Vol. 4, No. 2, Dec, 2025, pp. 52-64
e-ISSN: 2987-906X
https://ejournal.unib.ac.id/diophantine

Comparison of Robust Regression Methods: Least Trimmed Squares and Maximum Likelihood for Handling Outliers

Andro Kurniawan*, Cinta R. Oktarina, dan Sabarinsyah
Bachelor of Mathematics Program, Batam Institute of Technology, Indonesia

* Corresponding Author Email: andro@iteba.ac.id

Article Information

Article History:
Submitted: 27 November 2025
Accepted: 19 December 2025
Published: 31 December 2025

Key Words:
Robust Regression
Least Trimmed Squares
Maximum Likelihood
Outliers
Per Capita Expenditure

DOI:
https://doi.org/10.33369/diophantine.v4i2.46149

Abstract

This study investigates the determinants of per capita expenditure in 154 regencies and cities across Sumatra Island. The use of the Ordinary Least Squares method is deemed inappropriate due to violations of classical assumptions and the presence of outliers within the dataset. To address these issues, robust regression approaches are applied, specifically M-estimation and Least Trimmed Squares (LTS). The dependent variable in the analysis is per capita expenditure, while the explanatory variables include poverty line, human development index, average years of schooling, and expected years of schooling. The estimation procedures are performed using both raw and standardized data. The empirical results demonstrate that each independent variable significantly influences per capita expenditure under both robust estimation techniques. To determine the most reliable method, the residual standard error is used as the evaluation criterion. The outcomes indicate that the LTS estimator applied to standardized data provides the lowest error value, suggesting that it is the most suitable approach for estimating the regression parameters associated with per capita expenditure in Sumatra.

1. INTRODUCTION

Robust regression serves as an analytical framework that produces stable parameter estimates even when datasets contain outliers or exhibit violations of classical assumptions. Among the approaches frequently implemented within this framework are M-estimation and Least Trimmed Squares (LTS). Both techniques are specifically structured to minimize the impact of extreme observations, thereby generating parameter estimates that are generally more reliable than those obtained through the Ordinary Least Squares method, particularly when the data deviate from normality or include influential outliers. In this study, these two methods are applied to examine the factors associated with per capita expenditure, enabling a comparison of their performance based on the resulting residual standard error.

The application of robust regression is crucial because the Ordinary Least Squares estimator is highly susceptible to abnormal observations and non-normal error structures. Such conditions can produce biased or misleading parameter estimates that fail to reflect the actual socioeconomic conditions. Consequently, an analytical approach capable of accommodating data irregularities is required to assess more accurately the relationship between per capita expenditure and variables such as poverty line, human development index, average years of schooling, and expected years of schooling.

The socioeconomic landscape of Sumatra Island varies considerably across provinces, contributing to different determinants of per capita expenditure in each region. Disparities in infrastructure availability, educational attainment, and economic capacity among communities also influence variations in per capita expenditure levels and, ultimately, societal welfare.

Poverty remains a central indicator for measuring development outcomes. According to the Central Statistics Agency, poverty is defined through a basic-needs framework, which captures the inability of individuals or households to meet minimum food and non-food requirements as reflected in per capita expenditure levels. For this reason, per capita expenditure functions as a key metric for evaluating community welfare. Higher levels of per capita expenditure indicate an increased capacity to satisfy fundamental needs, thereby providing an essential depiction of economic progress throughout regions in Sumatra.

2. METHOD

2.1 Data

The data source for this study is secondary data obtained from the official website bps.go.id. The data obtained includes per capita expenditure, poverty line, human development index, average years of schooling, and expected years of schooling. This study will use cross-sectional data with observations of districts/cities on the island of Sumatra, 154 districts/cities in total.

2.2 Research Variables

The research variables used consist of 4 independent variables and 1 response variable, as shown in Table 1 as follows:

Table 1. Research's Variables.

No	Variable	Description
1	Y	Per Capita Expenditure (Thousand Rupiah/Person/Year)
2	X_1	Poverty Line (Rupiah/Per Capita/Month)
3	X_2	Human Development Index (Percent)
4	X_3	Average Years of Schooling (Years)
5	X_4	Expected Years of Schooling (Years)

Per capita expenditure refers to the total monthly consumption costs incurred by all members of a household—whether originating from purchases, received goods, or self-produced items—divided by the number of individuals within the household.

The Poverty Line (PL) represents the minimum monetary value required by an individual to fulfill essential monthly needs, encompassing both food and non-food components.

The Human Development Index (HDI) serves as a composite indicator that reflects achievements in life expectancy, education, and overall standard of living. This index illustrates the extent to which populations are able to access the outcomes of development in areas such as income, health, and education.

Average Years of Schooling (AYS) denotes the mean duration of formal education completed by individuals aged 15 years and older across all forms of schooling they have undertaken.

Expected Years of Schooling (EYS) represents the projected number of years of education that a child of a given age is anticipated to complete. This indicator provides insight into the expected performance and development trajectory of the education system at various stages.

2.3 Data Analysis

The following are the steps for data analysis:

a. Conduct data exploration.
b. Standardize the data.
c. The data used consists of 1 dependent variables, namely average per capita expenditure, and 4 independent variables, namely X_1, X_2, X_3, and X_4 with the following regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} \quad (1)$$

d. Implementing a series of classical assumption evaluations encompassing residual normality, heteroscedasticity, multicollinearity, and autocorrelation assessments. If the normality assumption test is not met, it is suspected that there is outlier data. The next process is to detect outliers.
e. Detect outliers using the *DFFITS* method.
f. Perform estimation on Robust regression using the LTS estimation algorithm.
g. Estimate robust regression using the M-estimation algorithm.
h. Conducting partial and simultaneous tests to see which factors are significant or influence average per capita expenditure.
i. Select the better estimation method by looking at the smallest standard error residual value.
j. Finally, interpret the results obtained.

2.4 Multiple Regression Analysis Using the Least Squares Method

The model for the multiple regression analysis can be expressed as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_{i2} X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \quad (2)$$

In this formulation, Y_i denotes the observed value of the dependent variable in observation-i, $\beta_0, \beta_1, \ldots, \beta_p$ is the parameter whose value is unknown, $X_{i1}, X_{i1}, \ldots X_{ip}$ is the value of the independent variable in observation-i, and ε_i is a random *error* distributed normally with a mean of zero and a variance σ^2 .

The estimation of the regression coefficients is carried out using the *Ordinary Least Square* (OLS) approach, in which the parameter β is obtained by minimizing the sum of squared residuals. The parameter estimation is as follows:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \tag{3}$$

Where $\widehat{\boldsymbol{\beta}}$ is the vector of estimated parameters of size $(p+1)\times 1$, X is the predictor variable matrix of size $n \times (p+1)$, and $\boldsymbol{y}$ is the observation vector of the response variable of size $n \times 1$.

2.4.1 Multicollinearity Test

According to[1], the purpose of multicollinearity assessment is to identify whether the explanatory variables in the regression model exhibit intercorrelation. An ideal regression model is expected to be free from multicollinearity. The presence of multicollinearity can be examined through tolerance values and the Variance Inflation Factor (VIF). A tolerance value not exceeding 0.10 or a VIF value greater than 10 generally indicates that multicollinearity is present among the predictors.

2.4.2 Heteroscedasticity Test

According to [1], heteroscedasticity test evaluates whether the variance of the residuals remains constant across observations. One of the fundamental assumptions in regression analysis is the absence of heteroscedasticity. In this study, heteroscedasticity is examined using the Glejser test, which investigates the relationship between the absolute residuals and the independent variables. If the resulting significance value exceeds the 5% confidence threshold, the data are considered free from heteroscedasticity problems.

2.4.3 Normality Test

According to [1], the normality test is conducted to determine whether the residuals of the regression model follow a normal distribution, as required for valid statistical inference. A properly specified regression model generally produces residuals that approximate a normal distribution. To evaluate this assumption, the Anderson–Darling test is applied at a 5% significance level. When the test yields a p-value greater than 5%, the residuals are regarded as normally distributed.

2.4.4 Autocorrelation Test

According to [1], the autocorrelation test examines whether residuals from one time period are correlated with residuals from another. Autocorrelation, if present, indicates a violation of classical regression assumptions. In this study, first-order autocorrelation is evaluated using the Durbin–Watson (DW) test, which is appropriate when the model includes a constant term. Additionally, the Run Test—a nonparametric method—can be employed to assess whether the sequence of residuals exhibits randomness or systematic correlation.

2.4.5 Significance Test

According to [2] , in multiple regression analysis, there are several significance tests that are useful for measuring the accuracy of the model, including the following:

1. Regression Model Significance Test
   This test is conducted to examine whether a linear association exists between the response variable Y and the predictor variables $X_1, X_2, \dots, X_k$ or not. When the dependent variable is linearly influenced by the predictor variables, the constructed regression model can be considered appropriate for describing the underlying relationship. The following are the steps:
   a. Hypothesis
      H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (The regression model is not appropriate)
      H_1: there is at least one$\beta_j \neq$, with j = 1, 2, …, k (The regression model is appropriate)
   b. Test statistic

$$\mathrm{F}_0 = \frac{\mathrm{SS_R}/k}{\mathrm{SS_E}/(n-k-1)} = \frac{\mathrm{MSR}}{\mathrm{MSE}}; \mathrm{SSR} = \sum_{i=1}^{n}\left(\hat{Y}_i - \hat{Y}\right)^2; \mathrm{SSE} = \sum_{i=1}^{n}\left(\hat{Y}_i - \hat{Y}_i\right)^2 \tag{4}$$

   c. Test criteria
      H_0 Rejected if $\mathrm{F}_0 > \mathrm{F}_{table} = \mathrm{F}_{(\alpha;k;n-k-1)}$, or $p_{value} < \alpha$
2. Test of Individual Regression Coefficient Significance
   The test assesses whether each predictor exerts a measurable effect on the response variable. The steps are:
   a. Hypothesis
      H_0: $\beta_j = 0$ (x_j t regression coefficient is not significant)
      H_1: $\beta_j \neq 0$, with j = 1, 2, …, k (x_j t regression coefficient is significant)
   b. Test statistic

$$t_0 = \frac{\hat{\beta}_j}{Se(\hat{\beta}_j)} \text{ with } Se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}} \quad (5)$$

Where C_{jj} is the diagonal element of$(X'X)^{-1}$ and $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-k-1}$

c. Test criteria
H_0 rejected if $|t_0| > t_{table} = t_{(\alpha/2;n-k-1)}$ *or* $p_{value} < \alpha$

2.5 *Outlier* Data

The effect of outliers in data analysis can be distinguished based on the origin of the outliers, namely those originating from the response variable (youtliers; *influence* points) or originating from the independent variable (*x-outliers*; *leverage* points) [3] .

In relation to regression analysis, *outliers* cause the following [4] :

1. Large residuals from the model
2. The variance in the data becomes large
3. The interval estimate has a wide range

In regression analysis, there are three types *of outliers* that affect the least squares estimation, namely:

a. *Vertical Outlier*
These observations represent cases that deviate substantially in the dependent variable while remaining within the expected range of the predictors. Such vertical outliers can distort the results produced by the least squares estimator.

b. *Good Leverage Point*
These observations exhibit extreme values in the predictor variables, yet they lie close to the fitted regression line. Although such good leverage points do not destabilize the least squares estimates, they may influence statistical inference by increasing the estimated standard errors.

c. *Bad Leverage Point*
This type of observation displays extreme values in the predictor variables and lies far from the regression line. Bad leverage points can substantially distort least squares estimates, influencing both the intercept and the slope of the regression model.

Outlier identification methods are divided into two types: graphical methods, which rely solely on visualization and are highly dependent on the researcher's perspective on the resulting graph, and statistical calculation methods. Several methods for identifying outliers in an analysis are as follows:

a. *Scatterplot*
This approach involves creating a plot of data for each observation i (i = 1, 2, ..., n). After a regression model is fitted, a residual plot–graphing the residual e_i against the predicted values $\hat{Y}_i$–may also be examined. The presence of one or more points that deviate markedly from the overall pattern of the data suggests the existence of outliers.

b. *Boxplot*
This method uses quartiles and range to detect outliers. Quartiles 1, 2, and 3 divide the previously sorted data into four parts. *The interquartile range (*IQR) is defined as the difference between quartiles 1 and 3, or IQR = $Q_3 - Q_1$. *Outliers* are values less than 1.5*IQR for quartile 1 and values greater than 1.5*IQR for quartile 3.

c. *Leverage Values Method*
This approach examines the impact that each observation may have on the resulting parameter estimates, allowing influential points to be identified. This can be seen from the distance of the X values of all observations. The *leverage* value for simple linear regression can be determined as follows [5] :

$$Laverage(h_{ii}) = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{(n-1)S_x^2} \quad (6)$$

where:
h_{ii} : *leverage* of the i-th case
n : number of data
X_i : value for case i
S_x^2 : the squares n The case consists of the deviation X_i s from the mean
$\bar{X}$: mean of X
For observations involving more than one explanatory variable, the leverage measure is computed using the matrix expression shown below:

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \quad (7)$$

Where $\boldsymbol{H}$ is the *hat* matrix, the elements of the$i$ diagonal of the *hat* matrix are *the leverage values*, and X is the matrix X. The *outlier* approach is based on *the cutoff* value, and if the value h_{ii} exceeds the *cutoff* value, it is detected

as *an outlier.* The specified *cutoff* value is $\frac{2p}{n}$, where n is the amount of data, and p is the number of parameters in the regression equation formed, including *the intercept* [6] .

d. *DFFITS* Method *(Difference Fitted Value FITS)*
This method displays the change in the predicted value when the i-th case is removed from the standardized research data [4] . The *DFFITS* calculation is as follows:

$$DFFITS = t_i \left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}} \quad (8)$$

wheret_i is *Rstudent* for case-i and h_{ii} is the *leverage* value for case-i. Data is considered *an outlier* if the value$|DfFITS| > 2\sqrt{\frac{p}{n}}$with p is the number of parameters and n is the number of data observations [7] .

2.6 *Robust* Regression

In his book, [8] explains that one of the deviations that occurs when there are outliers is a violation of the assumption of normality. Outliers should not be discarded without justification, as they may represent meaningful patterns or information that cannot be obtained from the remaining observations. If an observation is known to be an outlier in a study, the use of OLS will produce imperfect conclusions. *Robust* regression is used as an alternative. In general, *robust* means strong. [6] explains that *robust* regression is able to reduce the influence of outliers compared to using MKT, resulting in a strong estimator that is not affected by the presence of outliers. Using *robust* techniques, a regression model can be constructed that lessens the impact of data points with unusually large residuals. Instead of eliminating these observations, the method focuses on finding parameter estimates that align well with the bulk of the data, resulting in a more dependable model

In *robust* regression, there are several estimation methods for estimating regression parameters, one of which is M estimation introduced by Huber (1973) and *Least Trimmed Squares* (LTS) estimation introduced by Rousseeuw (1984). In M-estimation, the estimator is obtained by finding the parameter values that minimize a specified function ρ of the residuals, which serves as the basis for its robustness. Meanwhile, the LTS estimation method has the basic principle of minimizing the sum of trimmed residual squares.

2.7 *M-Estimation*

According to [8], as a likelihood-based estimation technique, robust M-Regression determines its parameter values through the minimization of a residual-related objective function. In the conventional MKT formulation, this objective simplifies to minimizing the total of the squared residuals. Equation (9) is the MKT estimator according to [9] .

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x_i'\beta})^2 = \sum_{i=1}^{n} e_i^2 \quad (9)$$

Robust-M estimator, replacing e_i^2 in equation (9) with $\rho(u_i)$ where the value of u_i can be seen in equation (10).

$$u_i = \frac{e_i}{s} \quad (10)$$

As a result, the estimation process for the Robust-M method consists of finding the parameter values that yield the smallest value of the objective function described in equation (11)

$$\sum_{i=1}^{n}\rho(u_i) = \sum_{i=1}^{n}\rho\left(\frac{e_i}{\sigma}\right) = \sum_{i=1}^{n}\rho\left(\frac{y_i - \boldsymbol{x_i'\beta}}{\sigma}\right) \quad (11)$$

The function contributes to each residual under the condition that it must satisfy the following property:

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 formulation of the Robust-M estimator is provided in equation (12) [10] .

$$\min\sum_{i=1}^{n}\rho(u_i) = \min\sum_{i=1}^{n}\rho\left(\frac{e_i}{\sigma}\right) = \min\sum_{i=1}^{n}\rho\left(\frac{y_i - \boldsymbol{x_i'\beta}}{\sigma}\right) \quad (12)$$

The M-estimator as a solution to equation (12) requires setting a scale to produce equation (13). The scale of the *Robust* estimator is *s*, with the following formula:

$$s = \frac{\text{median}|e_i - median(e_i)|}{0.6745} = \frac{MAD}{0.6745} \quad (13)$$

Tukey's bisquares weighting function can be utilized as the chosen ρfunction, with its formulation presented in equation (14).

$$\rho(u_i) = \begin{cases} \frac{u_i^2}{2} - \frac{u_i^4}{2c^2} + \frac{u_i^6}{6c^4}, & |u_i| \le c \\ \frac{c^2}{6}, & |u_i| > c \end{cases} \tag{14}$$

The Tukey's Bisquares weighting function provides better results in dealing with outliers than other weighting functions. In equation (13), the median is applied because of its robustness to the presence of outliers. By incorporating the constant 0.6745 into the computation, the resulting scale estimate $s$becomes approximately unbiased when the sample size $n$is sufficiently large.

To obtain a solution for equation (13), we compute the first partial derivative of ρ with respect to $\beta_j (j = 0,1,2 \dots, p)$ and impose the stationary conditions by equating these first-order partial derivatives to zero in equation (15). With $\Psi = \rho'$ and X_{ij} being the *i-th* observation at the *j-th* point.

$$\sum_{i=1}^{n} X_{ij} \Psi\left(\frac{y_i - \boldsymbol{x_i'\beta}}{s}\right) = 0 \tag{15}$$

To obtain the solution, a weighting function is first defined, as expressed in equation (16).

$$w(e_i) = \frac{\Psi\left(\frac{y_i - \boldsymbol{x_i'\beta}}{s}\right)}{\left(\frac{y_i - \boldsymbol{x_i'\beta}}{s}\right)} \tag{16}$$

It is known that $u_i = \frac{e_i}{s}$ so that equation (16) can be rewritten as equation (17).

$$w_i = \begin{cases} \left[1 - \left(\frac{u_i}{c}\right)^2\right]^2, & |u_i| \le c \\ 0, & |u_i| > c \end{cases} \tag{17}$$

In this study, the Tukey bisquares weight is applied using a tuning constant $c$specified as 4.685. After the *Tukey's Bisquares* weighting is substituted into equation (9), the equation can be rewritten as equation (18).

$$\sum_{i=1}^{n} x_{ij} w_i (y_i - \boldsymbol{x_i'\beta}) = 0 \tag{18}$$

Equation (18) can be written in matrix form as equation (19).

$$\boldsymbol{\hat{\beta}_{robust}} = (\boldsymbol{X'W_0X})^{-1}\boldsymbol{X'W_0y} \tag{19}$$

which refers to the diagonal matrix constructed from Tukey's bisquares weights, where each diagonal entry is given by. Equation (19) is known as the Weighted Least Squares (*WLS*) equation [9] .

The M-estimation algorithm can be seen from the description below:

a. Calculating the estimated $\hat{\beta}$ parameters using MKT
b. Identifying potential outliers within the dataset
c. Compute the residual values $e_i = (y_i - \hat{y}_i)$
d. Calculate the value $\hat{\sigma} = \frac{median|e_i - median(e_i)|}{0.6745} = \frac{MAD}{0.6745}$
e. Calculating the value $u_i = \frac{e_i}{\hat{\sigma}}$
f. Calculating the weighted value

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

with:
u_i : the value of the division of residual and the nthσ i
c : the tuning constant that has been set to determine the level of robustness
w_i : the weight value of the$i$

g. Calculate $\hat{\beta}$ the M-estimator residual using the Weighted Least Squares (WLS) method with weights w_i
h. Repeat steps (d) until (g) to obtain the convergent value of $\hat{\beta}$ the estimated M, meaning that at iteration-i, the estimated parameters will be the same as in the subsequent iteration.

i. Perform a test to determine whether the independent variables have a significant effect on the dependent variables using a simultaneous test that looks at the value $F_{calculated}$ and a partial test that looks at the value $t_{calculated}$

The algorithm above is taken from [11] .

2.8 *Least Trimmed Squares (LTS)* Estimation

The equation for the LTS estimation method is as follows:

$$min \sum_{i=1}^{h} e_{(i)}^2 \tag{20}$$

$$h = \left[\frac{n}{2}\right] + \left[\frac{p+1}{2}\right] \tag{21}$$

[12]

The LTS estimator possesses a high breakdown point, reaching up to 50%. The breakdown point indicates the maximum proportion of contaminated observations that the estimator can tolerate before the fitted model becomes unreliable. This method operates by minimizing the sum of the squared residuals for the best h observations.

The LTS estimation algorithm is as follows [13] :

a. Calculate the parameter estimates $\hat{\beta}_{initial}$ using the MKT method.
b. Calculate the residual valuese_i using $e_i = (y_i - \hat{y}_i)$ corresponding to $\hat{\beta}_{initial}$.
c. Calculate the observed h using equation (24) with the $e_{(i)}^2$ value.
d. Performing calculations using equation (23).
e. Calculate the parameter estimates $\hat{\beta}_{new}$ using the MKT method from the calculated h observations.
f. Determine the residual sum of squares $e_i^2 = (y_i - \hat{y}_i)^2$ corresponding to the obtained $\hat{\beta}_{new}$, then calculate the sum of h_{new} observations with the value $e_{(i)}^2$.
g. Perform the calculation from equation (24) with the value h_{new}.
h. Performing *C-steps* by repeating steps (e) to (g) until the objective function $\left(\sum_{i=1}^{h} e_{(i)}^2\right)$ converges, meaning that at iteration-i , the sum of residual squares and the parameter estimates will be the same as in the next iteration.

3. RESULTS AND DISCUSSION

3.1 Data Description

This study uses 2021 per capita expenditure data as the dependent variable and uses poverty line, human development index, average years of schooling, and expected years of schooling as independent variables. This data is from 154 districts/cities on the island of Sumatra, obtained from the official BPS website. Data description was performed to see the data profile for each variable, along with descriptive statistics of the research variables used in the case study.

Table 2. Descriptive Statistics of Data Variables.

Variable	Data Summary				
	Minimum	Maximum	Range	Average	Variance
Y	6152	18506	12354	10919	4280669
X_1	350452	860629	510177	509697	9316792976
X_2	62.19	86.28	24.09	71.61	19.83548
X_3	5.880	13.030	7.15	8.962	1.788573
X_4	11.43	17.81	6.38	13.36	1.088898

Table 3. Descriptive Statistics of Standardized Data Variables.

Variable	Data Summary				
	Minimum	Maximum	Range	Average	Variance
Y	-2.3039	3.6672	5.9711	0	1
X_1	-1.6498	3.6357	5.2855	0	1
X_2	-2.1156	3.2933	5.4089	0	1
X_3	-2.3046	3.0417	5.3463	0	1
X_4	-1.8484	4.2656	6.114	0	1

3.2 Least Squares Regression Analysis

The parameter estimation results are as follows:

Table 4. Original Data Estimation Results.

Parameter	Estimated Value
$Intercept$	-2.14×10^4
β_1	2.40×10^{-3}
β_2	6.51×10^2
β_3	-7.26×10^2
β_4	-6.72×10^2
$R - Squared = 85.11\%$	

Based on Table 4, parameter estimates were obtained to form an initial model using the least squares regression method, namely:

$$Y = -2.14 \times 10^4 + 2.40 \times 10^{-3} X_1 + 6.51 \times 10^2 X_2 - 7.26 \times 10^2 X_3 - 6.72 \times 10^2 X_4 + \varepsilon \quad (22)$$

After obtaining the initial model with an R^2 value of 85.11%, it means that the independent variables, namely the poverty line, human development index, average years of schooling, and expected years of schooling, can explain 85.11% of the variance in the dependent variable, while the remaining 14.89% is explained by other variables not studied.

Table 5. Standardized Data Estimation Results.

Parameter	Estimated Value
β_1	0.11210
β_2	1.40084
β_3	-0.46911
β_4	-0.33876
$R - Squared = 85.11\%$	

From the information presented in Table 5, the regression model is estimated as follows:

$$Y = 0.11210 X_1 + 1.40084 X_2 - 0.46911 X_3 - 0.33876 X_4 + \varepsilon \quad (23)$$

3.3 Classical Assumption Test

An assessment of the classical regression assumptions was performed, encompassing tests for normality, heteroscedasticity, multicollinearity, and autocorrelation. The findings from these evaluations are reported below:

Table 6. Results of Classical Assumption Tests.

Type of Test	Statistical Value	p-value	Decision	Conclusion
Normality	1.5515	0.0005	Reject H_0	Data is not normally distributed
Heteroscedasticity	2.6792	0.6129	Accept H_0	No heteroscedasticity
Multicollinearity	$X_1 = 1.578, X_2 = 6.902$ $X_3 = 6.909, X_4 = 2.407$	-	All $VIF < 10$	No multicollinearity
Autocorrelation	1.5391	0.0012	Reject H_0	There is autocorrelation

3.4 Outlier Identification

Outliers were identified using the *DFFITS* test, assuming that if the value is $|DFFITS| > 2\sqrt{\frac{p}{n}}$, then the data is considered an outlier. This study has 154 data points, resulting in $2\sqrt{\frac{5}{154}} = 0.3603$. The data that are outliers in the original data and standardized data are present in Table 7:

Table 7. Original Data and Standardized Data.

No	*DFFITS* Value	Absolute Value		Decision
19	1.2578	1.2578		
50	0.5139	0.5139		
146	-0.6214	0.6214		Data
149	0.4980	0.4980	> 0.3603	is
150	0.7018	0.7018		an outlier
151	0.5262	0.5262		
153	0.7683	0.7683		

Based on Table 7, using the original data and standardized data, it was found that observations 19, 50, 146, 149, 150, 151, and 153 were outliers. In accordance with the initial assumption, the 7 observations have a value of $DFFITS > 0.3603$. Outliers can also be seen from the plot in Figure 1 as follows:

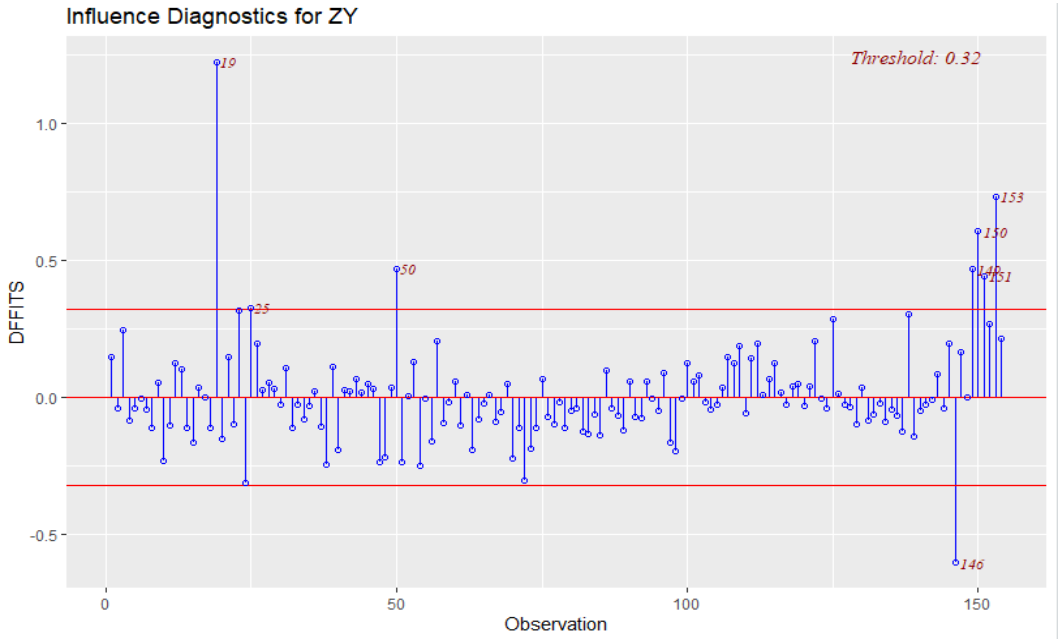

Figure 1. *DFFITS Plot*

3.5 *Robust* Regression Analysis with LTS Estimation

This study was conducted using *robust* regression with LTS estimation using *R-studio software* to facilitate the analysis. The results obtained with the help of *R-studio software* are shown in Table 8 as follows:

Table 8. Output of the Robust Regression Using the LTS Estimator.

Parameters	Original Data	Standardized Data
	Estimated Value	
Intercept	-1.80×10^{4}	
β_1	2.01×10^{-3}	0.0938
β_2	6.05×10^{2}	1.3029
β_3	-5.76×10^{2}	-0.3725
β_4	-7.76×10^{2}	-0.3914

Based on Table 8, the model for estimating LTS in the original data is:

$$Y = -0.000018 + 0.00201X_1 + 605X_2 - 576X_3 - 776X_4 + \varepsilon \quad (24)$$

Meanwhile, the model obtained by standardizing the data first is:

$$Y = 0.0938X_1 + 1.3029X_2 - 0.3725X_3 - 0.3914X_4 + \varepsilon \quad (25)$$

The model obtained must be tested through a simultaneous validation process, which is a simultaneous testing of all parameters in the regression model. Simultaneous test give value 2.2×10^{-16} and original data LTS significance test can see at the table below.

Table 9. Original Data LTS Significance Test.

Parameter	Partial Test		Decision
	t_{hitung}	p_{value}	
β_1	3.010	0.0031	Reject H_0
β_2	19.505	2×10^{-16}	Reject H_0
β_3	-5.613	1.03×10^{-7}	Reject H_0
β_4	-9.956	2×10^{-16}	Reject H_0

3.5.1 Simultaneous Test of Original Data

1. Formulating hypotheses
   $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ (There is no simultaneous effect of the independent variable on the dependent variable)
   $H_1: \beta_i \neq 0$, for $i = 1,2,3,4$ (There is a simultaneous effect of the independent variables on the dependent variable)
2. Test level
   The analysis is conducted using the F distribution at the 5% significance level.
3. Test statistic
$$F_{calculated} = \frac{KT_{Regression}}{KT_{Error}} = 291.6$$
4. Rejection criterion
   Reject H_0 if $F_{calculated} > F_{table}$ or $p_{value} < \alpha$
5. Conclusion
   The value of $F_{calculated} = 7.164 > F_{table}$ and $p_{value} < \alpha$, then H_0 is rejected. This means that there is an influence of the independent variable on the dependent variable simultaneously at the 5% significance level.

3.5.2 Partial Test of Original Data

All variables are found to significantly affect per capita expenditure based on the partial test results. The poverty line variable has $t_{calculated} = 3.010$ with $p_{value} = 0.0031$, the human development index variable has $t_{calculated} = 19.505$ with $p_{value} = 2 \times 10^{-16}$, the expected years of schooling variable has $t_{calculated} = 5.613$ with $p\_value = 1.03 \times 10^{-7}$, and the average years of schooling variable has $t_{calculated} = 9.956$ with $p\_value = 2 \times 10^{-16}$. All p_{value} are smaller than $\alpha = 5\%$ so that H_0 is rejected for all parameters. The significance of the four estimated parameters demonstrates that every predictor plays a meaningful role in explaining per capita expenditure. Consequently, the resulting model can be regarded as valid and appropriate for analytical purposes.

3.5.3 Simultaneous Test of Standardized Data

1. Formulating the hypothesis
   $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ (There is no simultaneous effect of independent variables on the dependent variable)
   $H_1: \exists \beta_j \neq 0$, for $j = 1,2,3,4$ (There is a simultaneous effect of independent variables on the dependent variable)
2. Test level
   The distribution used is the F distribution with a significance level of 5%.
3. Test statistics
$$F_{calculated} = \frac{KT_{Regression}}{KT_{Error}} = 276$$
4. Rejection criterion
   Reject H_0 if $F_{calculated} > F_{table}$ or $p_{value} < \alpha$
5. Conclusion
   The value of $F_{calculated} = 276 > F_{table}$ and $p_{value} < \alpha$, then H_0 is rejected. This means that there is an influence of the independent variable on the dependent variable simultaneously at the significance level of 5% .

3.5.4 Partial Test of Standardized Data

The four partial tests show that all variables, namely the poverty line ($p_{value} = 0{,}0031$), human development index ($p_{value} = 2 \times 10^{-16}$), expected years of schooling ($p_{value} = 1{,}03 \times 10^{-7}$), and average years of schooling ($p_{value} = 2 \times 10^{-16}$), have p_{value} smaller than α = 5%, so H_0 is rejected for each parameter. Thus, all regression coefficients ($\beta_1, \beta_2, \beta_3$, , and β_4) are significant, meaning that each variable has a significant effect on per capita expenditure and the regression model used is appropriate for explaining the relationship between these variables.

3.6 *Robust* Regression Analysis M Estimation

Table 10 below provides the summary of results obtained through the robust M-estimation regression procedure:

Table 10. Results of *Robust* M-Estimation Regression.

Parameter	Original Data	Standardized Data
	Estimated Value	
Intercept	−20505.03	
β_1	0.0023	0.1087
β_2	639.62	1.376
β_3	−678.39	−0.4353
β_4	−714.96	−0.3610

Based on Table 10, the model for estimating LTS in the original data is:

$$Y = -20505.03 + 0.0023X_1 + 639.62X_2 - 678.39X_3 - 714.96X_4 + \varepsilon \quad (26)$$

Meanwhile, the model obtained by standardizing the data first is:

$$Y = 0.1087X_1 + 1.376X_2 - 0.4353X_3 - 0.3610X_4 + \varepsilon \quad (27)$$

The model obtained must be tested through a simultaneous validation process, which is a simultaneous testing of all parameters in the regression model.

Table 11. M Significance Test of Original Data.

Parameter	Partial Test $t_{calculated}$	Decision
Intercept	12.3956	Reject H_0
β_1	3.0124	Reject H_0
β_2	18.3100	Reject H_0
β_3	5.8282	Reject H_0
β_4	8.1192	Reject H_0

3.6.1 Partial Test of Original Data

The results of the significance test using M-standardized data are presented below:

Table 12. Significance Test of Standardized Data M.

Parameter	Partial Test $t_{calculated}$	Decision
β_1	3.0071	Reject H_0
β_2	18.2157	Reject H_0
β_3	5.7584	Reject H_0
β_4	8.0899	Reject H_0

3.6.2 Partial Test of Standardized Data

The partial test results indicate that each independent variable exerts a statistically significant effect on per capita expenditure at the 5% significance level. The poverty line variable has a value of $t_{calculated} = 3.0071 > 1.98$, so that β_1 is significant. The human development index variable is also significant with $t_{calculated} = 18.2157 > 1.98$. The average years of schooling variable is also significant with $t_{calculated} = 8.0899 > 1.98$. Furthermore, the expected years of schooling variable has $t_{calculated} = 5.7584$, which exceeds t_{table} , so β_3 is significant. Therefore, given that all four independent variables exert a statistically significant effect on per capita expenditure, the regression model is deemed suitable for use.

3.7 Selection of the Best Estimation Model Using *Residual Standard Error*

Table 13. *Residual Standard Error* Values.

Residual Standard Error			
LTS Estimation		**M-Estimation**	
Original Data	Standardized Data	Original Data	Standardized Data
612.5	0.2961	722.6	0.332

Determination of the best estimation method from *Robust* regression related to solving outlier problems using the LTS estimation method and M-estimation, as well as using original data and standardized data by comparing the *Residual Standard Error* values of each method. The best estimation is the method that has the smallest *residual standard error* value. The data in Table 13 reveal that the lowest residual standard error originates from the LTS estimation using standardized data, among the two robust regression models. Therefore, LTS estimation based on standardized data proves to be a valid robust regression technique to use in estimating regression parameters for per capita expenditure data on the island of Sumatra in 2021.

3.8 Model Interpretation

Based on the process of selecting the best model, we obtain a *robust* regression with standardized data and use the selected LTS estimation as the best model. The model validation process using a partial test will form the following LTS estimation *robust* regression equation model:

$$Y = 0.1087X_1 + 1.376X_2 - 0.4353X_3 - 0.3610X_4 + \varepsilon \quad (28)$$

The standardized robust regression model yields an R^2 of 88.53%, indicating that 88.53% of the variation in the dependent variable—per capita expenditure—is accounted for by the poverty line, the human development index, average years of schooling, and expected years of schooling. The remaining 11.47% is influenced by other factors not included in this study.

The standardized data model shows that the regression coefficient for the poverty line variable (X_1) is 0.1087. This implies that, holding the other variables constant, a one-unit increase in the poverty line is associated with a 0.1087 increase in per capita expenditure. Since the coefficient is positive, it indicates a direct relationship between the poverty line and per capita expenditure—meaning that as the poverty line rises, per capita expenditure also tends to increase.

The standardized model indicates that the human development index variable (X_2) has a regression coefficient of 1.376. This suggests that, when all other variables are held constant, a one-unit rise in the human development index corresponds to a 1.376 increase in per capita expenditure. The positive sign of the coefficient shows a direct relationship between these variables—higher levels of the human development index are associated with higher per capita expenditure.

The standardized model shows that the regression coefficient for the average years of schooling variable (X_3) is –0.4353. This indicates that, with the other variables held constant, a one-unit increase in average years of schooling is associated with a 0.4353 decrease in per capita expenditure. The negative coefficient reflects an inverse relationship between average years of schooling and per capita expenditure—meaning that lower levels of average schooling correspond to higher per capita expenditure.

The standardized model indicates that the regression coefficient for the expected years of schooling variable (X_4) is –0.3610. This means that, when the other variables are held constant, a one-unit increase in expected years of schooling is associated with a 0.3610 decrease in per capita expenditure. The negative coefficient reflects an inverse relationship between expected years of schooling and per capita expenditure—where shorter expected schooling durations correspond to higher levels of per capita expenditure.

4. CONCLUSION

This study compares the Least Trimmed Squares (LTS) and M-estimation robust regression methods to model per capita expenditure in Sumatra. The results show that LTS estimation applied to standardized data yields the smallest residual standard error, indicating that it provides the most reliable parameter estimates in the presence of outliers.

The empirical findings reveal that the poverty line, human development index, average years of schooling, and expected years of schooling have statistically significant effects on per capita expenditure at the 5% significance level. This indicates that socioeconomic and educational factors play a crucial role in explaining regional variations in per capita expenditure levels across districts and cities in Sumatra.

These findings highlight the importance of applying robust regression techniques when socioeconomic data contain outliers. The proposed approach can be used as a reference for future studies involving regional welfare indicators and can assist policymakers in understanding the determinants of per capita expenditure.

REFERENCES

[1] I. Ghozali, *Multivariate Analysis Application with IBM SPSS 19 Program*. Semarang: UNDIP, 2011.

[2] K. Sambell, L. McDowell, and C. Montgomery, *Assessment for learning in higher education*. Abingdon: Routledge, 2012. doi: 10.4324/9780203818268.

[3] N. Nurdin, A. Islamiyati, and Raupong, "The Use of Robust Regression on Data Containing Outliers Using the Moment Method," *JMSK J. Mat. Stat. dan Komputasi*, vol. 10, no. 2, pp. 114–123, 2014, [Online]. Available at: journal.unhas.ac.id/index.php/jmsk

[4] Soemartini, *Outliers*. Bandung: Unpad Press, 2007.

[5] S. Wijaya, "Parameter Estimation in Robust Regression Models Using Huber Functions," University of Indonesia, 2009. [Online]. Available at: https://lib.ui.ac.id/file?file=digital/old23/20181962-010-09-Taksiran parameter.pdf

[6] M. H. Kutner and Christopher J. Nachtsheim, *Applied Linear Statistical Models*, vol. 29, no. 2. New York: New York: Mc.Graw-Hill Companies, 1997. doi: 10.1080/00224065.1997.11979760.

[7] D. Intan Perihatini, "COMPARISON OF LTS ESTIMATION, M ESTIMATION, AND S ESTIMATION METHODS IN ROBUST REGRESSION (Case Study: Car Financing at Company 'X' in 2016)," 2018.

[8] N. Draper and H. Smith, *Applied Regression Analysis (Second Edition, translated by Gramedia)*. Jakarta: PT. Gramedia Pustaka Utama, 1992.

[9] S. Yuliana, P. Hasih, H. Sri Sulistijowati, and L. Twenty, "M Estimation, S Estimation, and MM Estimation in Robust Regression," *Int. J. Pure Appl. Math.*, vol. 91, no. 3, pp. 349–360, 2014.

[10] D. C. Montgomery and E. A. Peck, *Introduction to Linear Regression Analysis*. New York: John Wiley & Sons, 1992.

[11] Y. D. K. Setyo Wira Rizki, "Robust M-Estimation Regression Analysis Using Tukey and Welsch Bisquare Weighting to Overcome Outlier Data," *Bimaster Bul. Ilm. Mat. Stat. dan Ter.*, vol. 8, no. 4, pp. 799–804, 2019, doi: 10.26418/bbimst.v8i4.36199.

[12] M. Y. Matdoan, "ROBUST LEAST TRIMMED SQUARE (LTS) REGRESSION MODELING (Case Study: Factors Factors Affecting the Spread of Malaria in Indonesia)," *Euclid*, vol. 7, no. 2, pp. 77, 2020, doi: 10.33603/e.v7i2.2926.

[13] Dina Rohmah, Y. Susanti, and E. Zukhronah, "Comparison of Robust Regression Models: M Estimation and Least Trimmed Squares (LTS) Estimation on the Number of Tuberculosis Cases in Indonesia," vol. 4, no. 2, pp. 136–146, 2020.