Volume 4 Issue 2 (2024) Pages 133-142 P-ISSN:2798-2696 E-ISSIN:2798-0057
JOSTECH
Journal of Science and Technology
Website: https://ejournal.uinib.ac.id/jurnal/index.php/jostech Email: jostech@uinib.ac.id

Quantile Regression Analysis; Simulation Study With Violation of Normality Assumption

Lilis Harianti Hasibuan ✉[1], Ferra Yanuar[2], Dodi Devianto[3], Maiyastri[4]
Department Matematics, UIN Imam Bonjol Padang, Indonesia[1]
Department Matematics, Universitas Andalas, Indonesia[2,3,4]
email: lilisharianti@uinib.ac.id[1], ferrayanuar@sci.unand.ac.id[2] , ddevianto@sci.unand.ac.id[3], maiyastri@sci.unand.ac.id[4]

Received 03 Agustus 2024, Accepted 21 September 2024, Published 30 September 2024

Abstrak

Regresi kuantil merupakan metode peluasan dari regresi linear sederhana yang sifat kerjanya memisahkan atau membagi data ke dalam kuantil kuantil tertentu. Metode ini meminimalkan asimetris sisaan absolut dan mengistimasi fungsi kuantil bersyarat. Estimasi parameter pada metode regresi kuantil tidak memerlukan asumsi parametrik kenormalan. Data pada penelitian ini adalah data bangkitan yang berasal dari distribusi yang berbeda. Distribusi dari variabel independent pada penelitian ini berasal dari distribusi t, Normal, dan distribusi eksponensial. Sedangkan distribusi error nya berasal dari distribusi chi kuadrat. Penelitian ini menghasilkan model yang beragam dari kuantil kuantil yang terpilih. Nilai estimasi parameter di setiap kuantil hampir mendekati nilai awal yang ditetapkan. Penelitian ini menemukan model terbaik pada kuantil 0.5 dengan melihat nilai MSE terkecil dari semua kuantil sebesar 1.2662. Sehingga model yang terbaik yang diperoleh $\hat{y} = 0.2653 + 0.9400X_1 + 1.1683X_2 + 0.9925X_3 + 1.0281X_4 + 0.9102X_5$.

Kata Kunci: Regresi Kuantil, Koefisien Regresi, MSE

Abstract

Quantile regression is an extension method of simple linear regression whose work is to separate or divide data into certain quantiles. This method minimizes the asymmetric absolute residual and estimates the conditional quantile function. Parameter estimation in the quantile regression method does not require the parametric assumption of normality. The data in this study are generated from different distributions. The distribution of the independent variables in this study comes from the *t* distribution, normal and exponential distribution. Meanwhile, the error distribution comes from the chi square distribution. This research produces various models of the selected quantiles. The estimated parameter values at each quantile are almost close to the initial values set. This research found the best model at quantile 0.5 by looking at the smallest MSE value of all quantiles of 1.2662. The best model obtained is $\hat{y} = 0.2653 + 0.9400X_1 + 1.1683X_2 + 0.9925X_3 + 1.0281X_4 + 0.9102X_5$.

Keywords: Quantile regression, coefficient regression, MSE

✉ Corresponding author

INTRODUCTION

Regression analysis is a statistical method that aims to model the relationship between a dependent variable (Y) and one or more independent variables (X) [1]. Based on the relationship pattern between independent variables and dependent variables, regression analysis is divided into linear regression analysis and non-linear regression analysis. A non-linear regression model is a model which, if differentiated, results are still a function of the model parameters [2]. Meanwhile, linear regression is an analysis used to obtain a linear relationship between independent variables and dependent variables. The relationship between these variables is usually expressed in a regression equation which is generally written as $Y=X'\beta+e$, with e representing the residual or error from an observation.

This regression model can connect independent variables and dependent variables through a regression parameter which is denoted as Beta (β). The overshadow the parameter values in the regression equation, the least squares method often called Ordinary Least Squares (OLS) is usually used. This method has the principle of minimizing the sum of the squared residuals (errors) of observation values relative to the average, because the OLS method is based on the mean distribution function.

The mean value is a measure of the centrality of a distribution so that only a little information is known about the entire distribution. This OLS method can be applied if several assumptions are made of normality, non-multicollinearity, homogeneity of residual variance and non-auto correlation. These assumptions must be met so that the value of the parameter (β) estimator is BLUE (Best Linear Unbiased Estimator). However, this method is very vulnerable to violations of these assumptions because not all data meets the assumption of normality, the data variance is not homogeneous (heteroscedasticity), there is data that contains multicollinearity, autocorrelation and so on.

Data that violates this assumption cannot be applied using the OLS method to find the estimated parameter values of the regression model. As a result, the mean is less appropriate to use as an estimator for the middle value of the data. However, if one of the assumptions are not met, the result could be misleading [3]. Then, median regression was developed with the LAD (Least Absolute Deviation) approach which was developed by replacing the mean approach in OLS with the median. The estimated parameter values using this method are obtained by minimizing the sum of the absolute values of the residuals. So that the parameter value leads to the median value of the data. This is done by considering whether the data is bell-shaped or asymmetrical. The same thing was also stated by [4] who said that this method was carried out by grouping differences in estimated values at certain quantiles. Quantile regression is a robust approach in situations where the limitations addressed above present for OLS estimator [5].

The quantile method is one of the regression modeling methods by dividing a batch of data into the same parts after the data is sorted from the smallest or largest [6]. Quantil regression is an approach in regression analysis instroduced by Koenker and Basset [7]. Quantile regression in his theory is able to overcome the violation of normality asumptions, heteroscedasticity, multicollinearity problems and so on. This methode uses the parameter estimation approach by separating or dividing the data

and minimizing the absolute asymmetry weigthed error and prespuses a conditional quantile function on a distribution of data [8].

Research related to quantile regression has been put forward by many researchers in different areas, including study by Ferra [4] who demonstrated the application of quantile regression techniques to determine indicators of health status and determine the characteristics of healthy and unhealthy individuals using survey data. Recently, several statitistical distribution heve been considered for quantile modeling with parametric quantile regression is formulated by first parameterizing the baseline distribution in term of a quantile [9]. Quantile regression also discussed [10] with introduce a new distribution on (0,1) with transformation of a positive random variable following Chen distribution with estimation parameter with quantile regression. The application of quantile regression was also introduced [3][11] with generate data to obtain the best model. Wu and Y. Liu [12] focused on variable selection using SCAD functions. Meanwhile, the research focuses on simulated data originating from a non-normal distribution to see the smallest MSE value at each quantile. Quantile regression can be an alternative if with multiple linear regression there is a violation of the assumption of homoscedasticity or the data is not normal because there is data that is very different from other data points and data diversity [13]. In this research, the quantile regression method will be applied using simulated data with an error distribution that violates the normality assumption.

METODOLOGY

Quantile Regression

Quantile regression is a statistical method used to analyze the relationship between an independent variable and a dependent variable by focusing on a particular quantile of the dependent variable distribution, after the data has been sorted from smallest to largest. The objectives of quantile regression include distribution analysis, addressing heteroscedasticity, exploring the effect of extremes, provides a more complete view, and flexible model.

Quantile Regression is a regression analysis method that was first introduced by Koenker and Bassett (1978) [14]. Regression is used to quantify the relationship between a response variable and some covariate [15]. Quantile regression method uses a parameter approach by dividing the data into certain quantile groups after the data is sorted. The quantile regression method can be used to overcome the limitations of linear regression in overcoming violations of assumptions in OLS. According to [5] the quantile regression method is able to overcome violations of the normality assumption. This quantile regression method is a regression method with an approach of separating or dividing data into certain quantiles, by minimizing weighted absolute residuals that are not symmetrical and estimating conditional quantile functions on a data distribution. This approach estimates various quantile functions of a distribution Y as a function of X. Quantile regression is very suitable for asymmetric data patterns, tails in the distribution, or truncated distributions [16].

The quantile function is denoted by Q_τ with $0 < \tau < 1$. Suppose a random variable Y with a probability density function is $f(y)$ and a cumulative distribution function $F(Y) = P(Y \leq y)$, where for every $0 < \tau < 1$. The quantile - τ function of Y is given by [5]:

$$Q_\tau(Y) = F_Y^{-1}(\tau) = inv\{y: F_Y(y)|X \geq \tau \quad (1)$$

If the sample average is the solution to the problem:

$$min_{\mu\epsilon\mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2 \quad (2)$$

So $\mu(x_i) = x_i'\beta$ which is the conditional average of y with x known, the value of β can be estimated by solving:

$$min_{\beta\,\epsilon\mathbb{R}} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \quad (3)$$

Next the model develops into a sample median which is stated:

$$min_{\beta\epsilon\mathbb{R}} \sum_{i=1}^{n} |y_i - x_i'\beta| \quad (4)$$

In general, a linear regression equation model for quantile τ with n samples and k predictors for i=1,2, ..., n is written in the form:

$$y_i = \beta_{0\tau} + \beta_{1\tau}x_{i1} + \beta_{2\tau}x_{i2} + \cdots + \beta_{k\tau}x_{ik} + e_i \quad (5)$$

The above equation can be expressed in the following matrix form:

$$y_i = X_i'\beta + e_i \quad (6)$$

Where:
y_i: Vector of dependent variables of size *n x 1*
X : matrics variable of independent of size *n x k*
β : vector coefisien quantil regression of size *1 x k* which depends on quantile τ
e_i : vector of error of size *n x 1*

The conditional quantile function τ in the quantile regression method define $Q_\tau(y_i|x_i) = X_i'\beta$. With $Q_\tau(y_i|x_i) = X_i'\beta$ is the to quantile - τ $(0 < \tau < 1)$ from y with a value x_i certain. Based on the median concept, the estimate for β from the τ- *th* quantile regression is obtained by minimizing the number of absolute values of the errors with weighting τ for positive errors and weighting to $1 - \tau$, for negative error [4][17]:

For

$$min_{\beta\epsilon\mathbb{R}} \sum_{i\epsilon i|y_i \geq x_i'\beta} \tau|y_i - x_i'\beta| + min_{\beta\epsilon\mathbb{R}} \sum_{i\epsilon i|y_i \geq x_i'\beta} (1-\tau)|y_i - x_i'\beta| \quad ((7)$$

Or it can be written again as:

$$min_{\beta \epsilon \mathbb{R}} \sum_{i \epsilon i | y_i \geq x_i' \beta} \rho_\tau (y_i - Q_\tau (Y|X)) \quad ((8)$$

With :
τ denotes the quantile index $\epsilon(0,1)$
ρ_τ expresses an asymmetric loss function
$Q_\tau(Y|X) = x_i'\beta$ is quantile function to τ with the provision X

Data

This research uses generated data with the residuals not having a normal distribution but a chi square distribution, in accordance with the main objective of this research applying quantile regression to a group of data that does not meet the assumption of normality of the residuals. This research data consists of five independent variables ($X_1, X_2, , X_3, X_4$ dan X_5) and one dependent variable (Y). The independent variable (X_1, X_2 , X_3, X_4 dan X_5) is denoted as variabel stocastic. Variable independent $X_1, X_2 \sim EXP(\theta = 1)$, $X_3 \sim N(0,1)$, and $X_4, X_5 \sim t(v = 20)$. So the dependent variable can be written as:

$$Y = X_1 + X_2 + X_3 + X_4 + X_5 + e \quad (9)$$

With error $e \sim \chi^2(v = 20)$. Then a simulation was carried out using 100 generated data for each variable. To generate and analyze data, R 4.2.2 software was used.

Mean Square Error (MSE)

The goodness of fit of a model is used to determine the best model for modeling a data set. The best model is the model that has the smallest value of MSE among the other models at different quantile. MSE is value of determined by calculating the average of the squared residuals of the resulting regression model. The MSE value can be determined in the following equation [12]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (10)$$

Where (y_i) is the i th observation value, and $\hat{y}_i$ is the predicted value of the i th estimation result.

RESULT AND DISCUSSION

Estimated Parameters with regression quantile

The results of data analysis using the quantile regression method as previously explained. The table below displays the results of estimating model parameters, namely $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ dan β_5 using the generated data for the quantiles - τ =0.05, 0.25,

0.50, 0.75 dan 0.95. The choice of quantile values depends on the researcher, in some articles choosing quantiles as already mentioned.

Table 1. Estimated Value of β for each quantile - τ

Quantile - τ	Parameter	Estimated Value of β	p value	Standar Error (SE)	Mean Square Error (MSE)
0.05	β_0	0.0324	0.2053	0.0253	1.6992
	β_1	0.9959	0.0000	0.0155	
	β_2	0.9955	0.0000	0.0149	
	β_3	1.0000	0.0000	0.0166	
	β_4	1.0094	0.0000	0.0125	
	β_5	0.9908	0.0000	0.0126	
0.25	β_0	0.1415	0.0107	0.0543	1.5340
	β_1	0.9687	0.0000	0.0334	
	β_2	1.0110	0.0000	0.0321	
	β_3	1.0311	0.0000	0.0357	
	β_4	1.0215	0.0000	0.0268	
	β_5	0.9837	0.0000	0.0270	
0.5	β_0	0.2652	0.0013	0.0800	**1.2662**
	β_1	0.9400	0.0000	0.0492	
	β_2	1.1683	0.0000	0.0472	
	β_3	0.9925	0.0000	0.0527	
	β_4	1.0281	0.0000	0.0395	
	β_5	0.9102	0.0000	0.0398	
0.75	β_0	1.3157	0.0059	0.4677	1.2698
	β_1	0.9271	0.0017	0.2875	
	β_2	0.9568	0.0008	0.2759	
	β_3	0.9478	0.0027	0.3079	
	β_4	1.1953	0.0000	0.2310	
	β_5	0.7413	0.0019	0.2326	
0.90	β_0	2.4286	0.0005	0.6726	3.1106
	β_1	0.8430	0.0443	0.4134	
	β_2	0.9009	0.0255	0.3968	
	β_3	1.0964	0.0151	0.4427	
	β_4	1.1489	0.0008	0.3322	
	β_5	0.8085	0.0176	0.3345	

Table 1 shows that all regression coefficients values for all quantiles are significant with a 95% confidence interval (p-value < 0.05), while the β_0 is not taken into account. This means that in the estimated model obtained, the five independent variables X_1, X_2 , X_3, X_4 and X_5 significantly affects the dependent variable (Y) at all selected quantile values. Table 1 also informs that the model coefficient values $\beta_1, \beta_2, \beta_3, \beta_4$ and β_5 at the 0.05 quantile are 0.9959, 09955, 1.0000, 1.0094, 0.9908. All five

values are very close to 1, even the value of is β_3 is equal to 1. The same results are also obtained at the 0.25, 0.5, 0.75, 0.90 quantiles.

The MSE values for each quantile can also be found in table 1. This quantile regression method produces various MSE values in each quantile. It is known that the best model can be seen from the smallest MSE value [5]. The smallest MSE value is found in quantile 0.5 which is 1.2662 compared to otherquantile.

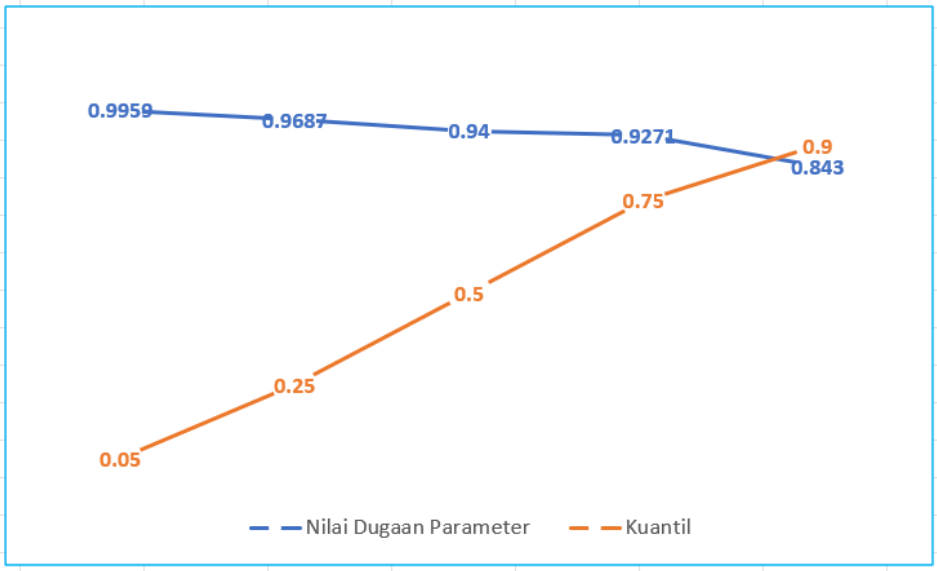

Figure 1. coefficient Variable (β_1)

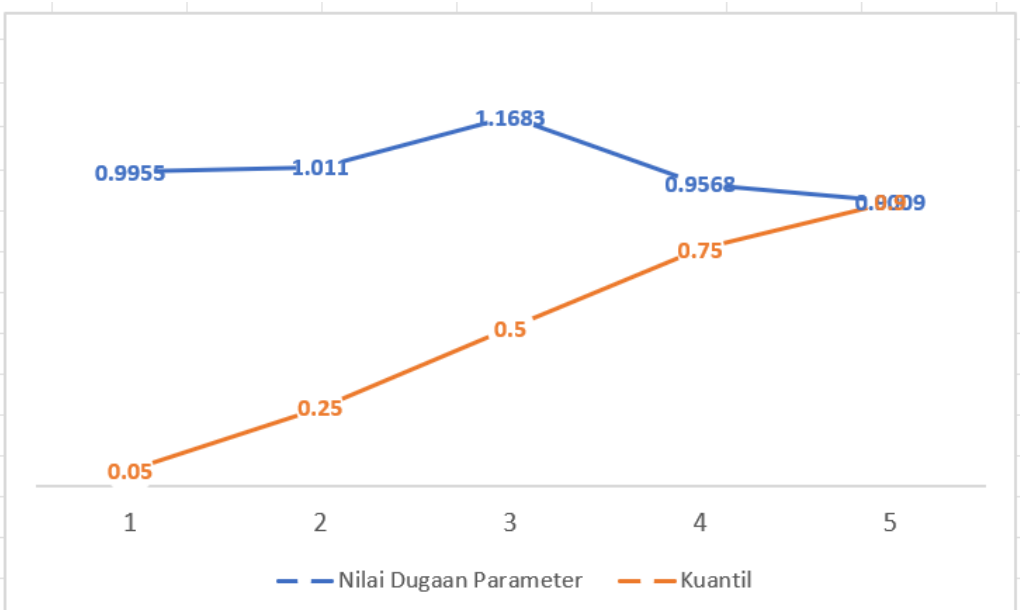

Figure 2. coefficient Variable (β_2)

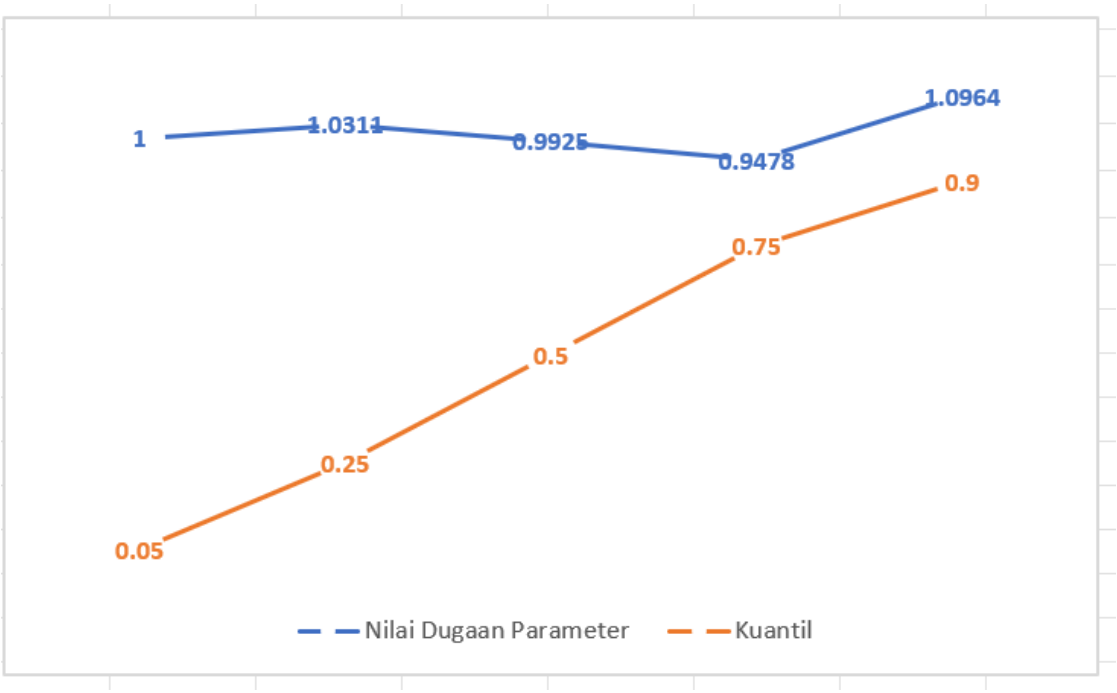

Figure 3. coefficient Variable (β_3)

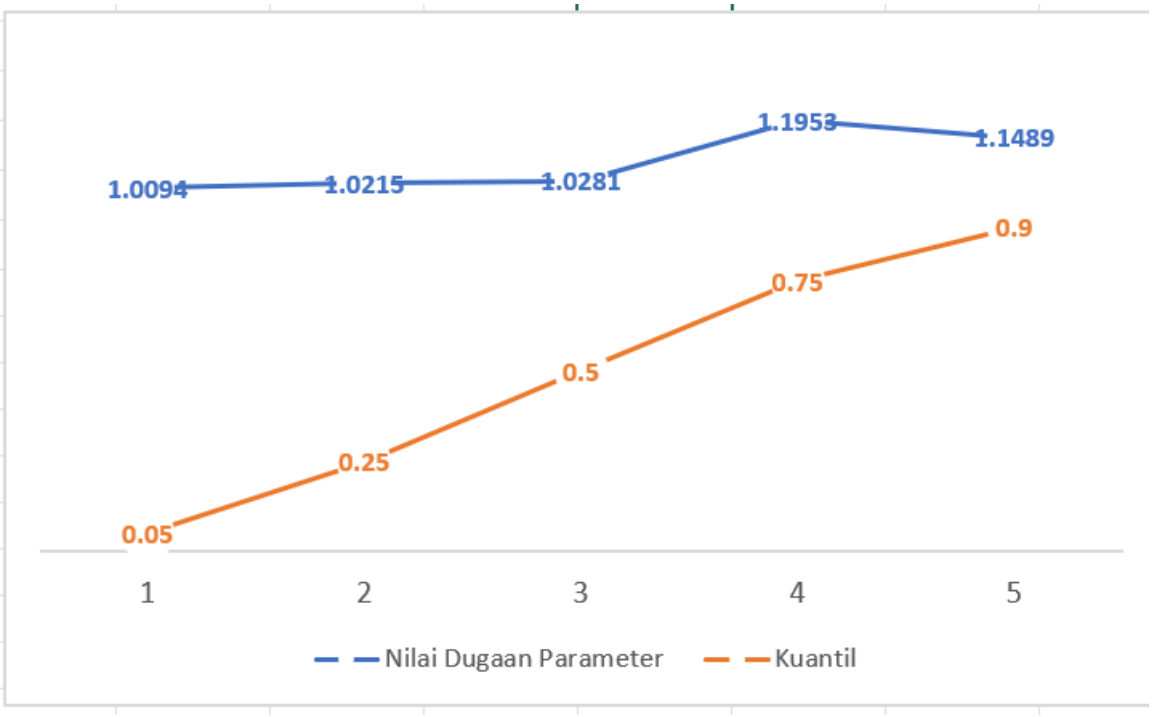

Figure 4. coefficient Variable (β_4)

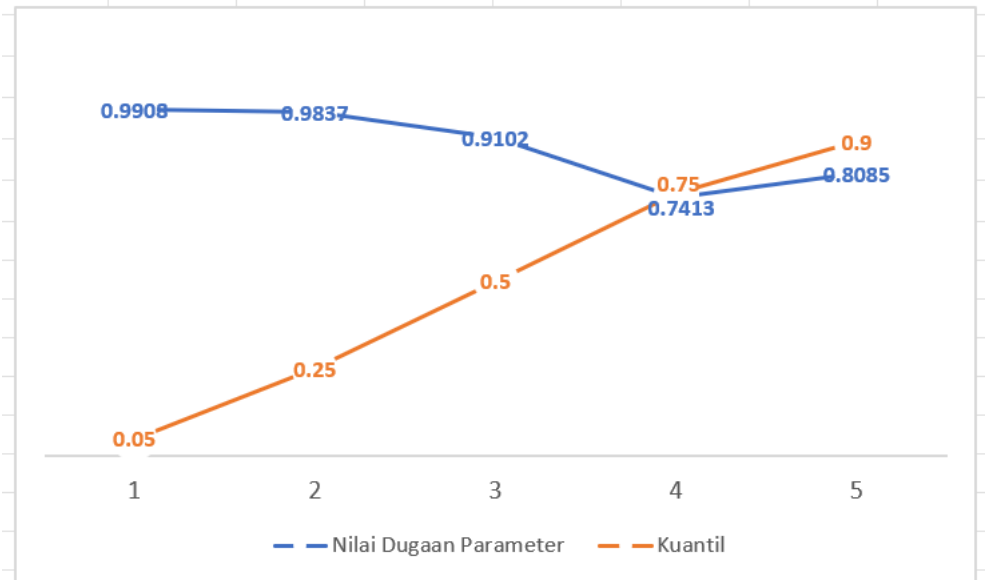

Figure 5. coefficient Variable (β_5)

The MSE values for each quantile can also be found in Table 1. This quantile regression method produces various MSE values in each quantile. It is known that the best model can be seen from the smallest MSE value [5]. The smallest MSE value is found in quantile 0.5, which is 1.2662 compared to other quantiles. The movement of changes in MSE values at each quantile can be seen in the figure below:

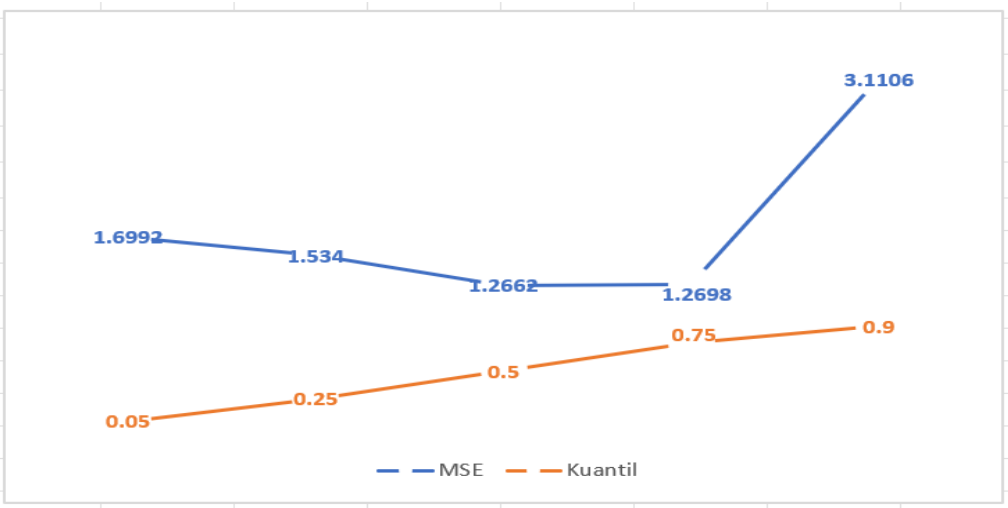

Figure 6. Trend of MSE value for each quantile

CONCLUSION

This paper aims to show analysis using quantile regression applied to non-normal errors. The nature of quantile regression is to divide the sorted data into certain quantiles, the size of the quantiles depends on the researcher. The approach used is to minimize the asymmetric weighted absolute residual and estimate the conditional quantile function on the data distribution. This research uses simulated data with diverse distributions of independent variables with residuals assumed to have a chi square distribution. This research produces several prediction models which are divided into selected quantiles. The resulting regression coefficient at each selected

quantile is almost close to the initial value set. For the selection the best model, it is from the smallest MSE value. The smallest MSE value is found in quantile 0.5. So, the best model from this research can be written as $\hat{y} = 0.2653 + 0.9400X_1 + 1.1683X_2 + 0.9925X_3 + 1.0281X_4 + 0.9102X_5$.

REFERENCES

[1] L. H. Hasibuan and S. Musthofa, "Penerapan Metode Regresi Linear Sederhana Untuk Prediksi Harga Beras di Kota Padang," *JOSTECH J. Sci. Technol.*, vol. 2, no. 1, pp. 85–95, 2022.

[2] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[3] F. Yanuar, H. Yozza, and I. Rahmi, "Penerapan Metode Regresi Kuantil pada Kasus Pelanggaran Asumsi Kenormalan Sisaan," *Eksakta*, vol. 1, pp. 33–37, 2016.

[4] F. Yanuar, "Quantile Regression Approach to Determine the Indicator of Health Status," *Sci. Res. Journal, I*, pp. 17–23, 2013.

[5] C. Davino, M. Furno, and D. Vistocco, *Quantile regression: theory and applications*, vol. 988. John Wiley & Sons, 2013.

[6] F. Yanuar, L. Hasnah, and D. Devianto, "The Simulation Study to Test the Performance of Quantile Regression Method With Heteroscedastic Error Variance," vol. 5, no. May, pp. 36–41, 2017.

[7] R. Koenker, "Quantile regression for longitudinal data," *J. Multivar. Anal.*, vol. 91, no. 1, pp. 74–89, 2004.

[8] H. Kozumi and G. Kobayashi, "Gibbs sampling methods for Bayesian quantile regression," *J. Stat. Comput. Simul.*, vol. 81, no. 11, pp. 1565–1578, 2011.

[9] F. Yanuar, "The use of Uniformative and informative prior distribution in Bayesian SEM," *Glob. J. Pure Appl. Math.*, vol. 11, no. 5, pp. 3259–3264, 2015.

[10] M. C. Korkmaz, E. Altun, C. Chesneau, and H. M. Yousof, "On the unit-Chen distribution with associated quantile regression and applications," *Math. Slovaca*, vol. 72, no. 3, pp. 765–786, 2022.

[11] F. Yanuar, C. Mukti, U. Andalas, and K. L. Manis, "Komparasi Model Pertambahan Tinggi Badan Balita Stunting Dengan Metode Regresi Kuantil dan Regresi Kuantil Bayesian," vol. 20, no. 2, pp. 165–177, 2023.

[12] Y. Wu and Y. Liu, "Variable selection in quantile regression," *Stat. Sin.*, pp. 801–817, 2009.

[13] I. A. P. Uthami, I. K. G. Sukarsa, and I. P. E. Kencana, "Regresi Kuantil Median untuk Mengatasi Heteroskedastisitas pada Analisis Regresi," *E-Jurnal Mat.*, vol. 2, no. 1, pp. 6–13, 2013.

[14] R. Koenker, V. Chernozhukov, X. He, and L. Peng, "Handbook of quantile regression," 2017.

[15] K. Yu, Z. Lu, and J. Stander, "Quantile regression: applications and current research areas," *J. R. Stat. Soc. Ser. D Stat.*, vol. 52, no. 3, pp. 331–350, 2003.

[16] D. N. Gujarati, *Essentials of econometrics*. Sage Publications, 2021.

[17] S. Bentzien and P. Friederichs, "Decomposition and graphical portrayal of the quantile score," *Q. J. R. Meteorol. Soc.*, vol. 140, no. 683, pp. 1924–1934, 2014.