JSDS: JOURNAL OF STATISTICS AND DATA SCIENCE
VOLUME 1, No 1, March 2022

https://ejournal.unib.ac.id/index.php/jsds/index

A Comparison of Weighted Least Square and Quantile Regression for Solving Heteroscedasticity in Simple Linear Regression

Welly Fransiska[1*], Sigit Nugroho[2], Ramya Rachmawati [2]

[1]IKIP Pagar Alam
[2]Department of Mathematics, Faculty of Mathematics and Natural Science, The University of Bengkulu

* Corresponding Author. Email: wellyfransiska@gmail.com

Article Info

Article History:
Received: December 23, 2021
Revised: January 15, 2022
Accepted: January 22, 2022
Available Online: March 1, 2022

Key Words:
BLUE
Heteroscedasticity
Quantile Regression
Wheigted Least Square

Abstract

Regression analysis is the study of the relationship between dependent variable and one or more independent variables. One of the important assumption that must be fulfilled to get the regression coefficient estimator Best Linear Unbiased Estimator (BLUE) is homoscedasticity. If the homoscedasticity assumption is violated then it is called heteroscedasticity. The consequences of heteroscedasticity are the estimator remain linear and unbiased, but it can cause estimator haven‘t a minimum variance so the estimator is no longer BLUE. The purpose of this study is to analyze and resolve the violation of heteroscedasticity assumption with Weighted Least Square(WLS) and Quantile Regression. Based on the results of the comparison between WLS and Quantile Regression obtained the most precise method used to overcome heteroscedasticity in this research is the WLS method because it produces that is greater (98%).

1. INTRODUCTION

Regression analysis is defined as a study of the functional relationship of one or more independent variables to one or more dependent variables. Regression is one of the most widely used statistical analysis techniques. The principles of regression analysis are then derived, various statistical analysis techniques that can be used in various fields ranging from science, economics and business, industry and so on. Regression itself was first introduced by Francis Galton in 1877, which then experienced many developments from year to year[1].

The Ordinary Least Square(OLS) method is one way to estimate the regression parameters in linear regression. To obtain a Best Linear Unbiased Estimator(BLUE), the following assumptions, such as Normal distribution, no multicollinearity, no autocorrelation, and homoscedasticity are all requirements for data[2]. Violation of the assumption of homoscedasticity is called heteroscedasticity, which means the error is not constant. When it occurs, the OLS estimates are still unbiased but become inefficient. Meaning that the variance tends to enlarge so that it is no longer a variance minimum. Estimates can even lead to incorrect conclusions; hence a solution for resolving the problem is required. The effect of heteroscedasticity must be lost if this assumption is violated[3].

The next problem is if there is a possibility that the slope of the data lies not in the median but a particular quantile piece. Approach with the median is considered inaccurate because it only sees two groups of data divided by the median value. So that the Quantile Regression method was developed. Quantile Regression is a regression method with an approach to separate or divide data into specific quantiles where it is suspected that there is a difference in the estimated value.

The Weighted Least Square (WLS) method is an alternative method that can overcome heteroscedasticity. The WLS method is the same as the OLS method by minimizing the number of residues, but the WLS method is weighted a relevant factor and then uses the OLS method on the weighted data.

This study will compare the Quantile Regression method and the WLS method to overcome the problem of heteroscedasticity and see which method is better in dealing with heteroscedasticity. This study follows the following criteria:

a. In this study used heteroscedasticity analysis that occurs in simple regression.

b. The WLS method uses b weighting (error variance proportional to X_{ij}) because, after investigating the data using graphs, it turns out that the pattern shows a linear relationship.

2. METHOD

The methods used in this research comprise three methods, those are Linear Regression Analysis, Weighted Least Square, and Quantile Regression.

a. Linear Regression Analysis

Regression analysis is a technique for examining the relationship between two or more variables. This relationship can be written as an equation, which connects the dependent variable to one or more independent variables[4]. This analysis is used to determine if the relationship between the independent and dependent variables is positive or negative and to predict the value of the dependent variable if the value of the independent variable increases or decreases. Typically, interval or ratio scale data is used.

In general, the simple regression equation is written as

$$\mathbf{y}_{ijk} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij} + \boldsymbol{\varepsilon}_{ijk} = \boldsymbol{\mu}_{ij} + \boldsymbol{\varepsilon}_{ijk} \tag{1}$$

We require a BLUE (Best Linear Unbiased Estimator) parameter estimate method in linear regression analysis, and the most commonly used method is the Ordinary Least Squares(OLS) method. Some assumptions must be met when using Ordinary Least Squares(OLS). The data must have a Normal distribution, no autocorrelation, and homoscedasticity. If all of these assumptions are met, OLS will be able to meet the BLUE estimator. If one or more assumptions are not met, the estimation results obtained will not meet the BLUE qualities. One of the assumptions that must be met when performing estimation is homoscedasticity. The term homoscedasticity refers to the fact that the error variance is constant[5]. The error variance that is not constant is called heteroscedasticity. Heteroscedasticity is a type of violation of the homoscedasticity assumption. When heteroscedasticity occured in line with estimating the Ordinary Least Square method, the estimate results obtained no longer meet the nature of BLUE, and therefore the other alternative approaches must be used to estimate parameters to overcome the heteroscedasticity.

1. Best

   A regression line is a way of understanding the pattern of relationship between two or more data series. The regression line is "Best" if it produces the smallest error. The error is difference between the observed value and the predicted value by regression line. Best accompanied with unbiased, the regression estimator is called efficient.
2. Linear

   An estimator β is called linear if it is a linear function of the sample. The mean of X is a linear estimator because it is a linear function of the values of X. The values of OLS is also a linear estimator [6]
3. Unbiased

   An estimator is said to be unbiased if the expected value is equal to the parameter ($E(\hat{\beta}) = \beta$). A Gaussian theorem which is the main concern in Econometrics is known as the classical assumption. Assumptions in the classical linear regression model has ideal properties known as the Gauss-Markov theorem. The OLS will produce a BLUE estimator.

The normality test is used to verify that the variables used in regression models are random and must be Normal distributed. Because only the dependent variable has randomness, Normality testing on the dependent variable is sufficient. The random variable is distributed independently and identically, then with a few exceptions distribution of numbers tends to a Normal distribution if the number of such variables increases unlimited[3]. However, if the assumption of Normal is not met, it can conclude that the theory does not apply because the data is not Normal distributed. One way to handle the issue of Normality is to add sample observation or to eliminate some samples that have extreme data values and are suspected of causing the non-fulfilment of the data's Normality. Therefore, before a further theory is used and conclusions are drawn based on the theory in which the assumption of normality is used, it needs to be investigated whether the assumption is fulfilled or not. Normality test is done by Shapiro Wilk test and a normal Q-Q plot graph [8]

The Shapiro Wilk test is a method or formula for calculating the distribution of data made by Shapiro Wilk. The requirements in the Shapiro Wilk test are interval or ratio scale data.Significance is compared with Shapiro Wilk's table to see the position of the T_3values compared with Shapiro Wilk's table value to see the position of the probability value (p). If the p-value is less than 5%, it is rejected H_0.

$$T_3 = \frac{1}{D}\left[\sum_{i=1}^{k} a_i (X_{(n-i+1)} - X_i\right]^2 \quad (2)$$

where

D : determinant ($D = \sum_{i=1}^{n}(X_i - \bar{X})^2$

a_i : the coefficient of Shapiro Wilk

$X_{(n-i+1)}$: the *n-i+1*-th observation

X_i : the *i*-th observation

According to Gasperz and Vincent, heteroscedasticity can lead to predictions where the parameter is inefficient and so without a minimal variance. Parameter estimation is considered efficient because it has a minimum variance or because the assumption of homoscedasticity is met. Heteroscedasticity can be solved by transforming the variables, both independent and dependent variables, to fulfil the homoscedasticity assumption.

If there is a heteroscedasticity condition, the impact is that it is difficult to measure the actual standard deviation, or in other words, it can yield a standard deviation that is too large or too small. As the error rate of the variance continues to increase, then the level of confidence will be narrower. Heteroscedastic detection is essential for determining whether or not the data is heteroscedastic. According to Manurung, there are two ways to determine the presence of heteroscedasticity, namely the informal method and the formal method. Informal methods are usually used by looking at the graph plot of the predictive value of an independent variable (ZPRED) with the residual (SRESID). The variable is declared not to have heteroscedasticity if there is no pattern and the point spread above and below zero on the Y-axis. The Park Test and Glejser are two formal methods for detecting the presence of heteroscedasticity.

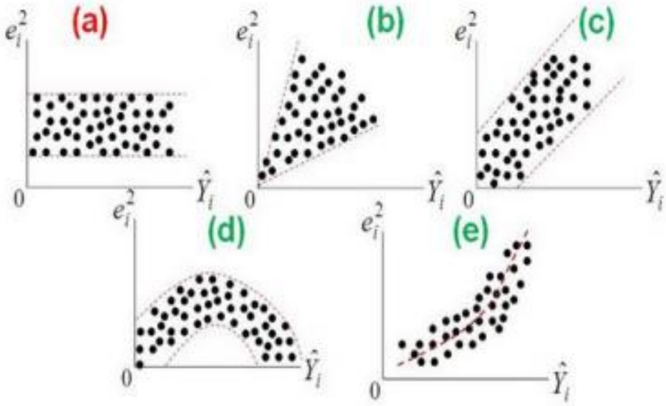

Figure 1. Plot of the square error and the variable

In Figure 1, the plot being compared is between the squared error and the variable. Figure 1.a, it can be seen that the error is relatively constant, so it does not show a pattern on the variance means homoscedastic. Figures 1.b, 1.c, 1.d, and 1.e shows their respective patterns, namely fluctuations, trends, and logarithms, so that is heteroscedastic.

The Park test involves regressing the residual value against each dependent variable. The hypothesis is as follows

$H_0: Homoscedasticity$

$H_1: Heteroscedasticity$

H_0 is accepted if $-t_{table} < t_{val} < t_{table}$. Hence homoscedasticity occurred. On the other hand, H_0 is rejected if $t_{val} < -t_{table}$ or $t_{val} > t_{table}$. It means that the data reflect heteroscedasticity.

b. Weighted Least Square

The Weighted Least Square (WLS) is similar to the Ordinary Least Square (OLS) method in theory; the difference is that the WLS method adds an additional w, which represents weight. The regression model $Y = X\beta + \varepsilon$ with $var(\varepsilon) = W\sigma^2$. The matrix **W** is a diagonal matrix containing w_i.

$$W = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix} \quad (3)$$

The weight is calculated through examining the pattern shown by the residual of the independent variable. These are some of the patterns:

1. The Error variance is proportional to ${X_{ij}}^2$, for $1 \le j \le k$

$$E(\varepsilon^2) = \sigma^2 X_{ij}^2 \quad (4)$$

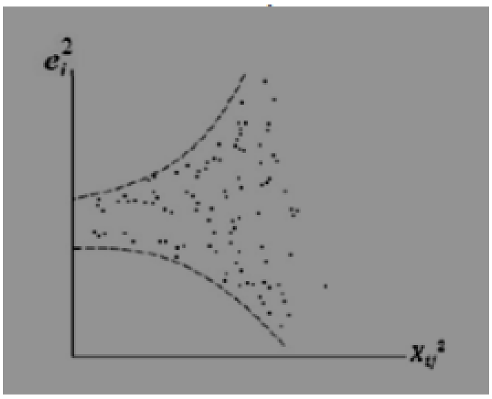

Figure 2. The error variance is proportional to ${X_{ij}}^2$

If the pattern shows a quadratic relationship as in Figure 2, it can be assumed that the error variance is proportional to ${X_{ij}}^2$, so that the weights used in the WLS method is $\frac{1}{{X_{ij}}^2},$

2. The Error Variance is proportional to X_{ij}

$$E(\varepsilon^2) = \sigma^2 X_{ij} \quad (5)$$

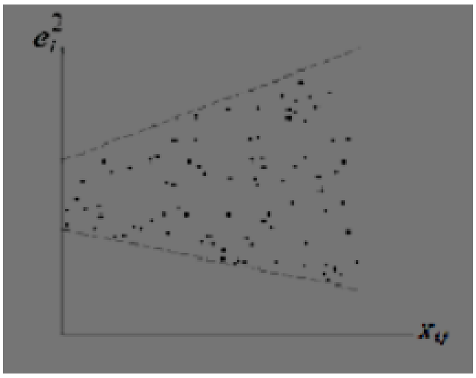

Figure 3. The error variance is proportional to X_{ij}

If the pattern shows a linear relationship as in Figure 3, it can be assumed that the error variance is proportional to X_{ij}, so that the weights used in the WLS method is $\frac{1}{\sqrt{X_{ij}}},$

3. The Error Variance is proportional to $[E(Y_i)]^2$

$$E(\varepsilon^2) = \sigma^2[E(Y_i)]^2 \tag{6}$$

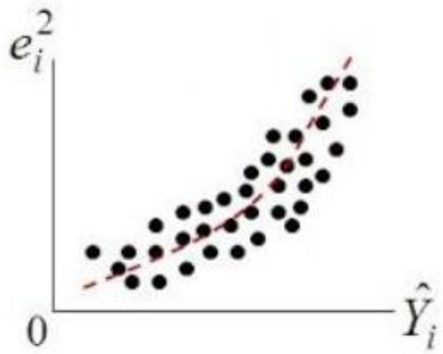

Figure 4. The Error Variance is proportional to $[E(Y_i)]^2$

If the Error variance is proportional to $[E(Y_i)]^2$, then the WLS method is carried out by performing OLS regression by ignoring heteroscedasticity to get the Y_i value which will be used as a weight so that the regression equation becomes

$$\frac{\beta_0}{\hat{Y}_i} + \beta_1\frac{X_{i1}}{\hat{Y}_i} + \beta_2\frac{X_{i2}}{\hat{Y}_i} + \cdots + \beta_k\frac{X_{ik}}{\hat{Y}_i} + \frac{\varepsilon_i}{\hat{Y}_i} \tag{7}$$

c. Quantile Regression

Koenker and Basset first introduced quantile regression in 1978. This method is an extension of the conditional quantile regression model in which the conditional quantile distribution of the dependent variable is expressed as a function of the covariates. This approach can make it possible to estimate the quantile function of the conditional distribution of the dependent variable on each quantile value according to the desired quantile [9]. Quantile regression is highly recommended to analyze several data that is not symmetrical and has an inhomogeneous distribution. Interval estimation in quantile regression can be done using the direct approach, rank score, and resampling [7]. The general equation for linear quantile regression specifically for the conditional quantile $Q_{yi}(\tau|X_{1i}, X_{2i}, \dots, X_{pi})$ of the dependent variable Y is $Y_i = \beta_0(\tau) + \beta_1(\tau)X_{1i} + \cdots + \beta_p(\tau)X_{pi} + \varepsilon_i(\tau)$

The quantile regression model is presented in the form of a matrix, the above equation can be written as follows:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{p1} \\ 1 & X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{pn} \end{bmatrix} \begin{bmatrix} \beta_0(\tau) \\ \beta_1(\tau) \\ \vdots \\ \beta_p(\tau) \end{bmatrix} + \begin{bmatrix} \varepsilon_1(\tau) \\ \varepsilon_2(\tau) \\ \vdots \\ \varepsilon_n(\tau) \end{bmatrix} \tag{8}$$

Furthermore, the equation can be written in the form of the following linear model:

$$y = X\beta(\tau) + \varepsilon \tag{9}$$

If the conditional function is from the τ –th quantile with a certain independent variable X, then the conditional function is defined as follows:

$$Q_{Y_i}(\tau | X_{1i}, X_{21}, \dots, X_{pi}) = Q_y(\tau|X) = X^T\beta(\tau), \;\; i = 1,2, \dots, n \tag{10}$$

3. RESULTS AND DISCUSSION

The data is generated with a value of n, starting from n = 20 to n = 1000. The results of the R program for simple linear regression with n = 500 obtained the parameter estimation results with the least squares method as follows:

Table 1. Parameter Estimation Results with Ordinary Least Squares Method

Variable	Estimation	The Standard Error	t_{hitung}	ρ – Value	sig
	1.37867	0.02605	52.91	<2e-16	***
x_1	1.99254	0.01145	173.95	<2e-16	***

From Table 1 the regression model can be written as follows:

$$y = 1.37867 + 1.99254x_1$$

When the p–value ≤ 0.01, the parameter is considered to be extremely significant (***), significant (**) when 0.01 <p-value ≤ 0.05, and not significant when p–value > 0.05. After being tested using the ordinary least squares approach, the model is assessed to see whether it fits the normality and heteroscedasticity assumptions.

The normality assumption test is done by observing the residual plot of the ordinary least squares method estimation model. In Figure 5, the results of the normality assumption test of the ordinary least squares method with data generated in the R program show that the QQ plot demonstrates the points which do not follow the necessary pattern and move away from the regression line, so it can be said that the error model does not follow the normal distribution.

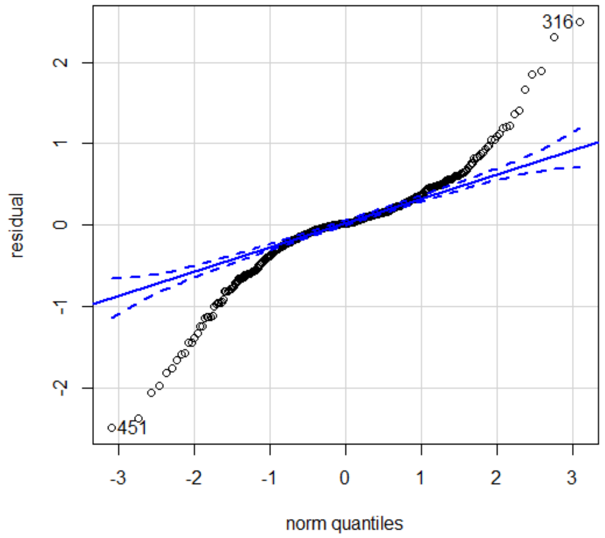

Figure 5. Normality test using QQ-plot

Furthermore, a normality test that is performed using the Shapiro Wilk test with the R program yields the following results: $w = 0{,}9051, \; \rho - value = 2{,}2e - 16$. The model is then tested by looking at the graph to check if it fits the heteroscedasticity assumption. The following is a graph showing the expected y against residual:

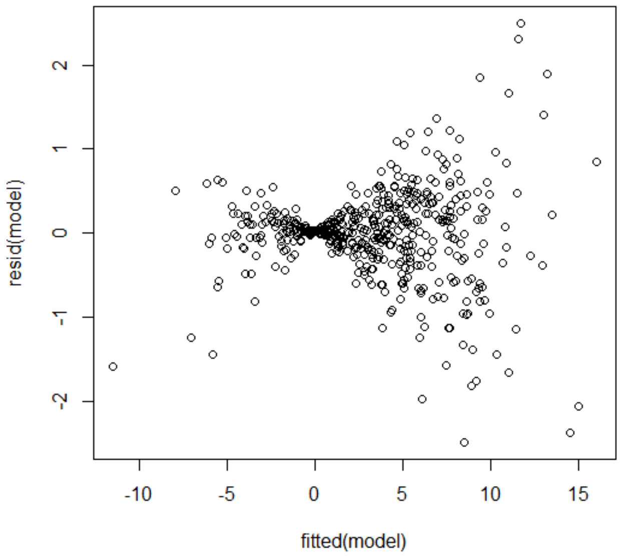

Figure 6. Plot of the prediction of y against it's residual

To confirm the heteroscedasticity test, the Breusch Pagan test has been carried out and has obtained the value BP = 70.393, df = 1, p-value < 2.2e-16, so it can be concluded that the p-value ≤ 0.05 indicates heteroscedasticity model. After the model is proven to be not normal distribution and heteroscedastic, the data will be transformed by two method.

The first method to transform the data to normal and homoscedasticity is the quantile regression method, with the R program, beginning with estimating each τ-th quantile so that the estimated value for each regression coefficient in each quantile is obtained.

Table 2. The Estimation Result of Intercept β_0 in Each Quantile

Parameter	Quantile	Coefficient	Standard Error	T-$Statistics$	$\rho - value$
β_0	0.05	0.69474	0.11715	5.93037	0.00000
	0.10	0.93718	0.05125	18.28803	0.00000
	0.15	1.05345	0.03828	27.51871	0.00000
	0.20	1.15559	0.03184	36.28961	0.00000
	0.25	1.23586	0.02575	47.98620	0.00000
	0.30	1.29143	0.02139	60.38882	0.00000
	0.35	1.33242	0.01739	76.62393	0.00000
	0.40	1.35717	0.01485	91.41322	0.00000
	0.45	1.37923	0.01223	112.74430	0.00000
	0.50	1.40314	0.01430	98.09206	0.00000
	0.55	1.43007	0.01375	103.96713	0.00000
	0.60	1.45375	0.01678	86.62688	0.00000
	0.65	1.48823	0.01697	87.71310	0.00000
	0.70	1.53400	0.01833	83.67014	0.00000
	0.75	1.47535	0.01923	79.86576	0.00000
	0.80	1.62836	0.02121	76.75720	0.00000
	0.85	1.69302	0.02467	68.63814	0.00000
	0.90	1.79698	0.04106	43.76947	0.00000
	0.95	2.00935	0.06859	29.29458	0.00000

The estimation results of intercept β_0 at quantile 0.50 yield the estimated intercept for β_0 of 1.40314 with a p-value of 0.00000, which is significant at the value of α = 0,05. Furthermore, the table below presents the estimation results for β_1.

Table 3. The Estimation Result of Slope β_1

Parameter	Quantile	Coefficient	Standard Error	$T-Statistics$	$\rho-value$
β_1	0.05	1.83699	0.11715	5.93037	0.00000
	0.10	1.87424	0.02395	18.28803	0.00000
	0.15	1.88735	0.01787	105.63945	0.00000
	0.20	1.89729	0.01443	131.45272	0.00000
	0.25	1.91018	0.01250	152.86895	0.00000
	0.30	1.9338	0.00939	205.85008	0.00000
	0.35	1.93964	0.01207	160.73754	0.00000
	0.40	1.96144	0.01197	163.91028	0.00000
	0.45	1.97864	0.01268	156.06094	0.00000
	0.50	2.00330	0.01018	196.80112	0.00000
	0.55	2.02006	0.01037	194.72209	0.00000
	0.60	2.02687	0.00911	222.42890	0.00000
	0.65	2.03897	0.00762	267.49301	0.00000
	0.70	2.05016	0.00832	246.33270	0.00000
	0.75	2.08765	0.00919	221.9865	0.00000
	0.80	2.06176	0.00951	216.74088	0.00000
	0.85	2.07563	0.01181	175.80879	0.00000
	0.90	2.08951	0.01715	121.85148	0.00000
	0.95	2.076531	0.01264	167.6254	0.00000

Based on Table 3, the estimation results of slope β_1 at quantile 0.50 of 2.00330 with a p-value of 0.00000, which is significant at the value of α = 0,05. The results of the coefficient above yield quantile regression models, $y = 1{,}40314 + 2{,}00330x$

The Weighted Least Square (WLS) is one of the data transformation methods, using weights to change the data. The data transformed by the WLS method will be re-estimated using the Least Square Method. Determination of weight is done by looking at the error variance plot. After detection using the graphical method, it is believed that the error variance is proportional to X_{ij}, as shown in the following Figure 7:

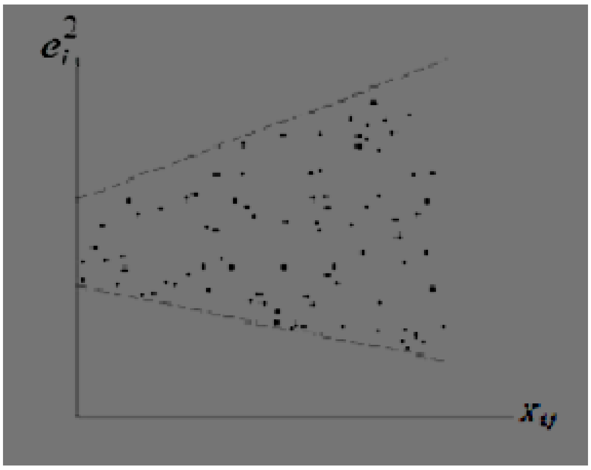

Figure 7. The Error Variance is proportional to X_{ij}

The estimation results using the WLS method in the R program lead the model: $y = 1.0086994 - 0,0242969x$. After obtaining a new model using the R program, both the quantile regression method and Weighted Least Square method, the normality test should be carried out using the QQ plot method and the Shapiro Wilk test, afterward the heteroscedasticity test is carried out with a plot between the predicted y value against the residual (y-axis) and perform the Breusch Pagan test.

The following are the results of the normality test for quatile regression and Weigted Least Square:

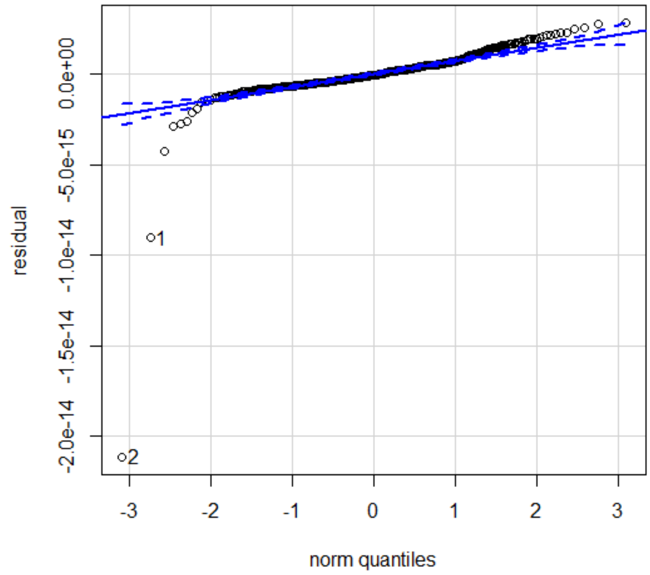

Figure 8. Normality Test using QQ-plot for Quantile Regression

In addition, the normality test is carried out with the Shapiro Wilk test, and its results are as follows: W = 0.54959, ρ-value < 2,2e-16. The p-value of 2.2e-16 shows that the data is normally distributed since it is greater than α.

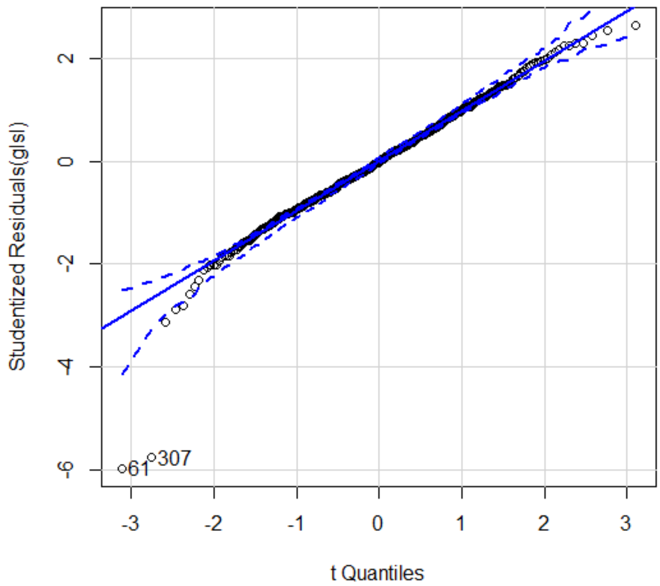

Figure 9. Normality Test using QQ-plot for WLS

In addition, the normality test is carried out with the Shapiro Wilk test and results in: W = 0.98762, p-value = 0.0003022. The p-value of 0.0003022 shows that the data is normally distributed because it is greater than α.

The next step is to determine whether or not the data is heteroscedastic. The predicted and residual y-value plot using the Quantile Regression and WLS methods are shown in the figures below.

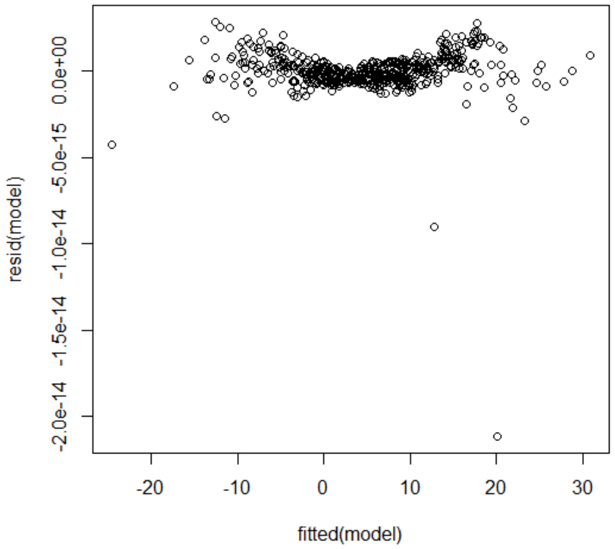

Figure 10. The Predicted Y against Residual using Quantile Regression

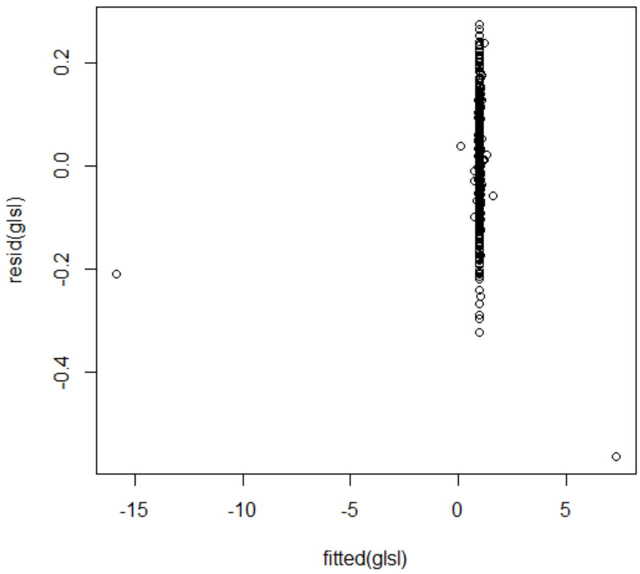

Figure 11. The Predicted Y against Residual using WLS

The data fulfill the assumption of homoscedasticity, as shown in Figure 10 and Figure 11; thus, the two approaches, Quantile Regression and WLS, have been compared to the value of the heteroscedasticity test with the Breush Pagan test.

The values for the heteroscedasticity test in the quantile regression model are BP = 1.9316, df = 1, p-value = 0.1646, so it can be concluded that the p-value greater than α = 0,05 indicates that the model is not heteroscedastic. Moreover the results of the heteroscedasticity test in the WLS model are BP = 15.265, df = 1, p-value = 9.341e-05. Hence it can be concluded that the p-value = 9.341e-05 is greater than α = 0,05, which means that the data is not heteroscedastic.

Based on the test results regarding the ability of quantile regression and WLS in overcoming heteroscedasticity in regression analysis, it can be said that these two methods can overcome the problem of heteroscedasticity in the data used in this research.

Table 4. The Determination Coefficient of Quantile Regression and WLS

n	R^2	
	WLS	**Reg. Quantil**
100	80.45%	54.34%
200	83.23%	60.28%
300	85.67%	63.43%
400	90.01%	70.19%
500	98.01%	71.13%
600	98.21%	71.88%
700	98.23%	72.97%
800	98.40%	65.29%
900	98.13%	67.89%
1000	98.55%	72.63%

The estimation result with quantile regression is the R^2 value of 71%. After testing the normality assumption, it is known that the error follows the normal distribution, and the homoscedasticity assumption is met. While the estimation result with WLS shows the R^2value of 98%, and after testing the normality assumption, it is known that the error follows the normal distribution and the assumption of homoscedasticity is met.

4. CONCLUSION

In this study, quantile regression and WLS have been proven to be able to overcome heteroscedasticity, but after several tests with different value of n, it can be said that the WLS method is better than quantile regression in overcoming the heteroscedasticity problem, because WLS can overcome the problem of normality as well as the problem of heteroscedasticity and has a higher determination coefficient, that is 98% (for n = 500).

REFERENCES

[1] Netter, J. W., Waserman, and Kutner, M. H., Applied Linear Statistical Models, Terjemahan Bambang Sumantri. Illinois. Homewood, 1997
[2] Koenker, R. and Basset, G., “Regression Quantile”. Econometrica. January. 46:1, pp. 33-50, 1978.
[3] Gujarati, Essentials Of Econometrics, 4th Edition, The McGraw-Hill Companies, New Jersey, 2003
[4] Nachrowi, Penggunaan Teknik Ekonometrika, Raja Grafindo Persada, Jakarta, 2008
[5] K Koenker, R. and D’Orey, V., “A remark on Computing Regression Quantiles”, Applied Statistics, Vol. 43, 410 – 414, 1993
[6] Koenker, R. & K. Hallock, “Quantile regression”, Journal of Economi Perspectives 15, 143—156, 2001
[7] Koenker, Qualtile Regression. University of Illinois, Urbana-Champign, 2005
[8] Ghozali, I., Aplikasi Analisis Multivariate dengan Program SPSS Edisi Kedua, Semarang: Badan Penerbit Universitas Diponegoro, 2001
[9] Chen, “Computational for Quantile Regression”, The Indian Journal of Statistics Vol 67, Part 2, Hal 399-417, 2005