Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754
P-ISSN : 2580-569X
Vol. 8, No. 2, 2024
DOI: 10.30829/zero.v8i2.21345
pp. 108-118

Comparison of OLS Regression and Robust Regression in Overcoming Outlier Problems

Susiana[1], Chairunisah[3], Nice Rejoice Refisis [3]
[1,3,4] Department of Mathematics, Universitas Negeri Medan, Medan, Indonesia

Article Info

Article history:

Received, 20 10 2024
Revised, 20 11 2024
Accepted, 21 12 2024

Keywords:

OLS Regression;
Robust Regression;
Cost of Living;

ABSTRACT

Multiple regression analysis in quantitative statistical studies describes the relationship between independent and dependent variables. On the other hand, outliers in a set of data can have an unfavorable influence on data analysis, such as high residuals, significant variances, and bias, and can even cause errors in decision-making. It can be done in several ways to overcome the outlier problem in multiple linear regression analysis, including using robust regression or Ordinary Least Square (OLS) Regression by removing data indicated as an outlier first. The OLS Regression method forms a regression model by minimizing the sum of squared residuals from the estimator of the regression equation. Meanwhile, robust regression is closer to the average parameters and variance-covariance of a particular estimator, namely by standardizing the estimator for the average parameters and variance-covariance in such a way as to produce a consistent estimator for these parameters. This research aims to compare the OLS Regression and robust regression methods as alternatives for dealing with outlier problems in data. The data used in this research is secondary data (cost of living) from the Cost of Living Survey conducted by The Central Statistics Agency of the Republic of Indonesia in 2018. The stages of this research method are literature study, data collection, descriptive analysis to see the characteristics of the data, forming a regression model using the OLS Regression method, testing classical assumptions, creating a new regression model OLS Regression, forming a regression model with Robust Regression, calculating the MSE (Mean Square Error) of each regression model formed, determining the best regression model, The results of the research show that for the cost of living data, the best regression model is obtained through the OLS Regression method with data without outliers, namely $\hat{Y} = -93626{,}521 + 0{,}949\,X_1 + 0{,}034X_2 + 0{,}930\,X_3 + 1{,}599\,X_4 + 0{,}956X_5 + 1{,}042X_6 + 1{,}195\,X_7 + 1{,}601\,X_8 + 0{,}416\,X_9 + 1{,}127X_{10} + 1{,}121\,X_{11} + 0{,}014\,X_{12} + 42049{,}864\,X_{13} + \varepsilon$.

Corresponding Author:

Susiana,
Department of Mathematics,
Universitas Negeri Medan, Medan, Indonesia
Email: susianafaisal313@gmail.com

1. INTRODUCTION

Statistical analysis for quantitative studies that describe the relationship between several variables can be carried out using multiple regression (multiple linear regression). Many methods in various regressions are adapted to the characteristics of the available data. The characteristics of the data in question are certain assumptions that the data set meets, such as normality, homoscedasticity, no extreme outliers, no multicollinearity, and no autocorrelation. One of the familiar and most straightforward methods for forming a regression model is Ordinary Least Square (OLS), obtained by minimizing the sum of the squared errors of the estimator of the regression equation[[1].

However, there are times when the data needs to meet the assumptions mentioned above, and alternative solutions are required to overcome these problems and obtain reliable results. For example, OLS Regression provides a solution for outliers by eliminating data indicated as outliers and then re-forming a new regression equation for data without outliers. Meanwhile, robust regression is also believed to overcome the problem of outliers in the regression model. As the name suggests, this method requires a weighting function that can minimize the influence of outliers on the model so that the best model resistant to outliers is obtained [2].

The difference in cost of living between regions is important for people, especially those who often travel to other regions [3]. The Central Bureau of Statistics regularly conducts cost of living surveys. This survey is done so that the movement of the cost of living can continue to be monitored, considering that the cost-of-living crisis can have a bad impact on the community because it affects other areas of life, such as health, which results in an increase in the poverty rate of citizens [4].

In addition, the cost-of-living survey is also one of the important variables used in calculating the Consumer Price Index (CPI), where the CPI describes price changes of a group of goods and services that are generally consumed by the community and also describes the level of inflation or deflation. The CPI value can be known as the level of increase in income, prices, and benchmarks for production costs and can be used as an economic indicator [5].

This research aims to compare the OLS Regression method with data without outliers and the robust regression method with the following details: 1) Form a regression estimation model for the cost of living in urban areas in Indonesia; 2) Determine the best regression model by comparing the MSE (Mean Square Error) value.

2. RESEARCH METHOD

This research is quantitative research with case studies. The population in this study were all urban households in Indonesia. The samples in the research were households domiciled in 90 cities in Indonesia as per the data in the cost of living survey 2018, namely: Banda Aceh, Meulaboh, Lhokseumawe, Medan, Sibolga, Pematang Siantar, Padangsidempuan, Gunungsitoli, Padang, Bukittinggi, Pekanbaru, Tembilahan, Dumai , Jambi, Bungo, Palembang, Lubuklinggau, Bengkulu, Bandar Lampung, Metro, Pangkal Pinang, Tanjung, Pandan, Tanjung Pinang, Batam, Jakarta, Bandung, Bogor, Sukabumi, Cirebon, Bekasi, Depok, Tasikmalaya, Semarang, Cilacap, Purwokerto, Kudus, Surakarta, Tegal, Yogyakarta, Surabaya, Jember, Banyuwangi, Sumenep, Kediri, Malang, Probolinggo, Madiun, Serang, Tangerang, Cilegon, Denpasar, Singaraja, Mataram, Bima, Kupang, Waingapu, Maumere, Pontianak, Sintang, Singkawang, Palangka Raya, Sampit, Banjarmasin, Kotabaru, Tanjung, Samarinda, Balikpapan, Tanjung Selor, Tarakan, Manado, Kotamobagu, Palu, Luwuk, Makassar, Bulukumba, Watampone, Pare-pare, Palopo, Kendari, Bau-Bau, Gorontalo, Mamuju, Ambon, Ternate, Tual, Manokwari, Sorong, Jayapura, Merauke and Timika.

In this study, the response variable, " Y " is the average household cost of living for one month (in rupiah units). Meanwhile, the independent variables in this research are factors that influence household expenditure, namely X_1 osts for food, drinks, and tobacco, X_2 costs for clothing and footwear, X_3 costs for housing, water, electricity, and home fuel, X_4 costs for supplies, equipment, and routine household maintenance, X_5 costs for health, X_6 costs for transportation, X_7 costs for information, communication and financial services, X_8 costs for recreation, sports and culture, X_9 costs for education, X_{10} costs for providing food and drinks/restaurants, X_{11} costs for personal care and other services, X_{12} household income in one month, and X_{13} number of family members per household. The data used in this study were taken from the results of a cost-of-living survey conducted by the Central Bureau of Statistics of the Republic of Indonesia in 2018 [6].

The stages in this research are literature study, data collection, descriptive analysis to see the characteristics of the data, forming a regression model using the OLS Regression method, testing classical assumptions, creating a new regression model OLS Regression, forming a regression model using Robust Regression, calculating MSE (Mean Square Error) of each regression model formed, determines the best regression model.

3. RESULT AND ANALYSIS

The OLS, called the least squares method, estimates the regression coefficients by minimizing the sum of the squared errors for unbiased regression coefficients. The OLS estimator gives quite good results when all the classical assumptions in regression are fulfilled, namely the assumptions of normality, homoscedasticity, no multicollinearity, and autocorrelation between the independent variables. This method aims to minimize the number of squared errors so that the regression equation can be $Y =\times \beta + \varepsilon,$ where Y and ε are matrix $n \times 1$, while X is matrix $n \times (k + 1)$. The allegation of regression equation can be $\hat{Y} =\times \hat{\beta} + \varepsilon$ [7].

For example, if sample Y is given, then the rule that allows the use of sample Y is to obtain an estimate from it by making $verepsilon = \hat{Y} - X\hat{\beta}$ as small as possible. The goal of the least squares method is to

determine the estimator of β_0 and β_1 which will minimize the sum of square errors. Based on this, parameters are needed β as small as possible. For instance, $S = \varepsilon'\varepsilon = (Y - X\beta)(Y - X\beta)$, is a scalar, so its components are also scalar. As a result, scalar transpose cannot change the scalar value. To minimize it can be obtained by doing the partial derivative of S with respect to:

$$\frac{\partial S}{\partial \beta'} = -2YX' + 2X'X\beta$$

By equating this equation to zero, it is obtained $XY' = X'X\beta$, which is called the normal equation, and $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$ as a least squares parameter estimate β[8].

Robust Regression (RR) is a regression analysis method not sensitive to outliers. RR is often used to overcome heteroskedasticity problems by using robust standard errors that are resistant to heteroskedasticity problems. This method is an important tool for analyzing data affected by outliers to produce robust models resistant to outliers. A relatively resistant estimate means that it is not affected by large changes in a small part of the data or small changes in a large part [9].

The RR method is more concerned with a particular estimator's mean and variance-covariance parameters, i.e., standardizing the estimator for the mean and variance-covariance parameters such that it produces an estimator consistent with these parameters. In this case, it takes the form of value restrictions on the parameter estimates. With RR, the estimates will stay within a reasonable range [10].

In general, RR has 4 estimation methods, including M estimation (Maximum likelihood type), LTS estimation (Least Trimmed Squares), S estimation (Scale), and MM estimation (Method of Moment). The M (Maximum likelihood type) estimation introduced by Huber is a simple method both in calculation and theoretically. This estimation analyzes the data by assuming most of the detected outliers in the independent variable. M-estimation is a frequently used robust regression method that minimizes a function (objective function) of the error.

$$\hat{\beta}_M = \min_{\beta} \sum_{i=1}^{n} p(e_i) = \min_{\beta} \sum_{i=1}^{n} p\left(Y_i - \sum_{j=1}^{k} X_{ij}\beta_j\right)$$

i: the number of observations with i=1, 2, 3,..., n and

j: the number of independent variables with j=1, 2, 3,..., n

The parameter estimation procedure in multiple linear regression models with M-estimation robust regression is as follows [9] :

a) Estimating regression coefficients on the data using the least squares method;

b) Detect outliers in the data; c) Calculating the initial parameters using the least squares method;

d) Calculating the error (residual) by using $e_i = y_i - \hat{y}_i$;

e) Calculate $\hat{\sigma}_i$ where MAD is the median of absolute deviation;

f) Calculating $u_i = \frac{e_i}{\hat{\sigma}_i}$; g) Calculating the Huber Weight function:

$$w_i = \begin{cases} \left[1 - \left(\frac{u_i}{4.685}\right)^2\right]^2, & |u_i| \le 4.685 \\ 0, & |u_i| > 4.685 \end{cases};$$

h) estimating the parameters of the weighted least squares method, b_M, with ;

i) Repeat steps 5-10 until b_M converged;

j) Test the model to determine whether the independent variables significantly affect the response variable.

3.1 OLS Regression Model

Cost of living analysis can be more in-depth by knowing the relationship between the cost of living and its related factors. Regression analysis is one method that can describe how the influence between one variable (independent) and another variable (dependent) [10]. In this study, the dependent variable, Y, is the average expenditure (cost of living) for the consumption category per household, with 13 independent variables.

Using SPSS software, through the OLS regression method, a regression model was constructed based on the B values listed in the following output display.

Table 1. SPSS Output: Coefficient of Regression Model with OLS Method

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-180691.971	111512.383		-1.620	.109
	X_1	.956	.035	.225	27.467	.000
	X_2	.017	.013	.007	1.247	.216

X_3	.925	.035	.251	26.311	.000
X_4	1.790	.138	.133	13.001	.000
X_5	.887	.194	.033	4.568	.000
X_6	1.024	.064	.167	15.927	.000
X_7	1.266	.136	.097	9.330	.000
X_8	1.656	.421	.051	3.933	.000
X_9	.433	.121	.046	3.570	.001
X_{10}	1.162	.085	.139	13.712	.000
X_{11}	.970	.107	.072	9.054	.000
X_{12}	.006	.006	.010	1.027	.308
X_{13}	70752.255	29152.245	.015	2.427	.018

a. Dependent Variable: cost of living

Based on the values listed in Table 1, the initial regression model for the problem in this study is :

$$\hat{Y}= -180691{,}971= 0{,}956\ X_1+ 0{,}017X_2+ 0{,}925\ X_3 + 1{,}790\ X_4 +0{,}887X_5+ 1{,}024X_6 + 1{,}266\ X_7+1{,}656\ X_8+0{,}433\ X_9+1{,}162\ X_{10}+ 0{,}970\ X_{11}+ 0{,}006\ X_{12}+70752{,}255\ X_{13}+ \varepsilon \quad (1)$$

Equation (1) is the initial regression model obtained by the OLS method. The next step is to perform several assumption tests on the regression model.

3.2 Classical Assumption Test

In regression analysis, several classical assumptions must be met so that the regression model obtained has good accuracy and precision and is unbiased or included in the BLUE (Best Linear Unbiased Estimator) category. Some regression analysis assumptions are a) normal distribution in residuals, b) no multicollinearity, c) no heteroscedasticity. After the regression model is formed, the next step is to conduct several tests of these classical assumptions. By using SPSS software, the following are the results of the classical assumption test on the problems in this study:

a) Residuals are normally distributed

The regression model formed is an approach to the regression model that occurs in nature because it will give rise to what is referred to as residuals (errors), namely the difference between the actual model and the estimated model formed. Regression analysis requires that the residuals in the regression model be normally distributed. Once the residual values for each possibility are obtained, a normality test can be performed using the Kolmogorov-Smirnov method. The SPSS output display for the normality test is presented in Table 2 below:

Table 2. SPSS Output: Residual Normality Test

		Unstandardized Residual
N		90
Normal Parameters[a,b]	Mean	.0000000
	Std. Deviation	68437.30258419
Most Extreme Differences	Absolute	.071
	Positive	.071
	Negative	-.059
Test Statistic		.071
Asymp. Sig. (2-tailed)		.200[c,d]

a. Test distribution is Normal.

Table 2 displays the values obtained by analyzing the residual data from the regression model formed. Based on the results of the normality test, it is known that the significance value is 0.200> 0.05, so it can be concluded that the residuals are normally distributed.

b) Multicollinearity Test

Multicollinearity indicates a strong relationship between the regression model's independent variables. Usually, this correlation can be easily characterized if one of the independent variables is expressed in other variables. The presence of multicollinearity in a regression model causes the model to lack precision [11] [12].

- The standard error value becomes large.
- Small changes in sample data cause large changes in the value of regression coefficients.
- The confidence interval value becomes so wid
- e that it will be difficult to reject the null hypothesis in a study.

Multicollinearity detection in the regression model can be done through the tolerance value (TOL) and VIF (Variance Inflation Factor), namely: TOL=1- R_j^2 and $VIF_j= \frac{1}{1-R_j^2}$; where VIF for coefficient j and R_j^2 is the coefficient of determination between $X_j$with other independent variables in the estimated

model with j = 1, 2, 3, ..., p. A TOL value of less than 0.01 and a VIF value of less than 10 indicates no multicollinearity in the model [13]. Table 3 shows the TOL and VIF values of the previously formed linear regression model.

Table 3. SPSS Output: VIF Values for Independent Variables

Model	Collinearity Statistics	
	Tolerance	VIF
X_{13}	.501	1.998
X_1	.293	3.419
X_2	.666	1.502
X_3	.216	4.631
X_4	.187	5.342
X_5	.370	2.701
X_6	.180	5.566
X_7	.182	5.485
X_8	.118	8.469
X_9	.119	8.387
X_{10}	.190	5.249
X_{11}	.308	3.248
X_{12}	.192	5.215

a. Dependent Variable: cost of living

Based on the SPSS output presented in Table 3, the TOL value is less than 0.01, and the VIF value for each independent variable is less than 10. Thus, there is no multicollinearity in the regression model. In this case, there is no correlation between the independent variables with one another.

3.3 Heteroscedasticity Test

Homoscedasticity and heteroscedasticity are two opposing terms. Judging from the word, homo means showing similarity, while hetero means showing unequally. Equality or inequality, in this case, refers to the variance of the residuals of all observations in the regression model. If the regression model indicates heteroscedasticity, then the regression model is declared invalid (less efficient) as a forecasting tool.

There are several ways to detect heteroscedasticity in a regression model, including the Park, geyser, spearman, and graph tests. In this study, the lesser test transforms residuals into absolute residual prices (abs_residual). Then, each independent variable is regressed on the abs_residual value. Conclusions are drawn by looking at the significance value, where the regression model identifies heteroscedasticity if the significance value is less than 0.05. Table 4 is the output of the Glejser test results using SPSS. From Table 4, it can be seen that there are significance values for several independent variables that are less than 0.05, namely variable X_2, X_7, and variable X_{11}, with significance values of 0.042, 0.018.

Table 4. SPSS Output: Glejser Test for Heteroscedasticity

Model	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta		Sig.
(Constant)	7575.849	58840.870		.129	.898
X_1	.002	.018	.019	.112	.911
X_2	.014	.007	.239	2.064	.042
X_3	.003	.019	.030	.147	.884
X_4	.118	.073	.354	1.622	.109
X_5	-.006	.102	-.009	-.056	.955
X_6	.033	.034	.216	.969	.336
X_7	-.173	.072	-.535	-2.421	.018
X_8	-.327	.222	-.404	-1.471	.145
X_9	-.052	.064	-.222	-.812	.419
X_{10}	.005	.045	.022	.102	.919
X_{11}	.163	.057	.491	2.885	.005
X_{12}	.001	.003	.067	.311	.757
X_{13}	-1361.416	15382.538	-.012	-.089	.930

a. Dependent Variable: abs_res

0.005, respectively. This indicates that there is potential heteroscedasticity in the regression model. In other words, the regression model that has been previously formed could be more efficient. Several things can cause heteroscedasticity symptoms to appear in a regression model, including extreme/outlier data [14]. In this study, before action is taken to overcome the problem of heteroscedasticity in the regression model, outliers will first be detected in the data used. If the suspicion of outlier data is proven, the next step is to overcome the outlier data.

From Table 4, it can be seen that there are significance values for several independent variables that are less than 0.05, namely variable X_2, X_7, and variable X_{11}, with significance values of 0.042, 0.018, and 0.005, respectively. This indicates that there is potential heteroscedasticity in the regression model. In other words, the regression model that has been previously formed is not efficient.

Several things cancause heteroscedasticity symptoms to appear in a regression model, including extreme/outlier data [14]. In this study, before action is taken to overcome the problem of heteroscedasticity in the regression model, outliers will first be detected in the data used. If the suspicion of outlier data is proven, the next step is to overcome the outlier data.

3.4 Outlier Detection

Outliers can be defined as data significantly different from the collected pattern. Outliers can be data that is very low or data that is very high compared to other data. The presence of outlier data in a research data set can affect the conclusions of the data set, for example, by involving the value of statistical measures such as mean, variance, and others so that the findings can be biased. The same thing can happen to regression models whose data sources contain outliers; the regression model becomes less accurate as a predictive tool. Therefore, it is essential to perform outlier detection.

Outlier detection on a data set in regression analysis can be done by observing the research data's histogram and scatter plot graphs [15].

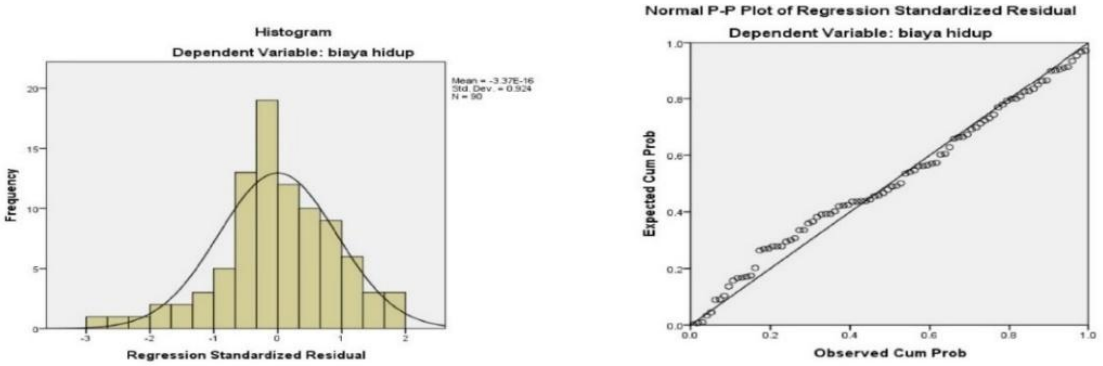

Figure 1. Outlier Detection with Histogram and Plots

The histogram of the data presented in Figure 1 looks relatively good but leaves a few potential outliers due to the visual histogram on the left, which needs to be more balanced with the right. The same thing is also seen in the data plot, where most of the data tends to be around the trend except at the beginning, relatively slightly away; potential outliers are also visible from the presented data plot. It can be done by observing the standardized residual value or using Cook's Distance method.

The standardized residual value is the residual value that is standardized or standardized. [16] state that outliers occur if the value of both standardized residuals is more than 3.3. Meanwhile, Cook's Distance is a method that detects outliers based on Cook's distance value, which shows the magnitude of the influence of outlier data on all regression coefficient estimators. Cook's Distance measurement value outliers occur if the Cook's Distance value is more than (4/n-p-1) with n as a lot of data and p as the number of independent variables [17].

Based on the standardized residual value, all data collected is less than 3.3, so according to the standardized residual value there are no outliers. Unlike the case using the Cook's Distance method, with a critical value of 0.0526, several data that are outliers have been identified, namely: 1st datum (Cook's Distance value = 0.05958), 5th datum (Cook's Distance value = 0.06710), 8th datum (Cook's Distance value = 0.05923), 10th datum (Cook's Distance value = 0.07925), 13th datum (Cook's Distance value = 0.06346), 21st datum (Cook's Distance value = 0.19861), 25th datum (Cook's Distance value = 0.08755), 30th datum (Cook's Distance value = 0.08562), 48th datum (Cook's Distance value = 0.05758), 74th datum (Cook's Distance value = 0.15175). There are a total of 10 datums that are outliers, if examined in relation to the Cook's Distance of each datum, it seems to be in line with what is shown by the previous data plot with a distance not so far from the trend.

Based on the standardized residual value, all data collected is less than 3.3, so according to the standardized residual value there are no outliers. Unlike the case using the Cook's Distance method, with a critical value of 0.0526, several data that are outliers have been identified, namely: 1st datum (Cook's Distance value = 0.05958), 5th datum (Cook's Distance value = 0.06710), 8th datum (Cook's Distance value

= 0.05923), 10th datum (Cook's Distance value = 0.07925), 13th datum (Cook's Distance value = 0.06346), 21st datum (Cook's Distance value = 0.19861), 25th datum (Cook's Distance value = 0.08755), 30th datum (Cook's Distance value = 0.08562), 48th datum (Cook's Distance value = 0.05758), 74th datum (Cook's Distance value = 0.15175). There are a total of 10 datums that are outliers, if examined in relation to the Cook's Distance of each datum, it seems to be in line with what is shown by the previous data plot with a distance not so far from the trend.

The presence of outliers in regression analysis can affect the regression model to be less accurate. Follow-up of outlier problems in the data can be overcome in several ways such as eliminating the outlier datum itself or forming a new regression model with other methods such as robust regression [18]

3.5 OLS Regression Model Without Outliers

As explained in the previous section, one way to deal with data containing outliers is to eliminate the outlier data. In this section, the regression model is formed using the OLS method without including outlier data. Ten datums were removed from the initial data of 90 datums, so the data used to form the OLS regression model was only 80. The SPSS output results for the data set are presented in Table 5.

Table 5. SPSS Output for Data without Outliers

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-93626.521	88518.174		-1.058	.294
x_1	.949	.029	.261	33.214	.000
x_2	.034	.011	.015	3.139	.003
x_3	.930	.028	.264	33.813	.000
x_4	1.599	.147	.101	10.855	.000
x_5	.956	.154	.042	6.202	.000
x_6	1.042	.053	.177	19.494	.000
x_7	1.195	.106	.097	11.224	.000
x_8	1.601	.352	.049	4.545	.000
x_9	.416	.099	.046	4.216	.000
x_{10}	1.127	.066	.142	17.031	.000
x_{11}	1.121	.103	.083	10.935	.000
x_{12}	.014	.005	.024	2.867	.006
x_{13}	42049.864	22026.699	.011	1.909	.061

a. Dependent Variable: Y

Based on the values presented in Table 5, a regression model can be formed for the data, namely:

$$\hat{Y} = -93626{,}521 + 0{,}949\ X_1 + 0{,}034 X_2 + 0{,}930\ X_3 + 1{,}599\ X_4 + 0{,}956 X_5 + 1{,}042 X_6 + 1{,}195\ X_7 + 1{,}601\ X_8 + 0{,}416\ X_9 + 1{,}127 X_{10} + 1{,}121\ X_{11} + 0{,}014\ X_{12} + 42049{,}864\ X_{13} + \varepsilon \quad (2)$$

As a follow-up to the regression equation (2), it is necessary to test for heteroscedasticity to see if the problem has been resolved. The steps to conduct a heteroscedasticity test using the Glejser test have been explained in the previous section, which regresses the absolute value of the residuals against all independent variables. The residual value of the data regression equation without outliers is obtained by calculating the difference between the dependent variable data (Y) and the dependent variable estimation data (Y). Furthermore, the residual data is solved (absRES_1), and then a new regression equation is formed from the dependent variable (absRES_1) and all independent variables. The SPSS output results show the following values:

Table 6. SPSS Output for Glejser Test for Data Without Outlines

Unstandardized Coefficients		Standardized Coefficients	T	Sig.
B	Std. Error	Beta		
58924.279	45069.210		1.307	.196
-.009	.015	-.135	-.650	.518
.006	.005	.130	1.053	.296
.011	.014	.159	.772	.443
.120	.075	.394	1.600	.114
-.012	.078	-.027	-.153	.879
-.001	.027	-.006	-.027	.979
-.039	.054	-.163	-.712	.479
-.184	.179	-.293	-1.029	.307
-.066	.050	-.375	-1.303	.197
-.018	.034	-.119	-.538	.592

.100	.052	.386	1.919	.059
-.001	.002	-.059	-.269	.789
-6927.331	11214.939	-.092	-.618	.539

a. Dependent Variable: absRES_1

The display presented in Table 6 shows that the significance value for each independent variable is not below 0.05, which leads to the conclusion that no heteroscedasticity is detected in the regression model obtained.

3.6 Regression Model with Robust Regression

In the previous section, a regression model with OLS Regression was obtained where the data detected as outliers were removed. In addition to eliminating outlier data, the Robust Regression method can be considered a solution to the outlier problem [19]. Using phyton, a display for regression coefficients is obtained, as in Figure 2.

```
                    Robust linear Model Regression Results
==============================================================================
Dep. Variable:                      Y   No. Observations:                   90
Model:                            RLM   Df Residuals:                       76
Method:                          IRLS   Df Model:                           13
Norm:                          HuberT
Scale Est.:                       mad
Cov Type:                          H1
Date:                Thu, 14 Dec 2023
Time:                        12:14:20
No. Iterations:                     2
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1.641e+05   1.09e+05     -1.511      0.131   -3.77e+05    4.87e+04
X1             0.9575      0.034     28.257      0.000       0.891       1.024
X2             0.0223      0.013      1.719      0.086      -0.003       0.048
X3             0.9220      0.034     26.927      0.000       0.855       0.989
X4             1.7141      0.134     12.785      0.000       1.451       1.977
X5             0.9570      0.189      5.064      0.000       0.587       1.327
X6             1.0500      0.063     16.774      0.000       0.927       1.173
X7             1.2126      0.132      9.179      0.000       0.954       1.472
X8             1.5301      0.410      3.733      0.000       0.727       2.334
X9             0.4465      0.118      3.781      0.000       0.215       0.678
X10            1.1457      0.083     13.885      0.000       0.984       1.307
X11            1.0011      0.104      9.593      0.000       0.797       1.206
X12            0.0085      0.006      1.420      0.155      -0.003       0.020
X13         6.721e+04   2.84e+04      2.368      0.018    1.16e+04    1.23e+05
==============================================================================

If the model instance has been used for another fit with different fit
parameters, then the fit options might not be the correct ones anymore.
Mean Squared Error (MSE): 4567683383.695796
```

Figure 2. Robust Regression Output with Python

Based on the coefficient values presented in Figure 2, the regression model can be formed as follows:

$$\hat{Y} = -164100 + 0{,}9575\ X_1 + 0{,}0223 X_2 + 0{,}9220\ X_3 + 1{,}6620\ X_4 + 0{,}9023 X_5 + 1{,}0620 X_6 + 1{,}2347\ X_7 + 1{,}4087\ X_8 + 0{,}4707\ X_9 + 1{,}1045\ X_{10} + 1{,}0463\ X_{11} + 0{,}0092\ X_{12} + 29930\ X_{13} + \varepsilon \quad (3)$$

The regression equations (2) and (3) are then compared to determine the best model. For this purpose, the MSE (Mean Square Error) values of the Robust Regression model and the OLS Regression model without outliers are compared. The regression model with the smallest MSE value will be selected as the solution.

3.7 MSE of Robust Regression and OLS Regression Models without outliers

MSE is a measure of error variance, which in this case will be estimated by the value s^2= (sum of squared residuals)/(n-k), where n is a lot of data and k is a lot of independent variables [6]. For regression models with robust regression methods, the MSE value can be seen in Figure 1.9, amounting to 4567683383,695796, while for regression models with OLS regression methods, the MSE value can be seen in Table 7 below:

Table 7. SPSS Output: MSE Value for Regression Model with OLS

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	171955425892942.000	13	13227340453303.230	4713.040	.000[b]
	Residual	185231724269.933	66	2806541276.817		
	Total	172140657617211.940	79			

a. Dependent Variable: Y

b. Predictors: (Constant), $x_{13}, x_3, x_2, x_{11}, x_9, x_1, x_5, x_7, x_{10}, x_{12}, x_6, x_4, x_8$

Based on Table 7, it can be seen that the MSE value for the regression model with the OLS method is 2806541276,817. By comparing the two MSE values, the regression model with OLS Regression is the best and worth considering as a solution.

Regression Model Interpretation

Some hypothesis tests are essential to be carried out on the regression model that has been built (equation 2.1) after previously the regression model meets the classical assumptions. This hypothesis test is

intended to maximize the interpretation of the obtained regression model. The accuracy of the regression function in estimating the actual value can be measured from the coefficient of determination, simultaneous significance test (F test), and partial significance test (t-test).

Coefficient of Determination (R^2)

The coefficient of determination measures how far the model can explain variations in the dependent variable. A small R^2 value (the proportion of variation in the sample described by the independent variable X) indicates that the ability of independent variables to explain variations in the dependent variable is minimal (Ghozali, 2013). However, for each additional independent variable in the regression model, the R^2 value will increase regardless of whether the variable affects the dependent variable. On the other hand, there is an adjusted $R^2=1-\left[\frac{n-1}{n-(n+k)}\right](1-R^2)$ which can be used as another alternative to see the significance of the model [6]. Unlike the R^2 value, if there are additional independent variables in the model, the adjusted R^2 value can increase and decrease.

Table 8. SPSS Output: Values R^2 and *adjusted* R^2

MModel	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.999[a]	.999	.999	52976.79942

a. Predictors: (Constant), x_1, x_3, x_2, x_{11}, x_9, x_1, x_5, x_7, x_{10}, x_{12}, x_6, x_4, x_8

Table 1.8, the value of R^2 by 0,999 and *adjusted* R^2 by 0,999, this indicates that the independent variable X can explain 99% of the variation in cost of living, Y.

Simultaneous Significance Test (F Test)

Furthermore, the significance of the influence of the independent variables on the dependent variable together (simultaneously) can be seen using the F test. The F test using SPSS can be seen from the Anova Table 8, where the calculated F value is 4713 > F table and sig. 0.000 <0.05 then H_0 is rejected. Based on this fact, it can be concluded that with α=0.05, together, the variables X_1 to X_{13} significantly influence the cost of living. Uji Signifikansi Parsial (Uji t)

a) Suppose the F test is conducted to see the significance of independent variables together. In that case, the t-test is used to see the effect of certain independent variables individually by ignoring others. Through SPSS, this t-test can be done by comparing the sig. value of each variable (Figure 1.7) with the value of α = 0.05, where H_0 (no effect) is rejected if the sig value. < α. Based on Figure 1.7, the t-test results are summarized as follows:
b) For the constant β_0; sig value. 0.294 > 0.05, so H_0 is accepted, meaning that β_0 does not need to be included in the model.
c) For variable β_1; sig value. 0.000 <0.05, so H_0 is rejected, meaning that variable X_1 (household expenditure on food, beverage, and tobacco groups) significantly affects the cost of living if variables X_2 to X_{13} are included in the model.
d) For variable β_2 ; sig. value 0,003 < 0,05, so H_0 ditolak, this means that variable X_2 (household expenditure on clothing and footwear group) significantly affects the cost of living if variable X_1 and variable X_3 up to variable X_{13} are included in the model.
e) For variable β_3; sig value. 0.000 <0.05, so H_0 is rejected, meaning that there is a significant effect of variable X_3 (household expenditure for housing, water, electricity, and household fuel groups.) on the cost of living if variables X_1, X_2 and variables X_4 up to variable X_{13} are included in the model.
f) For variable β_4; sig. 0.000 <0.05, so H_0 is rejected, meaning that there is a significant effect of variable X_4 (household expenditure for the group of equipment, equipment, and routine household maintenance) on the cost of living if variables X_1, X_2, X_3, and variables X_5 to variable X_{13} are included in the model.
g) For variable β_5; sig value. 0.000 <0.05, so H_0 is rejected, meaning that there is a significant effect of variable X_5 (household expenditure on health groups) on the cost of living if variables X_1, X_2, X_3, X_4, and variables X_6 to variable X_{13} are included in the model.
h) For variable β_6; sig. 0.000 <0.05, so H_0 is rejected, meaning that there is a significant effect of variable X_6 (household expenditure on transportation groups.) on the cost of living if variables X_1, X_2, X_3, X_4, and variables X_7 to variable X_{13} are included in the model.
i) For variable β_7; sig value. 0.000 <0.05, so H_0 is rejected, meaning that there is a significant effect of variable X_7 (household expenditure on information, communication, and financial services) on the cost of living if variables X_1, X_2, X_3, X_4, X_6, and variables X_8 to variable X_{13} are included in the model.

j) For variable β_8; sig value. 0.000 <0.05, so H_0 is rejected, meaning that there is a significant effect of variable X_8 (household expenditure on recreation, sports, and culture groups) on the cost of living if variables $X_1, X_2, X_3, X_4, X_6, X_7$, and variables X_9 to variable X_{13} are included in the model.

k) For variable β_9; sig. 0.000 <0.05, so H_0 is rejected, meaning that variable X_9 (household expenditure for the Education group) has a significant effect on the cost of living if variables X_1 to X_8 and X_{10} to variable X_{13} are included in the model.

l) For variable β_{10}, sig value. 0.000 <0.05, so H_0 is rejected, meaning that variable X_{10} (household expenditure for the Education group) has a significant effect on the cost of living if variables X_1 to X_9 and X_{11} to X_{15} are included in the model.

m) For variable β_{11}, sig value. 0.000 <0.05, so H_0 is rejected, meaning that variable X_{11} (household expenditure for personal care and other services) has a significant effect on the cost of living if variables X_1 to X_{10} and X_{12} to X_{13} are included in the model.

n) For variable β_{12}; sig value. 0.006 <0.05, so H_0 is rejected, meaning that variableX_{12} (household income in one month) has a significant effect on the cost of living if variables X_1 to X_{11} and X_{13} are included in the model. Untuk variabel β_{13} ; nilai sig. 0,061 > 0,05, sehingga H_0 diterima, artinya tidak terdapat pengaruh signifikan variabel X_{13} (rata-rata jumlah anggota keluarga per rumah tangga) terhadap *cost of living* jika variabel X_1 sampai dengan X_{12}dimasukkan kedalam model.

4. CONCLUSION

Based on the MSE value, the best regression model for estimating the cost of living is the regression model obtained using the OLS Regression method without outlier data, namely $\hat{Y}$= -93626,521+0,949 X_1+0,034 X_2+0,930 X_3+1,599 X_4+0,956X_5+ 1,042X_6 + 1,195 X_7+1,601 X_8+0,416 X_9+1,127X_{10}+ 1,121 X_{11}+ 0,014 X_{12}+42049,864 X_{13}+ ε

REFERENCES

[1] Fitrianto, A., & Xin, S. H. (2022). Comparisons between robust regression approaches in the presence of outliers and high leverage points. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, 16(1), 243–252. https://doi.org/10.30598/barekengvol16iss1pp243-252

[2] Susanti, Y., Pratiwi, H., & H., S. S., "Optimasi Model Regresi Robust Untuk Memprediksi Produksi Kedelai Di Indonesia", *FMIPA Universitas Negeri Yogyakarta (UNY)*, 2013.

[3] Halim, G.A, et.al., "Estimation of cost of living in a particular city using multiple regression analysis and correction of residual assumptions through appropriate methods", *Procedia Computer Science*, p. 613–619, 2023.

[4] Mulrenan C, Isobel Braithwaite, Anna Brook, Rachek Crossley, Emily Loud, Angelique Mavrodaris, "Comment: Asusutainable and Equitable Respons to the Cost of Living Crisis is Urgently Needed", *Public Health in Practice 5,* 2023.

[5] Sumantri Fazhar and Umi Latifah, "Faktor-faktor yang Mempengaruhi Indeks Harga Konsumen", *Widya Cipta*, Vol. 3 No.1. pp. 25-34, 2019.

[6] BPS, *Survey Biaya Hidup 2018*, Jakarta : Badan Pusat Statistik Indonesia.

[7] Suyono, *Analisis Regresi untuk Penelitian*, Yogyakarta: Deepublish, 2015

[8] Aprianto, Ade, Naomi N.D., and Nurfitri Imro'ah, "Metode *Cochrane-Orcutt* Untuk Mengatasi Autokorelasi Pada Estimasi Parameter *Ordinary Least Squares", Buletin Ilmiah, Stat, dan Terapannya (Bimaster)*, Vol. 09, No.1, pp. 95 – 102, 2020.

[9] Susanti, Y., Handayani, S. S., and Pratiwi, H. M Estimation, S Estimation, and MM Estimation in Robust Regression, *International Journal of Pure and Applied Mathematics*. 91(3): 349-360. 2014.

[10] Pratiwi, H., Susanti, Y., & Handajani, S. S. (2018). A Robust Regression by Using Huber Estimator and Tukey Bisquare Estimator for Predicting Availability of Corn in Karanganyar Regency, Indonesia. *Indonesian Journal of Applied Statistics*, *1*(1), 37-44.

[11] Setyaningsih, Y.D., and Noeryanti, "Penggunaan Metode Weighted Least Square untuk Mengatasi Masalah Heteroskedastisitas dalam Analisis Regresi (Studi Kasus pada Data Balita Gizi Buruk Tahun 2014 di Provinsi Jawa Tengah)", *Jurnal Statistika Industri dan Komputasi*, Vol. 2, No. 1, pp. 51 – 58, Jan 2017.

[12] Wasilaine, T.L., M.W. Talakua, and Y.A. Lesnussa, "Model Regresi Ridge untuk Mengatasi Model Regresi Linier Berganda yang Mengandung Multikolinieritas", *Jurnal Barekeng*, Vol. 8, No.1, pp. 31 – 37, 2014.

[13] Ohyver, Margaretha, (2011), Metode Regresi Ridge Untuk Mengatasi Kasus Multikolinear, *Comtech*, Vol.2 No. 1 Juni 2011: 451-457, Jurusan Matematika, Fakultas Sains Dan Teknologi, Binus University.

[14] Sriningsih, Mega, Djoni Hatidja, dan Jantje De Prang (2018) Penanganan Multikolinearitas Dengan cahyandariMenggunakan Analisis Regresi Komponen Utama Pada Kasus Impor Beras Di Provinsi Sulut, *Jurnal Ilmiah Sains*,Vol. 18 No. 1, April 2018.

[15] Ghozali, Imam and Dwi Ratmono (2013), *Analisis Multivariat dan Ekonometrika: Teori, Konsep dan Aplikasi dengan Eviews 8*, Semarang: UNDIP.

[16] Nurdin, N, Raupong, and Anna Islamiyati, "Penggunaan Regresi Robust pada Data yang Mengandung Pencilan dengan Metode Momen", *Jurnal Matematika, Statistika dan Komputasi*, Vol. 10, No.2, pp. 114 – 123, Jan 2014.

[17] Tabachnick, B. G., & Fidell, L. S. (2007). Using Multivariate Statistics (5th ed.). New York: Allyn and Bacon.

[18] Hair, J.F., Black, W.C., Babin, B.J. and Anderson, R.E. (2010) Multivariate Data Analysis. 7th Edition, Pearson, New York.

[19] Dewi, Elok Tri Kusuma, Arief Agoestanto dan Sunarmi, (2016), Metode Least Trimmed Square (Lts) Dan Mm-Estimation Untuk Mengestimasi Parameter Regresi Ketika Terdapat Outlier, *UNNES Journal of Mathematics*, Vol. 5 (1).

[20] Cahyandari, Rini and Nurul Hisani, "Model Regresi Linier Berganda Menggunakan Penaksir Parameter Regresi Robust M-Estimator (Studi Kasus: Produksi Padi di Provinsi Jawa Barat Tahun 2009)", *Jurnal ISTEK*, Vol. VI, No.1-2, pp. 85 – 92, 2012