JURNAL MATEMATIKA, STATISTIKA & KOMPUTASI E-ISSN: 2614-8811
https://journal.unhas.ac.id/index.php/jmsk/index
Vol. 18, No. 2, 251 -260, January, 2022
DOI: 10.20956/j.v18i2.18630

Comparison of M Estimation, S Estimation, with MM Estimation to Get the Best Estimation of Robust Regression in Criminal Cases in Indonesia

*1Malecita Nur Atala Singgih, *2Achmad Fauzan

Abstract

Crime incidents that occurred in Indonesia in 2019 based on Survey Based Data on criminal data sourced from the National Socio-Economic Survey and Village Potential Data Collection produced by the Central Statistics Agency recorded 269,324 cases. The high crime rate is caused by several factors, including poverty and population density. Determination of the most influential factors in criminal acts in Indonesia can be done with Regression Analysis. One method of Regression Analysis that is very commonly used is the Least Square Method. However, Regression Analysis can be used if the assumption test is met. If outliers are found, then the assumption test is not completed. The outlier problem can be overcome by using a robust estimation method. This study aims to determine the best estimation method between Maximum Likelihood Type (M) estimation, Scale (S) estimation, and Method of Moment (MM) estimation on Robust Regression. The best estimate of Robust Regression is the smallest Residual Standard Error (RSE) value and the largest Adjusted R-square. The analysis of case studies of criminal acts in Indonesia in 2019 showed that the best estimate was the S estimate with an RSE value of 4226 and an Adjusted R-square of 0.98

Keywords: Regression analysis, Robust regression, M Estimation, S Estimation, MM Estimation.

1. INTRODUCTION AND PRELIMINARIES

A crime is when a person commits an act that is prohibited and is against the law. Criminal acts can also be interpreted as all forms of prohibited actions and have been regulated in applicable law. Criminal acts have a broad scope, including immoral acts, corruption, fraud, persecution, theft, and so on. Anyone who violates the prohibition that has been regulated in applicable law can be threatened with a criminal [9].

Based on Survey Based Data on criminal data sourced from the National Socio-Economic Survey (SUSENAS) and the Village Potential Data Collection (PODES) produced by the Central Agency on Statistics Indonesia (BPS). The incidence of criminal acts that occurred in Indonesia in 2019 was recorded at 269,324 cases. Based on these data, out of 100,000 people in Indonesia, 103 of them are at risk of being hit by a crime; within 1 minute 57 seconds, there is one criminal act that occurred [4].

*Statistics Department, Faculty of Mathematics and Natural Science
Universitas Islam Indonesia
Email address: 1malecita.singgih@students.uii.ac.id, 2achmadfauzan@uii.ac.id

Looking at the incidence of criminal acts in Indonesia in 2019, the high number of criminal acts can be caused by several factors, including poverty and population density. In determining the most influential factors in criminal acts in Indonesia, it can be done with Regression Analysis. The Least Square Method is one of the most commonly used Regression Analysis methods [18]. Regression Analysis can be used if the assumption test is met, namely normality, homoscedasticity, no autocorrelation, and free of multicollinearity [8]. In some cases, regression analysis cannot be used to solve problems because outliers cause unfulfilled assumptions. The outlier problem can be overcome by using a robust estimation method [11]. Robust regression is a method used when the assumption test is not met and there are outliers. This method is very suitable to be used to analyze data that is affected by outliers to obtain a robust model or resistance to outliers [6]. The benefits of this research is to find out how to determine the factors that most influence crime in Indonesia in 2019. This can be a reference for the government to form policies to tackle criminal acts in Indonesia.

Several studies have compared the method of S estimation, LTS estimation, M estimation, and MM estimation on robust regression. Perihatini [12] conducted a comparative study of LTS estimation, S estimation, with M estimation for a case study of car financing at company "X" which aims to produce the best parameter estimation model seen from the Mean Square Error (MSE) and R^2 values. Widodo [19] compared the LTS estimate, the M estimate, with the MM estimate for a case study of farmer exchange rates. Comparison saw from the Residual Standard Error value.

Based on previous research, the methods used have advantages and disadvantages. Based on the characteristics of the data tested in this study, the authors conducted a study by comparing the M estimate, the S estimate, and the MM estimate on Robust Regression. The purpose of this research is to determine the best estimate to obtain the best model. The selection of the best estimate is based on the slightest Residual Standard Error (RSE) value and the largest R^2 value. The data will be processed using the help of the RStudio software

2. METHOD

This research was conducted at PT Kedata Indonesia Digital from January 18, 2021, to February 26, 2021. The data used is data on criminal acts in Indonesia in 2019. This data is survey-based secondary data sourced from the National Socio-Economics Survey (SUSENAS) and the Village Potential data collection (Podes) produced by the Central Statistics Agency (BPS). The variables used are the number of criminal acts (Y) with case units, the number of poor people (X_1) with a soul unit, and population density (X_2) with a person/km2 unit. This research was conducted using the Robust Regression analysis method to find out how to determine the factors that influence criminal acts in Indonesia in 2019 if the data contained outliers and assumptions were not met.

Fig 1 explains the research flow chart. The first thing to do in this research is to input data. The second step is the regression analysis is carried out using the Least Square Method. Furthermore, in the assumption test, if any assumptions are not met, outlier detection is carried out. If the data assumptions are not met, and there are outliers, proceed with Robust Regression analysis using M, S, and MM estimates. The three estimates are selected as the best estimate, and the best model is obtained

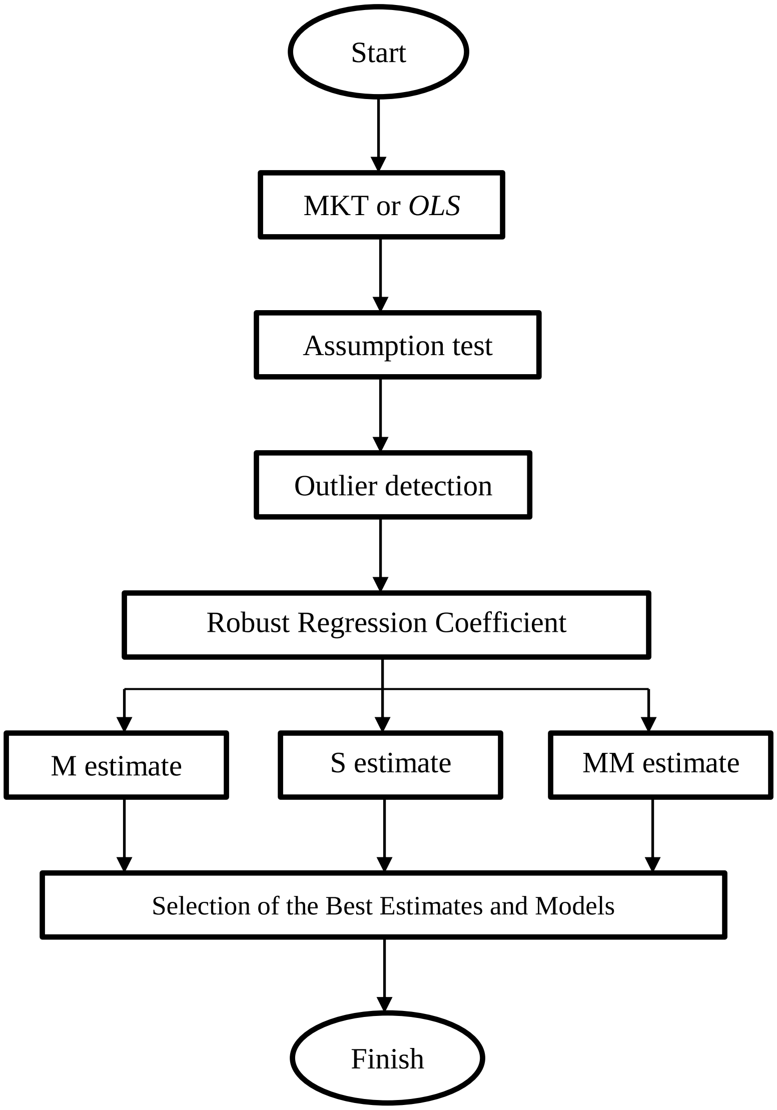

Fig. 1. Research flowchart.

2.1 Regression Analysis with Least Square Method

Ordinary Least Square (OLS) is an approach method for regression or equation formation in modeling, as well as measurement analysis in model validation [2]. In the Regression Analysis with the Least Squares Method, there are two validation tests: the overall and partial tests. The overall test or F test is used to determine whether the regression model is feasible or not to be used as a model [5]. In addition, the test is also used to determine simultaneously whether the independent variables are significant to the dependent variable. Reject H_0 if $p_{val} < \alpha$ means the model is feasible to use.

Partial test or t test is used to determine whether the independent variables have a significant effect on the model [20]. A partial test is also used to know whether the independent variables are significant to the dependent variable. Reject H_0 if $p_{val} < \alpha$ means that there is a partial influence of the independent variable on the dependent variable.

2.2 Assumption Test

The regression model obtained from OLS is a regression model with a regression coefficient that meets the characteristics of an unbiased linear estimator and the best, commonly referred to as the Best Linear Unbiased Estimator (BLUE) [1]. A normality test is conducted to test whether the regression model of the independent and dependent variables is normally distributed [7]. One of the methods used to test for normality is the Shapiro-Wilk test. This test has a good test power for small data samples or less than 50 [13]. The test statistic is formulated by Equation 1.

$$W = \frac{b^2}{(n-1)s^2} \tag{1}$$

$b^2 = \sum_{i=1}^{n/2} a_{n-i+1}(x_{(n-i+1)} - x_{(i)}$ and $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}$. If $p_{val} < 0.05$ or $w < w_{table}$ then rejected H_0 means that the data is not normally distributed. Then it must fail to reject H_0 so that the assumption test is met. A heteroscedasticity test is carried out to test whether the regression model has an inequality or similarity of residual variance from one observation to another observation [3]. The method used to test heteroscedasticity is the Breusch Pagan test. If $p_{val} <$ then reject H_0 means that heteroscedasticity occurs, then it must fail to reject H_0 so that the assumption is fulfilled.

Autocorrelation test is carried out to test the correlation between residuals in one observation and previous observations [15]. The method used to perform the autocorrelation test is Durbin-Watson. If the $p_{val} <$ then rejects H_0 meaning that there is autocorrelation in the residuals. For the assumption test to be fulfilled, the existing data obtained failed to reject H_0.

A multicollinearity test was conducted to see whether the independent variables had a significant relationship or not. One of the ways to determine the presence or absence of multicollinearity is by looking at the Variance Inflation Factor (VIF) value [17]. If the value of VIF < 10 fails to reject H_0, which means that there is no multicollinearity, so it can be said that the assumption is fulfilled.

2.3 Outlier detection

Outliers are data that does not follow the overall data pattern, or that does not follow the general pattern for the resulting regression model [16]. One outlier identification can be made using the Cook's Distance method, and the test statistic can be defined by Equation 2.

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

2.4 Robust Regression Analysis

Robust regression is an important tool for analyzing data affected by outlier data to produce a robust model or resistance to outliers. Robust regression aims to overcome deviations as an alternative to OLS [14]. There are three estimates used in Robust Regression Analysis. M estimation is a simple estimation method, both in theory and calculation. This M estimate can analyze the data assuming that most of the outliers detected are in the dependent variable. Estimation of M using Huber's weighting function [10].

Robust regression is essential for analyzing data affected by outlier data to produce a robust model or resistance to outliers. Robust regression aims to overcome deviations as an alternative to OLS [12]. MM estimation is a combination method between estimation with high breakdown point or estimation S with estimation M. This MM estimation has better performance than S estimate [21].

3. MAIN RESULTS

3.1. Descriptive statistics

The data obtained are the number of criminal acts (Y) with case units, the number of poor people (X_1) with a soul unit, and population density (X_2) with a person/km2 unit. The descriptive analysis is presented in Table 1.

Table 1. Descriptive analysis.

Variable	Statistical Measurement		
	Min	Average	Max
The number of criminal acts (case units)	718	7921	31934
The number of poor people (soul unit)	48780	739551	4112250
Population density (person/km2 unit)	9	742	15900

Based on Table 1, the average number of non-criminals in Indonesia is 7921 cases, with the highest case being 31934 cases and the lowest case being 718 cases. The average number of poor people is 739551 people, with the highest poor population being 4112250 people and the lowest poor being 48780 people. The average population density is 742 people/km2, with a maximum population density of 15900 people/km2 and a minimum population density of 9 people/km2.

3.2. Regression Analysis with Ordinary Least Square

Overall test is used for testing the feasibility of the model and testing the general parameters. The F test obtained p_{val} = 2.739×10^{-06} where this value is more significant than = 0.05 so that the conclusion obtained is that the model is feasible to use. The partial test (t-test) is used to determine whether the independent variable has a significant effect on the dependent variable.

Table 2. Partial test.

Coefficient	P_{val}	α
β_0 *(Intercept)*	0.002360	0.05
β_1	0.000235	0.05
β_2	6.03×10^{-05}	0.05

Based on Table 2, the p_{val} of the variables X_1(Poverty) and X_2(Population Density) are 0.000235 and 6.03×10^{-05}, respectively, less than $\alpha = 0.05$ so that it can be concluded that the variables X_1(Poverty) and X_2(Population Density) have an effect on significant to the variable Y (Criminal Act). Furthermore, the parameter estimation results for the Least Square Method will be obtained as shown in Table 3.

Table 3. Estimation of MKT Parameter.

Parameter	**Estimation value**	R^2
β₀ *(Intercept)*	3.886×10^{03}	0.5341
β₁	3.846×10^{-03}	
β₂	1.605	

Based on Table 3, the parameter estimation results will show the initial regression model using the Least Square Method defined in Equation 3.

$$\hat{Y} = 3886 + 0.003846X_1 + 1.605X_2 \tag{3}$$

The result of R^2 from the model is 0.5341, meaning that the independent variable X can explain the dependent variable Y in the model by 53.41%. In contrast, the rest is explained or influenced by other variables outside the model. The regression model obtained using OLS is said to meet the properties of an unbiased linear estimator. The best is also called the Best Linear Unbiased Estimator (BLUE) if the assumption test is completed.

When performing the analysis using the regression method, there are several assumption tests that must be met. They are normality test, homoscedasticity test, autocorrelation test, and multicollinearity test. Normality test can be done using the Shapiro-Wilk test. The normality test on the data obtained p_{val} = 3.305×10-05 where this value is greater than α= 0.05 so that the conclusion obtained is that the data is not normally distributed (the assumption is not met). Homoscedasticity test was carried out using the Breusch Pagan test. Homoscedasticity test obtained p_{val} = 0.004816 where this value is less than α = 0.05 so that the conclusion obtained is the assumption of residual homoscedasticity is not met (assumptions are not met). The autocorrelation test was performed using the Durbin-Watson test. Autocorrelation test obtained p_{val} = 0.004816 where this value is greater than α = 0.05 so that the conclusion obtained is that there is no autocorrelation (the assumption is satisfied).

Multicollinearity test by looking at the value of Variance Inflation Factor (VIF). The multicollinearity test obtained the VIF value of the variables X_1 and X_2 of 1,000667 where this value is less than 10 so that the conclusion obtained is that there is no multicollinearity (the assumption is fulfilled). The assumption test that has been carried out has several tests that are not met, namely the normality test and the homoscedasticity test.

The normality test and homoscedasticity test on the assumption test are not met, then outlier detection is carried out. Cook Distance (Cook's D) is used to measure the presence or absence of outliers, and the results are presented in Fig.2.

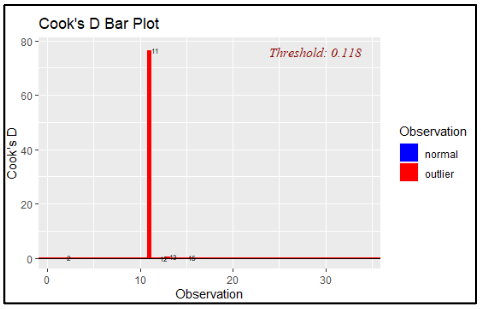

Fig. 2. Outlier detection.

Based on Fig. 2, it can be seen that there are five outliers in the data, namely the 2nd , 11th , 12th , 13th, and 15th observations.

3.3. Robust Regression

Robust Regression parameter estimation using M estimation is shown in Table 4.

Table 4. Estimating Parameter Estimation of M.

Parameter	Estimation value	*RSE*
β_0 *(Intercept)*	2.965089×10^{03}	
β_1	4.128021×10^{-03}	2422
β_2	1.692069	

From Table 4 it is known tha the value of RSE = 2422 and a Robust Regression model is obtained with an estimate of M defined in equation 4.

$$\hat{Y} = 2965.089 + 0.004128021X_1 + 1.692069X_2 \tag{4}$$

This regression model was obtained by convergent iteration 31 times. Equation 4 shows that if there is a one percent increase in X_1, then Y will increase by 0.004128021%, while if there is a one percent increase in X_2, then Y will increase by 1.692069%. Next, the approach for variables X_1 and X_2 is shown in Table 5.

Table 5. Procedure to variables X1 and X2.

Variable	P_{val}	α
X_1	1.175×10^{-06}	0.05
X_2	4.575×10^{-10}	0.05

Based on Table 5 the p_{val} of the variables X_1 and X_2, respectively 1.175×10^{-06} and 4.575×10^{-10}, is less than of $\alpha = 0.05$ so that it can be concluded that the variables X_1 and X_2 have a significant effect on the Y variable.

Robust Regression parameter estimation using S estimation is presented in Table 6.

Table 6. Estimating Parameter Estimation of *S*.

Parameter	Estimation value	*RSE*	$\bar{R}^2$
β_0 *(Intercept)*	1.708×10^{03}		
β_1	5.885×10^{-03}	2266	0.9826
β_2	1.764		

From Table 6, it is known that the value of R^2 = 0.9826 and the value of RSE = 2266, and the Robust Regression model is obtained with an estimate of S defined in Equation 5.

$$\hat{Y} = 1708 + 0.005885X_1 + 1.764X_2 \tag{5}$$

Equation 5 shows that if there is a one percent increase in X_1, then Y will increase by 0.005885%, while if there is a one percent increase in X_2, then Y will increase by 1.764%. Furthermore, the approach for variables X_1 and X_2 is presented in Table 7.

Table 7. Procedure to variables X_1 and X_2.

Variabel	*P-value*	α
X_1	2.03×10^{-15}	0.05
X_2	$<2 \times 10^{-16}$	0.05

Based on Table 7, the p_{val} of the variables X_1 and X_2, respectively 2.03×10^{-15} and < 2×10^{-16}, is less than $\alpha = 0.05$. It can be concluded that the variables X_1 and X_2 have a significant effect on the Y variable. Robust Regression parameter estimation using MM estimation is shown in Table 8.

Table 8. Estimating Parameter Estimation of MM.

Parameter	Estimation value	*RSE*	R^2
β_0 *(Intercept)*	2.171×10^{03}		
β_1	5.761×10^{-03}	2266	0.8986
β_2	1.722		

From Table 8, it is known that the value of R^2 = 0.8986 and the value of RSE = 2266, and the Robust Regression model is obtained with the MM estimate defined in the Equation. 6.

$$\hat{Y} = 2171 + 0.005761X_1 + 1.722X_2 \tag{6}$$

Equation 6 shows that if there is a one percent increase in X_1, then Y will increase by 0.005761%, while if there is a one percent increase in X_2, then Y will increase by 1.722%. Furthermore, the approach for variables X_1 and X_2 is presented in Table 9.

Table 9. Procedure to variables X_1 and X_2.

Variable	*P-value*	α
X_1	<2×10^{-16}	0.05
X_2	<2×10^{-16}	0.05

Based on Table 9, the p_{val} of the variables X_1 and X_2 of < 2×10^{-16} is less than $\alpha = 0.05$, it can be concluded that the variables X_1 and X_2 have a significant effect on the Y variable.

3.4. Selection of the Best Estimate

The best estimates are selected from the smallest RSE value, and the most significant R^2 value is presented in Table 10.

Table 10. Comparison of RSE and R^2 Each Estimate.

Estimation	Nilai *RSE*	R^2
M Estimation	2426	
S Estimation	2266	0.98
MM Estimation	2266	0.89

Based on Table 10, it can be concluded that the best estimate of Robust Regression is the S estimation with the regression model presented in Equation 7.

$$Y = 1708 + 0.005885\,X_1 + 1.764\,X_2 \tag{7}$$

Robust Regression Model with an estimated S value of R^2 of 0.98 means that the dependent variable Y can be explained by variable X in the model by 98% while the rest is explained or influenced by other variables outside the model.

4. CONCLUSION

This study was conducted to overcome the problem of regression analysis when the existing data assumptions are not met. There are data outliers by comparing the M estimate, the S estimate with the MM estimate from robust regression. Based on the analysis results, it is concluded that to determine the factors that most influence criminal acts in Indonesia in 2019 are to use the robust regression method because several test assumptions are not met, and there are outliers in the data. The robust Regression Model with S estimation is the best model. The R^2 value is 0.98, meaning that the dependent variable Y can be explained by variable X in the model by 98%. In contrast, the rest is explained or influenced by other variables outside the model.

ACKNOWLEDGEMENT

The researcher would like to thank PT Kedata Indonesia Digital for the opportunity and experience provided while carrying out the Practical Work at PT Kedata Indonesia Digital.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest

REFERENCES

[1] Algifari. 2000. *Analisis regresi: teori, kasus, dan solusi*. BPFE UGM.

[2] Andriani, D. P. 2014. Regresi linier berganda. *Brawijaya University*, 1–45.

[3] Andriani, S. 2017. Uji Park dan uji Breusch Pagan Godfrey dalam pendeteksian Heteroskedastisitas pada analisis Regresi. *Al-Jabar: Jurnal Pendidikan Matematika*, Vol 8, No.1, 63–72.

[4] BPS-Statistics Indonesia. 2020. *Criminal statistics 2020*. BPS-Statistics Indonesia.

[5] Briliant, E. H., & Kurniawan, M. H. S. (2019). Perbandingan Regresi Linier Berganda dan Regresi Buckley- James pada analisis survival data Tersensor Kanan. *Proceedings of The 1st STEEEM 2019*, Vol 1, No.1, 1–19.

[6] Chen, C. 2002. Statistics and Data Robust Regression and Outlier Detection with the ROBUSTREG Procedure. In *Statistics and Data Analysis* (Issue September).

[7] Ghozali, I. 2009. *Aplikasi analisis Multivariate dengan Program SPSS*. Badan Penerbit UNDIP.

[8] Mardiatmoko, G. 2020. Pentingnya uji Asumsi Klasik pada analisis Regresi Linier Berganda. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, Vol 14, No.3, 333–342.

[9] Maulana, A. 2020. *Hukum Online*. https://www.hukumonline.com/klinik/detail/ulasan/lt5236f79d8e4b4/mengenal-unsur-tindak-pidana-dan-syarat-pemenuhannya/. [Accessed April 25th, 2021

[10] Montgomery, D. C., Peck, E. A., & Vining, G. G. 2012. *Introduction to Linear Regression Analysis* (5th ed.). Wiley Series in Probability and Statistics.

[11] Nurdin, N., Raupong, & Islamiyati, A. 2014. Penggunaan Regresi Robust pada data yang mengandung pencilan dengan metode momen. *Jurnal Matematika, Statistika Dan Komputasi*, Vol 10, No.2, 115.

[12] Perihatini, D. I. 2018. *Perbandingan metode estimasi LTS , estimasi M , dan estimasi S pada Regresi Robust*. Universitas Islam Indonesia.

[13] Rini, D. S., & Faisal, F. 2015. Perbandingan Power of Test dari Uji Normalitas metode Bayesian, Uji Shapiro-Wilk, Uji Cramer-von Mises,dan Uji Anderson-Darling. *Jurnal Gradien*, Vol 11, No.2, 1–5.

[14] Safitri, D. 2015. *Perbandingan metode estimasi M dan estimasi MM (Method of Moment) pada Regresi Rrobust*. Universitas Islam Indonesia.

[15] Sari, I. M., Anugrah, R., & Nasir, A. 2020. Effect of Corporate Governance and Corporate Social Responsibility on Financial Performance. *Journal of Auditing, Finance, and Forensic Accounting*, Vol 8, No.22, 44–54.

[16] Seheult, A. H., Green, P. J., Rousseeuw, P. J., & Leroy, A. M. 1989. Robust regression and outlier detection. In *John Wiley & Sons*. John Wiley & Sons, Inc.

[17] Sriningsih, M., Hatidja, D., & Prang, J. D. 2018. Penanganan Multikolinearitas dengan menggunakan analisis Regresi Komponen Utama pada kasus impor beras di Provinsi Sulut. *Jurnal Ilmiah Sains*, Vol 18, No 1, 18.

[18] Tarno. 2007. Estimasi model Regresi Linier dengan metode Median Kuadrat Terkecil. *Jurnal Sains Dan Matematika (JSM)*, Vol 15, No.2, 69–72.

[19] Widodo, E., & Dewayanti, A. A. 2016. Perbandingan metode estimasi LTS, M, MM Pada Regresi robust. In *Direktorat Penelitian dan Pengabdian Masyarakat UII*.

[20] Widodo, E., Suriani, E., Putri, I., & Evi, G. 2019. Analisis regresi panel pada kasus kemiskinan di Indonesia. *PRISMA, Prosiding Seminar Nasional Penelitian*, Vol 2, 710–717.

[21] Zulkarnain, A., Setyo Wira, R., & Perdana, H. 2020. Analisis Regresi Robust estimasi-MM dalam mengatasi Pencilan pada Regresi Linear Berganda. *Bimaster : Buletin Ilmiah Matematika, Statistika Dan Terapannya*, Vol 9, No.1, 123–128.