WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Robust Regression M-Estimator with Huber Weight in Modeling Human Development Index (HDI) of Papua Province, Indonesia 2023

Erfani Nurul Hamidah[a], Suparti[b]*, Miftahul Jannah[c]

[a,b,c]*Department of Statistics, Faculty of Science and Mathematics, UniversitasDiponegoro, Semarang, Indonesia*
Email id :[b] suparti@live.undip.ac.id

A B S T R A C T

The Human Development Index (HDI) is used to measure the quality of life and community welfare. Indonesia's HDI continues to increase, but interregional disparities remain high. Papua Province has the lowest HDI value in Indonesia in 2023, which is 63.01 reflecting the low level of human development achievement in the region. The HDI value is calculated from the dimensions of health, education, and expenditure. Poverty has values that vary between regions so that the data used in the analysis have the potential to contain outliers that may affect the estimation results. This study aims to build a robust regression model of M-Estimation with Huber weighting on the HDI data of districts/cities in Papua Province in 2023 to obtain parameter estimates that are stable and efficient against outliers. The dependent variable used is HDI, while the independent variables consist of deterministic and stochastic variables. The deterministic variables include Life Expectancy, Expected Years of Schooling, Mean Years of Schooling, and Expenditure per Capita. The stochastic or random variable used is the Percentage of Poor Population. The estimation process was carried out using the Iteratively Reweighted Least Squares (IRLS) algorithm with initial estimates obtained from the Ordinary Least Squares (OLS) method. The robust regression model with Huber weighting produced a coefficient of determination (R^2) of 0.9962 indicating that 99.62% of the variation in HDI can be explained by the independent variables in the model.

Keywords: *Robust* Regression, M-estimation, Huber Weight, HDI, Outlier

Introduction

The Human Development Index is an indicator used to measure development achievements in health, education, and living standards (BPS, 2020). It has been widely adopted as a benchmark for evaluating regional development performance since the United Nations Development Programme introduced it in 1990. The HDI values range from zero to one hundred, with higher values indicating better development outcomes. Despite the national improvement in the HDI, regional disparities remain significant.

In 2023, Indonesia recorded an HDI of 74.39, showing a positive trend compared with previous years. However, regional inequalities are evident, with DKI Jakarta achieving an HDI of 83.55, the highest in the country, while Papua recorded the lowest value at 63.01 (BPS, 2023). This indicates that human development in Papua still lags behind that of other provinces and reflects limitations in the availability of health services, educational attainment, and economic well-being. Understanding the factors that contribute to this disparity is essential for strengthening regional development planning.

The HDI is determined by several key components, including life expectancy, expected years of schooling, mean years of schooling, and adjusted expenditure per capita (BPS, 2020). These indicators represent the main dimensions of human development and are commonly used to analyze variations in HDI across regions. In addition to these components, socioeconomic factors contribute to human development outcomes. One of the most influential factors is poverty. Papua has the highest poverty rate in Indonesia, reaching 26.03 percent in 2023, which limits access to education, health facilities, and income opportunities (BPS Papua, 2023). Poverty tends to lower human development achievements, creating a negative association between these two variables (Mirza, 2011).

Multiple linear regression is often used to examine the HDI determinants. However, the Ordinary Least Squares method relies on assumptions such as normality, homoscedasticity, and the absence of influential observations (Gujarati, 2009). Socioeconomic data, particularly from regions with high variability, often violate these assumptions. Outliers may distort the estimation process and lead to biased and inefficient results (Willems & Aelst, 2005). This issue is relevant for HDI modelling in Papua because the data contain substantial variability across districts and cities in the province. Therefore, classical regression methods may not be sufficient to produce stable estimates.

Robust regression provides an alternative approach that reduces sensitivity to extreme observations and produces more reliable parameter estimates in the presence of data irregularities (Olive, 2005). One of the most widely used robust methods is M-estimation, which limits the influence of large residuals by applying a weighting function. The Huber weighting function is particularly effective because it maintains efficiency under near-normal error distributions while controlling the influence of outliers (Huber, 1973). This makes it suitable for modeling socioeconomic data with heterogeneous patterns.

Although previous research has applied robust methods to socioeconomic indicators, studies that implement Huber M-Estimation specifically for modeling HDI in Papua are limited. The region's unique characteristics and high variability indicate the need for an estimation method that can handle the influential observations and provide stable results. This study aims to construct a robust regression model using Huber M-Estimation for the Human Development Index of districts and cities in Papua Province in 2023. The model incorporates both the main components of HDI and poverty as additional explanatory variables to provide a more comprehensive representation of human development in the region.

Structure Literature Review

The Human Development Index is a multidimensional indicator used to describe progress in health, education, and living standards. The construction of the index refers to the operational definitions published by the BPS, which outlines its components and measurement procedures (BPS, 2020). The health dimension is represented by life expectancy, which reflects the population mortality patterns and long-term health conditions (BPS, 2024). Educational attainment is captured through the expected years of schooling, which describes the projected duration of formal education for children under current conditions (BPS, 2015). The actual educational achievement of the adult population is reflected in the mean years of schooling, which measures completed schooling years among individuals aged twenty-five years and above (BPS, 2024). Economic welfare within the HDI framework is assessed using adjusted expenditure per capita, which reflects consumption capability after accounting for price differences across regions (BPS, 2024).

Human development outcomes are also influenced by broader socioeconomic conditions. Poverty has been identified as a factor that limits access to education, healthcare, and income opportunities. These limitations weaken human capital accumulation and suppress overall development indicators (Mirza, 2011). Empirical findings in development economics consistently show that higher poverty levels tend to coincide with lower human-development achievements. Multiple linear regression is one of the most widely used approaches for analyzing the relationships between human development indicators and their explanatory variables. The Ordinary Least Squares method is traditionally preferred because it yields efficient and unbiased estimators when its underlying assumptions are fulfilled (Gujarati, 2009). These assumptions require normal residuals, constant variance, independence of errors, and the absence of multicollinearity. However, socioeconomic datasets often violate these criteria owing to their variability and the presence of extreme observations. Outliers can substantially alter parameter estimates and compromise statistical inferences (Willems and Aelst, 2005). Further analysis demonstrated that influential observations may distort regression relationships and reduce the reliability of model conclusions (Indra et al., 2013).

To address the issues associated with outliers, robust regression techniques have been developed to provide parameter estimates that remain stable when classical assumptions are violated. Robust statistical methods reduce the effect of extreme observations and allow more reliable modeling of heterogeneous datasets (Olive, 2005). One commonly used approach is M-estimation, which modifies the contribution of residuals through a weight function that reduces the influence of large deviations (Huber, 1973). The theoretical properties of these influence functions and their robustness characteristics were formalized in the foundational work on robust statistics (Hampel et al., 1986). M-estimation is typically implemented using the Iteratively Reweighted Least Squares algorithm, which updates the observation weights iteratively until convergence is achieved (Montgomery et al., 2012).

The application of robust regression has been widely discussed in empirical research on socioeconomic indicators. Studies comparing estimation methods for HDI modeling show that M-estimation with Huber weights can provide more stable results than classical estimation techniques when the dataset includes variability or potential outliers (Atamia et al., 2021). These findings support the relevance of robust techniques for development-related analyses, particularly in contexts with heterogeneous conditions.

Material and Method

The data used in this study were secondary data obtained from BPS-Statistics Indonesia and the Papua Provincial Office. The dataset consists of the Human Development Index (HDI) and its associated socioeconomic indicators for 29 districts and cities in Papua Province for 2023. The response variable is the Human Development Index (HDI), while the predictor variables include life expectancy, expected years of schooling, mean years of schooling, adjusted expenditure per capita, and the percentage of the poor population.

Multiple linear regression was used to examine the relationship between the dependent variable and several independent variables. When more than one predictor is included, the multiple linear regression model is written as

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i,} \tag{1}$$

In matrix form, the regression model is expressed as

$$\boldsymbol{Y = X\beta + \varepsilon} \tag{2}$$

Before estimating the parameters, multicollinearity among the predictors must be checked. One commonly used diagnostic is the Variance Inflation Factor (VIF), defined as

$$\boldsymbol{VIF_j = \frac{1}{(1-R_j^2)}, j = 1, 2, \ldots, k} \quad (3)$$

where $\boldsymbol{R_j^2}$ is the coefficient of determination obtained by regressing variable $\boldsymbol{X_j}$ on the remaining predictors.

Parameter estimation begins with the Ordinary Least Squares (OLS) method, which minimizes the sum of the squared errors. The objective function is

$$\boldsymbol{S(\beta) = e^T e = (Y - X\beta)^T (Y - X\beta)} \quad (4)$$

Taking the derivative of $\boldsymbol{S(\beta)}$ with respect to $\boldsymbol{\beta}$ and setting it equal to zero yields

$$\boldsymbol{X^T X\hat{\beta} = X^T Y} \quad (5)$$

Solving the normal equations produces the OLS estimator

$$\boldsymbol{\hat{\beta} = (X^T X)^{-1} X^T Y} \quad (6)$$

Residuals from the initial OLS fit were then evaluated using classical regression assumptions, including normality, independence, and homoscedasticity.Normality was examined visually using a Q–Q plot and formally using the Kolmogorov–Smirnov test. The Kolmogorov–Smirnov statistic is

$$\boldsymbol{D = sup|S(x) - F_0(x)|} \quad (7)$$

where $\boldsymbol{S(x)}$ is the empirical distribution function and $\boldsymbol{F_0(x)}$ is the theoretical normal distribution function.

Independence of residuals is evaluated using the Runs test, with the test statistic

$$z = \frac{R-E(R)}{\sigma(R)} \quad (8)$$

where $\boldsymbol{R}$ is the number of runs in the residual sequence.

Homoscedasticity is examined using the Breusch–Pagan statistic,

$$\boldsymbol{BP = nR^2 \sim \chi^2_{df}} \quad (9)$$

obtained from an auxiliary regression of the squared residuals on the predictor variables.

The overall significance of the regression model is tested using the F-statistic,

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{(n-k-1)}} \quad (10)$$

while the significance of each predictor is assessed using the t-statistic

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \quad (11)$$

Model fit is summarized using the coefficient of determination

$$\boldsymbol{R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}} \quad (12)$$

Outliers and influential points are identified using the Difference in Fits Standardized (DFFITS) measure, defined as

$$(DFFITS_i) = t_i \times \sqrt{\frac{h_{ii}}{1-h_{ii}}} \quad (13)$$

The studentized deleted residual is

$$\boldsymbol{t_i = \frac{e_i}{\sqrt{S^2_{(i)}\,(1-h_{ii})}}}\ , \quad (14)$$

where the deleted residual variance is computed as

$$S^2_{(i)} = \frac{(n-p)MSE - \frac{e_i^2}{1-h_{ii}}}{n-p-1} \quad (15)$$

The leverage values are derived from the diagonal elements of the hat matrix:

$$\boldsymbol{h_{ii} = X_i^T (X^T X)^{-1} X_i} \quad (16)$$

with fitted values expressed as

$$\boldsymbol{\hat{Y} = X\hat{\beta} = HY}, \qquad \boldsymbol{H = X(X^T X)^{-1} X^T} \quad (17)$$

Observations with large DFFITS values were classified as influential.

To reduce the sensitivity of the regression estimates to outliers, a robust regression is applied by replacing the squared loss function with a more resistant objective function:

$$\min_{\beta} \sum_{i=1}^{n} \rho\,(e_i) = \min_{\beta} \sum_{i=1}^{n} \rho\,(Y_i - X_i^T \beta) \quad (18)$$

Standardized residuals are defined as

$$u_i = \frac{Y_i - X_i^T\beta}{s} = \frac{e_i}{s} \quad (19)$$

where the robust scale estimator is computed as

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

The Huber objective function is

$$\rho(u_i) = \begin{cases} \frac{1}{2}u_i^2 & , |u_i| \le c \\ c\left(|u_i| - \frac{1}{2}c^2\right), & |u_i| > c \end{cases} \quad (21)$$

with tuning constant $c = 1.345$.

The corresponding influence function is

$$\psi(u_i) = \begin{cases} u_i, & |u_i| \le c \\ c.\,\text{sign}\,(u_i), & |u_i| > c \end{cases} \quad (22)$$

Where

$$\text{sign}\,(u_i) = \begin{cases} 1, u_i > 0 \\ 0,\ u_i = 0 \\ -1,\ u_i < 0 \end{cases} \quad (23)$$

The weight function for M-estimation is

$$w(u_i) = \begin{cases} 1, & |u_i| \le c \\ \frac{c}{|u_i|}, & |u_i| > c \end{cases} \quad (24)$$

Parameter estimation in the robust regression model was performed using the Iteratively Reweighted Least Squares algorithm. The algorithm begins with initial coefficient estimates from the Ordinary Least Squares method, after which residuals and standardized residuals are computed. Weights are then assigned based on the Huber function, and the coefficients are updated using Weighted Least Squares according to

$$\hat{\beta}^{(m+1)} = \left(X^T W^{(m)} X\right)^{-1} X^T W^{(m)} Y \quad (25)$$

The procedure is repeated until convergence is achieved, defined as

$$\left|\hat{\beta}_j^{(m+1)} - \hat{\beta}_j^{(m)}\right| \le 10^{-4} \quad (26)$$

Result and Discussion

This study applies robust regression using an M-estimator with the Huber weighting function to examine the presence of influential observations in the dataset. This study aimed to examine the influence of life expectancy, expected years of schooling, average years of schooling, per capita expenditure, and the percentage of poor people on the Human Development Index (HDI) in Papua Province.

An initial multiple linear regression analysis was conducted using the Ordinary Least Squares (OLS) method to estimate the relationship between predictors and the HDI. The estimated OLS model is as follows:

$$\hat{Y} = -1.6969 + 0.2393X_1 + 1.3587X_2 + 1.3651X_3 + 0.0009X_4 - 0.0199X_5$$

Because the model violates the homoscedasticity assumption, the F-test, t-test, and standard errors in the OLS framework are not reliable for statistical inference. Although the OLS model produces an R^2 value of 0.9962, this measure cannot be used to assess model adequacy because the underlying assumptions are not satisfied. Therefore, robust regression is required to obtain stable parameter estimates and valid inference. Classical assumption testing was carried out to ensure the validity of the OLS model. The results are summarized in Table 1.

Table 1. Assumption Testing of the Multiple Linear Regression Model

Assumption	Method	Result	Conclusion
Normality	Kolmogorov–Smirnov Test	p-value > 0.05	Residuals are normally distributed
Autocorrelation	Runs Test	p-value > 0.05	No autocorrelation among residuals

Heteroscedasticity	Breusch–Pagan Test	p-value < 0.05	Heteroscedasticity is present
Multicollinearity	VIF	All VIF < 10	No multicollinearity detected

Although the residuals satisfy normality, independence, and multicollinearity assumptions, the Breusch–Pagan test indicates the presence of heteroscedasticity. Because this condition violates the OLS requirement of constant variance, OLS standard errors, t-tests, and F-tests cannot be considered reliable for inference. To assess the stability of the OLS model, influential observations were identified using the DFFITS statistic. Four observations Jayawijaya, Supiori, Nduga, and Jayapura City have absolute DFFITS values greater than 1, indicating strong influence on the model.

To overcome the weaknesses of the OLS method, robust regression using the M-estimator with the Huber weighting function was applied. The estimation was performed using the Iteratively Reweighted Least Squares (IRLS) algorithm, which reached convergence at the twelfth iteration. The final robust model obtained is:

$$\hat{Y} = -1.4082 + 0.2339X_1 + 1.3450X_2 + 1.3526X_3 + 0.0010X_4 - 0.0152X_5$$

The F-test confirms that the robust model is statistically significant. The t-test results show that life expectancy, expected years of schooling, mean years of schooling, and expenditure per capita remain significant predictors of HDI, while the poverty rate remains insignificant. The robust model achieves an R^2value of 0.9962, indicating excellent explanatory power.

From an economic perspective, the findings reflect the actual development conditions in Papua. Life expectancy, expected years of schooling, and mean years of schooling show strong positive effects on HDI because regions with better access to health services and education tend to achieve higher human development outcomes. Per capita expenditure also contributes positively, indicating that districts with stronger economic capacity are better able to support improvements in welfare. The poverty rate, however, is statistically insignificant. This may be due to the large variability of poverty levels across districts, as well as the fact that HDI already captures welfare dimensions through education and health indicators. As a result, the effect of poverty may be overshadowed by stronger predictors such as education variables and life expectancy.

Conclusion

This study examined the factors influencing the Human Development Index (HDI) in Papua Province in 2023 using a robust regression approach. The findings showed that the Ordinary Least Squares (OLS) method was not suitable for the dataset due to the presence of heteroskedasticity and influential observations, which reduced the efficiency of the parameter estimates. To address these limitations, the M-estimation method with Huber weighting was applied through the Iteratively Reweighted Least Squares (IRLS) algorithm. The algorithm converged successfully and produced stable and reliable parameter estimates. The robust regression model demonstrated that life expectancy, expected years of schooling, mean years of schooling, and per capita expenditure positively contributed to the improvement of HDI, while the percentage of poor population negatively affected human development outcomes. These results highlight the critical influence of health, education, and economic conditions on regional development disparities within Papua. The model achieved a coefficient of determination of 0.9962, indicating that the predictors explained nearly all of the variation in HDI across districts and cities in Papua. The Breusch–Pagan test further confirmed that the robust model satisfied the homoscedasticity assumption, showing a substantial improvement compared to the OLS model. Overall, the M-estimation with Huber weighting proved effective in mitigating the adverse effects of outliers and producing more consistent and efficient parameter estimates. This method provides a more reliable analytical framework for modeling socio-economic indicators such as HDI, particularly in regions characterized by high variability and extreme observations.

References

Atamia, F., Hakim, R., & Yulianto, D. (2021). Comparative analysis of S-Estimation and M-Estimation with Huber weighting in modeling the Human Development Index in Indonesia. *Gaussian Journal*, 10(4), 512–522.

Badan Pusat Statistik. (2020). *Indeks Pembangunan Manusia 2020*. Jakarta: BPS-Statistics Indonesia.

Badan Pusat Statistik. (2023). *Indeks Pembangunan Manusia menurut Provinsi 2023*. Jakarta: BPS-Statistics Indonesia.

Badan Pusat Statistik Provinsi Papua. (2023). *Profil Kemiskinan Provinsi Papua Maret 2023*. Jayapura: BPS-Statistics Papua.

Damayanti, R., Sutrisno, H., & Lestari, A. (2024). Robust regression using M-Estimation with Huber weights for poverty modeling in Indonesia. *Journal of Applied Statistics and Economics*, 8(1), 45–55.

Gujarati, D. N. (2009). *Basic Econometrics* (5th ed.). New York, NY: McGraw-Hill.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York, NY: Wiley.

Huber, P. J. (1973). Robust estimation of a location parameter. *Annals of Statistics*, 1(1), 73–101.

Indra, D., Sari, R. K., & Putra, A. (2013). The effect of outliers on parameter estimation in multiple regression models. *Journal of Mathematics and Statistics*, 6(2), 155–161.

Mirza, A. (2011). An analysis of poverty and human development in Indonesia. *Journal of Development Policy*, 12(3), 187–198.

Olive, D. J. (2005). *Applied Robust Statistics*. Southern Illinois University.

Salim, A., Wenda, E., & Kogoya, M. (2025). The relationship between poverty and Human Development Index in Papua Province. *Papua Development Review*, 4(1), 22–34.

Willems, G., & Van Aelst, S. (2005). Fast and robust estimation in linear regression. *Computational Statistics & Data Analysis*, 48(4), 703–720.