CAUCHY –Jurnal Matematika Murni dan Aplikasi
Volume 5(1)(2017), Pages 36-41
p-ISSN: 2086-0382; e-ISSN: 2477-3344

The Simulation Study to Test the Performance of Quantile Regression Method With Heteroscedastic Error Variance

Ferra Yanuar[1*], Laila Hasnah[2], Dodi Devianto[3]

[1,2,3]Jurusan Matematika FMIPA Universitas Andalas
Kampus Limau Manis 25163 Padang
* Corresponding Author. E-mail: ferrayanuar@yahoo.co.id

ABSTRACT

Least square estimator has many limitations. This estimator will not be a Best Linear Unbiased Estimator (BLUE) in the condition of the variance error term have heteroscedasticity problem. Quantile regression is a robust approach in situations where the limitation addressed above present for least square estimator. The purpose of this study is to describe the performance of quantile regression method in modeling a data set which contain the heteroscedasticity problem. To achieve the goal, a data set is generated and statistical framework quantile method then applied to the data. The consistency of the proposed model is then checked by doing a simulation study. This study proves that the quantile regression method is able to produce acceptable parameter model since the proposed models have large Pseudo R^2 and small mean square error (MSE) for all parameter estimated. It could be conclude here that quantile regression method is an unbiased estimator method and able to result acceptable model althought in the present of heteroscedasticity problem of variance error.

Keywords: Heteroscedasticity, quantile regression, simulation study, Pseudo R^2, mean square error (MSE).

INTRODUCTION

In modeling the relationship between covariates and responses, it need estimator method to estimate the parameter model. To estimate the value of parameters, it usually use the Ordinary Least Squares (OLS). The principle of this method is to minimize the sum of the squares of the error. This OLS is applied if all model assumptions are met (independent observations, linearity of conditional means, normality of response variable and homogeneity of error variance). In all model assumptions are met, the estimator method is called as BLUE (Best Linear Unbiased Estimator). However, if one or more of the assumptions are not met, the results could be misleading [1].

In regression, an error is how far a point deviates from the regression line. Ideally, our data should be homoscedastic (i.e. the variance of the errors should be constant). In many real world applications, this situation rarely happens. Most data is heteroscedastic by nature. Due to these limitatios, an alternative approach to classical linear regression is in demand. Quantile regression is a robust approach in situations where the limitations addressed above [2] present for ordinary least square estimator.

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 the largest [1, 3]. Quantile regression is an approach in regression analysis introduced by Koenker and Basset [4]. Quantile

Submitted: 15 May 2017 Reviewed: 24 July 2017 Accepted: 2 November 2017
DOI: http://dx.doi.org/10.18860/ca.v5i1.4209

regression in his theory is able to overcome the violation of normality assumptions, heteroscedasticity, multicollinearity problems and so on. This method uses the parameter estimation approach by separating or dividing the data into quantities, by assuming the conditional quantization function on a distribution of data and minimizing the absolute asymmetry of unsymmetric weighted error and presupposes a conditional quantile function on a distribution of data [5].

In this paper, we adopt the quantile regression approach to modeling groups of data with non homogeneity of error variance. Section 2 of the paper, describes the theoretical framework of quantile regression and its indicators to determine the goodness of fit of the proposed model. In section 3, we illustrate the implementation of quantile regression through a simulated case-study. We choose two covariates in our model hypothesis.as the predictors to the response variable. We end with a short discussion in Section 4.

FUNDAMENTAL THEORIES AND RELATED WORKS

In stricly linear models, a simple approach to estimating the conditional quantiles is suggested in Koenker and Basset [2]. Based on the classical regression model, we have:

$$y_i = \boldsymbol{x_i}'\boldsymbol{\beta} + \varepsilon_i, i = 1,2,\dots,n \quad (1)$$

where y_i are dependent variable for each data i , x_i are independent matrix nxp, with $\mu(y_i) = \boldsymbol{x_i}'\boldsymbol{\beta}$, $\boldsymbol{\beta}$ is vector of parameter model size $px1$, ε_i are error for each data i .

The parameter estimate by classical regression, by minimizing the sum of the error squares, is written as

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

where $\hat{y}_i$ are estimated value for each data i .

2.1. Quantile Regression Method

The prediction based on the median, that is, by minimizing the absolute number of errors can be written using following equation :

$$min \sum_{i=1}^{n} |y_i - \hat{y}_i|^2 \quad (3)$$

Furthermore, the linear equations for τ -th quantil can be written as follows :

$$y_i = \boldsymbol{x_i}'\boldsymbol{\beta_\tau} + \varepsilon_i, \quad i = 1,2,..,n \quad (4)$$

Noted that $\hat{y}_i = \boldsymbol{x_i}'\boldsymbol{\beta_\tau}$, the parameter estimate for τ th quantile is to minimize the absolute value of the error by weighting τ for the positive and weighted error (1- τ) for the negative error [4]. The τth $(0 < \tau < 1)$ quantile of ε_i is the value, Q_τ, for which $P\left(\varepsilon_i < Q_Y(\tau)\right) = \tau$. The τth conditional quantile of y_i given $\boldsymbol{x_i}$ is then simply [6,7] :

$$Q_\tau(y_i|\boldsymbol{x_i}) = \boldsymbol{x_i}'\boldsymbol{\beta_\tau},$$

where $\boldsymbol{\beta_\tau}$, is a vector of coefficients dependent of τ.

The τth regression quantile is defined as any solution, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}$, to the quantile regression minimasation problem :

$$min_{\beta \epsilon \mathcal{R}} \sum_{i=1}^{n} \rho_\tau (y_i - \boldsymbol{x_i}'\boldsymbol{\beta_\tau}), \quad (5)$$

where the loss function :

$$\rho_\tau(u) = u\big(\tau - I(u < 0)\big). \quad (6)$$

Equivalently, we may rewrite (5) as :

$$min_{\beta \epsilon \mathcal{R}} \{\sum_{i=1}^{n} \tau |y_i - \boldsymbol{x_i}'\boldsymbol{\beta_\tau}| + \sum_{i=1}^{n} (1-\tau)|y_i - \boldsymbol{x_i}'\boldsymbol{\beta_\tau}|\} \quad (7)$$

The meaning which can be added to explain the equation (7), namely that all observations greater than the quantile value, multiplied by the weighting τ and the observations whose value is less than the quantile multiplied by $1 - \tau$.

2.2 Goodness of Fit using Pseudo R^2

Simple quantile regression model with n independent variables can be formed as follows:

$$Q_\tau(\hat{\boldsymbol{y}}|\boldsymbol{x}) = \hat{\beta}_0(\tau) + \hat{\beta}_1(\tau)\boldsymbol{x} + \cdots + \hat{\beta}_n(\tau)\boldsymbol{x} \quad (8)$$

The indicator of the goodness of fit for the model can be predicted with Pseudo R^2 as defined below [2]:

$$\text{Pseudo } R_\tau^2 = 1 - \frac{RAWS_\tau}{TASW_\tau} \tag{9}$$

where :

$$RAWS_\tau = \sum_{y_i \geq \hat{\beta}_0(\tau)+\hat{\beta}_1(\tau)\boldsymbol{x_i}+\cdots+\hat{\beta}_n(\tau)\boldsymbol{x}} \tau\left|y_i - \hat{\beta}_0(\tau) - \hat{\beta}_1(\tau)\boldsymbol{x_i} - \cdots - \hat{\beta}_n(\tau)\boldsymbol{x_i}\right| + \sum_{y_i < \hat{\beta}_0(\tau)+\hat{\beta}_1(\tau)\boldsymbol{x_i}+\cdots+\hat{\beta}_n(\tau)\boldsymbol{x}} (1-\tau)\left|y_i - \hat{\beta}_0(\tau) - \hat{\beta}_1(\tau)\boldsymbol{x_i} - \hat{\beta}_n(\tau)\boldsymbol{x_i}\right| \tag{10}$$

and

$$TASW_\tau = \sum_{y_i \geq \tau} \tau|y_i - \hat{\tau}| + \sum_{y_i < \tau}(1-\tau)|y_i - \hat{\tau}| \tag{11}$$

The value of $RAWS_\tau$ (Residual Absolute Sum of Weighted) is always less than the value of $TASW_\tau$ (Total Absolute Sum of Weighted) so that the Pseudo R_τ^2 will be in the range 0 to 1. The closer the Pseudo R^2 value to one the model will be better. However, the virtues of Pseudo R^2 can not be used to test the overall goodness of fit for the model, it can only be used to test the merits of the selected quantile [1,2].

2.3 *Mean Square Error (MSE)*

The parameter estimate obtained is said to be good if it has a small bias and small variance. Therefore, to see the goodness of estimating the parameters based on the bias and variance values simultaneously, represented in the value of *Mean Square Error* (MSE) [8, 9, 10], formulated as follows :

$$MSE(\hat{\beta}_j(\tau_q)) = Var\left((\hat{\beta}_j(\tau_q)\right) + Bias\ (\hat{\beta}_j(\tau_q))^2 \tag{12}$$

where :

$MSE(\hat{\beta}_j(\tau_q))$: value of MSE for $j = 1,2,\ldots p$
$\tau_q = quantil\ q = 1,2,\ldots,k$

$Var\left((\hat{\beta}_j(\tau_q)\right)$: variance for selected quantile

$$Var\left((\hat{\beta}_j(\tau_q)\right) = \frac{n\sum_{i=1}^n(\hat{\beta}_j(\tau_q))^2 - \left(\sum_{i=1}^n(\hat{\beta}_j(\tau_q))\right)^2}{n(n-1)}$$

$Bias\ (\hat{\beta}_j(\tau_q))$: the value of bias for selected quantile is obtained from the mean of the difference of the expected value and the estimated value, or :

$$Bias\ (\hat{\beta}_j(\tau_q)) = \frac{1}{n}\sum_{i=1}^n\left((\hat{\beta}_j(\tau_q)\right) - \left(\beta_j(\tau_q)\right)$$

RESULT AND DISCUSSION

We describe our approach to quantile regression by conducting the simulation study. In this research, we design two covariates each measuring 100 samples. The response variable, y_i is generated from the model :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \qquad i = 1,\ldots,100 \tag{13}$$

where covariate x_{i1} is generated from a standard normal distribution and x_{i2} is generated from exponential with one degrees of freedom. The parameter β_0, β_1 and β_2 are set to 1.2, 1, and 1.7 respectively. The data for error is also generated by taking the mean at zero and its variances have heteroscedasticity problems. We consider the heteroscedastic normal, $N\left(0, \sqrt{0.01 \text{ x } (\boldsymbol{X\beta})^2}\right)$ for distribution of error term.

In this example, we choose τ = 0.10, 0.25, 0.50, 0.75 and 0.90 as the quantile points for estimated. Table 1 below shows the parameter estimated and its corresponding standard error.

Table 1. Quantile Regression Estimates

τth Quantile	Parameter	Estimates	Standard error
0.10	b0	1.0251*	0.0308
	b1	0.8004*	0.0220
	b2	1.4783*	0.0219
0.25	b0	1.0891*	0.0310
	b1	0.8438*	0.0222
	b2	1.5931*	0.0220
0.50	b0	1.1840*	0.0417
	b1	0.9955*	0.0298
	b2	1.7640*	0.0296
0.75	b0	1.2829*	0.0435
	b1	1.0871*	0.0312
	b2	1.8561*	0.0309
0.90	b0	1.3215*	0.0287
	b1	1.1514*	0.0205
	b2	2.0225*	0.0204

(* significant at α = 0,05)

Table 1 informs us that estimated coefficient ($\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$) for all quantile points are close to initial value, $b_0 = 1.2$ $b_1 = 1$ and $b_2 = 1.7$. For example, consider at 0.50th quantile, proposed parameter estimated are $\hat{\beta}_0 = 1.1840$, $\hat{\beta}_1$ = 0.9955 and $\hat{\beta}_2$= 1.7640.

The next analysis in the quantile regression is a consistency test of the proposed model to reveal the performance of the quantile approach and its associated algorithm in recovering the true parameters of the quantile regression analysis. Consistency test is done by doing simulation study. Simulation study does so by generating a set of new data set by sampling with replacement from the original data set, and fitting the model to each new data set [11, 12]. To compute standard errors for calculating the 95% confidence interval of all parameters in this study, roughly 25 model fits are determined. The goodness of fit of each model are also calculated. Table 2 presents the result taken from the simulation study.

Table 2. Simulation Results of 25 Data Sets Using Quantile Regression Approach

τ -th Quantile	Parameter	Paramter Estimated (Standard Error)	95% Interval Confidence	Pseudo R^2
0.10	b0	1.0317* (0.0439)	(1.0233 ; 1.0401)	0.8104
	b1	0.8974* (0.0312)	(0.8898 ; 0.9049)	
	b2	1.5091* (0.0321)	(1.4973 ; 1.5208)	
0.25	b0	1.0992* (0.0322)	(1.0928 ; 1.0992)	0.8288
	b1	0.9334* (0.0229)	(0.9270 ; 0.9398)	
	b2	1.1613* (0.0229)	(1.6000 ; 1.6226)	
0.50	b0	1.1912* (0.0301)	(1.1861 ; 1.1972)	0.8477
	b1	0.9949* (0.0220)	(0.9906 ; 0.9992)	
	b2	1.6973* (0.0218)	(1.6897 ; 1.7050)	
0.75	b0	1.2837* (0.0341)	(1.2747 ; 1.2928)	0.8655
	b1	1.0533* (0.0243)	(1.0478 ; 1.0589)	
	b2	1.8074* (0.0244)	(1.7925; 1.8223)	
0.90	b0	1.3693* (0.0414)	(1.3550 ; 1.3837)	0.8854
	b1	1.1064* (0.0299)	(1.0961 ; 1.1168)	
	b2	1.9164* (0.0295)	(1.8976 ; 1.9352)	

(* significant at α = 0,05)

Table 2 informs us that the estimated value of the model parameter coefficients ($\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$) on each quantile are close to the initial value ($b_0 = 1.2$ $b_1 = 1$ and $b_2 = 1.7$). For example,

consider the 50th quantile, the estimated value for $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ are 1.1916, 0.9949, and 1.6973 respectively. Based on Table 2, we also know that all parameter estimated values fall within 95% confidence intervals obtained from the simulation study. It means quantile 95% confidence interval seem to work well here and parameter estimated are acceptable.

The goodness of fit for the each quantile regression model is presented by the value of Pseudo R^2, shown in the last column in Table 2. All Pseudo R^2 values obtained here are more than 80%, indicating that all proposed model are adequate and could be accepted.

In this study we also determine the value of of MSE (Mean Square Error) to ensure that parameter estimated have small bias and small variance. Table 3 below presents the MSE value of the quantile regression method for all three parameter estimated at corresponding quantile points.

Table 3. The Value of MSE For Any Quantile Points

Parameter	MSE (Mean Square Error)				
	0.10	0.25	0.50	0.75	0.90
b0	0.0011	0.0011	0.0008	0.0022	0.0054
b1	0.0015	0.0012	0.0005	0.0008	0.0028
b2	0.0037	0.0033	0.0015	0.0058	0.0093

Table 3 above gives information that all parameter estimated have small MSE. These results indicate that quantile regression method is able to produce unbiased parameter model since it has small bias and small variance.

Based on any results here, we could believe that the power of our quantile regression method result the best fit for the model althought in the existence of heteroscedastic error variance.

CONCLUSIONS

This present study purposes to describe the performace of quantile regression method in modeling the data containing non-uniform variance problem (heteroscedasticity). A data set with two covariates is generated, each measuring 100 samples. The distribution for error term is heteroscedastic normal, $N\left(0, \sqrt{0.01 \text{ x } (\boldsymbol{X\beta})^2}\right)$ is designed to generate a response variable. Each parameter model are also set as initial values. A simulation study is done to check the power of quantile regression algorithm.

This study resulted that all parameter estimated are close to the initial values. The value of Pseudo R^2 for all proposed model at any selected quantile points are quite large, more than 80%. Based on simulation study, the value of parameter estimated are within 95% confidence intervals indicating that parameter estimated could be accepted. This study also result that quantile regression method is able to produce small value of MSE. Therefore, it could be concluded here that quantile regression methods is unbiased estimator method and could result the acceptable model although in the due of heteroscedasticity problem of error variance.

REFERENCES

[1] Ferra Y, Hazmira Y and Izzati R. 2016. Penerapan Metode Regresi Kuantil pada Kasus Pelanggaran Asumsi Kenormalan Sisaan. *Eksakta*, 1 (XVII) : 33 – 37.

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

[3] Arbia, G. 2006. *Spatial Econometrics: Statistical Foundation Application to Regional Convergence.* Springer, Berlin

[4] Koenkar.R. and Basset,G.Jr.1978. Quantiles Regression. *Econometrica*. **46**. 33-50

[5] Kozumi H and Kobayashi G. 2011. Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation*, **81** (11) : 1565 – 1578.

[6] Feng X and Zhu L. 2016. Estimation and Testing of Varying Coefficients in Quantile Regression. *Journal of the American Statistical Association* **111**, 266 – 274.

[7] Yanuar F. 2013. Quantile Regression Approach to Determine the Indicator of Health Status. *Scientific Research Journal* , I (IV): 17 – 23.

[8] Yang Y, Wang HJ and He X. 2015. Posterior inference in Bayesian quantile regression with asymmetric Laplace likelihood. *International Statistical Review*, 0 0, 1-18 doi: 10.1111.insr.12114

[9] He X. and Zhu LX. 2011. A lack of fit test for quantile regression. *Journal of the American Statistical Association*, **98** (464) : 1013 - 1022

[10] Feng, X., He, X., and Hu, J. 2011. Wild Boostrap for Quantile Regression. B*iometrica*, **98**, 995–999.

[11] Yanuar F, Ibrahim K and Jemain AA. 2013. Bayesian structural equation modeling index in the health model. *Journal of Applied Statistics*, **40** (6) : 1254–1269.

[12] Yanuar F. 2014. The Estimation Process in Bayesian Structural Equation Modeling Approach. *Journal of Physics : Conference Series*, **495,** 012047.