J-KOMA Journal of Computer Science and Applications

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

Open Access

***E-ISSN**: 2620 - 4872*
***Vol**.08, **No**.02*
***Doi**: 10.21009/JKOMA.082.01*

***Receivied**: 11 Dec 2025*
***Accepted**: 18 Dec 2025*
***Published**: 26 Dec 2025*

Keywords:
Robust Regression;
Outliers;
M-Estimations;
Least Trimmed Squares;
Poverty Modeling;

Correspondence Email:
cinta@iteba.ac.id

Outlier Handling in Applied Regression: Performance Comparison Between Least Trimmed Squares and Maximum Likelihood-Type Estimators

Cinta Rizki Oktarina[1*], Andini Setyo Anggraeni[2], Muhammad Arib Alwansyah[3], Reza Pahlepi[4]

[1,2] Department of Mathematics, Faculty of Information Technology, Institut Teknologi Batam, Indonesia

[3] Department of Mathematics Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Jakarta, Indonesia

[4] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Bengkulu, Indonesia

Abstract

Poverty analysis often relies on regression models whose performance can deteriorate in the presence of outliers, leading to biased estimates and unreliable conclusions. This study aims to evaluate the effectiveness of robust regression methods compared with Ordinary Least Squares (OLS) when modeling poverty levels across 154 regions in Sumatra. Four socioeconomic indicators were used as predictors, and outlier detection was conducted using the DFFITS approach. After identifying deviations from normality and the presence of influential observations, two robust estimation techniques M-estimation and Least Trimmed Squares (LTS) were applied to improve model stability. The results show that while all predictors significantly influence poverty, the LTS estimator provides the most accurate and robust performance, yielding the smallest Mean Squared Error (MSE) and an R-squared value of 53.37%. These findings demonstrate that LTS is better suited than OLS and M-estimation for handling data contamination and offers a more reliable approach for modeling poverty determinants.

INTRODUCTION

Poverty is one of the key indicators used to assess the success of development. Information on poverty is essential for the government in formulating policies aimed at poverty alleviation. Badan Pusat Statistik uses the *basic needs approach* to measure poverty. In this approach, poverty is viewed as an economic inability to meet basic food and non-food needs, which is assessed through household expenditure. A higher level of purchasing power reflects an increased ability of households to meet their essential needs, which in turn leads to improved social welfare. The poverty indicator measured as the percentage of the poor population is determined based on average per capita expenditure. Individuals whose average monthly per capita expenditure falls below the poverty line are categorized as poor [1].

Understanding poverty dynamics necessitates statistically reliable analytical tools, particularly when identifying socioeconomic determinants that contribute to variations in poverty levels. In practice, poverty-related datasets frequently exhibit deviations from classical regression assumptions, including the presence of extreme observations, non-normal error structures, and potential measurement inaccuracies. Such irregularities can severely compromise the performance of conventional estimation techniques like Ordinary Least Squares (OLS), resulting in biased parameter estimates, misleading inferential outcomes, and ultimately, ineffective policy interventions [2].

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

Robust regression serves as an alternative to the Ordinary Least Squares (OLS) method when the fundamental assumptions of linear regression particularly the normality of residuals are violated or when the data contain outliers that substantially distort model estimates [3]. The presence of such atypical observations can lead to biased parameter estimates, unreliable inference, and diminished predictive accuracy. Consequently, there is a methodological need for estimation procedures that maintain stability even under data contamination. Robust regression techniques are specifically designed to detect and mitigate the influence of outliers, thereby producing models that remain representative of the underlying data structure despite deviations from ideal conditions [4].

Among the most widely applied approaches in robust regression is M-estimation, which employs a weighting function to reduce the impact of outlying observations and yield more stable parameter estimates [5]. Another prominent method is the Least Trimmed Squares (LTS) estimator, which shares conceptual similarities with the least squares approach but operates on a subset of the data of size $h < n$ that yields the smallest sum of squared residuals [6]. By focusing on this selected subset, the LTS estimator effectively limits the influence of extreme values and enhances the robustness of the resulting model. Collectively, these methods offer distinct methodological advantages by providing parameter estimates that are more accurate and resilient when analyzing datasets prone to outliers.

METHODS

2.1 Research Data

This study uses cross-sectional data from 154 regencies and cities across the Sumatra region. The response variable (Y) is the percentage of the poor population, representing regional poverty levels. Four predictor variables are included in the analysis: X_1: Labor Force Participation Rate (LFPR / TPAK) the proportion of the working-age population engaged in the labor market. X_2: Human Development Index (HDI / IPM) a composite indicator reflecting education, health, and economic well-being. X_3: Percentage of Population with Access to Safe Drinking Water an indicator of basic service accessibility and living standards. X_4: Gross Regional Domestic Product (GRDP / PDRB) representing regional economic capacity and productivity. The dataset provides a comprehensive representation of socioeconomic and development conditions across Sumatra, enabling an empirical evaluation of factors associated with regional poverty levels.

2.2 Outlier

Outliers are observations that lie far from the center of the data distribution and may exert substantial influence on regression coefficients. Outliers may arise due to data entry errors, measurement inaccuracies, analytical mistakes, or other forms of procedural error. The impact of outliers in data analysis can be categorized based on their origin: those occurring in the response variable (y-outliers or influence points) and those occurring in the predictor variables (x-outliers or leverage points) [7]. Outliers are observations that do not follow the general pattern of the regression model or deviate substantially from the overall structure of the data. In many datasets, approximately 10% of the observations may be classified as outliers. The presence of outliers disrupts the data analysis process and should be carefully addressed in most statistical applications. In the context of regression analysis, outliers may lead to several undesirable effects, large residuals in the resulting model, increased variance within the dataset, and wider confidence interval estimates

2.3 Identification of Outlier Data

Outlier identification methods can generally be classified into two categories: graphical methods and statistical methods. Graphical methods rely solely on visual inspection and are highly dependent on the researcher's interpretation of the plotted data; therefore, they should be supplemented with statistical techniques to ensure more objective detection [8]. Several approaches for identifying outliers in a dataset include the following:

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

a. Scatterplot: This method identifies outliers by plotting each observation and examining whether certain points deviate markedly from the overall data pattern. After fitting a regression model, plotting residuals against predicted values can also reveal outliers when some observations lie far outside the general residual trend.
b. Boxplot: The boxplot detects outliers using quartiles and the interquartile range (IQR). Observations with values below $Q_1 - 1.5\,IQR$ or above $Q_3 + 1.5\,IQR$ are classified as outliers. This method is effective for identifying extreme values in univariate data.
c. Leverage Values: Leverage measures the influence of an observation on the estimation of regression parameters. An observation is considered a high-leverage outlier when its leverage value h_{ii} exceeds the cutoff $\frac{2p}{n}$, where $p$is the number of model parameters and $n$is the sample size.
d. DFFITS (Difference in Fitted Values): DFFITS evaluates how much the predicted value changes when a particular observation is removed. A data point is considered an outlier if

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

Data is considered an outlier if the value $|DFFITS| > 2\sqrt{\frac{p}{n}}$, where p is the number of parameters and n is the number of data observations.

2.4 Robust Regression

Robust regression is used when the presence of outliers violates key regression assumptions, particularly normality [9]. Removing outliers is not always appropriate because extreme observations may contain useful information. When outliers exist, the Ordinary Least Squares (OLS) method often produces biased and unreliable conclusions; therefore, robust regression offers a more suitable alternative. Robust regression reduces the influence of outliers and yields parameter estimates that remain stable, allowing the model to fit the majority of the data without discarding observations. Several estimation methods are commonly used in robust regression, including M-estimation introduced by Huber (1973) and the Least Trimmed Squares (LTS) estimator proposed by Rousseeuw (1984). M-estimation minimizes a robust loss function, whereas the LTS method minimizes the sum of the smallest squared residuals, providing strong resistance to outlier effects.

2.5 Robust Regression M-Estimation

M-estimation in robust regression was first introduced by Huber in 1973. Robust-M is one of the most widely used methods and is considered effective for estimating regression parameters in the presence of outliers. According to [9], Robust-M is a maximum likelihood type estimator. The objective of the OLS estimator is to minimize the sum of squared residuals, whereas the M-estimator modifies this objective by applying a robust loss function. The following equation represents the OLS estimator.

$$\sum_{i=1}^{n} (y_i - \boldsymbol{x_i'\beta})^2 = \sum_{i=1}^{n} e_i^2 \tag{2}$$

The M-estimator replaces e_i^2in Equation (2.6) with $\rho(u_i)$, where the value of $u_i$is defined in Equation (2):

$$u_i = \frac{e_i}{s} \tag{3}$$

Thus, the M-estimator minimizes the objective function shown in Equation (3):

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

$$\sum_{i=1}^{n}\rho(u_i)=\sum_{i=1}^{n}\rho\left(\frac{e_i}{\sigma}\right)=\sum_{i=1}^{n}\rho\left(\frac{y_i-\boldsymbol{x_i'\beta}}{\sigma}\right) \tag{4}$$

The function $\rho(\cdot)$ assigns contributions to each residual and must satisfy the following properties:

1. $\rho(u_i)\geq 0$
2. $\rho(0)=0$
3. $\rho(u_i)=\rho(-u_i)$
4. $\rho(u_i)\geq\rho(-u_i)$ for $|e_i|\geq|u_i|$

The equation for the Robust-M estimator is given in Equation (4)[10]:

$$\min\sum_{i=1}^{n}\rho(u_i)=\min\sum_{i=1}^{n}\rho\left(\frac{e_i}{\sigma}\right)=\min\sum_{i=1}^{n}\rho\left(\frac{y_i-\boldsymbol{x_i'\beta}}{\sigma}\right) \tag{5}$$

To obtain the M-estimator as the solution to Equation (5), a scale estimate is required, which leads to the formulation in Equation (6). The robust scale estimator is denoted by s, and is defined as follows:

$$s=\frac{\text{median}|e_i-median(e_i)|}{0.6745}=\frac{MAD}{0.6745} \tag{6}$$

For the function ρ, the Tukey's Bisquare weighting function may be used, as shown in Equation (5):

$$\rho(u_i)=\begin{cases}\frac{u_i^2}{2}-\frac{u_i^4}{2c^2}+\frac{u_i^6}{6c^4}, & |u_i|\leq c\\ \frac{c^2}{6}, & |u_i|>c\end{cases} \tag{7}$$

The Tukey's Bisquare weighting function provides better performance in handling outliers compared to other weighting functions. In Equation (2.6), the median is used because of its robustness to extreme observations; therefore, it is employed in computing the robust scale estimator. The constant 0.6745 ensures that the value of s becomes an approximately unbiased estimator when the sample size n is large.

2.6 Robust Regression Least Trimmed Squares Estimation

The LTS estimation method was first introduced by Rousseeuw in 1984. This method was developed to overcome the limitations of the Ordinary Least Squares (OLS) estimator [11]. LTS follows the same principle as OLS in minimizing the sum of squared residuals; however, instead of using all observations, LTS minimizes the sum of the smallest squared residuals from a subset of size h. Here, $h<n$, and the value of h represents the number of observations included in the subset with the smallest objective function value, where the objective function is defined in terms of the residuals [12]. The LTS estimator is defined as:

$$min\sum_{i=1}^{h}e_{(i)}^2 \tag{8}$$

With:

$$h=\left[\frac{n}{2}\right]+\left[\frac{p+1}{2}\right] \tag{9}$$

LTS estimation is characterized by its high breakdown point, which refers to the minimal proportion of contaminated observations required to distort the estimator. The LTS method has a breakdown point of up to 50%, making it highly robust against outliers. This breakdown point value reflects the general proportion

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

of outliers that must be present before the model becomes unstable. The principle of the LTS estimator is to minimize the sum of the h smallest squared residuals (the objective function).

RESULT AND DISCUSSION

4.1 Ordinary Least Squares

Least squares regression analysis is a method used to estimate parameters by minimizing the sum of residual squares. After conducting the analysis using R-studio software, the following parameter estimates were obtained:

TABLE 1. Parameter Estimates Of The Ordinary Least Squares Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	35.710	4.585	7.788	1.09E-12
X1	-0.730	0.088	-8.268	7.21E-14
X2	1.608	0.372	4.320	2.84E-05
X3	0.294	0.164	1.793	0.0749
X4	0.071	0.014	4.878	2.74E-06

Based on Table 1, all predictor variables (X1, X2, X3, and X4) are statistically significant at the 10% significance level. X1 has a negative coefficient, indicating a decreasing effect on Y, while X2, X3, and X4 show positive coefficients, meaning they contribute to increasing Y. Overall, the OLS model suggests that all predictors meaningfully influence the response variable. The multiple linear regression OLS model obtained can be written as follows:

$$y_i = 35.710 - 0.730x_{1i} + 1.608x_{2i} + 0.294x_{3i} + 0.071x_{4i} + \varepsilon_i \quad (10)$$

Following this, the classical assumption tests were performed to ensure that the OLS model satisfies the required statistical assumptions.

TABLE 2. Results of Classical Assumption Tests

Test	Statistic / Value	p-value	Conclusion (α = 10%)
Normality	A = 0.93732	0.0171	Residuals not normally distributed because p < 0.10
Heteroskedasticity	BP = 7.3114, df = 4	0.1203	No heteroskedasticity, p > 0.10
Multicollinearity	X1 = 2.077, X2 = 2.264, X3 = 1.029, X4 = 1.205	–	All VIF < 10 → No multicollinearity

Based on the results in Table 2, the residuals do not follow a normal distribution at the 10% significance level. One possible reason for this violation is the presence of outliers, which can distort the distribution of residuals and affect model accuracy. Therefore, it is necessary to perform outlier identification before proceeding with robust regression methods.

4.2 Identification Outlier

Based on the classical assumption tests that have been conducted, it can be observed that the data do not follow a normal distribution. This condition is suspected to be caused by the presence of outliers in the dataset. Outlier identification was carried out using the DFFITS method, where an observation is classified as an outlier if its value satisfies the criterion $|DFFITS| > 2\sqrt{\frac{p}{n}}$. In this study, the dataset consists of 154 observations with $p = 5$ parameters, resulting in a cutoff value of.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

$$2\sqrt{\frac{5}{154}} = 0.3603.$$

Observations exceeding this threshold are considered outliers. The detected outliers in both the original and standardized datasets are presented in Table 3.

TABLE 3. Outlier Detection Using Dffits Criterion

No.	*DFFITS*	| *DFFITS* |		Decision
17	0.438	0.438	> 0.3603	error is an outlier
39	-0.427	0.427		
47	0.366	0.366		
48	0.407	0.407		
65	-0.485	0.485		
85	0.469	0.469		
125	0.594	0.594		
143	-0.619	0.619		
147	0.637	0.637		
153	0.501	0.501		

The deviation can also be seen from the plot in Figure 1 as follows:

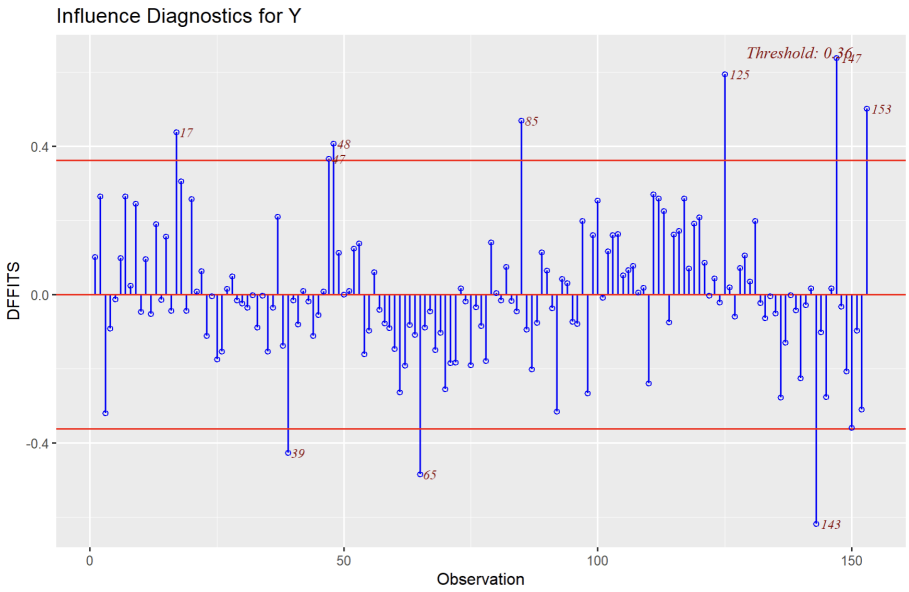

FIGURE 1. Influence Diagnostics for Y Using Dffits Method

4.3 M-Estimation

Robust regression using M-estimation is employed to obtain parameter estimates that are less sensitive to the presence of outliers. Unlike the OLS method, which minimizes the sum of squared residuals and is highly influenced by extreme observations, M-estimation minimizes a robust loss function such as Huber or Tukey's bisquare allowing undue influence from outlying points to be reduced. This approach results in parameter estimates that are more stable and reliable when the dataset contains atypical observations. After conducting the robust regression analysis using M-estimation in R, the following parameter estimates were obtained:

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

TABLE 4. Parameter Estimates Of The Ordinary Least Squares Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	35.890	4.794	7.487	2.86E-12
X1	-0.710	0.092	7.688	9.32E-13
X2	1.451	0.389	3.730	1.36E-04
X3	0.353	0.172	2.059	2.06E-02
X4	0.072	0.015	4.775	2.13E-06

Based on Table 4, all predictor variables (X1, X2, X3, and X4) are statistically significant at the 10% significance level. X1 has a negative coefficient, indicating a decreasing effect on Y, while X2, X3, and X4 show positive coefficients, meaning they contribute to increasing Y. Overall, the M Estimator model suggests that all predictors meaningfully influence the response variable. The multiple linear regression with M-Estimation model obtained can be written as follows:

$$y_i = 35.890 - 0.710x_{1i} + 1.451x_{2i} + 0.353x_{3i} + 0.072x_{4i} + \varepsilon_i \quad (11)$$

4.4 Least Trimmed Squares Estimation

Least Trimmed Squares (LTS) is a robust regression method designed to obtain parameter estimates that are resistant to the influence of outliers. Unlike OLS, which uses all residuals when estimating parameters, LTS minimizes the sum of a selected subset of the smallest squared residuals, effectively trimming observations that deviate excessively. This approach enhances the stability of the regression model when the dataset contains extreme or atypical values

TABLE 5. Parameter Estimates Of The Ordinary Least Squares Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	35.844	4.114	8.713	6.94E-15
X1	-0.689	0.079	-8.740	5.97E-15
X2	1.315	0.339	3.882	1.58E-04
X3	0.378	0.145	2.601	1.03E-02
X4	0.077	0.013	5.778	4.58E-08

Based on Table 5, all predictor variables (X1, X2, X3, and X4) are statistically significant at the 10% significance level. X1 has a negative coefficient, indicating a decreasing effect on Y, while X2, X3, and X4 show positive coefficients, meaning they contribute to increasing Y. Among them, X4 exhibits the strongest statistical significance, followed by X1 and X2. Overall, the LTS model suggests that all predictors meaningfully influence the response variable.

$$y_i = 35.844 - 0.689x_{1i} + 1.315x_{2i} + 0.378x_{3i} + 0.077x_{4i} + \varepsilon_i \quad (12)$$

4.5 Model Comparison and Selection

In this subsection, the performance of the three regression approaches Ordinary Least Squares (OLS), M-estimation, and Least Trimmed Squares (LTS) is compared to determine the most reliable model for the dataset. Using Mean Squared Error (MSE) as the evaluation metric, the results show that the LTS model produces the smallest MSE value, followed by the M-estimation model and the OLS model. These findings indicate that LTS offers the best predictive accuracy and robustness to outliers, making it the most suitable model for this analysis.

TABLE 6. Mean Squared Error (Mse) Comparison Of Regression Models

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

Model	MSE
Ordinary Least Squares (OLS)	10.132
M-Estimation	7.851
Least Trimmed Squares (LTS)	0.909

Based on Table 5, the LTS model yields the smallest MSE value, indicating that it provides the highest level of predictive accuracy among the three regression methods. The M-estimation model produces a lower MSE than OLS, reflecting its improved robustness in the presence of outliers. In contrast, the OLS model has the largest MSE value, showing reduced accuracy when the data contain influential or extreme observations. Overall, these results confirm that the LTS estimator is the most reliable and robust model for this dataset.

CONCLUSION

This study examined the impact of outliers on regression analysis and compared the performance of OLS, M-estimation, and LTS in modeling poverty levels across 154 regions in Sumatra. Applying robust regression methods, M-estimation and LTS produced more stable parameter estimates by minimizing the influence of extreme observations. Based on the Mean Squared Error (MSE), the LTS method demonstrated the best predictive accuracy and robustness, making it the most reliable model for analyzing poverty determinants in the presence of outliers. The LTS model showed that all predictors significantly influenced poverty, with an R-squared of 53.37%, indicating that just over half of the variation in poverty levels can be explained by the selected socioeconomic variables; however, violations of normality and the presence of outliers reduced the reliability of its estimates.

REFERENCES

[1] Badan Pusat Statistik Provinsi Bengkulu, Provinsi Bengkulu dalam Angka Tahun 2023, 2023.
[2] Y. Wen, Y. Tsai, D.B. Wu, P. Chen, The Impact of Outliers on Net-Benefit Regression Model in Cost-Effectiveness Analysis, 8 (2013) 1–9. https://doi.org/10.1371/journal.pone.0065930.
[3] G.B. Begashaw, Y.B. Yohannes, Review of Outlier Detection and Identifying Using Robust Regression Model, 5 (2020) 4–11. https://doi.org/10.11648/j.ijssam.20200501.12.
[4] M. Templ, Enhancing Precision in Large-Scale Data Analysis : An Innovative Robust Imputation Algorithm for Managing Outliers and Missing Values, (2023).
[5] W. Zhan, Y. Hu, W. Zeng, X. Fang, A robust M-estimation framework for spatial autoregressive models : loss function design and optimization strategies, 8816 (2025). https://doi.org/10.1080/13658816.2025.2478459.
[6] V. Berenguer-rico, S. Johansen, B. Nielsen, A model where the least trimmed squares estimator is maximum likelihood, J. R. Stat. Soc. Ser. B Stat. Methodol. 85 (2023) 886–912. https://doi.org/10.1093/jrsssb/qkad028.
[7] D.F. Filho, L. Silva, C. Malaquias, Living with outliers : How to detect extreme observations in data analysis, (2024) 1–24. https://doi.org/10.17666/bib9906/2023.
[8] L. Guan, R. Tibshirani, Prediction and outlier detection in classification problems, (2022) 524–546. https://doi.org/10.1111/rssb.12443.
[9] N.R. Draper, H. Smith, Applied regression analysis, Appl. Regres. Anal. (2014) 1–716. https://doi.org/10.1002/9781118625590.
[10] D.C. Montgomery, E.A. Peck, G.. Vinning, Linear Regression Analysis, 6th ed., John Wiley & Sons, Inc, 2021. https://doi.org/10.2307/1268395.
[11] Y. Zuo, H. Zuo, Computation of least squares trimmed regression – an alternative to least trimmed squares regression, (2023).
[12] B.A. Rasheed, R. Adnan, S.E. Saffari, K. Pati, Jurnal Teknologi Full paper Robust Weighted Least Squares Estimation of Regression Parameter in the Presence of Outliers and Heteroscedastic Errors, 1 (2014) 11–17.