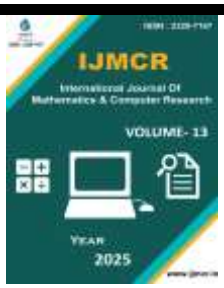
International Journal of Mathematics and Computer Research

ISSN: 2320-7167

Volume 13 Issue 06 June 2025, Page no. – 5292-5298

Index Copernicus ICV: 57.55, Impact Factor: 8.615

DOI: 10.47191/ijmcr/v13i6.03



Robust M-Estimator Modelling with Huber and Andrew Weighting Functions on Poverty Levels in North Sumatra

Agus Rusgiyono[1], Yasmine Orsini[2], Miftahul Jannah[3]

[1,2,3]Department of Statistics, Faculty of Science and Mathematics, Diponegoro University

ARTICLE INFO

Published Online:
04 June 2025

ABSTRACT

Poverty remains one of the major issues in developing countries such as Indonesia. In 2023, the poverty rate in North Sumatra Province reached 8.15%. This high percentage is influenced by various factors, prompting the need for a model to examine the effects of open unemployment, average years of schooling, life expectancy, and economic growth on poverty. Initial analysis using the Ordinary Least Squares (OLS) method identified two outliers that led to biased parameter estimates. This issue was addressed using robust regression with the M-estimator approach, applying Huber and Andrew weighting functions. This method is more resistant to outliers and does not require residual normality, although it still assumes linearity, no multicollinearity, autocorrelation, and homoscedasticity. The results show that the model using Huber weights performs best, with an Adjusted R^2 of 85.72% and a Mean Squared Error (MSE) of 2.509, while meeting all regression assumptions. Average years of schooling and economic growth have a significant effect on poverty. An additional year of schooling reduces the poverty rate by 1.184%, and a 1% increase in economic growth lowers it by 3.929%, both at the 5% significance level. Therefore, the robust regression model with Huber weights is considered the more optimal.

Corresponding Author:
Miftahul Jannah

KEYWORDS: Outliers; Robust Regression; M-Estimator; Huber; Andrew

I. INTRODUCTION

Poverty is a condition in which individuals or families have incomes that are insufficient to meet their basic needs, such as food, clothing, and shelter. It can be measured using the poverty line, which distinguishes between absolute and relative poverty [1]. The welfare of society is the primary goal of national development, which is determined by the fulfillment of both physical and spiritual needs, as well as the attainment of a decent standard of living. This standard of living comprises three main dimensions: health, education, and income level or economic well-being [2]. Poverty perpetuates a vicious cycle, where poor quality of education and health care limits economic opportunities, further deepening poverty itself [3].

In addition to causing economic hardship, poverty also affects health, education, and future prospects [4]. The percentage of the poor population is a commonly used indicator of poverty. Numerous studies on poverty in Indonesia, including those focusing on economic growth and welfare, have adopted this metric [5]. The proportion of poor individuals in a region is estimated by comparing the number of poor people to the total population [6].

According to the North Sumatra Provincial Statistics Agency [7], 8.15 percent of the population in North Sumatra lived in poverty in 2023, amounting to 1,239,710 individuals. Aini and Nugroho [8] found that open unemployment and average years of schooling significantly influence poverty, contributing to the high number of impoverished people. Suryawati [5] also identified health as a contributing factor to poverty. In addition, Susanto and Pangesti [9]discovered that regional economic expansion helps reduce poverty levels.

Regression analysis using the Ordinary Least Squares (OLS) method is commonly employed to model relationships between variables. However, this method has a notable weakness, particularly in the presence of outliers. Outliers can distort parameter estimates, leading to inaccurate models. Their presence may result from measurement errors, natural variability in the data, or unexpected external factors [10]. In regression analysis, outliers can undermine the stability of parameter estimation, producing biased and unreliable

estimates [11]. One approach to address this issue is robust regression, which is specifically designed to yield more stable estimates even in the presence of outliers [12].

Robust regression methods are more resistant to deviations compared to Ordinary Least Squares (OLS), thereby providing more reliable results even when classical assumptions are not fully met [13]. One widely used approach within robust regression is the M-estimator, which offers advantages in terms of computational and theoretical simplicity [14]. Several studies have examined the effectiveness of this method in addressing the presence of outliers. For instance, Damayanti and Susanti [15] demonstrated that robust regression using Huber and Tukey Bisquare weighting functions is effective in handling outliers in poverty data in Indonesia. Similarly, Deria et al. [16] found that the robust regression method using M-estimator with Andrew's weighting function yields more accurate parameter estimates compared to conventional regression methods.

Building upon previous research, the present study aims to apply the robust regression M-estimator using Huber and Andrew weighting functions to model poverty levels in North Sumatra for the year 2023. The objective of this study is to analyze poverty data containing outliers and to compare the parameter estimation results of the two weighting methods in order to identify the more robust and appropriate approach.

II. METHODS

Unemployment, education, health, and economic growth are several factors that can influence the level of poverty in a region. The percentage of people living in poverty reflects the proportion of the population whose expenditures fall below the poverty line, while the open unemployment rate (TPT) indicates the share of the labor force that remains unemployed. Average years of schooling serve as an indicator of educational attainment, whereas life expectancy reflects the overall quality of public health. Economic growth is measured by changes in GDP at constant prices, representing the increase in goods and services produced over a given period [6].

Many linear regression estimates predict the dependent variable using multiple independent variables. Montgomery et al. [17] formulated the general equation for multiple linear regression models.

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

In matrix form, Equation (1) can be expressed as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad (2)$$

where:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}; \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

According to Gujarati [18], the Ordinary Least Squares (OLS) method is a technique used to estimate regression parameters by minimizing the total sum of squared residuals. Montgomery et al. [17] proposed a method for finding the least squares vector $\hat{\beta}$ by minimizing:

$$\boldsymbol{S(\beta)} = \sum_{i=1}^{n} \boldsymbol{\varepsilon_i^2} = \boldsymbol{\varepsilon'\varepsilon} = \boldsymbol{(y - X\beta)'(y - X\beta)}$$

The least squares estimator ($\boldsymbol{\hat{\beta}}$) is obtained by satisfying:

$$\left.\frac{\partial s}{\partial \beta}\right|_{\beta=\hat{\beta}} = \boldsymbol{-2X'y + 2X'X\hat{\beta} = 0} \quad (3)$$

From Equation (3), the following equations are derived:

$$\begin{aligned} \boldsymbol{-2X'y + 2X'X\hat{\beta}} &= 0 \\ \boldsymbol{X'X\hat{\beta}} &= \boldsymbol{X'y} \quad (4) \\ \boldsymbol{\hat{\beta}} &= \boldsymbol{(X'X)^{-1}X'y} \quad (5) \end{aligned}$$

Hypothesis testing in multiple regression aims to determine the significance of the regression parameters. According to Montgomery et al. [17], the F-test is used to assess whether the predictor variables $x_1, x_2, \ldots, x_k$ collectively have a significant effect on the response variable y. The following are the steps in testing the significance of the regression using the F-test:

a. Hypotheses:
   H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
   H_1: There exists at least one $\beta_j \neq 0$, for $j = 1, 2, \ldots, k$
b. Test Statistic:
   $$F_{statistic} = \frac{SS_R/k}{SS_E/(n-p)} = \frac{MSR}{\text{MSE}} \quad (6)$$
   SS_R *(Regression Sum of Square)* $= \boldsymbol{\hat{\beta}'X'y} - \left(\frac{1}{n}\right)\boldsymbol{y'11'y}$.
   SS_E *(Residual Sum of Square)* $= \boldsymbol{y'y - \hat{\beta}'X'y}$.
c. Decision Rule:
   Reject H_0 if $F_{statistic} > F_{table}$ $(F_{\alpha,k,n-k-1})$ or *p-value* $< \alpha$.

The t-test in regression is used to evaluate the effect of each predictor variable in the linear regression model. According to Montgomery et al. [17], the following are the steps in conducting individual tests for regression coefficients.

a. Hypotheses:
   $H_0 : \beta_j = 0$, for j = 1, 2, …, k
   $H_1 : \beta_j \neq 0$, for j = 1, 2, …, k
b. Test Statistic:
   $$t_{statistic} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \quad ; \quad \text{where} \quad Se(\hat{\beta}_j) = \sqrt{var(\hat{\beta}_j)} = \sqrt{\hat{\sigma}^2 C_{jj}} \quad (7)$$
   $\hat{\sigma}^2 = MSE = \frac{SS_E}{(n-k-1)} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-k-1)}$
   C_{jj} is the diagonal element of $\boldsymbol{(X'X)^{-1}}$
c. Decision Rule:
   Reject H_0 if $|t_{statistic}| > t_{(\alpha/2, n-k-1)}$ or *p-value* $< \alpha$.

Multiple regression requires the assumptions of linearity, normality, homoscedasticity, no autocorrelation, and no multicollinearity. Gujarati [18] states that classical

linear regression assumes a linear relationship between the independent and dependent variables. The Ramsey RESET test is used to detect whether the linear regression model is correctly specified or if there are specification errors. This test is performed by adding the squared or higher-order terms of the fitted values into the regression model and testing their significance.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + \varepsilon_i \quad (8)$$

The test statistic for the Ramsey RESET test is as follows:

$$F = \frac{(R^2_{new} - R^2_{old})/q}{(1-R^2_{new})/(n-k-1-q)} \quad (9)$$

where R^2_{old} is the coefficient of determination of the original model, and R^2_{new} is the coefficient of determination after the additional variables are included. The variable q represents the number of new variables added. The assumption is considered violated when $F > F_{table}\left(F_{\alpha,df_1,df_2}\right)$ or when the *p-value* $< \alpha$.

Normality testing is used to evaluate whether the residuals in a regression model follow a normal distribution. According to Gujarati [18], in the classical linear regression assumptions, each ε_i is assumed to follow a normal distribution, i.e., $\varepsilon_i \sim N(0, \sigma^2)$. The residuals of a regression model must be normally distributed with a mean of 0 and constant variance (σ^2). One of the statistical techniques used to test for normality is the Kolmogorov–Smirnov test. Let the residuals be denoted as $e_1, e_2, \ldots, e_n$ with the distribution function to be tested denoted by $F(e)$, and the hypothesized distribution function denoted by $F_0(e)$. According to Conover [19], the test statistic is defined as:

$$D = \sup_e |F_0(e) - S(e)| \quad (10)$$

where $S(e)$ is the empirical cumulative distribution function (ECDF) or the sample distribution function, and $\sup_e$ denotes the supremum over all residual values, representing the maximum absolute difference between the two functions. $F_0(e)$ is the theoretical cumulative distribution function (theoretical CDF) at the residual value e. The residuals are considered not normally distributed if $D > D_{(n,\alpha)}$ or *p-value* $< \alpha$.

The homoscedasticity test is intended to evaluate whether the variance of the residuals in a regression model is constant, by assessing the presence or absence of differences in the dispersion of errors across observations in the data [20]. According to Gujarati [18], homoscedasticity is denoted by $E(\varepsilon_i^2) = \sigma^2$, where $i = 1, 2, \ldots, n$. The presence of heteroscedasticity can be assessed using the Glejser test, which involves regressing the absolute residual scores $|e_i|$ on the other predictor variables.

$$|\varepsilon_i| = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \upsilon_i \quad (11)$$

Autocorrelation refers to the relationship or correlation between elements in a series of observations that are ordered over time or space [18]. In regression analysis, one of the basic assumptions of the classical linear regression model is that the error components (ε_i) are uncorrelated with each other. The absence of autocorrelation can be denoted as $E(\varepsilon_i, \varepsilon_j) = 0$, for $i \neq j$. The Durbin-Watson test is an approach that can be used to detect the presence of autocorrelation in a regression model.

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{t=1}^{n} e_i^2} \quad (12)$$

When the results of the Durbin-Watson test fall into an inconclusive region, further testing is required. A commonly used approach is the Breusch-Godfrey test or the Lagrange Multiplier (LM) test. This test statistic is based on a regression of the residuals combined with lagged residuals. Autocorrelation among residuals is indicated if $LM > \chi^2_{\alpha,df}$ or if the *p-value* $< \alpha$.

$$LM = n \times R^2 \quad (13)$$

Multicollinearity testing is conducted to evaluate the presence of high linear correlations among the independent variables in a regression model [20]. In a regression model, multicollinearity can be identified using the Variance Inflation Factor (VIF), which is assessed using the following formula:

$$VIF_j = \frac{1}{(1-R_j^2)}, \quad j = 1, 2, \ldots, k \quad (14)$$

According to Montgomery et al. [17], a VIF score greater than 10 indicates serious multicollinearity, which can lead to unreliable estimates of regression coefficients.

Additionally, Montgomery et al. [17] state that the Adjusted R-Square (R^2_{Adj}) is used to assess the extent to which the model explains the variation in the response variable. It is considered superior to the R-Square (R^2) because it accounts for the number of independent variables, thereby helping to prevent overfitting.

$$R^2{}_{Adj} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} \quad (15)$$

Meanwhile, the Mean Squared Error (MSE) is used to measure the variance of the errors.

$$MSE = \frac{SS_E}{(n-k-1)}$$

A smaller MSE value indicates a more accurate model.

Montgomery et al. [12] state that an outlier is an extreme observation whose value differs significantly from the majority. The DFFITS (Difference in Fit Standardized), $DFFITS_i$ is used to assess the influence of the i-th observation on the prediction results when that observation is removed. The formula for $DFFITS_i$ is defined as:

$$DFFITS_i = t_i \left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}} \quad (17)$$

t_i is the studentized deleted residual (R-student) for the i-th case, defined by the formula:

$$t_i = e_i \sqrt{\frac{n-p-1}{SS_E(1-h_{ii}) - e_i^2}} \quad (18)$$

Robust regression is used to handle outliers in a model, offering greater resistance compared to the least squares method against violations of classical assumptions [13]. The model is similar to multiple linear regression but differs in the

use of weighting for parameter estimation. According to Chen [14], there are five estimation approaches in robust regression, among which is the M-estimator method developed by Huber in 1973. The M-estimator is a simple and widely used approach, especially when contamination is more dominant in the independent variables, based on the principle of minimizing an objective function.

$$\min_{\beta} \sum_{i=1}^{n} \rho(u_i) = \min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{e_i}{s}\right) = \min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta_j}{s}\right) \quad (19)$$

ρ represents the weighting function of the residuals, and s is the robust scale estimate. The applied estimate of s is:

$$s = \frac{MAD}{0.6745} = \frac{median|e_i - median(e_i)|}{0.6745} \quad (20)$$

and $u_i = \frac{e_i}{s}$, To minimize Equation (19), the first partial derivative of the function ρ with respect to $\beta_j (j = 0, 1, \cdots, k) = 0$, which leads to:

$$\sum_{i=1}^{n} x_{ij} \psi\left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta_j}{s}\right) = 0, \quad j = 0, 1, \cdots, k \quad (21)$$

According to Draper and Smith (1998), the weighting function is defined in Equation (22) as follows:

$$w_i = w(u_i) = \frac{\psi\left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta_j}{s}\right)}{\left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta_j}{s}\right)} \quad (22)$$

By substituting Equation (22) into Equation (21), the robust regression estimator can be expressed as shown in Equation (23).

$$\sum_{i=1}^{n} x_{ij} w_i \left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta_j}{s}\right) = 0, \quad j = 0, 1, \cdots, k \quad (23)$$

The Iteratively Reweighted Least Squares (IRLS) method is employed to solve Equation (20) through a weighted least squares iterative approach. In each iteration, the weights w_i are updated by recalculating the robust scale estimate and standardized residuals u_i based on the previous residuals, until the parameter estimate $\hat{\beta}$ converges. In matrix notation, the parameter estimation is given by:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y} \quad (24)$$

Thus, the robust regression parameter estimation using IRLS after *m+1* iterations become:

$$\hat{\boldsymbol{\beta}}^{(m+1)} = (\boldsymbol{X}'\boldsymbol{W}^{(m)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(m)}\boldsymbol{Y} \quad (25)$$

In robust regression, the weight function is derived from the derivative of the objective function, which results in an influence function $(\psi(u_i))$. By dividing $(\psi(u_i))$ by u_i, the weight function used in robust regression estimation is obtained [21].

Table 1. Objective Function and Weight Function of Huber and Andrew

Weighting Method	Objective Function	Weight Function	Interval
Huber	$\rho(u_i) = \begin{cases} \frac{u_i^2}{2} & \\ c\lvert u_i\rvert - \frac{c^2}{2} & \end{cases}$	$w(u_i) = \begin{cases} 1 & \lvert u_i\rvert \le c \\ \frac{c}{\lvert u_i\rvert} & \lvert u_i\rvert > c \end{cases}$	
Andrew	$\rho(u_i) = \begin{cases} c\left(1 - \cos\left(\frac{u_i}{c}\right)\right) & \\ 2c & \end{cases}$	$w(u_i) = \begin{cases} \frac{\sin\left(\frac{u_i}{c}\right)}{\frac{u_i}{c}} & \lvert u_i\rvert \le c\pi \\ 0 & \lvert u_i\rvert > c\pi \end{cases}$	

u_i represents the i-*th* standardized residual, while *c* is the tuning constant used to control the influence of outliers in robust regression. For the Huber weighting function, the value of *c* is 1.345 (Fox, 2016), and for the Andrew function, *c* = 1.339 (Montgomery et al., 2012).

III. METHODOLOGY

The data used in this study are secondary data obtained from BPS-Statistics of North Sumatra Province through the publication "Sumatera Utara Dalam Angka 2024". This study involves both response and predictor variables. The Poverty Rate based on data from districts/cities in North Sumatra Province for the year 2023, serves as the response variable (y). The predictor variables are open unemployment rate (x_1), average years of schooling (x_2), life expectancy (x_3), and economic growth (x_4).

The analysis steps to obtain the robust regression model using the M-Estimator method are as follows:

1. Collect data and define the response and predictor variables.
2. Perform initial estimation of multiple linear regression parameters using the Ordinary Least Squares (OLS) method.
3. Conduct assumption tests for multiple linear regression.
4. Perform F-test and individual significance test (t-test) on the regression coefficients.
5. Calculate Adjusted R^2 and Mean Squared Error (MSE).
6. Identify outlier observations formally using the DFFITS value.
7. Estimate regression parameters using the robust M-Estimator method.
8. Determine the best robust regression model based on the specified evaluation criteria.

IV. RESULTS

This study applies robust regression using the M-estimator with Huber and Andrew weighting functions to address outliers in the regression model assumptions, aiming to investigate the relationship between open unemployment, average years of schooling, life expectancy, and economic growth on the poverty rate in North Sumatra.

An initial multiple linear regression analysis using the Ordinary Least Squares (OLS) method was conducted to estimate the relationship between open unemployment, average years of schooling, life expectancy, and economic

growth on the poverty rate in West Sumatra. A regression model was obtained as follows:

$$\hat{y} = 26.803 - 0.221x_1 - 1.410x_2 + 0.240x_3 - 4.343x_4$$

Based on the F-test results, the regression model is significant with a *p-value* of 0.000. This indicates that at least one predictor variable has an effect on the response variable. The individual t-test results show that only the average years of schooling (x_2) with$t_{statistic} = -3.234$ and economic growth (x_4) dengan $t_{statistic} = -5.734$ are significant predictors of the poverty rate. Meanwhile x_1 dan x_3 do not have a significant effect. The goodness-of-fit analysis shows that the Adjusted R^2 is 0.751, meaning that 75.1% of the variation in the poverty rate can be explained by the independent variables in the model, while the remaining variation is explained by other factors.

Classical assumption testing in regression includes tests for linearity, residual normality, multicollinearity, homoscedasticity, and no autocorrelation. The following are the results of the assumption tests conducted on the multiple linear regression model using the Ordinary Least Squares (OLS) method.

Table 2.Assumption Testing of the Multiple Linear Regression Model Using OLS

Assumption	Method	Result	Conclusion
Linearity	Ramsey RESET Test	F(1.247) < $F_{5\%,2,26}$ = 3.369 and *p-value* = 0.304	No specification error in the model (linearity assumption is met).
Normality	Kolmogorov-Smirnov Test	D = 0.150 $D_{(33,5\%)}$ = 0.231 and *p-value* = 0.407	Residuals are normally distributed.
Autocorrelation	Durbin-Watson (DW) Test	$d_U(1.730)$ < $d(1.735)$ < $4 - d_U(2.270)$	No autocorrelation among residuals.
Heteroscedasticity	Glejser Test	All variables have *p-value* > 0.05	Tidak terjadi No heteroscedasticity detected (homoscedasticity assumption is met).
Multicolinearity	*Variance Inflation Factor* (VIF)	All variables have VIF < 10	No multicollinearity detected.

In regression analysis, the presence of outliers can cause distortion in parameter estimation, leading to a less reliable model and potentially inaccurate conclusions [22]. Several commonly used methods for identifying outliers in an analysis include graphical methods, such as residual plots and boxplots, as well as statistical-based methods like DFFITS.

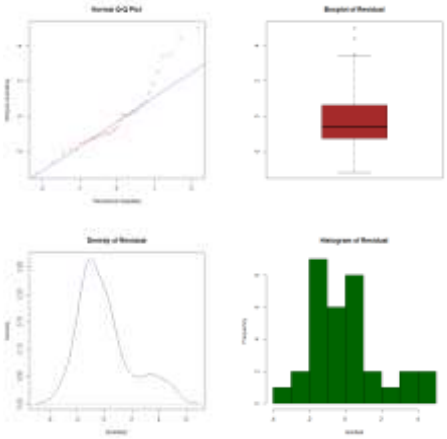


Figure 1.Residual Plot Using OLS Parameter Estimates

Figure 1 indicates the presence of potential outliers in the residual data. Formal detection of outliers was conducted using the DFFITS method to identify both the number and the specific observations that can be classified as outliers. According to Cohen et al. (2003), an observation is considered an outlier if its $|DFFITS| > 1$ for small to medium-sized samples. For large datasets, the threshold is defined by the formula $|DFFITS| > 2\sqrt{p/n}$, where $p = k + 1$, and n is the total number of observations analyzed. Given that the sample size $(n) = 33$ falls into the small to medium category, the threshold used in this study is $|DFFITS| > 1$. Based on this criterion, two outliers were identified: Observation 2 (North Nias Regency) and Observation 25 (West Nias Regency), with DFFITS values of 1.0953 and 1.4405, respectively. Therefore, the robust regression method using M-estimators was applied to address the presence of these outliers.

The robust M-estimator regression was applied using two weighting functions, namely Huber and Andrew, to compare the results of both models. In the initial iteration, the residual values from the OLS regression were used to calculate the Huber weights. The IRLS (Iteratively Reweighted Least Squares) method was then applied iteratively until the model converged at the 18th iteration. The resulting regression model is:

$$\hat{y} = 30.351 - 0.157x_1 - 1.184x_2 + 0.127x_3 - 3.929x_4.$$

The F-test results indicate that the overall model is statistically significant (*p-value* ≈ 0.000). The individual t-tests show that only variables (x_2) and (x_4) are significantly associated with the percentage of the poor population (*p-value* < 0.05). The adjusted R^2 value of 0.857 reflects an improvement in the model's ability to explain the response variable, while the MSE is 2.509. Based on these results, it is

necessary to evaluate the regression assumptions to ensure the reliability of the model used.

The Ramsey RESET test for linearity resulted in an F-value of 2.629 and a *p-value* of 0.091. Since the *p-value* is greater than 0.05, the model does not exhibit specification errors and thus satisfies the linearity assumption. Normality was tested using the Kolmogorov-Smirnov test, which produced a D statistic of 0.221 with a *p-value* of 0.068. A *p-value* greater than 0.05 indicates that the residuals are normally distributed. Due to uncertainty in the Durbin-Watson (DW) value, an alternative autocorrelation test was conducted using the Breusch-Godfrey (LM Test), which yielded an LM statistic of 3.815 with a *p-value* of 0.051. As the *p-value* exceeds 0.05, there is no evidence of autocorrelation among the residuals.

The Glejser test was applied to examine heteroscedasticity, and all variables showed *p-value* greater than 0.05. This indicates the absence of heteroscedasticity, confirming that the model meets the homoscedasticity assumption. Lastly, multicollinearity was assessed using the Variance Inflation Factor (VIF), where all variables had VIF values below 10. Therefore, multicollinearity is not present in the model.

The robust M-estimator regression using the Andrew weighting function converged at the 30th iteration. The resulting regression model is:

$\hat{y} = 28{,}734 - 0{,}077x_1 - 1{,}010x_2 + 0{,}095x_3 - 3{,}544x_4.$

The F-test results indicate that the overall model is statistically significant (*p-value* $\approx 0{,}000$). Individual t-tests show that only variables (x_2) and (x_4) are significantly associated with the percentage of the poor population (*p-value* < 0.05). The adjusted R^2 of 0.814 indicates a strong improvement in the model's explanatory power, with a Mean Squared Error (MSE) of 0.695.

The Andrew approach showed that the Ramsey RESET test for linearity yielded an F-value of 5.841 and a *p-value* of 0.008 for the robust M-estimator regression model. This indicates that the model failed to meet the linearity assumption due to specification errors (*p-value* < 0.05). The Kolmogorov-Smirnov test for normality produced a D statistic of 0.233 with a *p-value* of 0.046, suggesting that the residuals are not normally distributed. The Durbin-Watson (DW) test indicated questionable results, prompting the use of the Breusch-Godfrey (LM Test), which yielded an LM statistic of 9.449 and a *p-value* of 0.002. This confirms the presence of autocorrelation among the residuals *p-value* < 0.05. However, the Glejser test showed that all variables had *p-value* greater than 0.05, indicating the model satisfies the homoscedasticity assumption. The Variance Inflation Factor (VIF) values were all below 10, suggesting that multicollinearity is not an issue in the model.

The best model in robust M-estimator regression was selected based on assumption tests, adjusted R^2, and Mean Squared Error (MSE). The model using the Huber weighting function outperformed the Andrew model, as it had a higher adjusted R^2 (0.857203) and met all regression assumptions, despite having a slightly higher MSE. Among the explanatory variables, average years of schooling (x_2) and economic growth (x_4) had significant effects on the percentage of poor population (y). The variable average years of schooling had a coefficient of -1.184, indicating that each additional year of education reduces the poverty rate by 1.184%. The variable economic growth had a coefficient of -3.929, meaning that a 1% increase in economic growth can lower the poverty rate by 3.929%, which aligns with economic theory. Therefore, the robust M-estimator regression with the Huber weighting function is selected as the best model for estimating the poverty rate in North Sumatra in 2023.

V. CONLUSION

The analysis results indicate that the use of the Ordinary Least Squares (OLS) method in regression revealed two outliers, which led to inefficient and biased parameter estimates. To address this issue, robust M-estimator regression with Huber and Andrew weighting functions was employed. A comparison of the two methods, based on adjusted R^2 and Mean Squared Error (MSE) values, showed that the best model was obtained using the robust M-estimator regression with the Huber weighting function.

$\hat{y} = 30{,}351 - 0{,}157x_1 - 1{,}184x_2 + 0{,}127x_3 - 3{,}929x_4.$

Based on the robust M-estimator regression using the Huber weighting function, it can be concluded that the percentage of poor population in North Sumatra in 2023 is significantly explained by variable x_2 (average years of schooling) with a $t_{statistic}$ of -3.2339, and variable x_4 (economic growth) with a $t_{statistic}$ of -5.7344. The model achieved an adjusted R^2 of 85.7203% and a Mean Squared Error (MSE) of 2.5093. This indicates that each additional year of schooling reduces poverty by approximately 1.184%, while a 1% increase in economic growth reduces poverty by about 3.929%.

REFERENCES

1. N. G. Mankiw, Principles of Economics (10th ed.), Cengage Learning , 2019.
2. I. S. Chaudhry, S. Malik dan A. ul Hassan, "The Impact of Socioeconomic and Demographic Variables on Poverty: A Village Study," The Lahore Journal of Economics, vol. 1, no. 14, pp. 39-68, 2009.
3. M. P. Todaro dan S. C. Smith, Economic Development (12th ed), New York: Pearson, 2015.
4. L. O. Mardiyana dan H. M. Ani, "The Effect of Education and Unemployement on Poverty in East Java Province," dalam In IOP Conference Series: Earth and Environmental Science, 2019.

5. Suryawati, "Memahami Kemiskinan Secara Multidimensional," Jurnal Manajemen Pelayanan Kesehatan , vol. 3, no. 8, pp. 121-129, 2005.
6. Badan Pusat Statistik, Statistik Indonesia 2023, Badan Pusat Statistik.
7. Badan Pusat Statistik Provinsi Sumatera Utara, Provinsi Sumatera Utara Dalam Angka 2024, Badan Pusat Statistik Provinsi Sumatera Utara, 2024.
8. S. N. Aini, "Pengaruh Pertumbuhan Ekonomi, Pendidikan, Pengangguran, dan Ketimpangan Pandapatan Terhadap Kemiskinan di Provinsi Jawa Timur," Buletin Ekonomika Pembangunan , vol. 1, no. 4, pp. 20-36, 2023.
9. R. Susanto dan I. Pangesti, "Pengaruh inflasi dan pertumbuhan ekonomi terhadap tingkat kemiskinan di Indonesia," Journal of Applied Bussiness and Economics (JABE) , vol. 2, no. 7, pp. 271-278, 2020.
10. P. J. Rousseeuw dan A. M. Leroy, Robust Regression and Outlier Detection, New York: John Wiley & Sons, 1987.
11. P. J. Huber, Robust Statistics, New York: John Wiley & Sons, 1981.
12. R. A. Maronna, R. D. Martin, V. J. Yohai dan M. Salibian-Barrera, Robust Statistics: Theory and Methods (with R), New York: John Wiley & Sons, 2019.
13. N. R. Draper dan S. H. H., Applied Regression Analysis (3rd ed.), New York: Wiley, 1998.
14. C. Chen, "Robust Regression and Outlier Detection with The ROBUSTREG Procedure," ASA Institue Inc., 2002.
15. Damayanti dan M. Susanti, "Analisis Regresi Robust Estimasi-M Pembobot Huber Dan Tukey Bisquare Pada Tingkat Kemiskinan Indonesia," Jurnal Kajian dan Terapan Matematika, vol. 2, no. 8, pp. 377-388, 2019.
16. D. Deria, A. Hoyyi dan Mustafid, "Regresi Robust Estimasi-M Dengan Pembobot Andrew, Pembobot Ramsay dan Pembobot Welsch Menggunakan Software R," Jurnal Gaussian , vol. 2, no. 10, pp. 130-141, 2019.
17. D. C. Montgomery, E. A. Peck dan G. G. Vining, Introduction to Linear Regression Analysis (5th ed.), Hoboken, NJ: John Wiley & Sons, 2012.
18. D. N. Gujarati, Basic Econometrics (4th ed.), McGraw-Hill, 2003.
19. W. J. Conover, Practical Nonparametric Statistics (3rd ed), Wiley, 1999.
20. I. Ghozali, Aplikasi Analisis Multivariate dengan Program IBM SPSS 25 (Edisi ke-9), Semarang: Badan Penerbit Universitas Diponegoro, 2018.
21. J. Fox, Applied Regression Analysis and Generalized Linear Models (3rd ed.), SAGE Publication, 2016.
22. I. G. A. M. Srinadi, "Pengaruh Outlier Terhadap Estimator Parameter Regresi dan Metode Regresi Robust," Prosiding Konferensi Nasional Matematika XVIII , vol. 1, no. 1, pp. 1259-1266, 2014.