

SIMULATION STUDY OF ROBUST REGULATORS PARAMETER AND SIDER-S ROBUST PARAMETERS IN LONGITUDINAL DATA WITH A VARIETY OF LEVELS

Waego Hadi Nugroho¹, Ni Wayan Suryawardhani², Adji Achmad Rinaldo Fernandes³

Abstract---Robust regression is used to obtain the right model when the data contains outliers and are not normally distributed. Robust regression has several kinds of estimators, one of which is using M-estimator and S-estimator. The M-estimator robust regression is the simplest approach both computationally and theoretically while the S-estimator is an estimator that has a high breakdown point for estimating error scales. This study wants to find out the comparison of M-estimator and S-estimator robust estimation regression that is more efficient by comparing the variance between estimators using relative efficiency in longitudinal simulation data. The results showed that the model with parameter estimators obtained from the S-estimator robust regression method was more effectively used to predict malnutrition in East Java Regency / City in 2013 - 2018 compared to the M-estimator robust regression method.

Keywords---Longitudinal Data, Robust Regression, Outliers, M-Estimator, S-Estimator

I. INTRODUCTION

Regression analysis is a statistical method used to determine the relationship between predictor variables and response variables (Draper and Smith, 1992). There are three data that can be used in regression analysis, namely cross-section, time-series and longitudinal. Longitudinal data is data obtained from observations that are of concern more than once at different times (Danardono, 2018).

The robust regression method is used to obtain the right model when the data contains outliers and are not normally distributed. Robust regression has several kinds of estimators, one of which is using M-estimator and S-estimator. M-estimator robust regression was introduced by Huber in 1973, this estimator is the simplest approach both computationally and theoretically. According to Drapper and Smith (1992), M-estimator robust regression is a type of maximum likelihood estimator. While the S-estimator robust regression, was introduced by Rousseeuw and Yohai in 1984. The S-estimator is an estimator that has a high breakdown point for estimating error scales (Hutahayan *et al*, 2019).

The M-estimator is recommended for simulation data because it provides the smallest Total Absolute Bias (TAB) and Total Mean Square Error (TMSE) compared to other estimators in robust regression, while the S-estimator is the best model for corn production data in Indonesia in 2011 compared to M-estimators (Susanti, et al. 2014).

¹University of Brawijaya, Malang, Indonesia, Email: waego@ub.ac.id

In this study, we want to know the comparison of M-estimator and S-estimator robust regression that is more efficient by comparing the variance between estimators using relative efficiency. A good parameter estimator is estimator which has a small variety. In contrast to previous studies, this study wanted to find out the efficiency of the M-estimator and the S-estimator on longitudinal simulation data. Based on research by Mentari (2019), this study wants to modify the conditions of the regression coefficient and outliers level. Modified outliers rates are 0%, 3% and 5% on longitudinal data on the effect of malnutrition on 3 variables, namely, percentage of misin (%), population density (/ km²) and number of health facilities (units) in districts / cities located in the Province of East Java in 2013 to 2018.

II. METHODOLOGY

The data used in this study, namely simulation data based on actual data. Simulation data meet the following criteria.

- 1) The actual data used are data from the Central Statistics Agency of East Java Province and the Department of Health of East Java Province about malnutrition in 2013 to 2018.
- 2) The data are Malang City, Batu City, Blitar City, Pasuruan City, Pasuruan Regency and Lumajang Regency. That city / regency can represent East Java Province.
- 3) Residual values are obtained from normal distribution trip generation data based on outliers level
- 4) Perform calculations to get a new response variable

Robust Regression

Robust regression is a regression method used when there are outliers of observational data that affect the model (Lainun, et al, 2018). The main purpose of robust regression is to obtain a consistent parameter estimate on the data containing outliers. Robust regression has the advantage of a fairly simple calculation using iteration to obtain the minimum parameter estimator (Fernandes *et al*, 2015).

Regresi Robust Penduga-M

The M-estimator was first introduced as a result of the approach of a robust least-squared estimator (Huber, 1973). The M-estimator is one of the most commonly used methods and is considered good for estimating parameters caused by outliers (Fernandes *et al*, 2019).. M-estimator regression is a Maximum likelihood type estimator. The robust-M estimator equation is as follows (Montgomery and Peck, 1992).

$$\min \sum_{i=1}^N \sum_{t=1}^T \rho(u_{it}) = \min \sum_{i=1}^N \sum_{t=1}^T \rho\left(\frac{e_{it}}{\sigma}\right) = \min \sum_{i=1}^N \sum_{t=1}^T \rho\left(\frac{y_{it} - \hat{y}_{it}}{\sigma}\right) \quad (2.14)$$

when: $i=1,2,3,\dots,N$; $t = 1,2,3,\dots,T$

Calculates robust-M regression estimators using iterations of weighted MKT or IRLS (Iteratively Reweighted Least Square). Following are the steps to calculate robust-M regression.

Calculate the value of the estimator using MKT, then calculate the residual (e_{it}).

Calculates the initial s value and initial weighting.

Calculates $\hat{\beta}_m$ using the least weighted square based on the weighted value W_i .

$$\hat{\beta}_m = (X'WX)^{-1}X'Wy$$

Change the parameter estimator in step 3 as

$\hat{\beta}_0$ in step 1 and get a new residual, new s value and a pombobot.

Repeating steps 2 and 3 repeatedly to converge. End of iteration the difference between the amount of residual muklak with the absolute amount of residual iteration previously less than

$$1 \times 10^{-6}, \quad \text{atau} \quad \|\hat{\beta}_{(k)} - \hat{\beta}_{(k-1)}\| < 1 \times 10^{-6}, \quad \text{dengan} \quad \|\hat{\beta}_{(k)} - \hat{\beta}_{(k-1)}\| = \sqrt{\sum_{j=0}^{p-1} (\hat{\beta}_{i(k)} - \hat{\beta}_{i(k-1)})^2},$$

where k is the iteration index and the maximum iteration specified is 10. Tukey's Bisquare Weighting has a constant value or tuning constant (c) of 4,685 which is owned by each weighter (Kutner et al., 2004).

S-estimator Robust Regression

Rousseeuw and Yohai (1984) were the first people to introduce S-estimator robust regression with a breakdown point that reached up to 50%. The steps taken in estimating parameters with the S-estimator are:

Menaksirkan Estimate the initial β that is $\hat{\beta}_{(0)}$ using the least squares method.

Calculating the residual value $e_{it} = y_{it} - \hat{y}_{it}$

Calculates the value of $u_{it} = \frac{e_{it}}{\hat{\sigma}_{it}}$

Calculate the weighting value (W_{it}) using the Tukey's bisquare weighting function with a constant tuning value $c = 1.547$ so that a breakdown point of 50% is obtained

$$w_{it} = \frac{\psi(u_{it})}{u_{it}} = \begin{cases} \left(1 - \left(\frac{u_{it}}{c}\right)^2\right)^2, & \text{iterasi} = 1 \\ \frac{\rho(u)}{u^2}, & \text{iterasi} > 1 \end{cases}$$

Calculate $\hat{\beta}_s$ using the least weighted square based on the weighting value W_i .

$$\hat{\beta}_s = (X'WX)^{-1}X'Wy$$

Repeat steps 1-4 until you get the converging $\hat{\beta}_s$. End of iteration difference between the amount of residual muklak with the absolute number of residual iterations of less than 1×10^{-6} , or $\|\hat{\beta}_{(k)} - \hat{\beta}_{(k-1)}\| < 1 \times 10^{-6}$, when

$$\|\hat{\beta}_{(k)} - \hat{\beta}_{(k-1)}\| = \sqrt{\sum_{j=0}^{p-1} (\hat{\beta}_{i(k)} - \hat{\beta}_{i(k-1)})^2}$$

where k is the iteration index and the maximum iteration specified is 10. Tukey's Bisquare Weighting has a constant value or tuning constant (c) of 4,685 which is owned by each weighter (Kutner et al., 2004).

Outlier Detection

Outliers provide information that cannot be explained by other observations, so outliers detection is needed in order to determine an observation can be categorized as outliers. Outlier detection uses TRES values (Fernandes *et al*, 2018). TRES is used to identify outliers on response variables (Cousineau and Chartier, 2010). The following hypothesis is used.

H0: The i-th observation is not outlier

H1: The i-th observation is outlier

$$TRES_{it} = \varepsilon_{it} \left[\frac{NT - Np - 2}{JKG(1 - h_{it}) - \varepsilon_{it}^2} \right]^{\frac{1}{2}}$$

$$JKG = \sum_{i=1}^N \sum_{t=1}^T (Y_{it} - \hat{Y}_{it})^2$$

$$|TRES_{it}| = \begin{cases} \leq t_{N(T-k-2)}^{\alpha/2} ; H_0 \text{ accepted} \\ > t_{N(T-k-2)}^{\alpha/2} ; H_1 \text{ rejected} \end{cases}$$

If the value of $|TRES_{it}|$ get the results received H_0 , it can be concluded that the observation to the t-period is not an outlier. Conversely, if the result of rejecting H_0 , it can be concluded that the i-th observation is outlier.

Relative Efficiency

Relative efficiency is used to compare two predictors. The relative efficient comparison of two estimators (bS) relative to (bM) can be defined as follows (Wackerly, et al., 2008).

$$eff(\hat{\beta}_{i,M}, \hat{\beta}_{i,S}) = \frac{V(\hat{\beta}_{i,S})}{V(\hat{\beta}_{i,M})}$$

If $eff > 1$, then it can be stated that the estimator $\hat{\beta}_{i,M}$ is an unbiased estimator than estimator $\hat{\beta}_{i,S}$. In another hand, if $eff < 1$, then it can be stated that the estimator is an unbiased estimator than the $\hat{\beta}_{i,M}$ estimator because it has a smaller variety.

III. RESULTS AND DISCUSSION

Simulation Function of Secondary Data

In the study, simulation functions are obtained from actual data. The actual data used are from the Central Statistics Agency of East Java Province and the Department of Health of East Java Province regarding malnutrition in 2013 to 2018.

Outlier detection

Outliers detection in this study there are three outliers conditions, namely 0%, 3% and 5%. The data generated are residual values using the normal distribution. Based on the generation data, will get a new response variable value. The new response variable value is used for detection of outliers using TRES detection. The TRES detection results are compared with the critical point value $t_{((t-p-1))}^{\alpha/2}$. The research hypothesis is used as follows.

H_0 : The i-th observation is not an outlier Vs

H_1 : The i-th observation is outlier

The results of TRES detection obtained outliers at each level of outliers. The number of outliers can be seen in Appendix 9. Based on the results of the detection of TRES a robust regression method is needed to handle the outliers problem in the data. Outlier detection uses the M-estimator and the S-estimator to obtain a more efficient method.

Estimating Parameters and Testing the M-Regulator Robust Regression Hypothesis

M-estimator robust regression analysis using the IRLS procedure. IRLS uses MKT estimator parameters which are then iterated across all regression slope of each data until robust parameter estimation is obtained (Fernandes *et al*, 2014). The weighting used on IRLS is the Tukey's Bisquare weighting which is always updated with each iteration.

Calculation of parameter estimation and hypothesis testing based on outliers using R software, with the following results.

0% Outliers

The outreach rate of 0% gets a new model to explain the amount of malnutrition in Malang City, Batu City, Blitar Regency, Pasuruan City, Pasuruan Regency and Lumajang Regency, East Java from 2013 to 2018 with a total determination coefficient of 99%. The results obtained can explain that the level of malnutrition is influenced by the level of population poverty, the level of population density and health facilities by 99%. While 1% can be explained by other variables that are not contained in the model.

3% Outliers

Outlier level of 3% with a coefficient of determination of 98% explains that the level of malnutrition in Malang City, Batu City, Blitar Regency, Pasuruan City, Pasuruan Regency and Lumajang Regency in 2013 to 2018 were influenced by the level of population poverty, the level of population density and facilities health. Whereas 2% can be explained by other variables not contained in the model.

5% Outliers

An outreach rate of 5% can explain 98% of the level of malnutrition affected by poverty levels, population density and the level of health facilities in Malang City, Batu City, Blitar Regency, Pasuruan City, Pasuruan Regency and Lumajang Regency in 2013 to 2018.

Testing of M-estimator Robust Regression Hypothesis

Parameter testing is performed after getting the M-estimator robust regression results. Simultaneous test and partial test are used to test the results of M-estimator robust regression. Following the testing of the M-estimator robust regression hypothesis based on the level of outliers.

0% outliers

The M-estimator robust regression results with the help of R software, can be concluded as follows.

Regency / City in East Java in 2013 to 2018 was declared significant because the value of F table (2.901) < Fcount (163.87). From these results it can be concluded that overall there is an influence between the level of malnutrition with the level of poverty, the level of population density and the level of health facilities.

Partial test results with a table value (1.69) on the parameters can be concluded that:

In Pasuruan Regency, the level of health facilities did not significantly influence the level of malnutrition in 2013 to 2018 compared to other variables.

All variables, namely poverty level, population density and health facility level, have a significant effect on the level of malnutrition in Malang City, Batu City, Blitar Regency, Pasuruan City and Lumajang Regency from 2013 to 2018 on the level of malnutrition.

3% Outliers

The M-estimator robust regression results with the help of R software, can be concluded as follows.

Based on the simultaneous test states that $F_{table} (2.901) < F_{calculate} (76.66)$ which means it has a significant effect. Simultaneous test results can be concluded that districts / cities in East Java in 2013 to 2018 as a whole have an influence between the level of malnutrition with the level of poverty, the level of population density and the level of health facilities.

The partial test results with the table value (1.69) on the parameters can be concluded that the Regency / City in East Java represented by Malang City, Batu City, Blitar Regency, Pasuruan Regency, Pasuruan City and Lumajang Regency all variables have significant effect on the level malnutrition except in Pasuruan Regency the level of health facilities does not affect the level of malnutrition. And in the Lumajang Regency the poverty level did not affect the level of malnutrition in 2013 to 2018.

5% Outliers

The M-estimator robust regression results with the help of R software, can be concluded as follows.

District / City in East Java in 2013 to 2018 overall there was an influence between the level of poverty, the level of population density and the level of health facilities on the level of malnutrition. From these conclusions, based on simultaneous tests stated that $F_{table} (2.901) < F_{calculate} (52.99)$, which means significantly.

Partial test results with a table value (1.69) on the parameters can be concluded that:

In Malang City, Batu City and Blitar Regency, all variables significantly influence the level of malnutrition in 2013 to 2018.

In Pasuruan City the poverty level did not affect the level of malnutrition in 2013 to 2018.

Same is the case with Pasuruan Regency but the level of population density also does not affect the level of malnutrition in 2013 to 2018.

Inversely related to Lumajang Regency, there are no variables that affect the level of malnutrition in 2013 to 2018.

Parameter Estimation and Testing Hypothesis of the S-Estimator Robust Regression

The S-estimator is based on the residual scale of the M-estimator. Robust regression analysis using S-assignment can use iteration of weighted OLS or IRLS (Fernandes *et al*, 2014). IRLS uses OLS parameter estimator, which is then iterated across all regression slope of each malnutrition data until Robust parameter estimation is obtained. Using R software, the estimation results for malnutrition data in Malang City in East Java from 2013 to 2018 will be obtained with various outline levels as follows.

0% Outliers

Outlier level of 0% can explain 99% of the level of malnutrition which is influenced by poverty level, population density and the level of health facilities in the districts / cities in East Java from 2013 to 2018 represented by Malang City, Batu City, Blitar Regency, Pasuruan City and Lumajang Regency. Whereas 1% can be explained by other factors.

3% Outliers

Outlier level 3% The S-estimator robust regression results were only able to explain 98% of the effects of poverty levels, population density and health facility levels on levels of malnutrition in the Districts / Cities, East Java in 2013 to 2018 represented by Malang City, Batu City, Blitar Regency, Pasuruan City, Pasuruan Regency and Lumajang Regency. While other influences can explain 2% of the level of malnutrition.

5% Outliers

Outlier level 5% The S-estimator robust regression results explain 99% of the poverty rate, population density and health facility levels against malnutrition levels from 2013 to 2018 in districts / cities in East Java. Meanwhile, 1% is explained by other influences.

Testing Hypothesis of the S-Estimator Robust Regression

After obtaining the S-estimator robust regression data, then re-testing the parameters, namely simultaneous test and partial test of the S-estimator robust regression results based on the outlier level as follows.

0% Outliers

The results of S-estimator robust regression with the help of R software, can be concluded as follows.

Regencies / cities in East Java Province from 2013 to 2018 overall did not have an effect between poverty level, population density and health facility level on the level of malnutrition. From these conclusions, based on simultaneous tests stated that $F_{table} (2.901) < F_{count} (18.17)$ which means significantly influence.

Partial test results with a table value (1.69) on the parameters can be concluded that:

In Malang City, Batu City, Pasuruan City and Pasuruan Regency, all variables affect the level of malnutrition in 2013 to 2018.

Whereas the best comparison with Blitar and Lumajang Regencies, the level of poverty and the level of health facilities did not affect the level of malnutrition in 2013 to 2018.

3% Outliers

The results of S-estimator robust regression with the help of R software, can be concluded as follows.

Regencies / cities in East Java from 2013 to 2018 overall did not have an influence between poverty level, population density and the level of health facilities on the level of malnutrition. From these conclusions, based on simultaneous tests stated that $F_{table} (2.901) < F_{calculate} (80.07)$, which means significantly influence.

Partial test results with a table value (1.69) on the parameters can be concluded that:

Malang City, Batu City, Blitar Regency, Pasuruan City and Lumajang Regency on each variable affected the level of malnutrition in 2013 to 2018.

Pasuruan Regency, there is one independent variable, namely the level of health facilities does not affect the level of malnutrition in 2013 to 2018.

5% Outliers

The results of S-estimator robust regression with the help of R software, can be concluded as follows.

There are no independent variables that will affect the level of malnutrition in the Regency / City, East Java represented by Malang City, Batu City, Blitar Regency, Pasuruan City, Pasuruan Regency and Lumajang Regency. Value F table (2.901) < Fcalculate (25.30) which means that it has a significant effect.

Partial test results with a table value (1.69) on the parameters can be concluded that:

Only Pasuruan Regency on the variable level of health facilities does not affect the variable level of malnutrition in 2013 to 2018.

Whereas in Malang City, Batu City, Blitar Regency, Pasuruan City and Lumajang Regency each independent variable can explain the effect on the level of malnutrition in 2013 to 2018.

Relative Efficiency

By using R software, the value of the relative efficiency of the model is obtained based on the level of the M-robust estimator and S-robust estimator regression in malnutrition data in East Java Regency / City represented by Malang City, Batu City, Blitar Regency, Pasuruan City, Pasuruan Regency and Lumajang Regency from 2013 to 2018 are as follows.

Outcome level of 0% gets the value of relative efficiency between the M-estimator and the S-estimator of 1.88. Because the relative efficiency value is more than 1, it can be concluded that the S-estimator is more effective than the M-estimator.

Outlier level of 3% has an S-estimator more efficient than an M-estimator. Because the relative efficiency value obtained is 1.04.

Not much different results obtained for outliers level of 5%, namely the S-estimator is more effective than the M-estimator. This is supported by the calculation of the value of relative efficiency to get results of 4.14 which is more than one.

Based on the comparison of the M-estimator value and the S-estimator value, the M-estimator produces a middle squared error value that is greater than the S-estimator value. The M-estimator uses a constant tuning value (c) = 4.685 so that an efficiency of 95% is obtained. While the S-estimator uses a constant tuning value (c) = 1.547 to obtain a breakdown point of 50%. Based on this, if the M-estimator is compared with the S-estimator then the S-estimator is more efficiently slammed with the M-estimator.

IV. CONCLUSION AND SUGGESTIONS

Conclusion

Based on the results of the analysis that has been done, it can be concluded that, the model with parameter estimators obtained from the S-estimator robust regression method is more effective to be used to predict malnutrition in East Java Regency / City from 2013 to 2018 compared to the M-robust estimator regression method . The results of the

calculation of the relative efficiency of the M-estimator and the S-estimator, support the S-estimator more effectively because the value obtained is more than one.

V. Suggestion

Based on the results of the analysis that has been done, it can be concluded that, the model with parameter estimators obtained from the S-estimator robust regression method is more effective to be used to predict malnutrition in East Java Regency / City from 2013 to 2018 compared to the M-robust estimator regression method . The results of the calculation of the relative efficiency of the M-estimator and the S-estimator, support the S-estimator more effectively because the value obtained is more than one.

REFERENCES

- [1] BPS Jatim. 2018. Badan Pusat Statistik Jawa Timur. Diakses di <http://www.bpsjatim.com/> (diakses 1 Oktober 2019).
- [2] Cousineau, D. dan Chartier, S. 2010. Outlier detection and treatment. *International Journal of Psychological Research*, 3(1), 58–67.
- [3] Danardono. 2018. *Analisis Data Longitudinal*. Yogyakarta: Gadjah Mada University Press
- [4] Drapper, N. dan H. Smith. 1992. *Analisis Regresi Terapan (Edisi Kedua)*. Jakarta: Gramedia Pustaka Utama.
- [5] Fernandes, A.A.R.,Nyoman Budiantara, I.,Otok, B.W.,Suhartono, (2014), “Spline estimator for bi-responses nonparametric regression model for longitudinal data ”, *Applied Mathematical Sciences*, Vol 8 No 114, pp 5653-5665.
- [6] Fernandes, A.A.R.,Nyoman Budiantara, I.,Otok, B.W.,Suhartono, (2014), “Reproducing Kernel Hilbert space for penalized regression multi-predictors: Case in longitudinal data”, *International Journal of Mathematical Analysis*, Vol 8 No 40, pp 1951-1961.
- [7] Fernandes, A.A.R, Budiantara, I.N, Otok, B.W., and Suhartono. (2015). “Spline Estimator for Bi-Responses and Multi-Predictors Nonparametric Regression Model in Case of Longitudinal Data”, *Journal of Mathematics and Statistics*, Vol 11, No 2, pp. 61-69.
- [8] Fernandes, A.A.R, Jansen, P., Sa’adah, U, Solimun, Nurdjannah, Amaliana, L., and Efendi, A. (2018). “Comparison of Spline Estimator at Various Levels of Autocorrelation in Smoothing Spline Nonparametric Regression For Longitudinal Data”, *Communications in Statistics – Theory and Methods*, Vol 46, No 24, pp 12401-12424.
- [9] Fernandes, A.A.R., Hutahayan, B., Solimun, Arisoesilarningsih, E., Yanti, I., Astuti, A.B., Nurjannah, & Amaliana, L, (2019), “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.
- [10] Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1(5), 799-821.
- [11] Hutahayan, B., Solimun, Fernandes, A.A.R., Arisoesilarningsih, E., Yanti, I., Astuti, A.B., Nurjannah, & Amaliana, L, (2019), “Mixed Second Order Indicator Model: The First Order Using Principal Component Analysis and The Second Order Using Factor Analysis”, *IOP Conference Series: Materials Science and Engineering*, Forthcoming Issue.
- [12] Kutner, M. H., Nachtsheim, C. J., Neter, J. dan Li, W. 2004. *Applied Linier Regression Models*. Fifth Edition. McGraw-Hill Companies, Inc. New York.
- [13] Lainun, H., Tinungki, G. M., & Amran, A. (2018). Perbandingan Penduga M, S, dan MM pada Regresi Linier dalam Menangani Keberadaan Outlier. *Jurnal Matematika, Statistika dan Komputasi*, 15(1), 88-96.
- [14] Mentari, H. W. 2019. *Pendugaan Parameter Analisis Regresi Robust Penduga-M dan Penduga-S Pada Data Simulasi dengan Berbagai Tingkat Pencilan*. Skripsi: Universitas Brawijaya.
- [15] Rousseeuw, P. and Yohai, V. 1984. *Robust regression by means of S-estimators*. In *Robust and nonlinear time series analysis* (pp. 256-272). Springer. New York, NY.
- [16] Susanti, Y., Pratiwi, H., Sulistijowati, H., and Liana, T. 2014. *M Estimation, S Estimation, and MM Estimation In Robust Regression*. *International Journal of Pure and Applied Mathematics*. 91(3), 349-360.
- [17] Wackerly, D., et al. 2008. *Mathematical Statistics with Applications 7th issue*. Thomson Brooks/Cole. Florida.

