BAREKENG: Jurnal Ilmu Matematika dan Terapan
June 2020 Vol. 14 No. 2 Page 301–308
P-ISSN: 1978-7227 E-ISSN: 2615-3017
: https://doi.org/10.30598/barekengvol14iss2pp301-308
National Accredited in SINTA 3, Decree No.: 28/E/KPT/2019

EFEKTIVITAS REGRESI KUANTIL DALAM MENGATASI PONTENSIAL PENCILAN

The Effectiveness of Quantile Regression in Dealing with Potential Outliers

Netti Herawati [1*]

[1] *Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Lampung, Bandar Lampung, 35141, Indonesia*

e-mail: netti.herawati@fmipa.unila.ac.id
*Corresponding Author**

Abstraks

Regresi kuantil sebagai metode regresi robust dapat digunakan untuk mengatasi dampak kasus yang tidak biasa pada estimasi regresi seperti keberadaan pencilan (*outlier*) pada data. Tujuan dari penelitian ini adalah untuk mengevaluasi efektivitas regresi kuantil untuk menangani pencilan potensial dalam regresi linear berganda dibandingkan dengan metode kuadrat terkecil (MKT). Penelitian ini menggunakan data simulasi pada model regresi berganda dengan jumlah variabel independen (p=3) pada ukuran sampel yang berbeda (n = 20, 40, 60, 100, 200) dan $\beta_0 = 0$ and $\beta_1 = \beta_2 = \beta_3 = 1$ diulang 1000 kali. Efektivitas metode regresi kuantil dan MKT dalam pendugaan parameter β diukur dengan *Mean Square Error* (MSE) dan model terbaik dipilih berdasarkan nilai *Akaike Information Criterion* (AIC) terkecil. Hasil penelitian menunjukkan bahwa berbeda dengan OLS, regresi kuantil mampu menangani potensial pencilan (*outlier*) dan memberikan estimator yang lebih baik dengan nilai MSE yang lebih kecil. Dibandingkan dengan MKT dan kuantil lainnya, studi ini juga memberikan hasil yang cukup untuk memastikan bahwa kuantil 0,5 memberikan estimasi parameter terbaik dan model terbaik berdasarkan nilai MSE dan AIC terkecil.

Kata Kunci : *AIC, MSE, pencilan, regresi kuantil*

Abstract

Quantile regression as a robust regression method can be used to overcome the impact of unusual cases on regression estimates such as the presence of potential outliers in the data. The purpose of this study was to evaluate the effectiveness of quantile regression in dealing with potential outliers in multiple linear regression compared to ordinary least square (OLS). This study used simulation data in multiple regression model with the number of independent variables (p=3) for different sample sizes (n = 20, 40, 60, 100, 200) and $\beta_0 = 0$ and $\beta_1 = \beta_2 = \beta_3 = 1$ repeated 1000 times. The effectiveness of the quantile regression method and OLS in estimating β parameters was measured by Mean square error (MSE) and the best model is chosen based on the smallest Akaike Information Criterion (AIC) value. The results showed that in contrast to OLS, quantile regression was able to deal with potential outliers and provided a better estimator with a smaller mean mean square error. Compared to OLS and other quantiles, this study also provides sufficient results that quantile 0.5 provides the best parameter estimate and the best model based on the smallest MSE and AIC values.

Keywords: *AIC, MSE, outliers, quantile regression*

Submitted: 02nd April 2020 *Accepted: 29th May 2020*

This is an open access article under the CC–BY-SA license

1. INTRODUCTION

Classical linear regression estimates the mean response of the dependent variable dependent on the independent variables. The method usually use to estimates in classical linear regression is ordinary least square [1, 2, 3]. However, it is a parametric model and relies on assumptions of certain distribution in residuals that are often not met. There are many cases that the conditional mean behavior fails to entirely capture the patterns in the data when the data is skewed, multimodal, or contains outliers. In this condition, the residual distribution assumptions in the classical linear model will not be met, especially the normality distribution assumption. For this type of data is better estimated using methods that do not require any residual distribution. One such method is quantile regression.

Quantile regression has been used in many studies [4, 5, 6]. Quantile regression method propose a technique for estimating models for the conditional median function and the full range of other conditional quantile function. Just like the least squares regression, quantile regression is interested in studying the linear relationship between a response variable and one or more independent or explanatory variables. However, the main purpose of the least squares regression is to determine the conditional mean of the response variable Y while the quantile regression models is related to the conditional $\tau \in (0,1)$ with τ is quantile level of Y. In addition, quantile regression allows multiple quantiles to be modelled. Quantile regression is offering more comprehensive analysis of the data to be carried out compared to OLS where only the mean is considered. Quantile regression makes no assumptions about the distribution of the residuals [7, 8, 9]. It also lets you explore different aspects of the relationship between the dependent variable and the independent variables.

The study of handling outliers has been done by many researches [10, 11, 12, 13]. In this study we will investigate the behavior of quantile regression in handling outliers compare to OLS using simulated data based on MSE and AIC.

2. RESEARCH METHOD

The quantile regression is an extension of ordinary quantiles ideas. Classical linear regression method is based on minimizing the sum of squared residuals to model the conditional mean of the target variable against the covariates. On the other hand, quantile regression provides estimates of a range of conditional quantiles to model conditional percentiles of the target variable against the covariates. It is a useful tool for estimating not only upper or lower tail but also the center of the conditional distribution of interest [4, 8, 9, 14, 15].

Consider a general regression function with y is the response variable and $\boldsymbol{x} = (x_1, x_2, \dots, x_p)^{\tau}$is a set of predictors. To obtain the sample mean, least square regression model solve

$$\min_{\mu \in \Re} \sum_{i=1}^{n} (y_i - \mu)^2 \qquad (1)$$

as an estimate of the unconditional population mean, EY. By replacing the scalar μ with a parametric function μ (x, β), we can solve

$$\min_{\beta \in \Re p} \sum_{i=1}^{n} (y_i - \mu(x_i, \beta))^2 \qquad (2)$$

to find the estimate of the conditional expectation function $E(Y|x)$. The least square estimate for (2) is given by $\hat{\beta} = (X'X)^{-1}X'Y$ [16].

In term of quantile regression, consider a continuous real valued random variable Y characterized by the following distribution function in the ordinary quantile

$$F_Y(y) = P(Y \leq y) \qquad (3)$$

The τ-th quantile of Y for any $\tau \in (0,1)$ is defined as

$$Q(\tau) = \inf(y: F_Y(y) \geq \tau) \qquad (4)$$

When Q(1/2), it is equal to median [17, 18]. Just like the distribution function F, the quantiles function provides a complete characterization of the random variable Y [8]. When estimating quantiles, the value of y in the sample data corresponding to a given probability τ has to be determined. The τ^{th} quantile in a sample of data refers to the probability of τ for a value y, such that $F_Y(y_\tau) = \tau$. It can also write as $y_\tau = F_Y^{-1}(\tau)$where y_τ is such that an inverse of the function $F_Y(\tau)$ for a probability τ.

The 100 τ% quantile (say τ =0.5) of the conditional distribution of the response (y) given covariates (**x**) based on independent observations $(\boldsymbol{x}_i, y_i)_{i=1}^{n}$, the conditional τ-quantile is estimated by minimizing

$$\sum_{i=1}^{n} \rho_\tau\left(y_i - \beta_0 - \boldsymbol{x}_i^T \beta\right) \quad (5)$$

Where $\rho_\tau(t) = \tau t_+ + (1-\tau)t_-$ is the check function with subscript s '+' and '-' stand for the positive and negative parts, respectively [9]. The estimation of selected significant predictors in quantile regression use L_1 [8]. The L_1 quantile regression model is estimated by

$$\widehat{\boldsymbol{\beta}}(L_1 norm) = \arg\min_{\beta_0 \beta} \sum_{i=1}^{n} \rho_\tau\left(y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta}\right) + \lambda \|\boldsymbol{\beta}\|_1 \quad (6)$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ is the L_1-norm penalty (or lasso penalty) on $\boldsymbol{\beta}$. When λ is chosen appropriately, some components of $\widehat{\boldsymbol{\beta}}$ will be shrunk to exact zero. Since the check loss function is piecewise linear, the quantile regression estimator is inherently robust in handling extreme value point and outliers.

To get a measure of how close the regression line was to a set of points, the Mean Square Error (MSE) of the regression coefficient $\widehat{\boldsymbol{\beta}}$ was examined. The MSE is defined by $MSE(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{l=1}^{m} \left\|\widehat{\boldsymbol{\beta}}^{(l)} - \boldsymbol{\beta}\right\|^2$ where $\hat{\beta}^{(l)}$is the estimated parameter in the *l*-th simulation. The slope and intercept are correctly estimated when MSE approaches to zero. To determine the performance of the proposed estimate, the Akaike Information Criterion (AIC) was used. AIC can be written as $AIC_C = 2k - 2\ln(\hat{L})$ where $\hat{L} = p\left(x|\hat{\theta}, M\right)$, $\hat{\theta}$ is the value that maximize the likelihood function, n = sample size, and k = the number of parameters [19, 20]. A good estimation model was indicated by the lowest AIC value.

Simulated data was carried out in this study with five different sample sizes (n=20, 40, 60, 100, 200) for three independent variables *(p*=3) using a package for quantile regression developed by [21]. Dependent variable ($\boldsymbol{Y}$) for each p independent variables was from $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ with x_i~N(0, 1) and ε~N(0, 1) contaminated with various number of outliers (10%, 15%, 20%). β parameters were chosen with β_0=0, and $\beta_1, \beta_2, \beta_3$=1. After simulating each data, we fitted the OLS and quantile regression and measured the efficiency of both methods in estimating the regression coefficient and determined the best estimation model using AIC.

3. RESULTS AND DISCUSSION

The simulation results for identifying the effectiveness of quantile regression presented in Table 1 in terms of standard error of parameter estimates. It shows that for n=20, 40, 60, 100, 200 with 10% outliers, standard error of estimates using quantile regression was lower than OLS. Similar results were obtained for n= 20, 40, 60, 100, 200 with 20% and 30% outliers. It indicates that quantile regression provides better parameter estimates than OLS for all sample sizes and various number of outliers being studied. We can also see that quantile 0.5 gives the lowest standard error of parameter estimates compare to quantile 0.25 and quantile 0.75 for n=20, 40, 60, 100, 200 with various number of outliers. It proves that the quantile 0.5 is the most accurate parameter estimates than quantile 0.25 and quantile 0.75.

Table 1. $\widehat{\boldsymbol{\beta}}$ and SE ($\widehat{\boldsymbol{\beta}}_i$) for different sample sizes and various number of outliers

Sample size	Method	10% outliers			20% outliers			30% outliers		
		SE ($\hat{\beta}_1$)	SE ($\hat{\beta}_2$)	SE ($\hat{\beta}_3$)	SE ($\hat{\beta}_1$)	SE ($\hat{\beta}_2$)	SE ($\hat{\beta}_3$)	SE ($\hat{\beta}_1$)	SE ($\hat{\beta}_2$)	SE ($\hat{\beta}_3$)
n=20	OLS	2.8046	2.3233	2.4495	2.3998	3.3882	4.0557	4.8388	2.8077	4.2070
	QR 0.25	2.1241	2.1766	2.0195	3.1991	3.1465	3.1412	3.6997	3.7711	3.8961
	QR 0.50	2.0933	2.1944	2.0252	3.1989	3.1462	3.1402	3.6990	3.7710	3.8961
	QR 0.75	2.1241	2.1766	2.0195	3.3116	3.3660	3.3260	3.9607	3.7982	4.0457
n=40	OLS	1.9844	1.2343	1.8596	2.3226	2.4937	2.9881	3.0259	3.2395	2.0931
	QR 0.25	1.6162	1.6174	1.6894	2.2147	2.2566	2.2404	2.7026	2.7457	2.7248
	QR 0.50	1.6160	1.6170	1.6894	2.2140	2.2489	2.2400	2.7020	2.7455	2.7240
	QR 0.75	1.7222	1.7072	1.7262	2.2405	2.2566	2.2610	2.7287	2.7476	2.7589
n=60	OLS	1.3450	1.2208	1.1639	1.7725	1.8507	1.77421	2.3037	2.3443	1.9384
	QR 0.25	1.2146	1.2215	1.1880	1.7267	1.7341	1.7091	2.0941	2.1605	2.0852
	QR 0.50	1.2144	1.2210	1.1880	1.7266	1.7339	1.7081	2.0931	2.1600	2.0850
	QR 0.75	1.2387	1.2343	1.2070	1.7623	1.7585	1.7591	2.1920	2.2037	2.1573
n=100	OLS	1.0329	1.0482	1.1334	1.2240	1.4546	1.3552	1.6024	1.7452	1.7047
	QR 0.25	1.0125	1.0382	1.0609	1.3018	1.3199	1.3123	1.6564	1.6474	1.6322
	QR 0.50	1.0125	1.0272	1.0125	1.3011	1.3189	1.3120	1.6562	1.6470	1.6321
	QR 0.75	1.0382	1.0480	1.0468	1.3314	1.3636	1.3405	1.6763	1.6894	1.6728

n=200	OLS	0.6429	0.7979	0.6867	0.9394	0.9394	0.9394	1.1853	1.3932	1.1794
	QR 0.25	0.6835	0.6978	0.6921	0.9826	0.9826	0.9826	1.1404	1.1552	1.1425
	QR 0.50	0.6825	0.6970	0.6911	0.9394	0.9394	0.9394	1.1400	1.1542	1.1422
	QR 0.75	0.7075	0.7111	0.7156	0.9826	0.9826	0.9826	1.1799	1.1748	1.1776

The results of analyzing the effectiveness of quantile regression compares to OLS in estimating parameter β using simulated data for n=20, 40, 60, 100, 200 and contaminated by 10%, 20%, 30% outliers repeated 1000 times in terms of MSE are given in Table 2.

Table 2. MSE of OLS and QR for different sample sizes and various number of outliers

Sample size	Method	MSE		
		10% outliers	20% outliers	30% outliers
n-20	OLS	0.1531	0.5307	0.5175
	QR 0.25	0.0738	0.0153	0.0086
	QR 0.50	0.0067	0.0061	0.0067
	QR 0.75	0.0103	0.0133	0.0142
n=40	OLS	0.5280	0.1199	0.4259
	QR 0.25	0.0042	0.0052	0.0052
	QR 0.50	0.0005	0.0008	0.0015
	QR 0.75	0.0309	0.0034	0.0048
n=60	OLS	0.0288	0.0191	0.0731
	QR 0.25	0.0088	0.0085	0.0422
	QR 0.50	0.0032	0.0008	0.0004
	QR 0.75	0.0123	0.0108	0.0150
n=100	OLS	0.0850	0.0603	0.0167
	QR 0.25	0.0019	0.0245	0.0063
	QR 0.50	0.0006	0.0022	0.0001
	QR 0.75	0.0145	0.0034	0.0046
n=200	OLS	0.1095	0.1366	0.2466
	QR 0.25	0.0153	0.0157	0.0336
	QR 0.50	0.0005	0.0001	0.0008
	QR 0.75	0.0012	0.0013	0.0047

It can be seen from Table 2 that for 10 % outliers, OLS gives MSE =0.1531 for *n*=20, MSE= 0.5280, for *n*=40, MSE= 0.0288, for *n*=60, MSE= 0.0850, for *n*=100 and MSE= 0.1095, for *n*=200, respectively. These values are much higher than the MSE of quantile 0.25, 0.50, and 0.75. Similarly, when there is 20% and 30% outliers in the data, the MSE of OLS all sample sizes being studied is much higher than MSE of quantile regression. The MSE of OLS appears to be increasing due to the increasing of the outliers in smaller sample sizes (n=20, 40, 60). However, it reverse for larger sample sizes (n=100, 200). Whereas MSE of quantile 0.25, 0.5 and 0.75 decreased as the number of samples increased. When there are potential outliers, quantile regression does not appear to be affected. This phenomenon occurs in all number of outliers studied. Among all quantiles, the MSE value for quantile 0.5 is the smallest compared to MSE for quantile 0.25 and quantile 0.75 for all sample sizes and various number of outliers. As it shows in Table 2, for 10% outliers, quantile 0.5 has MSE =0.0067 for *n*=20, MSE= 0.0005, for *n*=40, MSE= 0.0032, for *n*=60, MSE= 0.0006, for *n*=100 and MSE= 0.0005, for *n*=200, respectively. The MSE values at quantile 0.5 are also the smallest in all sample sizes containing 20% and 30% outliers compared to quantile 0.25 and quantile 0.75 as shown in Table 2. To present a more comprehensive result, the MSE values of both methods and each quantile are displayed in Figure 1-5.

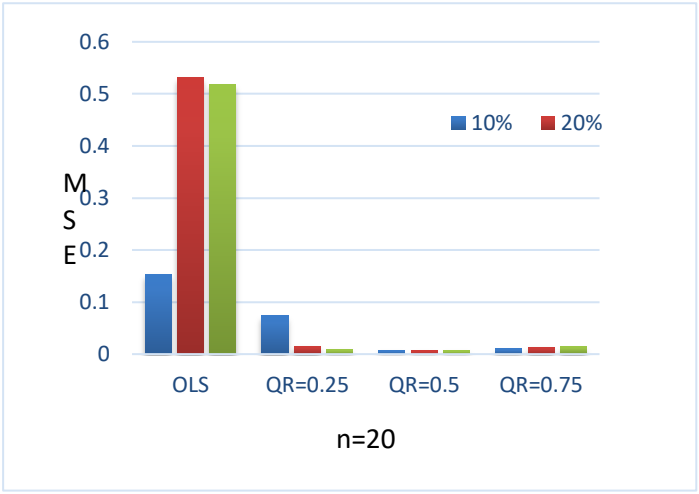

Figure 1. MSE for n=20 contain various number of outliers

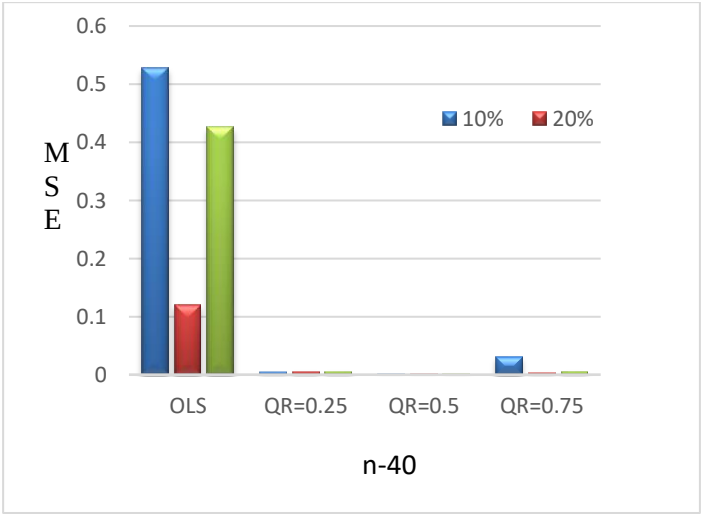

Figure 2. MSE for n=40 contain various number of outliers

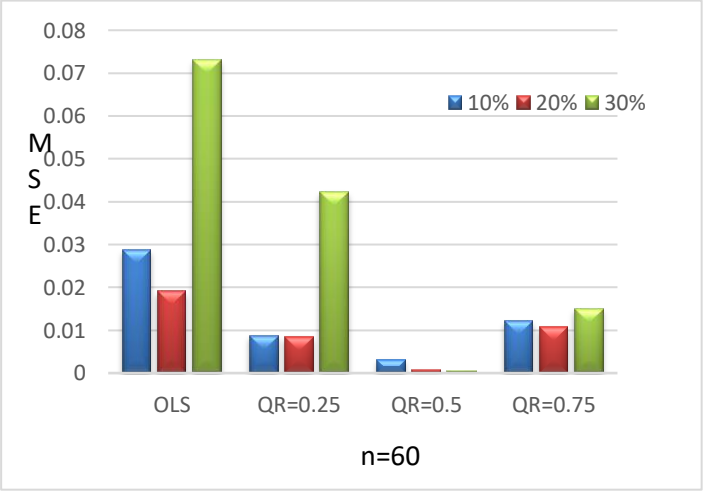

Figure 3. MSE for n=60 contain various number of outliers

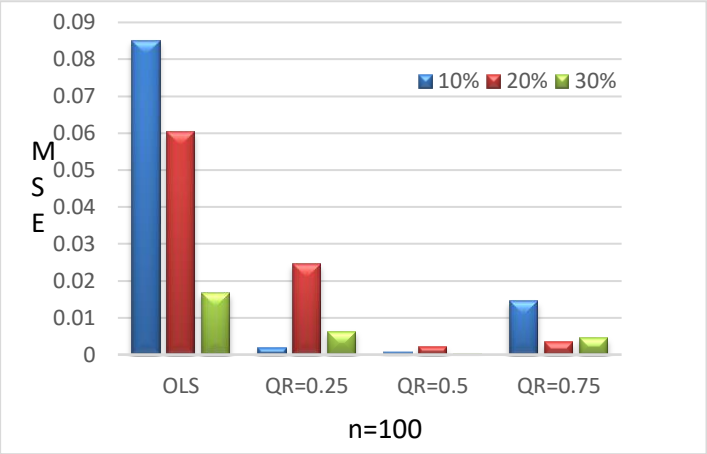

Figure 4. MSE for n=100 contain various number of outliers

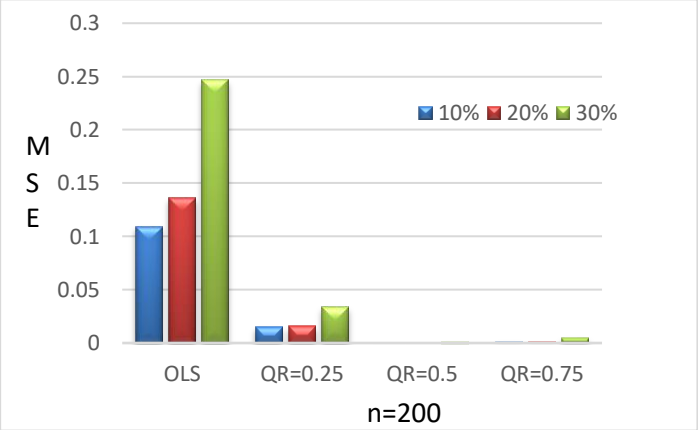

Figure 5. MSE for n=200 contain various number of outliers

In the clear view of Figure 1-5, we can see that the MSE of quantile regression 0.25, 0.5, and 0.75 are significantly lower than OLS for all sample sizes (n=20, 40, 60, 100, 200) and various number of outliers (10%, 20%, 30%). It is clear that MSE of quantile regression does not to be affected by outliers. This indicates that quantile regression is able to handle potential outliers very well and robust to potential outliers up to 30% of the data.

Moreover, if we compare the quantiles it becomes clear that the quantile 0.5 gives the lowest MSE value compared to the quantile 0.25 and quantile 0.75. This provides evidence that quantile 0.5 gives better parameter estimates than quantile 0.25 and quantile 0.75.

To find the best estimation model, the AIC value for OLS and quantile regression in each number of sample sizes contaminated by outliers was measured. Table 3 shows the AIC values for n = 20, 40, 60, 100, 200 and were contaminated by 10%, 20%, 30% outliers which were repeated 1000 times.

Table 3. AIC of OLS and QR contain various number of potential outliers

Sample size	Method	AIC		
		10% outliers	20% outliers	30% outliers
n=20	OLS	-1.5767	-0.3334	-0.3586
	QR 0.25	-2.3064	-3.8767	-4.4456
	QR 0.50	-4.7075	-4.7910	-4.7027
	QR 0.75	-4.2701	-4.0184	-3.9476
n=40	OLS	-0.4884	-1.9706	-0.7035
	QR 0.25	-5.3203	-5.0946	-5.0946
	QR 0.50	-7.3308	-6.9583	-6.2911
	QR 0.75	-3.3266	-5.5247	-5.1728
n=60	OLS	-3.4450	-3.8574	-2.5154
	QR 0.25	-4.6247	-4.6674	-3.0630
	QR 0.50	-5.6176	-7.0122	-7.5020
	QR 0.75	-4.2980	-4.4191	-4.0996

n=100	OLS	-2.4040	-2.7482	-4.0269
	QR 0.25	-6.1631	-3.6482	-4.9946
	QR 0.50	-7.3252	-6.0498	-8.7135
	QR 0.75	-4.1727	-5.6019	-5.3185
n=200	OLS	-2.1814	-1.9603	-1.3696
	QR 0.25	-4.1446	-4.1241	-3.3624
	QR 0.50	-9.8145	-10.8246	-7.0512
	QR 0.75	-4.3161	-4.2661	-5.3121

From Table 3 it is visible that there is a significant difference between OLS and quantile regression in the AIC value. Quantile regression produces AIC value much lower than OLS in all sample sizes and number of outliers. That is, compared to OLS, quantile regression provides a better regression model if there are potential outliers. It proves that quantile regression model is robust to potential outliers than OLS. The study also shows that quantile regression 0.5 gives the best estimates than other quantiles. That is one of the advantages of using quantile regression where one can estimate parameters using different quantiles which are not provided in OLS. In this study, quantile 0.5 provided the best parameter estimate and the best model compared to OLS and the other quantiles for all sample sizes and various number of outliers.

4. CONCLUSION

The study showed that quantile regression performanced far better than OLS based on MSE and AIC for n=20, 40, 60, 100, 200 and number of potential outliers 10%, 20%, 30%. We concluded that in contrast to OLS, quantile regression model was more effective in dealing with potential outliers for different sample sizes and various number of potential outliers. In addition, quantile 0.5 gives the best parameter estimate and the best model based on the smallest MSE and AIC values compared to quantile 0.25 and quantile 0.75 for all sample sizes and number of outliers studied.

REFERENCES

[1] M.H. Kutner, C.J. Nachtsheim and J. Neter, *Applied Linear Regression Models*, 4th Ed., New York: McGraw-Hill Co. Inc., 2004.

[2] D.C. Montgomery, E.A. Peck and G.G. Vining, *Introduction to Linear Regression Analysis*, 3rd Ed., USA: John Wiley & Sons, Inc., 2001.

[3] N.R. Draper and H. Smith, *Applied Regression Analysis*, 3rd Ed., USA: John Wiley & Sons, Inc., 1998.

[4] R. Koenker and K.F. Hallock, "Quantile regression," *Journal of Economic Perspectives*, vol. 15, no.4, pp 143-156, 2001.

[5] G.A.Haile and A.N. Nguyen, "Determinant of academic attainment in the United States: A quantile regression analysis of test scores, " *Education Economics*, vol. 16, pp. 29-57, 2008.

[6] J. Staruss, K. Beege, B. Sikoki, A. Dwiyanto, Y. Herawati and F. Witoelar, *The third wave of the Indonesia family life survey: Overview and field report,* Santa Monica, CA: RAND Corp, 2004.

[7] L. Hao and D.Q. Naiman, *Quantile Regression*, Thousand Oaks: Sage Publications Inc., 2007.

[8] R. Koenker, *Quantile regression*, New York: Cambridge University Press, 2005.

[9] R. Koenker and G. Bassett, "Regression quantile," *Econometrica,* vol. 46, pp.33-50, 1978.

[10] K. Nisa and N. Herawati, . 'Robust Estimation of Generalized Estimating Equation when data contain outliers,' *Insist*, vol. 2, No. 1, pp., 1-5, 2017.

[11] P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection*, New York: Wiley, 1987.

[12] E. Setiawan, N. Herawati, K. Nisa, Nyusirwan, and S. Saidi, "Handling full multicollinearity and various numbers of outliers using robust ridge regression," *Sci.Int.(Lahore)*, vol 31, no. 2, pp. 201-204, 2019.

[13] V. Barnett and T Lewis, *Outlier in Statistical Data*, New York :John Wiley & Sons, 1984.

[14] W. Gilchrist, *Statistical modelling with quantile functions*, FL: Chapman and Hall/CRC, Boca Raton, 2000.

[15] C.M. Kuan, *An Introduction to Quantile Regression*. Institute of Economics, Canada: Academia Sinica, 2007.

[16] W.H. Greene, *Econometric Analysis*, 5th ed., New Jersey: Prentice Hall, 2002.

[17] B. Fitzenverger, R. Koenker, J.A.F. Machado, (Eds.), *Economic Applications of Quantile Regression*, Berlin Heidelberg GmbH: Springer-Verlag, 2002.

[18] M. Furno, *Parameter Instability in Quantile Regressions*, NewYork: Elsevier, 2007.

[19] H. Akaike, Information theory and an extension of the maximum likelihood principle, In B.N. Petrow and F. Csaki (eds), Second International symposium on information theory (pp.267-281), Budapest: Academiai Kiado, 1973.

[20] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp.716-723, 1974.
[21] R. Koenker, *Quantreg: Quantile Regression*, Version R package version 4.22, 2008.