Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 3 No. 2 2023, 84 - 92
DOI: *https://doi.org/10.22487/27765660.2023.v3.i2.16923*
eISSN: 2776-5660

REGRESSION ANALYSIS OF ROBUST ESTIMATION-S WITH TUKEY BISQUARE WEIGHTING ON POVERTY LEVEL ON SULAWESI ISLAND

Sandra Saputri , Nur'eni , Sitti Masyitah Meliyana R[3*]

[1,2]Statistics Study Program, Tadulako University
[3]Statistics Study Program, State University of Makassar

***e-mail** : *sittimasyitahmr@unm.ac.id*

ABSTRACT

Poverty is a situation where a person experiences difficulty in meeting basic needs. There are several factors that influence poverty, including population, unemployment, gross regional domestic product, human development index, average years of schooling and labor force participation rate. Therefore, it is necessary to carry out regression analysis to determine the relationship between one variable and other variables. One method for estimating regression parameters is the least squares method. Some classic assumptions are not met because there are outlier data. Outliers are data that do not follow the overall distribution pattern, so a method is used that can overcome outliers, namely the S-estimation robust regression method with the Tukey bisquare weighting function. The results of the research show that the best model was obtained from robust S-estimation regression with Tukey bisquare weighting, namely $Y = -0{,}21023 + 0{,}46522\,x_1 + 0{,}16551\,x_4 - 0{,}33444\,x_5 + 0{,}15864\,x_6$ *. factors that influence the level of poverty on the island of Sulawesi, namely Population Number* (X_1)*, Human Development Index (* X_4*), Average Years of Schooling (* X_5*) and, Force Participation Level. Work* (X_6).

Keywords: Regression Analysis, Poverty, Outliers, *Robust Regression*

Cite: Saputri, S., Nur'eni, & Meliyana, R. M. S., (2023). *Regression Analysis of Robust Estimation-S With Tukey Bisquare Weighting On Poverty Level On Sulawesi Island*. *Parameter: Journal of Statistics, 3*(2), 84-92, https://doi.org/10.22487/27765660.2023.v3.i2.16923.

*Received: Dec 2023; Reviewed: Dec 2023; Published: Dec 2023

INTRODUCTION

National development is an effort to make society just and prosperous. Various efforts have been made for development, especially in areas with relatively high levels of poverty which continue to increase every year. Another factor that causes poverty was presented by Mirah et al (2020) and Hasanah et al (2021) who said that the level of labor force participation and the average number of years of schooling have an influence on poverty.

A method called regression analysis is used to determine the relationship between the independent variable and the dependent variable. One of the methods used to estimate regression coefficients is the least squares method (MKT). The least squares method can estimate parameters by minimizing the sum of residual squares. The MKT model requires several assumptions that must be met. However, it is often the case that assumptions are not met due to outliers, so the use of the least squares method is not appropriate. Therefore, robust regression is used to overcome the presence of outliers.

Robust Regression is a method used to analyze data that is influenced by outliers so as to produce a model that is *robust* against *outliers.* One of the methods used in robust regression is S-estimation which is one of the estimation techniques that has the highest breakdown point value of up to 50%, which means that S-estimation can overcome half of the outliers. This method uses the Tukey Bisquare weighting function to produce a weighting scale with iteration until the estimator obtained converges. The iteration process used in the Tukey bisquare weighting is also less compared to other weightings.

Several previous studies have been carried out using robust regression analysis, including Susanti et al (2014) regarding M-Estimation, S-Estimation, and MM-Estimation in Robust Regression analysis on corn production data in Indonesia. It was found that S-estimation produces a better model. Good. Based on the description above, this research uses the Robust Estimation-S Regression method with Tukey Bisquare weighting to analyze the factors that influence the level of poverty on Sulawesi Island.

DFFITS Test

DFFITS test is a measurement that provides information regarding the influence of the first case *on* the overall regression equation. The following is the DFFITS test hypothesis as follows:

H_0 : e_i= 0(outliers have no effect)
H_1 : $e_i \neq 0$(influential outliers)
So, outlier data detection is used from the *DFFITS value* which is calculated using the formula:

$$|DFFTIS| = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}} e_i \left[\frac{n-p-1}{JKG(1-h_{ii})-e_i^2}\right]^{\frac{1}{2}} \quad (1)$$

Where:
JKG : Sum of squared errors
K : Number of independent variables
n : Number of samples
X_i: Data of the i-th variable *X*

The test criteria for H_0rejection are values $|DFFITS| > 1$ for small data clusters $(n \leq 30)$,and large data clusters $(n > 30)$use values$|DFFITS| > 2\sqrt{\frac{p}{n}}$, Where $p = k + 1$.

S-Estimation

The S-estimation was first introduced by Rousseew and Yohai (1984) as a *robust estimate* that can reach *a breakdown point* of up to 50%. *Breakdown points* are used to address outlier problems before observations affect the model. *Scale* estimates can be seen using the formula.

$$\beta_s = \arg\min_{\beta} \sigma_s \ (\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n) (1.7)$$

S-estimation has a breakdown *point value* written as $\frac{b}{\max \rho(e)} = 0{,}5.$ With is the σ_s minimum *robust* scale estimator value and satisfies:

$$\min \sum_{i=1}^{n} \rho \left(\frac{Y_i \sum_{j=0}^{k} X_{ij}\beta}{S_s}\right) \quad (2)$$

Where:
β_s: Regression parameter coefficients
e_i: Residual value
σ_s*Robust* scale estimator value
X_{ij}: Value of the i- *th observation*
n : Number of observations
i : The number of observation units
Y_i: Dependent variable *i*

With a *robust scale* σ_s:

$$\sigma_s = \sqrt{\frac{n\sum_{i=1}^{n}(e_i)^2-(\sum_{i=1}^{n}(e_i)^2}{n(n-1)}} \quad (3)$$

The initial estimates used are as follows:

$$\sigma_s = \frac{median\,|\varepsilon_i - median(\varepsilon_i)|}{0{,}6745} \quad (4)$$

Solve equation (1.10) by finding its derivative $\beta$to obtain:

$$\sum_{i=1}^{n} X_{ij}\psi\left(\frac{Y_i-\sum_{j=0}^{k} X_{ij}\beta}{\sigma_s}\right) = 0 \qquad j = 0,1,2,\dots,k \quad (5)$$

ψ is called the influence function which is the derivative of $\rho(\rho' = \psi)$. The derivative of the function $\rho$is:

$$\psi(u_i) = \rho'(u_i) = \begin{cases} u_i\left(1-\left(\frac{u_i}{c}\right)^2\right)^2 & |u_i| \le c \\ 0 & |u_i| > c \end{cases}$$

Iteratively Reweighted Least Square (IRLS) is used to solve equation (11) which can be written as follows:

$$\sum_{i=1}^{n} X_{ij}w_i\left(Y_i\sum_{j=0}^{k} X_{ij}\beta\right) = 0, \qquad j = 0,1,2,\dots,k \quad (6)$$

It is assumed that an initial estimate β^0exists $\hat{\sigma}_i$when using IRLS. j is the number of parameters to be estimated, then the equation can be written as:

$$\sum_{i=1}^{n} X_{ij}w_i^0\left(Y_i - \sum_{j=0}^{k} X_{ij}\beta^0\right) = 0, \qquad j = 0,1,2,\dots,k \quad (7)$$

The estimated parameters in the first iteration are β^0 and the weight values in the initial iteration are w_i^0. Then the equation can be written as:

$$X^TWX\beta = X^TWY \quad (8)$$

Where *W* is an nxn matrix with diagonal elements containing weights. By providing an estimator $\beta$is:

$$\beta = (X^TWX)^{-1}(X^TWY)$$

***Tukey Bisquare* Weighting**

Tukey bisquare weighting function can be defined as follows [11].

$$\rho(u_i) = \begin{cases} \frac{c^2}{6}\left\{1-\left[1-\left(\frac{u_i}{c}\right)^2\right]^3\right\} & |u_i| \le c \\ \frac{c^2}{6} & |u_i| > c \end{cases}$$

Since $\rho(u_i)$it is the first derivative of the *bisquare Tukey influence function*, the following equation can be obtained:

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

Tukey Bisquare weighting function is:

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

Where the residual scale in the i-th observation *is* the value $u_i$and the c value is the *constant tuning value* that has been determined to determine the level of robustness

.

Parameter Testing

F test

The F test is a test used to determine whether the independent variables as a whole have a significant effect simultaneously on the dependent variable, with the following hypothesis [4].

$H_0: \beta_j = 0$(all independent variables have no significant effect simultaneously on the dependent variable)

$H_1: \beta_j \neq 0$(all independent variables have a significant effect simultaneously on the dependent variable)

The test criteria are if the value $F_{hitung} > F_{(\propto,k,n-k-1)}$or significant value is <0.05, then reject it H_0, which means that all independent variables have a significant effect simultaneously on the dependent variable.

T test

The T test was carried out to see the effect of each independent variable on the dependent variable individually with hypothesis [6].

$H_0 : \beta_j = 0$ (The independent variable has no effect on the dependent variable)

$H_1 : \beta_j \neq 0$**(** The independent variable has an effect on the dependent variable)

If the test criteria $|t_{hitung}| > t_{\left(1-\left(\frac{\alpha}{2}\right),n-k-1\right)}$or significant value is <0.05, it is said to be rejected H_0, which means the independent variable has an effect on the dependent variable.

MATERIALS AND METHODS

Data

The data in this research is secondary data obtained from the Central Statistics Agency. The variables that will be used in this research are the response variable (*y*) and the predictor variable (*x*) which can be seen in the table as follows:

Table 1. Research data

Variable	Operational definition	Unit
Poverty (Y)	The percentage of the population that is below the poverty line	Percent
Total population (X_1)	The number of people who live in a region or region	Thousand Souls
Unemployment (X_2)	People entering the workforce who are looking for work	Percent
Gross Regional Domestic Product (X_3)	Knowing the economic conditions of a region or region in a certain period	Billions of Rupiah
Human Development Index (X_4)	A comparative measure of life expectancy, literacy, education and living standards	Percent
Average Years of Schooling (X_5)	The average number of years spent by the population in all types of education	Percent

Labor Force Participation Rate (X_6)	The percentage of the working age population who are in the labor force	Percent

Method

The methodology in this research uses *Robust Estimation-S Regression with Tukey Bisquare* Weighting Function which is carried out using *R.4.1 software.* The stages of analysis carried out are as follows:

1. Data collection.
2. Descriptive analysis.
3. Estimating regression parameters using MKT.
4. Carrying out classical assumption testing.
5. Detecting outliers with the DFFITS Test.
6. Estimating *robust regression coefficients* using *scale estimation* with the *Tukey bisquare weighting function.*
   a. Calculate the residual value(ε_i)
   b. Calculating value(σ_i)
   c. Calculating value(u_i)
   d. Calculating weighting $w(u_i)$
   e. Calculating the β *parameter*
   f. Repeat steps b to e until a convergent β *value is obtained.* Which is a difference β^{m+1}that is β^mclose to or equal to zero.
7. Carry out simultaneous tests and partial tests to find out whether the independent variable has an effect on the dependent variable.
8. Determining the goodness of the model (*Adjusted R square)*.
9. Conclusion
10. Finished

RESEARCH RESULT

Least Squares Method parameter estimation

The aim of the least squares method (MKT) is to minimize the residual sum of squares. With the parameter estimation results written in the table as follows:

Table 2. MKT Parameter Estimation

Parameter	Estimated Value	*Adjusted R Square*
β_0	-2,303× 10^{-16}	
β_1	1,074	
β_2	7,132× 10^{-2}	
β_3	-3,715	0.6249
β_4	-4,017× 10^{-2}	
β_5	-3,095	
β_6	7,469× 10^{-2}	

In Table 2, the initial regression model using the least squares method (MKT) is obtained as follows:

$$\bar{Y} = -2{,}303 \times 10^{-16} + 1{,}074\,X_1 + 7{,}132 \times 10^{-2}\,X_2 - 3{,}715\,X_3 - 4{,}017 \times 10^{-2}\,X_4 - 3{,}095\,X_5 + 7{,}469 \times 10^{-2}\,X_6$$

With *adjusted R square* is 0.6349, which means that the variables of population (X_1), gross regional domestic product (X_3), and average years of schooling) X_5)have an effect on the poverty level on the island of Sulawesi by 63.5% while the remaining 36.5% is explained by other variables.

Classic assumption test

Normality test

One of the tests used to test the normality assumption is the *Kolmogorov-Smirnov test* which can be seen in the Table 3.

Table 3. Normality Test Results

Kolmogorov-Smirnov	***p-value***
0.10195	0.03675

Table 3 above shows that the *p-value* < α, so it can be concluded that reject H_0means the residual is not normally distributed.

Multicollinearity Test

To test multicollinearity, it can be done by looking at the *VIF value* seen in the table as follows:

Table 4. Normality Test Results

Variable	X_1	X_2	X_3	X_4	X_5	X_6
VIF	4,180	2,334	4,322	5,483	4,701	1,347

From Table 4, it can be seen that the VIF value is < 10, so it can be concluded that accept H_0means there is no multicollinearity problem.

Autocorrelation Test

To test for autocorrelation problems, the *Durbin-Watson test can be done* as seen in the following table :

Table 5. Autocorrelation Test Results

d_L	d_W	d_U	***p-value***
1.4842	1,991	1.8008	0.4216

In Table 5 the values are obtained $d_W < d_L$so it can be concluded that the rejection H_0 means there are symptoms of autocorrelation.

Heteroscedasticity Test

One test to detect heteroscedasticity is the *Breusch-Pagan test* . It can be seen in table 4.7 as follows:

Table 6. Heteroscedasticity Test Results

Breusch-Pagan	***p-value***
16,321	0.01213

Table 6 above shows that the *p-value* < α, so it can be concluded that rejecting H_0means there are symptoms of heteroscedasticity.

Identify Outliers

To see whether there are outliers in the data, you can plot the data against the i-th observation , as seen in the following picture:

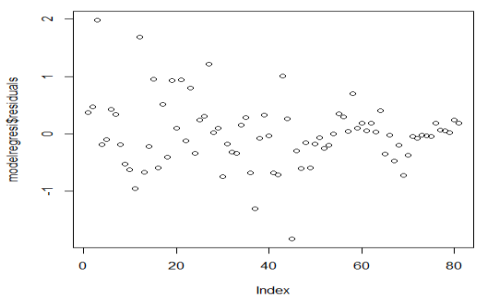

Figure 1. Residual Plot

In Figure 1, it can be seen that there are several points that are far from the data set pattern, meaning this indicates the presence of outliers.. Residual plot above cannot provide information on which data are outliers, therefore to identify outliers needs to be done by testing DFFITS. The following are the results of the DFFITS test values as seen in the table as follows:

Table 7. DFFITS Value Results

Data	DFFITS Value	Decision
3	0.761	Outlier Data
12	1,093	
15	0.868	
24	-3,717	
43	1,244	
45	-1,034	

Table 7 shows that, of the 81 data, there are several that have values |DFFITS|> 0.587. So it can be concluded that the observation data has outliers.

S-Estimation *Robust* Regression with *Tukey Bisquare Weighting*

The iterative S-estimation calculation process begins by determining the initial estimate of the regression coefficient, then based on the S-estimation algorithm, residual and residual values are calculated. This iteration process uses *Tukey bisquare weighting* which is carried out repeatedly until a convergent value is obtained. The calculation results for each S-estimation iteration are as follows:

Table 8. Iterated *Tukey Bisquare* S-Weighted Estimation

Iteration	Y	X_1	⋯	X_6
1	-30.4300235	3585283	⋯	0.5002306
2	-0.2102310	4652189	⋯	0.1586419
3	-21.02307	4.652190	⋯	0.1586421
4	-0.1845852	5.051842	⋯	0.1584502
5	-0.2102308	4.652190	⋯	0.1586421
⋮	⋮	⋮	⋱	⋮
37	-0.2102308	4,652,190	⋯	0.1586421
38	-0.2102307	4,652,190	⋯	0.1586421
39	-0.2102308	4,652,190	⋯	0.1586421
40	-0.2102307	4,652,190	⋯	0.1586421
41	**-0.2102308**	**4,652,190**	⋯	**0.1586421**

Based on table 8, a regression model is obtained converges at the 41st iteration as follows:

$$Y = -0{,}2102308 + 4{,}652190\,X_1 + 0{,}05346983\,X_2 + 0{,}06430887\,X_3 + 0{,}1655135\,X_4 - 0{,}3344431\,X_5 + 0{,}1586421\,X_6$$

Parameter Testing

F test

The F test aims to determine the influence of the relationship between the independent variable and the dependent variable as a whole. It can be seen in the table as follows:

Table 9. Results of F Test Statistical Values

F test	DF_1	DF_2	P-value
30,165	6	74	$2{,}2\times 10^{-16}$

In Table 9 above, the values are obtained $F_{hitung} > F_{tabel}$ or *P- value* < 0.05, so it can be concluded that reject H_0means that all independent variables have a significant effect simultaneously on the poverty level on the island of Sulawesi.

T test

The T test aims to determine the effect of each independent variable on the dependent variable. It can be seen in the table as follows:

Table 10. Results of T Test Statistical Values

Var/Coef	Estimate	*t value*	Pr($\|t\|$)	t_{tabel}	Conclusion
Intercept	-0.21023	-6,665	4.11×10^{-9}		RejectH_0
X_1	0.46522	5,830	1.36×10^{-7}		RejectH_0
X_2	0.05347	1,261	0.2112		AcceptH_0
X_3	0.06431	0.958	0.3414	1,995	AcceptH_0
X_4	0.16551	2,410	0.0184		RejectH_0
X_5	-0.33444	-5,185	1.81×10^{-6}		RejectH_0
X_6	0.15864	5,034	3.28×10^{-6}		RejectH_0

Based on Table 10, it can be seen that $t_{hitung} > t_{tabel}$*the P-value* is < 0.05 . So it can be concluded that H_0it means that there is a partial influence of the independent variable on the level of poverty on the island of Sulawesi.

CONCLUSION

Based on the results and discussions that have been carried out previously, the best model obtained from *robust* S-estimation regression with *Tukey bisquare weighting* is:

$$Y = -0{,}21023 + 0{,}46522\, x_1 + 0{,}16551\, x_4 - 0{,}33444\, x_5 + 0{,}15864\, x_6$$

Factors that influence the level of poverty on the island of Sulawesi using the S-estimation *robust regression analysis method with Tukey bisquare weighting* include Population Number (X_1), Human Development Index (X_4), Average Years of Schooling (X_5) and, Labor Force Participation Rate (X_6).

REFERENCE

Ghozali, I. (2016). *Multivariate Analysis Application with the IBM SPSS 23 Program.* Semarang: BPFE Diponegoro University.

Harlik, Amir, A., & Hardiani. (2013). Factors that Influence Poverty and Unemployment in Jambi City. *Journal of Regional Financing and Development Perspectives, 2* (1) , 109-120.

Hasanah, R., Syaparuddin, & Rosmeli. (2021). The Influence of Life Expectancy, Average Years of Schooling and Per Capita Expenditure on Poverty Levels in Districts/Cities in Jambi Province. *Journal of Regional Economic and Development Perspectives, 10* (3), 223-232.

Hendri, & Setiawan, R. (2017). The Influence of Work Motivation and Compensation on Employee Performance at PT. Main Ocean. *AGORA Journal , 1* (50), 1-8

Hidayatulloh, FP, Yuniarti, D., & Wahyuningsih, S. (2015). Robust Regression with S-Estimation Method. Exponential Journal, 6(2), 163-170.

Indrasetianingsih, A., & Wasik, TK (2020). Panel Data Regression Model to Find Out Factors That Influence Poverty Levels on Madura Island. *Gaussian Journal, 3* (9) , 355-363.

Mirah, MR, Kindangen, P., & F. Rorong, IP (2020). The Influence of Labor Force Participation Levels on Economic Growth and Poverty in North Sulawesi Province. *Journal of Regional Economic and Financial Development, 21* (1), 85-100.

Setiarini, Z., & Listyani, E. (2017). S-Estimation Robust Regression Analysis Using Welsch and Tukey Bisquare Weighting. *Journal of Mathematics, 1* (6) , 48-55.

Setiawan, W., Debatarajaya, NN, & Sulistianingsih, E. (2019). S-Estimation Method in Robust Regression Analysis with Tukey Bisquare Weighting. *Mat, Stat and Applications Scientific Bulletin (Bimaster), 2* (8) , 289-296.

Susanti, Y., Pratiwi, H., & H, SS (2014). M Estimation, S Estimation, and MM Estimation In Robust Regression. *International Journal of Pure and Applied Mathematics, 91* (3), 349-360.

Utami, FP, Lubis, I., & Rahmanta. (2022). Analysis of Factors Affecting Poverty in Eastern Aceh. *Journal of Samudra Economics, 1* (6) , 1-9.

Widyaningsing A, SM (2014). Estimation of the Seemingly Unrelated Regression (SUR) Model using the Generalized Least Square (GLS) Method. *Journal of Mathematics , 4* (2), 102-110.