401-12424, 2018.

[7] Fernandes, A.A.R., Hutahayan, B., Solimun, Arisoesilaningsih, E., Yanti, I., Astuti, A.B., Nurjannah, & Amaliana, L. Comparison of Curve Estimation of the Smoothing Spline Nonparametric Function Path Based on PLS and PWLS In Various Levels of Heteroscedasticity. IOP Conference Series: Materials Science and Engineering, Forthcoming Issue, 2019.

[8] Gujarati, N Damodar. Basic Econometrics (3rd ed.). New York: McGraw-Hill. 2003.

[9] Hidayat, M.F., Fernandes, A.A.R., & Solimun. Estimation of Truncated Spline Function in Nonparametric Path Analysis Based on Weighted Least Square (WLS). IOP Conference Series: Materials Science and Engineering, Forthcoming Issue, 2019.

[10] Kutner, M.H., Nachtsheim, C. J. and Neter, J. Applied Linear Statistical Model. New York: McGraw-Hill. 2004.

[11] Montgomery D.C. and Peck E.A. Introduction to Linear Regression Analysis (2nd ed.). New York: John Wiley. 1992.

[12] Risnawati, Fernandes, A.A.R., and Nurjannah. The Estimation Function Approach Smoothing Spline Regression Analysis for Longitudinal Data. IOP Conference Series: Materials Science and Engineering, Forthcoming Issue, 2019.

[13] Rousseeuw, P.J. and Leroy, A.M. Robust Regression and Outlier Detection. New York: Wiley Interscience. 1987.

[14] Yuliana S., Hasih P. and Sri S. H. M-Estimation, S-Estimation, and MM-Estimation in Robust Regression. International Journal of Pure And Applied Mathematics. 91(3), 349-360, 2014.