Science and Technology Indonesia
e-ISSN:2580-4391 *p*-ISSN:2580-4405
Vol. 5, No. 3, July 2020

Research Paper

Quantile Regression Approach To Model Censored Data

Sarmada[1], Fera Yanuar[1]*

[1] *Department of Mathematics, Faculty of Mathematics and Natural Sciences, Andalas University, Kampus Limau Manis, 25163, Padang - Indonesia*
*__Corresponding author__: ferrayanuar@sci.unand.ac.id

Abstract
The censored quantile regression model is derived from the censored model. This method is used to overcome problems in modeling censored data as well as to overcome the assumptions of linear models that are not met. The purpose of this study is to compare the results of the analysis of the quantile regression method with the censored quantile regression method for censored data. Both methods were applied to generated data of 150, 500, and 3000 sample size. The best model is then chosen based on the smallest absolute bias and the smallest standard error as an indicator of the goodness of the model. This study proves that the censored quantile regression method tends to produce smaller absolute bias and a smaller standard error than the quantile regression method for all three group data specified. Thus it can be concluded that the censored quantile regression method could result in an acceptable model for censored data.

Keywords
Censored data, quantile regression, censored quantile regression, standard error, absolute bias

Received: 20 May 2020, Accepted: 11 July 2020
https://doi.org/10.26554/sti.2020.5.3.79-84

1. INTRODUCTION

A good parameter estimator is a BLUE (Best Linear Unbias Estimator). A BLUE will be obtained if it meets all the assumptions of the linear model. The Ordinary Least Squares (OLS), as an estimator method that is often used, is not always a BLUE estimator. In real events, there are some cases where the assumptions of normality and homoscedasticity are not met when there is zero value data or outlier data. In such cases, the OLS cannot be used (Delviyanti et al., 2018).

Then the quantile method appears as one of the parameter estimation methods that can overcome unfulfilled assumptions such as normality and homoscedasticity assumptions. This method uses a parameter estimation approach by separating or dividing data into quantiles, namely by using conditional quantile functions on data distribution and minimizing the asymmetrical absolute weighted remainder.

The next problem has often encountered data where the independent variable is available, while the value of the response variable y is below c (left sensor), and the value of y is above c (right sensor). Constant c is the threshold (sensor point). In many cases, the sensor threshold is zero. Multiple censorships are also possible if the value of the response variable is only available for observations where $c < y < d$, c and d become two thresholds (Davino et al., 2013). The parameter estimation method used to overcome problems with censored data and the BLUE assumptions that are not met is the censored quantile regression method.

Several studies that have been carried out relating to censored quantile regression include Gustavsen and Rickertsen (2013) approaching censored quantile regression in research on adjusting Value Added Tax (VAT) rates to promote healthy diets in Norway. The study of this important and interesting, Buchinsky (1994) and Fitzenberger et al. (2002) model the conditional wage distribution using censored quantile regression. Machado et al. (2006) and Fitzenberger and Wilke (2006) consider using CQR (censored Regression Quantile) for the duration analysis. Quantile regression with censoring and endogeneity has been studied by (Chernozhukov et al., 2015). The censored regression model has also been carried out by Leiker (2012) who studies the comparison of estimators of censored regression methods with Maximum Likelihood Estimation (MLE).

This study aims to illustrate the comparison between classical quantile regression method and censored quantile method to the censored data where normality assumption is violated. The indicators for the best method are based on the lowest of absolute bias value and a standard error value. This study used three sets of generated data of size 150, 500 and 3000. Such of this study never been done by any researchers before.

2. EXPERIMENTAL SECTION

2.1 Materials

2.1.1 Quatile Regression

Regression analysis is a tool in the development of statistics used in many fields of life (Muharisa et al., 2018). Quantile regression introduced by Koenker and Bassett Jr (1978). Quantile regression is the development of statistical tools used to explain the relationship between response and predictor variables (Yanuar et al., 2019a). The quantile method is a technique of dividing a group of data into several parts after the data is sorted from the smallest to the largest Yanuar et al. (2017).

Consider the following linear model (Lin et al. (2012); Yu and Moyeed (2001); Oh et al. (2016).

$$y_i = x_i^T \beta + \varepsilon_i, i = 1, 2, ..., n \quad (1)$$

where y_i is the response variable, x_i is the independent variable, and ε_i is an error, and $\theta\ \epsilon(0,1)$ is a certain quantile level. Conditional quantile regression is as follows (Yanuar et al. (2019b); Chernozhukov and Hong (2002)):

$$Q_\theta(y_i|x_i) = x_i^T \beta(\theta), i = 1, 2, ..., n \quad (2)$$

where $Q_\theta\ (y_i|x_i)$ represents the conditional quantile $\theta - th$ of the Y response given x and the parameter $\beta(\theta)$ is an unknown functional vector. Estimated point $\beta(\theta)$ from parameter $\beta\ \beta(\theta)$ is obtained by minimizing the objective function (Feng et al. (2015); Yanuar et al. (2016); Buhai (2005)):

$$min_{\beta(\theta)} \sum_{i=1}^{n} \rho_\theta(y_i - x_i^T \beta(\theta)) \quad (3)$$

whereρ_θ is the loss function, the loss function of quantile regression is as follows:

$$\rho_\theta(\varepsilon) = \varepsilon\theta, \text{ if } u \le 0, \text{ and } \rho_\theta(\varepsilon) = \varepsilon(\theta - 1), \text{ if } \varepsilon<0 \quad (4)$$

2.1.2 Censored Regression Quantile

The regression model based on the preceding discussion is referred to as the censored regression model or the Tobit model (Greene, 2008).To overcome censored data a special model is called a Tobit model which is stated as (Odah et al. (2017); Wang et al. (2009); Koenker (2008):

$$y^* = x_i^T \beta + \epsilon_i, i = 1, 2, 3..., n \quad (5)$$

where y^* is a latent variable containing the true value of the response variable for all observations, x is the explanatory variable and ϵ_i is error. Relationship analysis of censored data occurs if explanatory variables are available. Still, the value of the dependent variable is only known in observations where the dependent variable is larger (right censored) or smaller (left-censored) than the threshold value c (Davino et al., 2013).

For left-censored cases, it is defined as follows:

$$y_i = y_i^*, \text{ if } y_i^*>c$$

$$y_i = c, \text{ if } y_i^* \le c$$

It can be written as $\mathbf{y} = \max\ (c, \mathbf{y}^*)$

$$y = max(c, x_i^T \beta) \quad (6)$$

The CQR (Censored Regression Quantile) model is derived from the Tobit model. The left-censored LAD (Least Absolute Deviation) estimate is the β value by minimizing:

$$S(\beta) = \frac{1}{n} \sum_{i=1}^{n} |y_i - max(0, x_i^T \beta)| \quad (7)$$

The CQR model can be derived by extending the model in Equation (7) to other interesting quantiles and taking into account that max(0, β_0+$\beta_1\mathbf{x}$) is a monotonous transformation of y.

The CQR model for left-censored cases becomes:

$$Q_\theta(\hat{y}|x) = max[0, x_i^T \beta(\theta)] \quad (8)$$

As a result, the CQR estimator is obtained by minimizing:

$$S(\theta) = \sum_{i=1}^{n} \rho_\theta|y_i - max[0, x_i^T \beta(\theta)]| \quad (9)$$

2.1.3 The Best Indicator of Prediction Model

The standard error is defined as the standard deviation of the sample mean. This statistical measure can see the accuracy of the sample estimator against population parameters. The smaller the standard error value, the more probable the sample shows its accuracy, and the number of samples influences the standard error value. This statement means that the more samples the smaller the standard error, the more samples can represent the population. Mathematically the formula of the standard error is as follows:

$$SE(\beta) = \frac{SD(\beta)}{\sqrt{n}} \quad (10)$$

where $SE\ (\beta)$ is standard error of β, $SD\ (\beta)$ is standard deviation of (β) and n is the number of study samples.

2.2 Data

This study uses generated data of 150, 500, and 3000 sample sizes by using Software R version 1.1463. The data used in this study consisted of two independent variables (X_1 and X_2) and dependent variables (Y). Data for the independent variables (X_1 and X_2) are generated from a normal distribution with ($X_1 \sim N$ (0.2)) and ($X_2 \sim N$ (0.1)). Whereas the response variable (Y), is the value set $Y = 0.5\ X_1 + 1.5\ X_2 + \varepsilon$, where $\varepsilon \sim chisq$ (0.5) means the normality assumption is violated. The response variable is generated randomly and meet scenarios as a variable containing censored data. Censored data scenarios are engineered by assigning data below zero (left-censored) as censored data.

3. RESULTS AND DISCUSSION

The estimation results of each model parameter for each quantile selected using the censored and quantile regression methods are presented in Tables 1,2 and 3 as follows:

In Table 1 it can be seen that the estimated parameters β_1 and β_2 for quantile regression have been significant at the α = 0.05 level except for the quintile 0.9 where the β_1 α parameter is not significant. Then for the 4th and 5th columns are the results of guesses with the censored quantile regression method for both parameters. From these two columns, it is known that the alleged parameters β_1 and β_2 in all quantiles are significant at the level of α = 0.05 except for β_1 at quantile θ 0.9. Based on the results of the alleged parameters with the two methods, it can be concluded here that the quantile regression method and the censored quantile regression method tend to produce the same guessed results.

Next, the two methods are applied for the 500 exchange data generation. The estimated parameters are presented in Table 2.

In Table 2 the second and third columns can be obtained information that the estimated parameters β_1 and β_2 for quantile regression has been significant at the level α = 0.05 for each selected quartile. The same results were also obtained for the alleged parameter β_1 and β_2 allegations with the censored quantile regression method. It can be seen in columns 4 and 5 in table 2 above.

Then the two methods are applied for 3000 exchange trip generation data. The estimated parameters are presented in Table 3.

Based on table 3 it can be seen that the estimated parameters β_1 and β_2 for quantile regression has been significant at the level of α = 0.05 in each quantile this can be seen in column 2 and column 3. Then, the estimated parameter results for censored quantile regression which show that the parameters β_1 and β_2 have been significant at the level α = 0.05 in each quantile. It can be shown in column 4 and column 5.

Based on Tables 1,2 and 3, the β_1 parameters for quantile and censored quantile regression have been significant in each quantile by selecting a 5% significance level, except in the table for 150 generation data, β_1 in the quantile regression and censored quantile is not significant, can be seen in Table 1, column 2, and 4 at quantile 0.9. then the parameter β_2 for quantile regression and censored quantile is significant at each quantile by selecting a 5% significance level and a positive value which means that all independent variables x affect y significantly.

Then both methods are applied for 150-degree trip generation data. The goodness of fit of the model is based on the absolute value of bias and the standard error. The result of both criteria is presented in Table 4.

Based on Table 4 it is known that the absolute bias values for β_1 and β_2 in all selected quantiles are smaller than the quantile regression results. These results inform that the estimated parameter in the censored quantile regression is closer to the true value. Then, for the alleged standard error β_1 and β_2 the censored quantile regression results tend to be smaller than the quantile regression results. Based on the alleged absolute bias value and the estimated standard error value, it can be concluded that the censored quantile regression is better than the quantile regression for 150 pieces of data.

Furthermore, both methods are applied for 500-degree trip generation data. The absolute value of bias and the standard error for both methods are presented in Table 5.

Based on Table 5 it is known that the absolute bias values for β_1 and β_2 on all selected quantiles are smaller than the quantile regression results. This informs that the estimated parameter in the censored quantile regression is closer to the true value. Then, for the alleged standard error β_1 and $\beta_2$the censored quantile regression results tend to be smaller than the quantile regression results. Based on the alleged absolute bias value and the estimated standard error value, it can be concluded that the censored quantile regression is better than the quantile regression for 500 data.

Then the two methods are applied for 3000 trip generation data. To see the goodness of the model based on the absolute value of bias and the standard error presented in Table 6.

Based on Table 6, it is known that the absolute bias values for β_1 and β_2 in all selected quantiles are smaller than the quantile regression results. This result informs that the estimated parameter in the censored quantile regression is closer to the true value. Then, for the alleged standard error β_1 and β_2 the censored quantile regression results tend to be smaller than the quantile regression results. Based on the alleged absolute bias value and the estimated standard error value, it can be concluded that the censored quantile regression is better than the quantile regression for 3000 pieces of data.

Based on Tables 4,5 and 6 by taking samples of 150, 500 and 3000, respectively, information can be obtained about the goodness of the model in the trip generation data, by looking at the absolute bias value where absolute bias is the difference between the estimated parameter values and the actual parameter values. At each quantile, the absolute bias values for the β_1 and β_2 parameters of the censored quantile regression are smaller than the quantile regression results. These results mean that the value of the censored quantile regression parameter is more likely to produce an estimated value close to the actual value. Furthermore, the alleged error standard for each selected sample size also produces linear results. The default value of errors in

Table 1. Parameter estimation for n=150

Quantile	Quantile Regression		Censored Regression Quantile	
	β_1	β_2	β_1	β_2
0.1	0.34463*	0.96472*	0.49998*	1.49995*
0.25	0.32798*	0.81871*	0.49874*	1.49517*
0.5	0.32192*	0.85485*	0.49439*	1.46964*
0.75	0.26694*	0.84395*	0.35997*	1.28251*
0.9	0.20029	0.81145*	0.24335	1.13395*

* Significant at the 5% level

Table 2. Parameter estimation for n=500

Quantile	Regression Quantile		Censored Regression Quantile	
	β_1	β_2	β_1	β_2
0.1	0.22300*	0.71296*	0.50003*	1.50010*
0.25	0.25805*	0.78768*	0.50133*	1.50726*
0.5	0.26287*	0.76243*	0.50108*	1.56972*
0.75	0.28410*	0.78003*	0.46517*	1.57564*
0.9	0.31266*	0.82605*	0.64391*	1.70754*

* Significant at the 5% level

Table 3. Parameter estimation for n=3000

Quantile	Regression Quantile		Censored Regression Quantile	
	β_1	β_2	β_1	β_2
0.1	0.27620*	0.82913*	0.50000*	1.50004*
0.25	0.27221*	0.81115*	0.50041*	1.49978*
0.5	0.27217*	0.81092*	0.50158*	1.49584*
0.75	0.30504*	0.82298*	0.51545*	1.52908*
0.9	0.33883*	0.88136*	0.55131*	1.53972*

* Significant at the 5% level

Table 4. Absolute bias and standard error n=150

Quantile	Absolute Bias				Standard Error			
	Quantile		Censored Quantile		Quantile		Censored Quantile	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
0.1	0.15537	0.53528	0.00002	0.00005	0.01799	0.03327	0.00101	0.00145
0.25	0.17202	0.68129	0.00126	0.00483	0.02938	0.05434	0.00256	0.00722
0.5	0.17808	0.64515	0.00561	0.03036	0.04898	0.09057	0.01956	0.03572
0.75	0.23306	0.65605	0.14003	0.21749	0.07183	0.13282	0.06149	0.16509
0.9	0.29971	0.68855	0.25665	0.36605	0.11212	0.20733	0.15635	0.27046

Table 5. Absolute bias and standard error n=500

Quantile	Absolute Bias				Standard Error			
	Quantile		Censored Quantile		Quantile		Censored Quantile	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
0.1	0.277	0.78704	0.00003	0.0001	0.008	0.01727	0.00022	0.00076
0.25	0.24195	0.71232	0.00133	0.00726	0.01602	0.03459	0.00212	0.00479
0.5	0.23713	0.73757	0.00108	0.06972	0.02108	0.0455	0.0085	0.02776
0.75	0.2159	0.71997	0.03483	0.07564	0.0288	0.06219	0.04272	0.09691
0.9	0.18734	0.67395	0.14391	0.20754	0.06351	0.13711	0.12388	0.26797

Table 6. Absolute bias and standard error n=3000

Quantile	Absolute Bias				Standard Error			
	Quantile		Censored Quantile		Quantile		Censored Quantile	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
0.1	0.2238	0.67087	0	0.00004	0,00417	0.00848	0.00002	0.00007
0.25	0.22779	0.68885	0.00041	0.00022	0,00635	0.01292	0.00059	0.00154
0.5	0.22783	0.68908	0.00158	0.00416	0,00743	0.01513	0.00557	0.01145
0.75	0.19496	0.67762	0.01545	0.02908	0,00918	0.01867	0.02193	0.04434
0.9	0.16117	0.61864	0.05131	0.03972	0,02280	0.04641	0.02681	0.08018

censored quantile regression tends to be smaller than the results of quantile regression. It can be concluded here that censored quantile regression (CQR) is a better estimating method than quantile regression. Especially in the case of censored data.

4. CONCLUSIONS

This study aims to find the estimated value of the parameters by using quantile regression and CQR with censored data, and this study is applied to 3 defined data sizes, namely 150 data generated, respectively, and 3000 data. In this study, the independent variables generated are normally distributed and with abnormal errors. The results obtained on the three data sizes generated indicate that the estimated value of the parameters in the CQR tends to be better than the quantile regression. This result is indicated by the prediction value of the standard error and, the absolute value of the bias in censored quantile regression tends to be smaller. In other words, it can be concluded that the scattered quantile regression method is better than the quantile method.

REFERENCES

Buchinsky, M. (1994). Changes in the US wage structure 1963-1987: Application of quantile regression. *Econometrica: Journal of the Econometric Society*; 405–458

Buhai, S. (2005). Quantile regression: Overview and selected applications. *Ad Astra*, **4**(2005); 1–17

Chernozhukov, V., I. Fernández-Val, and A. E. Kowalski (2015). Quantile regression with censoring and endogeneity. *Journal of Econometrics*, **186**(1); 201–221

Chernozhukov, V. and H. Hong (2002). Three-Step Censored Quantile Regression and Extramarital Affairs. *Journal of the American statistical Association*, **97**(459); 872–882

Davino, C., M. Furno, and D. Vistocco (2013). *Quantile Regression: Theory and Applications*, volume 988. John Wiley & Sons

Delviyanti, S. O., F. Yanuar, and D. Devianto (2018). Simulation Study The Implementation of Quantile Bootstrap Method on Autocorrelated Error. *Cauchy: Jurnal Matematika Murni dan Aplikasi*, **5**(3); 95–101

Feng, Y., Y. Chen, X. He, et al. (2015). Bayesian Quantile Regression with Approximate Likelihood. *Bernoulli*, **21**(2); 832–850

Fitzenberger, B., R. Hujer, T. E. MaCurdy, and R. Schnabel (2002). Testing for Uniform Wage Trends in West-Germany: A Cohort Analysis Using Quantile Regressions for Censored Data. In *Economic Applications of Quantile Regression*. Springer, pages 41–86

Fitzenberger, B. and R. A. Wilke (2006). Using Quantile Regression for Duration Analysis. In *Modern Econometric Analysis*. Springer, pages 103–118

Greene, W. H. (2008). Econometrics Analysis, 6th Edition. *Prentice - Hall, New Jersey*

Gustavsen, G. W. and K. Rickertsen (2013). Adjusting VAT Rates to Promote Healthier Diets in Norway: A Censored Quantile Regression Approach. *Food Policy*, **42**; 88–95

Koenker, R. (2008). Censored Quantile Regression Redux. *Journal of Statistical Software*, **27**(6); 1–25

Koenker, R. and G. Bassett Jr (1978). Regression Quantiles. *Econometrica: journal of the Econometric Society*; 33–50

Leiker, A. (2012). A Comparison Study on The Estimation in Tobit Regression Models. *Thesis Master of Science, Kansas State University, Kansas.*

Lin, G., X. He, and S. Portnoy (2012). Quantile Regression with Doubly Censored Data. *Computational Statistics & Data Analysis*, **56**(4); 797–812

Machado, J. A., P. Portugal, and J. Guimaraes (2006). U.S Unemployment Duration: Has Long Become Longer or Short Become Shorter. *IZA discussion paper, No. 2174*

Muharisa, C., F. Yanuar, and D. Devianto (2018). Simulation Study The Using of Bayesian Quantile Regression in Nonnormal Error. *Cauchy: Jurnal Matematika Murni dan Aplikasi*, **5**(3); 121–126

Odah, M. H., A. S. M. Bager, and B. K. Mohammed (2017). Tobit Regression Analysis Applied on Iraqi Bank Loans. *American Journal of Mathematics and Statistics, 7 (4), 179*, **182**

Oh, M.-S., J. Choi, and E. S. Park (2016). Bayesian Variable Selection in Quantile Regression Using The Savage-Dickey Density Ratio. *Journal of The Korean Statistical Society*, **45**(3); 466–476

Wang, H. J., M. Fygenson, et al. (2009). Inference for Censored Quantile Regression Models in Longitudinal Studies. *The Annals of Statistics*, **37**(2); 756–781

Yanuar, F., H. Laila, and D. Devianto (2017). The Simulation Study to Test the Performance of Quantile Regression Method With Heteroscedastic Error Variance. *Cauchy: Jurnal Matematika Murni dan Aplikasi*, **5**(1); 36–41

Yanuar, F., H. Yozza, F. Firdawati, I. Rahmi, and A. Zetra (2019a). Applying Bootstrap Quantile Regression for The Construction of a Low Birth Weight Model. *Makara Journal of Health Research*, **23**(2); 5

Yanuar, F., H. Yozza, and I. Rahmi (2016). Penerapan Metode Regresi Kuantil pada Kasus Pelanggaran Asumsi Kenormalan Sisaan. *Eksakta*, **1**; 33–37

Yanuar, F., H. Yozza, and A. Zetra (2019b). Bayesian Quantile Regression Methods in Handling Non-normal and Heterogeneous Error Term. *Asian Journal of Scientific Research*, **12**(3); 346–351

Yu, K. and R. A. Moyeed (2001). Bayesian Quantile Regression. *Statistics & Probability Letters*, **54**(4); 437–447

