4th ICRIEMS Proceedings
Published by The Faculty Of Mathematics And Natural Sciences
Yogyakarta State University, ISBN 978-602-74529-2-3

Application of Robust M-Estimator Regression in Handling Data Outliers

Julita Nahar [a)] and Sri Purwani [b)]

Department of Mathematics, Faculty of Mathematics and Natural Science, Padjadjaran University
Jalan Raya Bandung Sumedang Km 21, Jatinangor, Indonesia

[a)] Corresponding author: julita.nahar@unpad.ac.id
[b)] sri.purwani@unpad.ac.id

Abstract. The population growth in Indonesia is significantly high. Indonesia became the fourth country with the largest population in the world, whereas in the region of ASEAN, Indonesia was rated as the first. In 2007, the total population of Indonesia was amounted to 231 million. This population will continue to grow over time. Factors which basically influence to the increasing number of population are death rate (mortality), birth rate, and migration (mobility). This research discusses about the effect of those factors to the total population.

Multiple linear regression analysis is aimed to study relationship between a dependent variable with more than one independent variable. An estimation method is used for the ordinary least squares method. However, such a method in linear regression is very sensitive to outliers in the data. If this is the case, and an ordinary least squares estimation method used, this will generate an estimated parameter which is not appropriate for the data. This can result in significant error. Therefore, robust regression analysis can be used to improve it. We use robust M-estimator regression method using a weighting function Huber and Tukey bisquare. The case studies used were data of the number of people in Indonesia with a fertility rate and migration in 2010 regarded as variable. The results obtained contained residual value of the standard error by Tukey function smaller than that by Huber function. Thus, the best method of estimating the parameters or the regression coefficients in this study is a robust M-estimator regression method using Tukey bisquare weighting function.

Keywords: *multiple linear regression, Outliers, M-estimator Robust Regression*

INTRODUCTION

The population growth in Indonesia is so fast, Indonesia became the fourth country with the largest population in the world. As for the ASEAN region, Indonesia was ranked as the first. In 2007, the number of people in Indonesia amounted to 231 million. The population of this course will continue to grow over time. Factors increasing population is basically influenced by death (mortality), birth (birthrate) and migration (mobility). In looking at the relationship factors such population growth with a population can be performed using linear regression models.

Linear regression models are used to study the relationship between a dependent variable and more than one independent variable [5]. According to Montgomery and Peck [7], the estimation method that can be used in a linear regression analysis is the least squares method. In the least squares method, there are assumptions that must be satisfied, such as linearity, no autocorrelation, no multikollinearity, homoskedastisitas, and having normally distributed errors [6].

However, the method of least squares for linear regression model is very sensitive to outliers in the data. When there are outliers in the data, and a least squares estimation method is used, the results of parameter estimation will not provide accurate information for the data, as these will result in a significant value of error. Hence, we develop robust regression analysis to estimate the improvement in the least squares linear regression.

Robust regression analysis is an important method to analyse data containing outliers. The methods discussed in this paper use robust M-estimator (iteratively reweighted least squares) with a weighting function Huber and Tukey bisquare.

LITERATURE

Multiple Linear Regression Model

Linear regression models synchronize between a dependent variable and more than one independent variable. In a multiple linear regression model analysis, we aim to explain the dependent variable by using more than one independent variable. With k independent variables, this has the form [7] as,

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i \quad (1)$$

where :

y_i : dependent variable
$X_{i1}, X_{i2}, \ldots, X_{ik}$: independent variable values on the the i-th observation
$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$: regression parameters
ε_i : error at the i-th observation

Least Squares Method

The least squares method was proposed by Carl Friedrich Gauss, a Germany mathematician. With certain assumptions, the least squares method has some very interesting statistical properties that make it one of the most powerful analytical method and popular [6].

In general the method of least squares estimator is obtained by minimizing the residual/error ($\sum e_i^2$) (Montgomery and Peck, 1992).

$$\sum e_i^2 = \sum \left(y_i - \hat{y}_i\right)^2 = \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \quad (2)$$

where :

x_i : independent variable
$\hat{\beta}_0$: Estimated coefficient intercept
$\hat{\beta}$: regression coefficient estimates
e_i : residual (error)

To obtain the estimated value for each parameter, Eqn (2) is on a decline against each parameetr will be assessed later equated to zero. This results in

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \sum x_i y_i / n}{\sum x_i^2 - \left(\sum x_i\right)^2 / n} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \text{ or} \quad (3)$$

$$\hat{\beta} = \left(X'X\right)^{-1} X'y \quad (4)$$

Linear Regression Model Assumptions

The assumptions applied to the classical linear regression model [5] are:

a. Errors follow a normal distribution
   Error (ε_i) assumed to be identical, independent, following normal distribution with a mean of zero and constant variance or ordinary symbolized $\varepsilon_i \sim IID\ N\ (0, \sigma^2)$.
b. There are multicollinearity
   Multikolinearity is linear relationship between the complete or some or all of the independent variables of a regression model.
c. Homoskedastisitas
   Homoskedastisitas is the assumption that must be satisfied in linear regression, where each error (ε_i) has certain common variance (σ^2).
d. Non autocorrelation
   Autocorrelation or self correlation is one particular form of correlation in which the elements of disorder associated with an observation influenced by other observations. Or it can mean error (ε_i) did not correlate with any other error (ε_j) or $cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$ or so-called absence of autocorrelation.

Outlier

Outlier is an extreme observation. The absolute value of residual belonging to the outlier is much larger than the other ones. The outlier hence is a totally different data point from the other ones [7]. Sometimes outliers can provide information that can not be given by other data points, such as outliers arising from the unusual combination of circumstances which might be very important, and hence needs to be investigated further (Draper and Smith, 1998). Various rules are proposed to reject outliers, or to eliminate them from observational data, and then re-analyze the observational data. However, outlier rejection procedure is certainly not wise.

In general, new outliers are rejected if after being traced the result is incorrect such as entering the wrong size or analysis, inaccurate recording of data, and inaccurate measurement. If it turns out not as a result of such errors, a thorough investigation must be carried out (Montgomery and Peck, 1992). The problems that arise due to outliers are as follows:

a. A great error of the model form $E(e_i) \neq 0$.
b. Variations of the data will be larger.
c. Estimated interval will have a greater range.

Outliers can be determined by chart methods and some statistical tools. Graphical method is to create a graph between X and Y, or create a residual graph that plots residual (e_i) with $\hat{y}$. If the plot of the data between X and Y is stray far from the average of the mean X or Y, this then indicates the data contain outliers. If done residual plot (e_i) with extreme data $\hat{y}$ is found then it is likely that data is data outliers. Besides using a graph, there are several statistical tools that can be used as a reference to determine whether data outlier exists or not, that is Studentized Residual, DFFITS, and Cook's Distance.

a. *Studentized Residual*

Studentized residuals is studentized residual value which i-th observation is removed from the calculation. The formula is:

$$t_i = \frac{e_i}{\sqrt{S^2_{e(i)}(1-h_{ii})}} \quad i = 1, 2, \ldots, n \tag{5}$$

where

t_i : R-student for the data to-i
e_i : residual (error)
h_{ii} : diagonal elements of the hat matrix (H) with $H = X(X'X)^{-1}X$
$S^2_{e(i)}$: MSE (Mean Square Error) without observation i-th

If the value t_i has a value greater than t table so that it can be concluded there are observations that have the effect of significantly different from other observations to y in the model.

b. *DFFITS*

To assess the influence of a point to changes that occur with the use DFFITS the following formula:

$$DFFITS_i = t_i \left(\frac{h_{ii}}{1-h_{ii}} \right)^{\frac{1}{2}} \tag{6}$$

Called outliers when the value of $|\text{DFFITS}| > 1$ for a range of data that is small till medium and value $|\text{DFFITS}| > 2\sqrt{p/n}$ for a cluster of data is large, with $p = k + 1$, and n is the number of observations [8].

c. *Cook's Distance*

Cook's Distance (D) shows the difference between the value of the regression coefficient by incorporating the i-th observation and the regression coefficients without the i-th observation. Cook's distance is calculated by the following formula:

$$D_i = \frac{e_i}{s^2 p}\left[\frac{h_{ii}}{(1-h_{ii})^2} \right] \tag{7}$$

Where e_i is residual, p is the number of predictors k plus 1 $(k + 1)$. Observations with D_i great value enables the outliers. Statistically value *Di* can be compared with F table with $\alpha = 0{,}05$ and $df = p, n - p$.

Robust Regression M-Estimator

According to Chen [2], robust regression is an important tool to analyze data contaminated with outliers. The main objective is to provide a robust regression having stable results due to the presence of outliers. Robust Regression consists of five estimation methods, i.e. M-estimators, Least Median Square (LMS) -estimator, Least Trimmed Square (LTS) -estimator, S-estimator, and MM-estimator. M-estimator is a commonly used robust regression method. M-estimator is deemed well to estimate the parameters caused by outliers. In general, robust regression M-estimator is done by minimizing the objective function:

$$\min \sum_{i=1}^{n} \rho(e_i) = \min \sum_{i=1}^{n} \rho\left(y_i - \sum_{j=0}^{k} x_{ij}\beta_j \right) \tag{8}$$

To obtain a scale invariant at this estimator, it is usually done by completing

$$\min \sum_{i=1}^{n} \rho(u_i) = \min \sum_{i=1}^{n} \rho\left(\frac{e_i}{s} \right) = \min \sum_{i=1}^{n} \rho\left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta_j}{s} \right) \tag{9}$$

where s is the scale of robust estimation. S estimator that is often used is

$$s = \frac{median\{|\, e_i - median(e_i, \ldots, e_n)\,|\}}{0,6745}$$

To minimize Eqn (9), the first partial derivatives of ρ to the same $\beta_j$same with zero, then

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta_j}{s} \right) = 0,\ j = 0,1,\ldots,k \tag{10}$$

Where $\psi = \rho'$ and x_{ij} is the i-th observation of the j-th parameter. Definition and the weighting function $w(e_i^*) = \frac{\psi(e_i^*)}{e_i^*}$ and $w(e_i^*) = w_i$, so that Eqn (10) can be written

$$\sum_{i=1}^{n} x_{ij} w_i \left(y_i - \sum_{j=0}^{k} x_{ij}\beta_j \right) = 0 \tag{11}$$

In matrix notation Eqn (10) can be written as follows,

$$\boldsymbol{X^T W_0 X b = X^T W_0 Y} \tag{12}$$

where $\boldsymbol{W_0}$ is $n \times n$ diagonal matrix of weights. X is the independent variable matrix size $(n \times (p+1))$, b is the estimator of outlier values (β) and Y is the dependent variable matrix size $(n \times n)$. So that the robust regression estimator with M-Estimator (IRLS) for β is:

$$b_{l+1} = (X^T W_l X)^{-1}(X^T W_l Y) \tag{13}$$

The weighting function for the method M-estimator with Huber and Tukey bisquare functions are as follows

TABLE 1.

Method	Weighting Function	Interval
Huber	$w(e_i^*)\begin{cases} 1 \\ \frac{c}{\lvert e_i^* \rvert} \end{cases}$	$\lvert e_i^* \rvert \le c$ $\lvert e_i^* \rvert > c$ $c = 1{,}345$
Tukey Bisquare	$w(e_i^*)\begin{cases} \left[1 - \left(\frac{e_i^*}{c}\right)^2\right]^2 \\ 0 \end{cases}$	$\lvert e_i^* \rvert \le c$ $\lvert e_i^* \rvert > c$ $c = 4{,}685$

METHODOLOGY

Research Data

In this study, the research data is secondary data taken from the official website of BPS www.bps.go.id in 2010 with the retrieved data is data as variable X_1 fertility rate and migration of data entry as X_2, and the data on the number of people as the response variable (Y).

Stages of Research

In this study, there are several stages in analyzing the data, namely:

a. Determining the value of the parameter estimation using the least squares method.
b. Namely regression test classic assumption test the assumption of normality, non multicolinearity, non autocorrelation and homoskedastisitas, to detect irregularities classical regression assumptions or not.
c. If there is a deviation assumption of normality, do surveillance detection outliers (outliers).
d. Determining the value of the parameter estimation using robust regression M-estimator.
e. Determine the best model between functions bisquare Huber and Tukey function based on the value of the smallest standard error.

RESULTS AND DISCUSSION

Estimation Model with Least Squares Method

Estimated regression models were conducted to determine the linear relationship between the independent variables fertility rate (X_1) and in-migration (X_2) on the dependent variable is the number of residents (Y). Based on the analysis using software *R.3.0.2* output results was obtained as follows:

$$\hat{Y} = 6577e^{3} - 1781e^{3}X_1 + 3289X_2$$

The regression model shown above can not be defined as the best regression model. Therefore, we need to see whether it satisfies the assumptions of regression.

Assumptions of Regression Test

Assuming a linear regression test results are shown in the table below. Data processing used SPSS and R software:

TABLE 2.

No.	Assumption	Method	Result	Conclusion
1	Normality	*Kolmogorov Smirnov Test*	Test p-value obtauned by 2,2e-16 < α = 0.05	Violation Occurs Assumptions
2	Multicollinearity	ValueVIF$_j$ (*Varians Inflation Factor*)	Value VIF for X_1 and X_2 is 1,244	There is no assumption Violations
3	Homoskedastisitas	*Breusch-Pagan Test*	$\chi2_{(0.05,3)}$ = 7.81 BP value 6.7394 < 7.81 ($\chi2_{(0.05,3)}$), then H_0 is accepted	Assuming there is no violation
4	Autocorrelation	*Durbin Watson Test*	Value d_H=2,354 and p-value=0,809, then H_0 is accepted	Assuming there is no violation

From the above results for the assumption of normality test p-value obtained by 2,2e-16 <α = 0.05. This means that the null hypothesis is rejected so that we can conclude the residue assumption does not follow a normal distribution. Therefore, there is a violation of the normality assumption in regression models. To test the assumption of multicollinearity, VIFj calculation results for the variables X_1 and X_2 respectively at 1.244 which

is less than 10. Therefore, it can be concluded that the 95% confidence level, there are no multicollinearity between independent variables. As for the test homoskedastisitas obtain BP value 6.7394 <7.81 ($\chi 2_{(0.05,3)}$), this means that the null hypothesis is accepted so that we can conclude there is no breach homoskedastisitas significant assumptions. And for autocorrelation using Durbin Watson statistical test, the obtained value d_H = 2,354 are within the interval d_U=1,5770 and 4-d_U=2,423, then H_0 is accepted. Therefore, it can be concluded that the 95% confidence level there is no autocorrelation in the residuals of the linear regression model.

Outlier Detection (Outliers)

To check the cause of the violation classical regression assumptions about normality, do a graphical method for analyzing the residuals of the regression model. The method used is the residual graph vs. fitted, normal qq plot, scale location and the cook's distance.

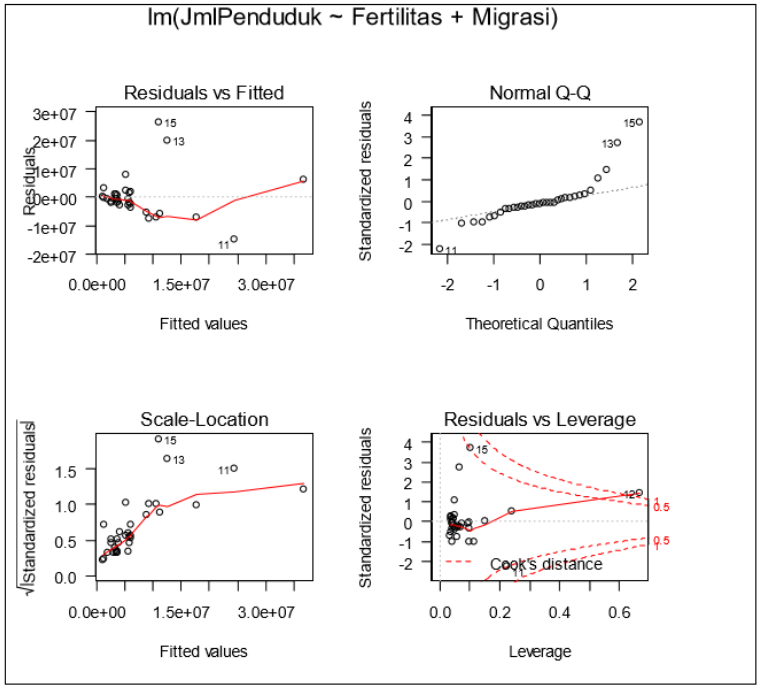

FIGURE 1.

- Residual vs fitted, Model valid if the dots spread around 0. The plot above shows the data is not entirely spread around 0, so that the model is not valid.
- Normal QQ plot, QQ plot of the picture above shows residual normal spread not because there is a point not spread around that line. Points 11, 13 and 15 is most likely an outlier of data.
- Scale-location, points 11, 13 and 15 had a great residual value, this is indicated with a point away from the line.
- Cook's distance, the cook's distance value is calculated by using leverage values and standardized residuals, and consider whether an observation is a strange observation / outstanding with respect to x and y. Cook's distance indicates the difference between the value of the regression coefficient by incorporating the i-th observation and the regression coefficients without the i-th observation. From the plot above shows that the observations 11, 12, and 15 very large impact on the regression line.

From the plot above, we see that the observations 11, 12, 13, and 15 may give you problems with the model. Province represented by these observations are: Jakarta, West Java, Central Java and East Java. So it can be concluded that the four provinces are observations that have a major influence on the regression model.

Robust Regression Estimates of M-Estimator

From the result of the detection of outliers above can be deduced that the observed data are outliers (outliers) or are the observational data that affect the model. The regression analysis for robust M-estimator to estimate the parameters contained outliers (outliers).

By using statistical software R, obtained regression analysis robust M-estimator with a Huber function is $\hat{Y} = -972669{,}1533 + 714832{,}8793\ X_1 + 29{,}0456\ X_2$, with the residual standard error of regression analysis 3122000. And robust M-estimator with a Tukey function Bisquare is $\hat{Y} = -4589554{,}6169 + 1718098{,}2933X_1 + 35{,}9072X_2$ with the residual standard error of 2984000.

Best Model

From the above results obtained standard value error residual for each masung function Huber and Tukey Bisquare is 3122000 and 2984000 in which the value of a standard error of residuals to function Tukey Bisquare smaller than the function Huber, so that a good method for use in estimating the parameters is using regression robust M-estimator with a Tukey Bisquare weighting function.

CONCLUSION

The value of the residual standard error of regression robust M-estimator with a Tukey function Bisquare smaller in the amount of 2984000 robust regression M-estimator with a Huber function that is equal to 3122000. Therefore, the best method to estimate the parameters or the regression coefficients of the factors affect the population in Indonesia in 2010 is a robust regression method M-estimator using Tukey Bisquare weighting function with the linear regression model is as follows:

$$\hat{Y} = -4589554,6169 + 1718098,2933X_1 + 35,9072X_2$$

From the above regression model, mean that every 1% increase in the fertility rate (X_1) will increase the population of 1,718,098.2933%. While in case of a 1% increase in-migration (X_2) will increase the 35.9072% of the population.

REFERENCE

1. Candraningtyas, Sherli., dkk. 2013. Regresi Robust MM-Estimator Untuk Penanganan Pencilan Pada Regresi Linear Berganda. *Jurnal Gaussian Vol.2 No.4 Tahun 2013, Hal. 395-404*. UNDIP.
2. Chen, C. 2002. *Robust Regression and Outlier Detection with the ROBUSTREG Procedure*. Paper 265-27. North Carolina: SAS Institute.
3. Draper, N.R., Smith, H. 1998. *Applied Regression Analysis*. USA: John Wiley & Sons Inc
4. Fox, J. 2002. *Robust Regression*. Appendix to An R and S-Plus Companion to Applied Regression.
5. Greene, W. H. 1951. *Econometric Analysis, 5th edition*. New York: Pearson Education.
6. Gujarati, D. 1997. *Ekonometrika Dasar*. Dra. Ak. Sumarno Zain, MBA, penerjemah. Jakarta: Erlangga. Terjemahan dari: Basic Econometrics.
7. Montgomery, D. C., Peck, E. A. 1992. *Introduction to Linear Regression Analysis*, 2nd edition. New York: John Wiley & Sons, Inc.
8. Neter. J., Wasserman, W., Kutner M. H. 1997. *Model Linear Terapan. Bambang Sumantri, penerjemah*. Jurusan Statistika FMIPA-IPB. Terjemahan dari: Applied Linear Model