International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.
© 2022 Scholar AI LLC.

Vol. 33 No. 2 July 2022, pp. 178-187

Bayesian Tobit Quantile Regression Modeling In The Case of Length Hospital Stay of COVID-19 Patients

Chyntia Dwi Yan, Ferra Yanuar*, Dodi Devianto

Mathematics Department,
Faculty of Mathematics and Natural Sciences, Andalas University,
Campus of UNAND Limau Manis Padang-25163, Indonesia
*ferrayanuar@sci.unand.ac.id

***Abstract*— COVID-19 *(Coronavirus Disease 2019)* is an acute respiratory syndrome infectious disease caused by SARS-CoV-2. The high number of positive cases of COVID-19 in West Sumatra resulted in many patients undergoing isolation and treatment in hospitals. The length of stay of patients varies for each patient because it is triggered by several factors. Data on length of stay for COVID-19 patients is a type of censored data. Therefore, this study aims to model censored data and identify factors that influence the length of stay of COVID-19 patients using the tobit quantile regression method and the Bayesian tobit quantile regression method. This study will also evaluate the goodness of the model using the RMSE and *PseudoR*^{2}model goodness evaluation methods. The results showed that the Bayesian tobit quantile regression method was a better method in estimating the parameters of the model of length of stay for COVID-19 patients. Meanwhile, it was found that the age of the patient, the diagnosis of the patient in the Positive category and the number of comorbidities had a significant influence on the length of stay of COVID-19 patients.**

Keywords—Censored Data; Tobit Quantile Regression; Bayesian Tobit Quantile Regression; COVID-19

I. INTRODUCTION

In regression modeling, data with dependent variable which has no value in some of its observations is often encountered. This data is known as censored data. In some parts of the observation value, It is also often encountered that the data have certain values that vary or are below a certain threshold (sensor point) (left censored data), or above a certain threshold (right censored data) [4]. If the censored data is modeled using linear regression, it will experience bias condition, or allow the possibility of autocorrelation, heteroscedasticity and other problems [7].

Tobit regression method (censored regression) is an approach for modeling censored data. The use of tobit regression will reduce the effect of bias because data that is assumed to be constant can be processed simultaneously with continuous data so that there will be no loss of information from discrete data [6]. Furthermore, the tobit quantile regression method was developed to overcome the problem of the error distribution such as heteroscedasticity, abnormal and asymmetric data [13]. Then, the tobit quantile regression method was hybridized with the Bayesian method. Bayes introduces a method for estimating parameters that utilizes initial information called the prior distribution. This method is known as the Bayesian method. The Bayesian method uses the MCMC (Markov Chain Monte Carlo) algorithm to estimate the posterior distribution of parameters that have complex formulations that are difficult to estimate analytically the posterior mean and posterior variance.

COVID-19 (Coronavirus Disease 2019) is an acute respiratory syndrome infectious disease caused by SARS-CoV-2. This disease initially broke out in Wuhan, China in December 2019 and has now spread throughout the world [20]. According to West Sumatra Provincial Health Office data as of November 30, 2020, there were 20,036 positive cases of COVID-19 infection in West Sumatra [12]. The high number of positive cases infected with COVID-19 in West Sumatra has resulted in many COVID-19 patients undergoing isolation and treatment at home and in hospitals. The length of stay of patients varies for each patient because it is triggered by several factors [12]. Therefore, it is necessary to study the factors that influence the length of stay of COVID-19 patients in the hospital. Modeling the length of stay is important for the hospital to provide the number of beds, the provision of medical services, staff and other medical equipment in the hospital.

The problem raised in this study is how to model the censored data contained in the length of hospitalization stay for COVID-19 patients and identify the factors that affect the length of stay of COVID-19 patients using the tobit quantile regression method and the Bayesian tobit quantile regression method. This study will also evaluate the goodness of the model generated by the tobit quantile regression method and the Bayesian tobit quantile regression method using the RMSE and PseudoR^2 model goodness evaluation methods.

II. Review of Related Literature

2.1 Tobit Regression

Suppose there are n observational data consisting of one dependent variable Y, m independent variables $X_1, X_2, \dots, X_m$ and a threshold (sensor point) p, then the dependent variable $Y = y_1, y_2, \dots, y_n$ is said to be left censored if for every $i = 1,2,\dots,n$ the equation is:

$$Y_i = \begin{cases} p\,, y_i^* & \leq p \quad (censored\ data) \\ y_i^*\,, y_i^* & > p\ (uncensored\ data) \end{cases}$$

The observation value of y_i^* can be expressed as follows.

$$y_i^* = f(X_i) + e_i, i = 1,2,\dots,n,$$

in which $f(X_i) = X_i^T \beta$

Thus, the Tobit regression model is obtained as:

$$Y_i = \begin{cases} p & , y_i^* \leq p \quad (censored\ data) \\ X_i^T \beta + e_i & , y_i^* > p\ (uncensored\ data) \end{cases}$$

2.2 Quantile Tobit Regression Method

Tobit regression method is commonly used to model censored data. However, because the marginal effect may be different at the lower or higher quantile conditions in comparison with the conditional mean, the Tobit regression model is less accurate to use. To overcome this deficiency, the Tobit quantile regression method is used as a parameter estimator in the problem of heteroscedasticity and abnormal error distribution problems [13]. Tobit quantile regression method begins with formulating a model which is stated as follows:

$$y_i^* = X_i^T \beta + e_i, i = 1,2,\dots,n,$$

In general, parameter estimates can be obtained by minimizing the number of squares of errors as follows:

$$\min_{\beta \in R} \sum_{i=1}^{n} (y_i - \max(p, X_i^T \beta))^2$$

Suppose we will estimate β at a certain θ. Thus, there are θ percent of y in which value is less than or equal to max $(p, X_i^T\beta)$ and there is another $1-\theta$ percent that has a value greater than or equal to max $(p, X_i^T\beta)$. This concept is used as the basis for solving the following problem:

$$\min_{\beta \in R} \left[\sum_{i=1}^{n} \theta \,|y_i - \max(p, X_i^T\beta\ (\theta))| + \sum_{i=1}^{n}(1-\theta)|y_i - \max(p, X_i^T\beta(\theta))|\right] \qquad (1)$$

The general estimate for β for the θ^{th} quantile containing censored observations based on equation (1) can be written as follows:

$$\hat{\beta}(\theta) = \min_{\beta \in R} \sum_{i=1}^{n} \rho\theta\left(y_i - \max\left(p, X_i^T\beta(\theta)\right)\right).$$

2.3 The Bayesian Method

The Bayesian method is a parameter estimation method based on the Bayes theorem. Bayes introduced a method by which we need to obtain information about the distribution of the parameters to be estimated which is known as the prior distribution. This prior distribution will be used together with the likelihood function to determine the posterior distribution [18].

Posterior Distribution

The posterior distribution is the probability density function of the random variable β if it is known that the observed value is *x*. It can be written as follows [5]

$$f(\beta|x) \propto f(x|\beta)f(\beta)$$

Likelihood function

Definition 2.1. [2] *The joint probability density function of n random variables $X_1, X_2, \dots, X_n$ which is assigned the value $x_1, x_2, \dots, x_n$ is denoted by $f(x_1, x_2, \dots, x_n; \beta)$ which is likelihood function. For certain values of $x_1, x_2, \dots, x_n$, the likelihood function is a function of the parameter β . If $X_1, X_2, \dots, X_n$ are independent random samples, then*

$$\begin{aligned} L(x_i;\beta) &= f(x_1,\beta)f(x_2,\beta)\dots f(x_n,\beta) \\ &= \prod_{i=1}^{n} f(x_i,\beta) \\ &= L(\beta) \end{aligned}$$

2.4 Bayesian Quantile Tobit Regression Method

Bayesian tobit quantile regression method is a combination of Tobit quantile regression method and Bayesian method. The Bayesian method uses *Markov Chain Monte Carlo (MCMC)* to estimate the posterior distribution. The Bayesian approach assumes that y follows the *Asymmetric Laplace Distribution (ALD)*, so that the random variable $y \sim ALD(\theta;\mu;\sigma)$ with the addition of the location parameter $\mu$and the scale parameter $\sigma > 0$, the probability density function of ALD is given as follows: [1], [5].

$$f_\theta(y;\mu,\sigma) = \frac{\theta(1-\theta)}{\sigma} exp\left(-\rho_\theta\left(\frac{y-\mu}{\sigma}\right)\right) ,$$

in which ρ_θ is a loss function defined as follows:

$$\rho_\theta(\varepsilon) = \varepsilon\big(\theta - 1(\epsilon < 0)\big)$$

Bayesian tobit quantile regression method begins by formulating a model which is stated as follows:

$$y_i^* = X_i^T\beta + e_i, i = 1, 2, \dots, n,$$

The general estimate for β for the θth quantile containing the censored observations can be written as:

$$\hat{\beta}(\theta) = \min_{\beta \in R} \sum_{i=1}^{n} \rho\theta \left(y_i - \max\left(p, x_i^T \beta(\theta)\right)\right).$$

Koenker and Machado [10] found that minimizing the loss function of the tobit quantile regression is equivalent to maximizing the likelihood function formed from the data assumed to have an ALD distribution. The Bayesian approach assumes that y follows the *Asymmetric Laplace Distribution (ALD),* this method is very helpful in the next parameter estimation process in the Bayesian method. By assuming that y has an ALD distribution, it will be easy to construct the likelihood function [11], but because the formula is not simple, it is easier to solve using a numerical approach [19].

2.5 Root Mean Square Error (RMSE)

To evaluate the goodness of the model that is often used is the Root Mean Square Error (RMSE). The smaller the residual, the better the model [3].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$$

2.6 Factors Affecting Length of Hospitalization for COVID-19 Patients

In general, people who have mild, moderate to severe symptoms of COVID-19 infection must seek treatment and services in hospitals [9]. COVID-19 patients have different lengths of hospitalization depending on the conditions and symptoms of each patient [17]. The following is an explanation of the factors that are assumed to affect the length of stay of COVID-19 patients:

1. Age

   Age is one of the factors that influence a person to be infected with COVID-19. According to the latest national data from KPC PEN, it shows that those aged 31-45 years occupy the first position for positive diagnoses of COVID-19 [16]. The research [?] states that for age ≥ 60 have a longer length of stay and a fairly high mortality rate among other hospitalized patients.

2. Gender

   In general, males are more susceptible to COVID-19 infection than females [14]. There are several factors that make males more susceptible to being infected with COVID-19, one of which is because males are more active outside of the house for working than females [14].

3. Comorbid

   The number of comorbids for each person is different, the presence or absence of comorbidities can affect the length of time when the patient is hospitalized/isolated [8].

4. Diagnosis

   With several symptoms of COVID-19 infection, COVID-19 patients are distinguished based on the following categories: People Without Symptoms (OTG), People Under Monitoring (ODP), Patients Under Supervision (PDP), and Positive [9]. So that, COVID-19 patients with different diagnoses will have different lengths of hospitalization [15].

III. Methodology of the research

3.1 Research Data

The data used in this study is data on COVID-19 patients at the Central General Hospital (RSUP) Dr. M. Djamil Padang and

Andalas University Hospital (UNAND Hospital) from March 2020 to November 2020. The variables in this study can be divided into two types, namely dependent and independent variables as follows.

1. Independent Variable (Y), namely the length of hospitalization for COVID-19 patients in days.

   $y_i = 0$ if there is no data on the length of hospitalization for COVID-19 patients

   $y_i = y_i^*$ if there is data on the length of hospitalization for COVID-19 patients

2. Independent variables, there are five independent variables used in, such as:

   a. X_1 = Patient's age in years

   b. X_2 = Gender of the patient

   $$(X_{2D1}) = \begin{cases} 1, & laki-laki, \\ 0 & lainnya \end{cases}$$

   The category of women as a comparison category.

   c. X_3 = Patient diagnosis related to the status of being infected with COVID-19 with the categories of OTG, ODP, PDP, and Positive COVID-19.

   $$(X_{2D1}) = \begin{cases} 1 & ,ODP, \\ 0 & lainnya \end{cases}$$

   $$(X_{2D2}) = \begin{cases} 1 & ,PDP, \\ 0 & lainnya \end{cases}$$

   $$(X_{2D3}) = \begin{cases} 1 & ,Positif, \\ 0 & lainnya \end{cases}$$

   OTG category as comparison category.

   d. X_4 = The number of comorbids or the number of comorbidities suffered by the patient.

IV. RESULT AND DISCUSSION

4.1 Exploration of Research Data

The length of stay for COVID-19 patients is a type of censored data. The number of censored and uncensored data on data on length of stay of COVID-19 patients is presented in Figure 1

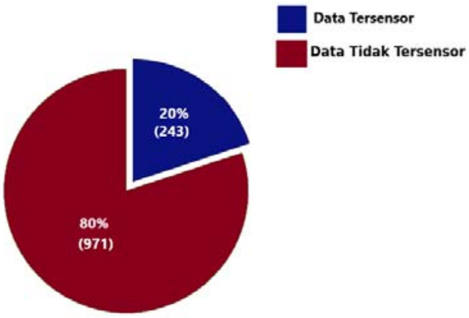

Figure 1: Pie Diagram of Percentage of Data on Length of Hospitalization for COVID-19 Patients in Hospitals

Table 1 presents descriptive statistics of independent variables on a categorical scale, namely Gender (X_2) and Diagnosis (X_3). Meanwhile, the descriptive statistics for the independent variables Age (X_1) and Comorbid (X_4) which have a numerical scale are presented in Table 2:

Table 1: Descriptive Statistics of Independent Variables Category Scale

Variable	Category	Frequency	Percentage (%)
Gender (X_2)	Male	620	51,1
	Female	594	48,9
Diagnoses (X_3)	OTG	1	0,1
	ODP	6	0,5
	PDP	912	75,1
	Positive	295	24,3

Table 2: Descriptive Statistics of Independent Variables Numerical Scale

Variable	Minimum	Maximum	Average	Standard Deviation
Age (X_1)	0	88	38,90	23,697
Comorbid (X_4)	0	10	1,96	1,724

4.2 Parameter Estimation of Quantile Tobit Regression Model

Table 3: Estimation Results of Model Parameters through Quantile Tobit Regression Method

Independent Variables	**Estimation of the n - th quantile parameter**			
	0,25	**0,5**	**0,75**	**0,9**
Age (X_1)	-0,230	-0,184	**-0,0175***	**-0,0439***
Diagnoses (X_3)				
ODP (X_{3D1})	0,600	2,000	6,8772	8,1491
PDP (X_{3D2})	**1,200***	3,000	4,1228	5,7281
Positive (X_{3D3})	**4,000***	**9,000***	**13,6316***	**19,2807***
Number of Comorbids	**-0,400***	-0,0112	0,0702	0,3596
***level of significance $\alpha = 0,05$**				

Acoording to the data in Table 3, it can be seen that the independent variables Number of Comorbids (X_4), Age (X_1), and diagnoses for the PDP (X_{3D2}) and Positive (X_{3D3}) categories were significant at a significant level of $\alpha = 0{,}05$ in influencing the length of hospitalization stay of COVID-19 patients. After the tobit quantile regression modeling is obtained, it is necessary to test how accurate the resulting model is by calculating the PseudoR^2 value. The results of the calculation of the PseudoR^2 value for each quantile can be seen in Table 4. In table 4, the model generated at quantile 0.90 gives the largest PseudoR^2 value among the other selected quantiles. Thus, it can be said that the model produced at quantile 0.90 is the best model. So that the estimated data regression model for the length of stay of COVID-19 patients at quantile 0.90 is obtained as follows:

$$\hat{y}_{(0,90)} = -0{,}0439X_1 + 19{,}2807X_{3D3}$$

Table 4: Quantile Tobit Regression PseudoR^2 Value

The *n-th* Quantile	**PseudoR^2**
0,25	0,2714
0,5	0,4631
0,75	0,6459
0,9	**0,7614**

4.3 Parameter Estimation of Bayesian Quantile Tobit Regression Model

The estimation results of the Bayesian tobit quantile regression method performed with the help of software R with 30000 iterations and burn-in of 100 can be seen in the table below.

Table 5: Estimation Results of Model Parameters through Bayesian Quantile Tobit Regression Method

Independent Variables	**Estimation of the *n* - th quantile parameter**			
	0,25	**0,5**	**0,75**	**0,9**
Age (X_1)	-0,0174	-0,0256	**-0,0567***	**-0,0477***
Diagnoses (X_3)				
ODP (X_{3D1})	-1,0501	-0,5460	0,4899	0,8934
PDP (X_{3D2})	**-7,5369***	**-6,3606***	-3,3569	-1,7147
Positive (X_{3D3})	3,1791	**4,5731***	**8,3423***	**11,7880***
Number of Comorbids (X_4)	**-0,2090***	**-0,0179***	0,4427	**0,4767***

Based on Table 5 it can be seen that the independent variables Age (X_1), PDP patient diagnosis (X_{3D2}) and Positive (X_{3D3}) and the number of comorbidities (X_4) were significant at a significant level $\alpha = 0{,}05$ in influencing the length of stay of COVID-19 patients in each different quantile. Next, the estimation of the accuracy of the resulting model will be calculated by calculating the PseudoR^2 value.

Table 6: Bayesian Quantile Tobit Regression PseudoR^2 Value

The *n-th* Quantile	**PseudoR^2**
0,25	0,5875
0,5	0,9761
0,75	0,9869
0,9	0,9669

It can be seen in Table 6 that the 0.75 quantile is the best predictor model because it produces the highest PseudoR2 value. Furthermore, the convergence test of the model parameters was carried out at the 0.75 quantile.

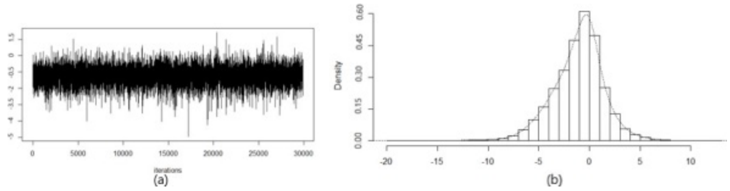

Figure 2: (a)Trace-Plot, (b) Density-Plot on the quantile $\theta = 0{,}75$ for the parameter of the variable Age (X_1).

In Figure 2 (a) it can be seen that the trace plot has formed a pattern that converges to a value so that the Age parameter has converged to the 0.75 quantile using the Bayesian quantile tobit regression method. In Figure 1 (b) it can be seen that the density plot for the Age parameter generated at the 0.75 quantile has formed a curve resembling a normal distribution curve. Therefore, based on the convergence test of these parameters, it can be concluded that the estimated value for the parameter of the Age variable has met the convergence criteria so that the estimated value obtained is acceptable.

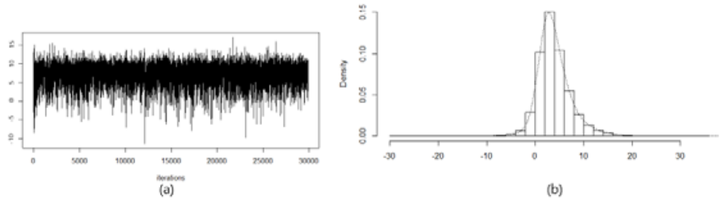

Figure 3: (a)Trace-Plot, (b) Density-Plot on the quantile $\theta = 0{,}75$ for the parameter of the variable Positive (X_{3D3}).

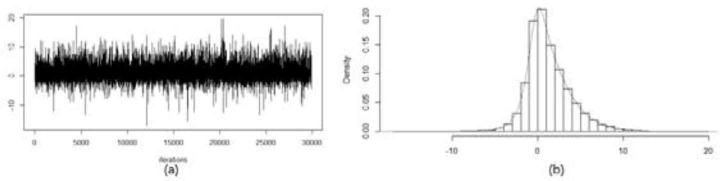

Figure 4: (a)Trace-Plot, (b) Density-Plot on the quantile $\theta = 0{,}75$ for the parameter of the variable Comorbid (X_4).

Based on the convergence test of the model parameters, it can be concluded that the estimates of the three parameters have converged and the estimated model at the 0.75th quantile is acceptable. The estimated model for the 0.75 quantile obtained is:

$$\hat{y}_{(0{,}75)} = -0{,}0567X_1 + 8{,}3423X_{3D3} + 0{,}4427X_4$$

4.4 Comparison of Quantile Tobit Regression Analysis and Bayesian Quantile Tobit Regression

Comparison of the results of the tobit quantile regression method and the Bayesian tobit quantile regression method was carried out to determine the best method in producing a model with the largest PseudoR^2 value and the smallest RMSE value. The results of the comparison are presented in Table 7 below.

Table 7: Comparison of PseudoR^2 Values for Quantile Tobit Regression through Bayesian Tobit Regression

The *n-th* Quartile	**PseudoR^2**	
	Quantile Tobit Regression	**Bayesian Quantile Tobit Regression**
0,25	0,2714	**0,5875**
0,50	0,4631	**0,9761**
0,75	0,6459	**0,9869**
0,90	0,7614	**0,9669**

Furthermore, the results of the comparison of the RMSE value the tobit quantile regression method and the Bayesian tobit quantile regression method are presented in Table 8.

Table 8: RMSE Comparison Results of Quantile Tobit Regression and Bayesian Quantile Tobit Regression

The *n-th* Quartile	**RMSE**	
	Quantile Tobit Regression	**Bayesian Quantile Tobit Regression**
0,25	15,0476	**14,1789**
0,50	12,3968	**6,5291**
0,75	8,4007	**5,4628**
0,90	**6,1676**	7,6066

Overall, it was found that the tobit quantile Bayesian regression method tends to produce a predictive model with better model goodness than the tobit quantile regression method because it has a smaller RMSE value and a PseudoR2 value. the greater one.

V. Conclusion

After analyzing the data on the length of stay of COVID-19 patients in West Sumatra Province, the following conclusions were obtained:

1) By using the tobit quantile regression method, it was found that the patient's age (X_1) and the diagnosis of the patient with the Positive category (X_{3D3}) had an influence on the length of stay of COVID-19 patients. Meanwhile, using the Bayesian tobit quantile regression method, it was found that the patient's age (X_1), the diagnosis of the patient in the Positive category (X_{3D3}) and the number of comorbidities (X_4) had an influence on the length of stay of COVID-19 patients.

2) The results of the comparison of the estimation of model parameters based on the PseudoR^2 value and the RMSE value showed that the Bayesian tobit quantile regression method is a better method in estimating the parameters of the length of stay for COVID-19 patients in West Sumatra. Bayesian tobit quantile regression method produces a larger PseudoR^2 value and a smaller RMSE value.

REFERENCES

[1] Alhamzawi, R., dan Yu, K. 2012. Variable selection in quantile regression via Gibbs Sampling. *Journal of Applied Statistics*, 39(4), 799-813.

[2] Bain, L.J. dan Engelhardt, M. 1992. *Introduction to Probability and Mathematical Statistics*. United States of America: Brooks/Cole.

[3] Chai, T., dan Draxler, R. R. 2014: Root Mean Square Error (RMSE) or Mean Absolute Error (MAE): Arguments Against Avoiding RMSE in The Literature. *Geoscientific Model Development*, 7, 1247-1250.

[4] Davino, C., Furno, M. dan Vistocco, D. 2014. *Quantile Regression Theory and Applications*. John Wiley dan Sons, Ltd.

[5] Feng, Y., Chen, Y., dan He, X. 2015. Bayesian quantile regression with approximate likelihood. *Bernoulli*, 21(2), 832-850.

[6] Greene, W.H. 2008. *Econometrics Analysis, 6th edition*. Prentice Hall, New Jersey.

[7] Gujarati, D.N. 1995. *Basic Econometrics Third edition*. McGraw-Hill International Editions, Economic Series

[8] Handayani, Diah. dkk. 2020. Penyakit Virus Corona 2019. *Jurnal Respirologi Indonesia*, 40(2), 119-129.

[9] Keputusan Menteri Kesehatan Republik Indonesia Nomor HK.01.07/MENKES/413/2020 tentang Pedoman Pencegahan dan Pengendalian Corona Virus Disease 2019 (COVID-19).

[10] Koenker, R., dan Machado, dan Jose A.F. 1999. Goodness of Fit and Related inference Processes for Quatile Regression. *Journal of The American Statistical Association*. 94, 1296-1310.

[11] Kozumi, H. dan G. Kobayashi. 2011. Gibbs Sampling Methods for Bayesian Quantile Regression. *Journal of Statistical Computation and Simulation*, 81, 1565-1578.

[12] Portal Resmi Provinsi Sumatera Barat. 2020. Informasi Covid-19 Provinsi Sumatera Barat. https://sumbarprov.go.id/home/news. Diakses tanggal 10 Mei 2021.

[13] Powell, J. 1986. Censored Regression Quantiles. *Journal of Econometrics*, 32, 143-155.

[14] Sari, A. R., Rahman F., Wulandari A., dkk. 2020. Perilaku Pencegahan Covid-19 Ditinjau dari Karakteristik Individu dan Sikap Masyarakat. *Jurnal Penelitian dan Pengembangan Kesehatan Masyarakat Indonesia*. 32-37.

[15] Satuan Pusat Penanganan COVID-19. 2020. Peta Sebaran. https://covid19.go.id/peta-sebaran. Diakses pada 10 Mei 2021.

[16] Souza, F dkk. 2021. On The Analysis of Mortality Risk Factors for Hospitalized COVID-19 Patients : A Data-Driven Study using The Major Brazilian Database. *PLoS ONE*, 16, 1-21.

[17] World Health Organization. 2020. Coronavirus. www.who.int. Diakses pada 10 Mei 2021.

[18] Yanuar, F., Saputri, C., dan Devianto, D. 2020. Bayesian Inference for Pareto Distribution with Prior Conjugate and Prior Non Congjugate. *Jurnnal Matematika dan Komputasi* . 16(3), 382-390.

[19] Yu, K., dan Stander, J. 2007. Bayesian Analysis of Tobit Quantile Regression Model. *Journal of Econometrics*, Vol. 137, 260-276.

[20] Zhai, P., dkk. 2020. The Epidemiology, Diagnosis, and Treatment of COVID-19. *Elsevier B.V. and International Society of Chemotherapy*, 55, 1-12.