Available online at https://ejournal.corespub.com/index.php/ijmsc/index

International Journal of Mathematics, Statistics, and Computing

e-ISSN 3025-0803

Vol. 2, No. 2, pp. 60-67, 2024

Comparison of Weights in Weighted Least Square Method For Handling Heteroscedasticity on Multiple Regression Model

Fitria Virgantari[1*], Maya Widyastiti[2], Nathalia Media Lifa[3]

[1,2]*Pakuan University, Bogor, Indonesia*
[3]*Al Fath Islamic School, Bogor, Indonesia*

**Corresponding author email: fitria.virgantari@unpak.ac.id*

Abstract

Regression analysis is the most popular and commonly used to determine causality between two or more variables. In regression analysis there are several assumptions that must be held, so that the property of the best linear unbiased estimator (BLUE) is still guaranteed. In fact, we often found violations of the assumptions. One of them was violations of the homoscedasticity or occurs heteroscedasticity. The impact of heteroscedasticity in the regression model is that the ordinary least square (OLS) estimator no longer has a minimum variance although still linear and unbiased. To handle this, weighted least square (WLS) regression is used instead, which giving weights on the observations. But the problem often encountered is choosing which the best weight in WLS method. This paper aimed to compare and determine the best weight among 1/X, $1/\sqrt{X}$, 1/Y and $1/\sigma$ in multiple regression model. Human development index factors data, which were obtained from the Indonesian Central Bureau of Statistics, were used. The results showed that the best weight on human development index data was $1/\sigma$. The coefficient of determination was 98.7% indicating that the model was very good.

Keywords: multiple regression model, heteroscedasticity, weighted least square, coefficient of determination

1. Introduction

Regression analysis is most popular and commonly used to determine causality between two or more variables. There are several assumptions in regression analysis that must be considered, namely normally distributed of error, no multicollinearity, and homogeneity of variance which is called homoscedasticity. All these assumptions must be met so that the estimated parameters in the regression analysis meet the BLUE (Best Linear Unbiased Estimator) property (Montgomery, 2001).

When there is violation of the assumption of homoscedasticity or occurs heteroscedasticity, it means that the error is not constant. The impact of heteroscedasticity in the regression model is that the ordinary least square (OLS) estimator no longer has a minimum variance although is still linear and unbiased. It also causes the calculation of the standard error/standard error of the OLS method to be unreliable. In addition, hypothesis testing based on the t and F distributions can no longer be trusted for evaluating regression results (Nisa et al, 2020). The Weighted Least Square (WLS) method is an alternative method that can be used to overcome the problem of heteroscedasticity.

Research on overcoming the violation of the homoscedasticity assumption has been done before by several researchers. The research carried out by Setiyoko (2018) stated that for panel data regression that is heteroscedastic, the weighted least square method can be used as an alternative solution. Another study by Rustam & Aisyah (2022) stated that factors of health, education, and standard of living were very influential in assessing the Human Development Index.

This study aims to analyze the symptom of heteroscedasticity in multiple linear regression models using the weighted least square method. The data used is data on factors that affect the human development index in 34 provinces in Indonesia in 2021 .

2. Literature Review

2.1 Multiple Regression Analysis

Linear regression analysis is an analysis that used to investigate the causal relationship between two or more variables (Riyanto &Wikarya, 2018). Regression analysis was first named by San Francis Galton (a geneticist) in 1885 in his studies to investigate the relationship between a child's height and his father's height. Furthermore, this regression analysis is widely used to analyze variable relationships in the natural sciences, such as the study of the average oxygen consumption of each person with variables that influence it, namely age, body weight, average resting heart rate, the average heart rate immediately after a person runs and the running time for a certain distance.

In further developments, social sciences began to use regression analysis to analyze the relationship between socio-economic variables and business. For example, analyzing the effect of advertising spending on sales, analyzing the effect of income on consumption, and many more.

Regression analysis is concerned with the study of the dependence of one variable, namely the dependent variable, to one or more independent variables. If there is only one dependent variable and one independent variable it is called simple regression analysis, whereas if there are several independent variables it is called multiple regression analysis. In general, a multiple linear regression model with the dependent variable Y and the independent variable $X_1, X_2, \dots, X_i$ can be written as follows:

$$Y = \beta_0 + X_i\beta_1 + X_2\beta_2 + \dots + X_p\beta_p + \varepsilon_i \quad (1)$$

where:

Y	: dependent variable
$X_1, X_2, \dots, X_p$	: independent variable
$\beta_0, \beta_1, \beta_2, \dots, \beta_p$	: regression parameters/coefficients
ε	: error

The above equation can also be written simply as matrix notation as follows :

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

where

Y : observation vector of dependent variable with size $n \times 1$

X : independent variable with size $n \times (p + 1)$

β : coefficient vector of independent variables with size $(p + 1) \times 1$

ε : error vector of size $n \times 1$

The assumptions of classical linear regression according to Stang (2017), are as follows:

1. The average value of the error is zero, $E(\varepsilon) = 0$.
2. The error range is the same, namely a constant value of $\sigma 2$.
3. The error is normally distributed, $\varepsilon \sim N(0, \sigma 2)$.
4. There is no autocorrelation between errors so the covariance is zero, $cov(\varepsilon i\ ,\ \varepsilon j) = 0,\ i \neq j$.
5. There is no correlation between the independent variables X or there is no multicollinearity between the independent variables X.

2.2. Ordinary Least Square

The Ordinary Least Square (OLS) method is a method used to estimate classical regression coefficients by minimizing sum of squared errors (Sum of Square of Error/SSE) or $\sum \varepsilon^2$ (Maziyya et al., 2015). If the SSE is small, then the error for the data as a whole will also be small and this means that the regression line that has been created is the regression line that is closest to the pair of observations. The Ordinary Least Square (OLS) method was discovered by a mathematician from Germany named Carl Friedrich Gaus, where multiple linear regression analysis was used to see the influence of indicators on collectability both individually and together.

The estimator in the OLS method is obtained by minimizing the sum of the squared errors, as in the following equation

$$\varepsilon = \boldsymbol{Y} - \boldsymbol{X}\beta \quad (3)$$

the sum of the squares of the errors becomes :

$$\begin{aligned} \varepsilon_i^T\varepsilon_i &= (\boldsymbol{Y} - \boldsymbol{X}\beta)^T(\boldsymbol{Y} - \boldsymbol{X}\beta) \\ &= (\boldsymbol{Y^T} - \beta^T\boldsymbol{X^T})(\boldsymbol{Y} - \boldsymbol{X}\beta) \\ &= \boldsymbol{Y^TY} - \boldsymbol{Y^TX}\beta - \beta^T\boldsymbol{X^TY} + \beta^T\boldsymbol{X^TX}\beta \end{aligned} \quad (4)$$

To minimize the sum of squared errors (SSE) then $\varepsilon^T\varepsilon i$ in equation (4) is differentiated with respect to β so that we obtain the equation

$$\begin{aligned} -2\boldsymbol{X^TY} + 2\boldsymbol{X^TX}\beta &= \boldsymbol{0} \\ 2\boldsymbol{X^TX}\beta &= 2\boldsymbol{X^TY} \\ \boldsymbol{X^TX}\beta &= \boldsymbol{X^TY} \end{aligned} \quad (5)$$

Both sides of equation (5) are multiplied by $(\boldsymbol{X^T X})^{-1}$ to become

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

The OLS estimator contains ideal or optimum properties described in the Gauss-Markov theory (Gujarati & Porter, 2013). This theory considers the properties of BLUE (Best Linear Unbiased Estimator) or the best linear unbiased estimator. These characteristics include:
1. Linear, which is a linear function of a random variable, such as the dependent variable Y in a regression model.
2. Unbiased, where the average value or expected value is the same as the actual value.
3. Has minimum variance from all groups of linear and unbiased estimates; An unbiased estimator with the smallest variance is known as an efficient estimator.

2.3. Weighted Least Square

The Weighted Least Square (WLS) method is a form of development of OLS which is used to overcome heteroscedasticity problems (Nisa et al., 2020). One of the classic assumptions that must be met in OLS estimation so that the estimation results are reliable, namely the homogeneous residual variance $E(\varepsilon^2) = \sigma^2$ (homoscedasticity). Montgomery (2001) said that overcoming the heteroscedasticity problem in the regression model can be done using the method Weighted Least Squares (WLS). In this method, weighting is used which is proportional to the inverse of the various response variables so that an error is obtained according to the nature of the regression with ordinary least squares.

The WLS method is in principle the same as the least squares method, the difference is that in the WLS method there is the addition of a new variable, namely W as a weight. Parameter estimates β_0, β_1, β_2, … , βp for multiple linear regression using the WLS method are as follows:

$$\beta = (\boldsymbol{X^T W X})^{-1}\boldsymbol{X^T W Y} \quad (7)$$

Determining the weighting can be done by looking at the pattern shown by the residuals of the independent variables (Gujarati & Porter, 2013). These patterns include

a) The variance error is proportional to X_i^2

$$E(\varepsilon^2) = \sigma^2\ X_i^2 \quad (8)$$

If the heteroscedasticity test uses the graphical or Glejser method, it is believed that the error variance is proportional to the square value of the variable Xi as in Figure 1 below.

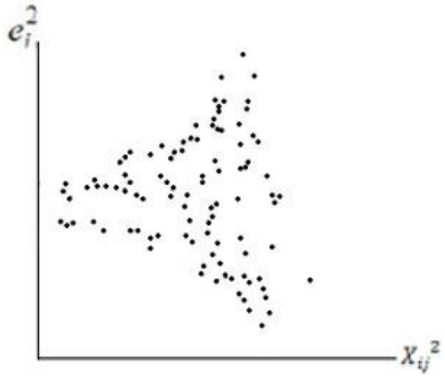

Figure 1 : Variance error is proportional to X_i^2

If the pattern shows a quadratic relationship as in Figure 1, then it can be assumed that the error variance is proportional to X^2, the weight used in the WLS method is $1/Xi$ so the regression equation will become :

$$\frac{Y}{X_i} = \frac{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p+\varepsilon}{X_i} \quad (9)$$

b) The variance error is proportional to X_i

$$E(\varepsilon^2) = \sigma^2\ X_i$$

If the heteroscedasticity test uses the graphic or Glejser method, it is believed that the error variance is proportional to the variable Xi as in Figure 2 below.

Figure 2 : Variance error is proportional to X_i

If the pattern shows a relationship like in Figure 2 then it can be assumed that the error variance is proportional to Xi so that the weighting used in the WLS method is $\frac{1}{\sqrt{X_i}}$ so the equation will become :

$$\frac{Y}{\sqrt{X_i}} = \frac{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon}{\sqrt{X_i}} \tag{10}$$

c) The variance of the error is proportional to the mean square value

$$E(\varepsilon^2) = \sigma^2 \ [E(Y)]^2 \tag{11}$$

If the error variance is proportional to $[E(Y)]2$, as illustrated in Figure 3 below :

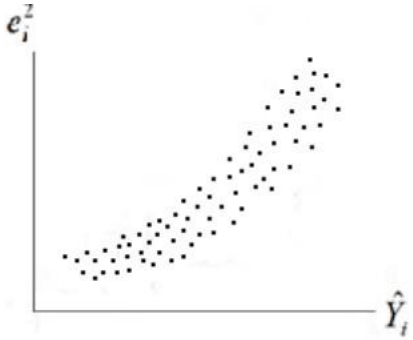

Figure 3. : Variance error is proportional to mean square value

If the error variance is proportional to $[E(Y)]2$ then the WLS method is carried out by OLS regression by ignoring heteroscedasticity to get the value of $\hat{Y}$ (Y predicted) which will be used as a weight, so that the regression equation becomes :

$$\frac{Y}{\hat{Y}} = \frac{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon}{\hat{Y}} \tag{12}$$

d) If it is assumed that σ^2 is known or can be estimated then that will be used as a weighting is $\frac{1}{\sigma^2}$ (squared residual) so that σ the regression equation becomes:

$$\frac{Y}{\sigma} = \frac{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon}{\sigma} \tag{13}$$

3. Materials and Methods

3.1. Materials

The data used in this study is secondary data obtained from the official website of the Indonesian Central Bureau of Statistics (BPS), namely https://bps.go.id. The variables used in this study were variables which affect the human development index in 34 provinces in Indonesia. The list of variables can be seen in Table 1.

Table 1 : List of Research Variables

No.	Variables	Name of Variables
1.	Y	Score of Human Development Index
2.	X_1	Life expectancy index (year)
3.	X_2	Expected length of school (year)
4.	X_3	Average length of school (year)
5.	X_4	Expenditure for food and non-food (IDR/capita/month)

3.2. Methods

Method of analysis is carried out as follow

1. Create the initial equation for the multiple linear regression model according to the variables used in the research. The model analyzed ini this paper was multiple regression model as can be seen in equation (1).

$$Y_{\mathrm{i}} = \beta_0 + X_{11}\beta_1 + X_{12}\beta_2 + X_{13}\beta_3 + X_{14}\beta_4 + \varepsilon_{\mathrm{i}} \quad (14)$$

   where:
   Y_1 : Score of Human Development Index
   X_1 : Life expectancy index (year)
   X_2 : Expected length of school (year)
   X_3 : Average length of school (year)
   X_4 : Expenditure for food and non-food (IDR/capita/month)
   β_{i} : regression coefficient
   ε_1 : error term
2. Estimating parameters using the OLS method as shown on equation (6).
3. Classical assumption test is carried out consisting of the normality test, multicollinearity test and homoscedasticity test.
4. If homoscedasticity assumption was violated, then heteroscedasticity occurs, WLS regression can be used instead (Fransiska et al., 2022), which giving weights W. The weight that will be tested are $\frac{1}{X}, \frac{1}{\sqrt{X}}, \frac{1}{\hat{Y}}, \frac{1}{\sigma}$.
5. Estimating parameters using the WLS method as shown on equation (7).
6. Test the classical assumptions again on the model that has been improved using the WLS method
7. Testing parameter significance.
8. The best weight will be chosen according to signification test of the model, assumptions test and coefficient of determination (Xu, 2019).
9. Interpreting the final model.

4. Results and Discussion

The first step in data analysis is to form a multiple linear regression as equation (1) using the OLS method and assumptions test. Table 2 shows that all coefficients were significant at 5% level. All VIF were under 10, indicating no multicollinearity among the independent variables.

Table 2 : Estimated Parameter and Significance Test of The Model Using Ordinary Least Square

		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
	Model	β	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-16.344	7.199		-2.270	.031		
	X_1	.767	.104	.491	7.396	.000	.773	1.293
	X_2	1.336	.392	.251	3.406	.002	.627	1.595
	X_3	1.397	.371	.329	3.767	.001	.445	2.245
	X_4	3.197E-6	.000	.235	2.966	.006	.542	1.845

Figure 4 shows that the error was normally distributed as the Jarque-Bera was 2.49 with probability more than 0.05.

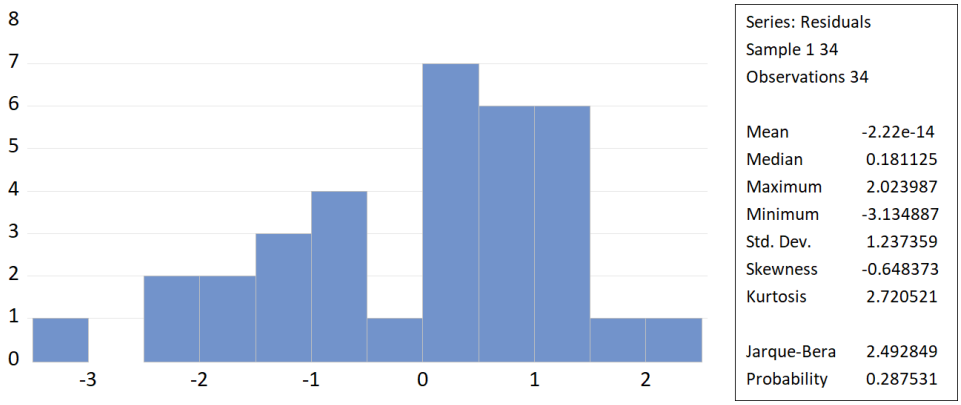

Figure 4 : Result of normality test Using Ordinary Least Square

The homoscedasticity test is performed using Glejser test. Table 3 shows that heteroscedasticity occurred in variable X_4 so WLS methods will be used.

Table 3 : Result of homoscedasticity test Using Ordinary Least Square

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model	β	Std. Error	Beta		
(Constant)	7.623	3.494		2.182	.037
X_1	-.082	.050	-.291	-1.621	.116
X_2	.002	.190	.002	.010	.992
X_3	-.303	.180	-.399	-1.686	.102
X_4	1.330E-6	.000	.545	2.543	.017

Weighted Least Square Model

Based on the Glejser test result, the data contains heteroscedasticity problems in variable X_4. Therefore, appropriate weighting will be used to handle the heteroscedasticity problem. Results of estimation using $\frac{1}{X}$, $\frac{1}{\sqrt{X}}$, $\frac{1}{Y}$, and $\frac{1}{\sigma}$ weights are presented at Table 4.

Table 4 : Estimated Parameters and Significance Test of the Model Using Weighted Least Square Methods with $\frac{1}{X}$, $\frac{1}{\sqrt{X}}$, $\frac{1}{Y}$ and $\frac{1}{\sigma}$ Weights

Variables	$\frac{1}{X}$	$\frac{1}{\sqrt{X}}$	$\frac{1}{Y}$	$\frac{1}{\sigma}$
C	-16.344	-16.413	-17.354	-10.027
	(7.064)	(7.127)	(7.283)	(2.649)
X1	**0.776****	0.772**	0.776**	0.718**
	(0.099)	(0.101)	(0.104)	(0.037)
X2	1.344**	1.342**	1.356**	1.257**
	(0.388)	(0.390)	(0.400)	(0.088)
X3	1.321**	1.357**	1.420**	1.042**
	(0.360)	(0.365)	(0.374)	(0.123)
X4	**3.091E-6****	**3.143E-6****	**3.058E-6****	4.58E-6**
	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.880	0.883	0.90	0.987

*Note : **significance at $\alpha = 5\%$*

Table 4 shows that for the weight $\frac{1}{X}$, all independent variables were significant at 5% level, all assumption were fulfilled but still there was heteroscedasticity on X_1 and X_4, while coefficient determination was 88%.

For the weight $\frac{1}{\sqrt{X}}$ all independent variables were also significant at 5% level, all assumption were fulfilled but still there was heteroscedasticity on X_4, while coefficient determination was 88.3%.

For the weight $\frac{1}{Y}$ all independent variables were also significant at 5% level, all assumption were fulfilled but still there was heteroscedasticity on X_4, while coefficient determination was increased to 90%.

For weight $1/\sigma$, all independent variables were also significant at 5% level, all assumption were fulfilled, no more heteroscedasticity, while coefficient determination was the highest 98.7%.

According to those results, it can be said that the best weight for regression model tested was $1/\sigma$ with all assumptions were fulfilled. The regression model using the best weight was as follows:

$$Y = \text{-}10.027 + 0.718X1 + 1.257X2 + 1.042X3 + 4.58E\text{-}6X4 \quad (15)$$

where Y is the human development index, X_1 is life expectancy, X_2 is the expected length of schooling, X_3 is the average length of schooling, and X_4 is the monthly expenditure per capita per month for food and non-food.

The model is very good because it produces coefficient determination value of 98.7%, which means that the variability in the coefficient of the human development index is significantly influenced by life expectancy, the expected length of schooling, the average length of schooling, and expenditure per capita per month for food and non-food. The plot of prediction and residual is presented at Figure 5.

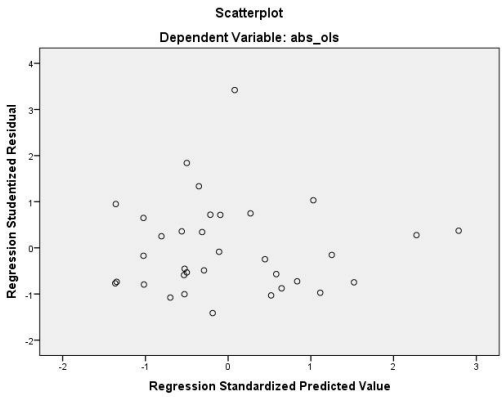

Figure 5 : Plot of prediction and residual

Figure 5 shows that the plot of prediction and residual has no pattern, it strengthens that the model estimated using weighted least square was fit to the data.

5. Conclussion

According to those results, it can be said that the best wight for regression model tested was $1/\sigma$ with all assumptions were fulfilled. The model estimated using weighted least square is very good because it produces coefficient determination value of 98.7%, which means that the variability in the coefficient of the human development index is significantly influenced by life expectancy, the expected length of schooling, the average length of schooling, and expenditure per capita per month for food and non-food.

References

Fransiska W., Racmawati, and S. Nugroho. (2022). A Comparison of Wighted Least Square and Quantile Regression for Solving Heteroscedasticity in Simple Linear Regression. *Journal of Statistics and Data Science* Vol 1 (1) : 19-29.

Gujarati, D. N. & Porter D. C. (2013). Basic Econometrics 5th Edition. New York: McGraw Hill Companies Inc.

Maziyya, P. A., Sukarsa, I.K.G., and Asih, N.M. (2015). Overcoming Heteroscedasticity on Regression Using Weighted Least Square. *E-Jurnal Matematika*. (4)1: 20- 25.

Montgomery, D.C. (2001). Introduction to Linear Regression Analysis 5th ed. New Jersey: John Wiley& Sons, Inc.

Nisa, H., Kusnandar, D., & Martha, S. (2020). Parameter Estimation of Weighted Least Square Method in Overcoming Heteroscedasticity Problem. *E-Jurnal Matematika.* (9)1: 65–70.

Riyanto & Wikarya, U. (2018). Economic and Business Statistics. Fisrt Edition. Jakarta: Mitra Wacana Media.

Rustam, D. & Aisyah, S. (2022). Analysis of Human Development Index in West Sumatera Districts Using Panel Data Analysis. *Jurnal Pundi.* (6)1: 205-206.

Setiyoko, R. (2018). Heteroscedasticity Analysis with The Weighted Least Square Method on Panel Data. Thesis. Faculty of Matemathics and Natural Science. Semarang: Semarang State University.

Xu, P. (2019). Improving The Weighted Least Square Estimation of Parameters in Errors-in Variables Models. *Journal of The Franklin Institute* Vol 356 (15) : 8785-8802.