Applying bootstrap quantile regression for the construction of a low birth weight model

Ferra Yanuar
Department of Mathematics, Faculty of Mathematics and Natural Sciences, Andalas University, Indonesia,
ferrayanuar@sci.unand.ac.id

Hazmira Yozza
Department of Mathematics, Faculty of Mathematics and Natural Sciences, Andalas University, Indonesia

Firdawati Firdawati
Department of Public Health, Faculty of Medicine, Andalas University, Indonesia

Izzati Rahmi
Department of Mathematics, Faculty of Mathematics and Natural Sciences, Andalas University, Indonesia

Aidinil Zetra
Department of Political Science, Faculty of Social and Political Sciences, Andalas University, Indonesia

Applying bootstrap quantile regression for the construction of a low birth weight model

Ferra Yanuar[1*], Hazmira Yozza[1], Firdawati[2], Izzati Rahmi[1], Aidinil Zetra[3]

1. Department of Mathematics, Faculty of Mathematics and Natural Sciences, Andalas University, Padang 25163, Indonesia
2. Department of Public Health, Faculty of Medicine, Andalas University, Padang 25163, Indonesia
3. Department of Political Science, Faculty of Social and Political Sciences, Andalas University, Padang 25163, Indonesia

*E-mail: ferrayanuar@sci.unand.ac.id

Abstract

Background: Most investigators use ordinary least squares (OLS) methods to model low birth weight. When the data are non-normal or contain outliers, OLS become ineffective. However, the quantile method of forecasting low birth weight has not been fully evaluated, although it has good potential for overcoming problems associated with linear regression. **Methods**: The present study reports our comparison between the OLS and quantile regression methods for modeling low birth weight when the data are right skewed and outliers are presented. Additionally, we evaluated the performance of the associated algorithm in recovering the true parameter using the bootstrap method. **Results**: Our study found that a mother's education level, the number of maternal parities, and the last birth interval significantly impacted low birth weight at any selected low quantile. Based on the bootstrap simulation study, the proposed model was considered to be acceptable since both methods generated nearly identical estimates of the parameter model. An accuracy test proved that the quantile method was an unbiased estimator. **Conclusions**: The present study found that low birth weight is significantly affected by the mother's educational level, the number of maternal parities, and the last birth interval.

Keywords: bootstrap approach, low birth weight, quantile regression

Introduction

Birth weight in humans is described as the weight of an infant obtained within the first 60 minutes after birth. Birth weight is determined by two major processes: length of gestation and intrauterine growth rate.[1] Low birth weight is defined by the World Health Organization as a birth weight < 2500 grams. Low birth weight can be caused by either a short gestation period, retarded intrauterine growth, or a combination of the two.[1,2] It is considered to delay child development and carries a greater risk of early childhood mortality. Moreover, infants with low birth weight also have a significantly greater risk of infection, decreased chances of survival, higher susceptibility to childhood illnesses, and difficulties associated with psychosocial development, behavior, and learning during childhood.[3] Over the last few decades, many studies have investigated the causes of low birth weight. Recently, low birth weight and its determinants have come under intense global scrutiny.

Conventional regression methods, such as the ordinary least squares (OLS) method, are typically used to model the factors affecting birth weight. OLS is based on a central tendency, which may not appropriately represent the reality in cases where the dependent variable ranges between the lower and upper values; hence, the relationship may not be homogenous across different percentiles of the dependent variables. Thus, using OLS to estimate the mean may not accurately reflect or represent heterogeneity in the estimated relationship. However, studies have shown that the resulting estimates of various effects on the conditional mean of birth weight do not necessarily indicate the size and nature of these effects on the lower tail of the birth weight distribution.[4,5]

A more complete picture of the covariate effects can be seen by estimating a family of conditional quantile functions. Estimates of conditional quantiles can be used overcome any problem associated with the classical method (OLS), such as outlier data or heteroscedasticity cases, as long as the error distribution of the data has a continuous, symmetric, and unimodal density.[6,7] The quantile regression method is used to estimate the relationship at any point of conditional distribution of the dependent variable, which generates various estimated coefficients at certain quantiles of the dependent variable.[8–10] The objective of this study is to identify the determinants of low birth weight using quantile regression. We report a quantile regression model for

settings in which significant variables indicate scientifically determined variables. Quantile regression attempts to divide the response distribution into many parts, including estimating conditional quantile functions.[11−13]

Methods

This study utilized primary data collected by questionnaire distributions from March through July 2016. Our sample was limited to mothers who just delivered a singleton live birth and were living in West Sumatera, Indonesia. In total, 92 respondents with complete information were included in the analysis. The response variable was the child's birth weight recorded in kilograms. Eleven indicator variables were used in this study, including continuous and categorical types, i.e., mother's education, mother's job, residence, number of pregnancy problems, mother's age, number of parities, number of prenatal care visits, mother's weight gain during pregnancy, mother's hemoglobin (Hb) level, last birth interval, and sex of the baby.[1]

Mother's education was divided into three levels: low, middle, and high level, with the low level considered a reference category for interpreting coefficients. Mother's job was classified into three categories: government employee, housewife, and other, and residence was categorized as urban or rural. The number of pregnancy problems was categorized into three types: > 1 problem (reference category), 1 problem, and no problem. Meanwhile, mother's age, the number of parities, the number of prenatal care visits, mother's weight gain during pregnancy, mother's Hb level, and last birth interval were represented by continuous variables. This research has been conducted in full accordance with the World Medical Association Declaration of Helsinki.

Figure 1 (a) presents a histogram of the dependent variable of 92 birth weights. Distribution of data is skewed to the right. Figure 1 (b) demonstrates a normal Q-Q plot of the data, indicating a violation of the normality assumption in the birth weight data. Summary statistics were calculated for all of the selected independent variables. Table 1 presents the descriptive statistics for all of the continuous independent variables, and Table 2 shows the percentage of each category of qualitative variables. In the present study, the quantile regression approach was used to model low birth weight based on the following ideas:

Considering a linear model,[14]

$$y_i = x_i^T \beta + e_i, i = 1, \ldots \ldots, n \tag{1}$$

where y_i is the ith observation, x_i is the ith independent variable, and e_i is an independent error variable with probability density f_i. For identifiability, we assume that,

for a quantile level of interest $\tau \in (0,1)$, the conditional τth quantile of e_i given x_i is zero. The conditional quantile regression is as follows:
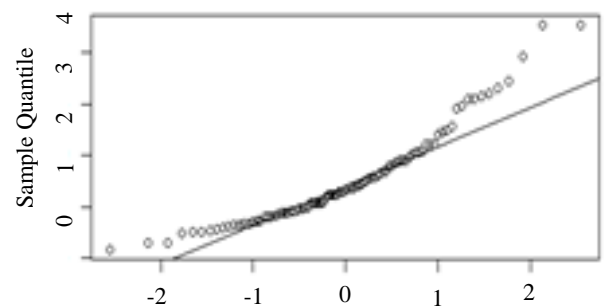
$$Q_Y(\tau|\mathbf{x}) = x^T \beta(\tau) \tag{2}$$

where $Q_Y(\tau|\mathbf{x})$ represents the τth conditional quantile of the response Y given $\mathbf{x}$ and parameter $\beta(\tau)$ is an unknown functional vector. A point estimate $\hat{\beta}(\tau)$ of the parameter $\beta(\tau)$ is obtained by minimizing the objective function:

$$\sum_{i=1}^{n} \rho_\tau(y_i - x_i^{T\prime}\beta) \tag{3}$$
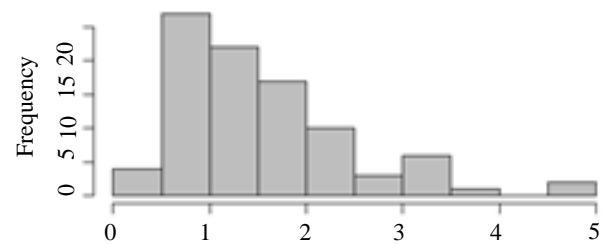
(a)

(b)

Figure 1. (a) Histogram and (b) normal QQ plot for birth weight data

Table 1. Summary statistics for continuous independent variables of Birth Weight Data

Variables	Mean	Standard Deviation	Min.	Max.
Mother's age	29.87	6.58	17.00	45.00
Parity	2.18	1.19	1.00	7.00
Last birth interval (year)	2.77	2.86	0.00	14.00
Mother's weight gain (kg)	12.65	5.06	2.00	28.00
Hb	11.27	1.25	7.40	14.30
Prenatal care	7.64	2.68	1.00	17.00
Birth weight (kg)	3.06	0.67	1.10	4.50

Table 2. Summary statistics for continuous independent variables of Birth Weight Data

Variable	Frequency	Percentage (%)
Maternal education		
Low (Junior high school and below)	20	21.7
Low (Junior high school and below)	20	21.7
Middle (Senior high school)	41	44.6
High (University)	31	33.7
Mother's job		
Government employee	13	14.1
Housewife	59	64.2
Others	20	21.7
Residence		
Urban	66	71.7
Rural	26	28.3
The number of pregnancy		
> 1 problem	4	4.3
One problem	31	33.7
No problem	57	62.0
Sex of the child		
Female	42	45.7
Male	50	54.3

where $\rho_\tau(.)$ denotes the following loss function:

$$\begin{aligned}\rho_\tau(u) &= u\big(\tau - I(u < 0)\big) \\ &= |u|\big((1 - \tau)I(u \le 0) + \tau I(u > 0)\big)\end{aligned} \quad (4)$$

and $I.$ is the usual indicator function. Such loss function is then an asymmetric absolute loss function, i.e., a weighted sum of absolute deviations, where a $(1 - \tau)$ weight is assigned to the negative deviations and a τ weight is used for the positive deviations.[6]

We evaluated goodness of fit for these quantile regressions using R^2 values. The R^2 index formulation for quantile regression differs from OLS regression since it is based on the minimization of an absolute weighted sum (not an unweighted sum of squares as in OLS). The R^2 formulation for quantile regression is represented by what is typically called a pseudo-R^2, which is formulated as follows:

$$\text{Pseudo} - R^2 = 1 - \frac{RASW_\tau}{TASW_\tau} \quad (5)$$

where $RASW_\tau$ is the residual absolute sum of weighted differences between the observed dependent variable and the estimated quantile of conditional distribution in the more complex model, and $TASW_\tau$ is the total absolute sum of weighted differences between the observed dependent variable and the estimated quantile of conditional distribution in the simplest model.[4]

We evaluated the performance of the quantile method and its associated algorithm in recovering the true parameter using a simulation study, which was performed by applying the bootstrap method.[8] The bootstrap resampling method is a fully nonparametric

procedure that is suitable for use in a wide range of models and easy to implement. In this method, a new data set is generated by sampling with replacement from the original data set, and the hypothesis model is fitted to the new data set.[8,15,16]

The estimation of standard errors for parameters was obtained by fitting the hypothesis model to the new data set.

Previous study[7] presented the following procedure to perform bootstrap sampling as follows: (1) Generate a random sample of size n from the original data denoted by $X^* = (X_1^*, X_2^*, \ X_n^*)$; (2) For this one bootstrap sampling, apply the quantile function estimator to each element of X^* to obtain $U^* = (U_1^*, U_2^*, ..., U_n^*)$ where $U_1^* = \widehat{Q}(X_1^*)$; (3) Calculate the statistic of interest $S_i(U^*)$; (4) Repeat steps 1–3 for B times in order to obtain the empirical bootstrap distribution for S(U), for B = 1,…,reB; (5) Calculate bootstrap parameter average value using

$$\bar{\bar{\beta}}(\tau) = \frac{1}{B}\Sigma_{b=1}^B \hat{\beta}_b(\tau), \quad (6)$$

and bootstrap variance as follows:

$$\widehat{V}_{q,j} = \frac{1}{B}\Sigma_{b=1}^B \left(\hat{\beta}_{b,j}(\tau_q) - \bar{\bar{\beta}}_{b,j}(\tau_q)\right)\left(\hat{\beta}_{b,j}(\tau_q) - \bar{\bar{\beta}}_{b,j}(\tau_q)\right)^T, \quad (7)$$

where $j = 1, \text{rep}; q = 1, \text{rek}$; (6) Next, construct a confidence interval for each conditional quantile parameter for the generic jth parameter and the qth quantile using the following formula:

$$\bar{\bar{\beta}}_j(\tau_q) \pm z_{\alpha/2}\text{SD}\left(\hat{\beta}_j(\tau_q)\right), \quad (8)$$

where $\text{SD}\left(\hat{\beta}_j(\tau_q)\right)$ is a standard deviation of $\hat{\beta}_j(\tau_q)$ or the square root of bootstrap variance, $\widehat{V}_{q,j}$.

Results

In this study, the model hypothesis was presented in the birth weight equation as follows:

$$\begin{aligned}\text{Birth weight}_i = \; &\beta_1 \text{Age}_i + \beta_2\text{Education (Middle)}_i + \\ &\beta_3\text{Education (High)}_i + \beta_4\text{Parity}_i + \\ &\beta_5\text{Last birth interval}_i + \\ &\beta_6\text{Weight gain}_i + \\ &\beta_7\text{Problems (One problem)}_i + \\ &\beta_8\text{Problems (No problem)}_i + \\ &\beta_9\text{Hb}_i + \beta_{10}\text{Rural}_i + \\ &\beta_{11}\text{Female}_i + e_i; \quad (9)\end{aligned}$$

Next, the model hypothesis was fitted to the birth weight data. After fitting, four indicator variables were found to indicate a statistically significant effect on the response. The variable "problems" were excluded from the model because they were not statistically significant in any of the constructed equations. Table 3 presents the results of

conditional quantile regression of low quantiles (at quantiles 0.10, 0.20, 0.30, 0.40, and 0.50) and the OLS approach for significant variables only.

Our results demonstrate that standard errors from the OLS were the highest among the other standard errors for the corresponding indicators. At lower quantiles, 0.10,

0.20, and 0.30, all four indicator variables in the model were statistically significant. Meanwhile, quantiles 0.40 and 0.50 had three significant variables only, with the OLS method yielding the same results. In spite of this, we could not accept the OLS model as a proposed model as it violated the normality assumption and presence of outliers in the empirical data.

Table 3. Comparison of coefficients estimated for low birth weight model using quantile regression

Indicator Variable	Estimate of QR (Standard error)					Estimate of OLS (Standard error)
	$\tau = 0.10$	$\tau = 0.20$	$\tau = 0.30$	$\tau = 0.40$	$\tau = 0.50$	
β_2 (Middle)	0.391 (0.071)**	0.374 (0.128)**	0.416 (0.167)**	0.186 (0.144)	0.121 (0.143)	0.117 (0.250)
β_3 (High)	0.494 (0.075)**	0.432 (0.137)**	0.518 (0.178)**	0.492 (0.153)**	0.407 (0.152)**	0.531 (0.266)**
β_4 (Parity)	0.207 (0.030)**	0.249 (0.054)**	0.320 (0.071)**	0.401 (0.061)**	0.556 (0.060)**	0.491 (0.106)**
β_5 (Last birth interval)	−0.042 (0.012)**−	−0.052 (0.022)**	−0.059 (0.029)**	−0.077 (0.025)*	−0.092 (0.024)**	−0.091 (0.043)**

* Significant at 10% level
** Significant at 5% level

Table 4. Pseudo-R^2 for selected quantile for low birth weight

Quantiles	Pseudo-R^2
0.10	0.794
0.20	0.849
0.30	0.861
0.40	0.909
0.50	0.881

Table 5. Bootstrap simulation study results

Parameter Model	β_2 (Middle)	β_3 (High)	β_4 (Parity)	β_5 (Last birth interval)
$\tau = 0.10$				
Quantile Method	0.391 (0.071)**	0.494 (0.075)**	0.207 (0.030)**	−0.042 (0.012)**
Bootstrap Method	0.389 (0.056)**	0.478 (0.066)**	0.244 (0.043)**	−0.053 (0.025)**
95% CI Boot	(0.297;0.471)	(0.383; 0.572)	(0.187; 0.321)	(−0.105; −0.021)
Length of 95% CI Boot	0.174	0.189	0.134	0.083
$\tau = 0.20$				
Quantile Method	0.374 (0.128)**	0.432 (0.137)**	0.249 (0.054)**	−0.052 (0.022)**
Bootstrap Method	0.397 (0.085)**	0.489 (0.107)**	0.295 (0.056)**	−0.062 (0.026)**
95% CI Boot	(0.265; 0.509)	(0.331; 0.666)	(0.213; 0.404)	(−0.121; −0.025)
Length of 95% CI Boot	0.243	0.335	0.191	0.096
$\tau = 0.30$				
Quantile Method	0.416 (0.167)**	0.518 (0.178)**	0.320 (0.071)**	−0.059 (0.029)**
Bootstrap Method	0.357 (0.148)**	0.527 (0.149)**	0.375 (0.105)**	−0.070 (0.034)**
95% CI Boot	(0.097; 0.588)	(0.273; 0.732)	(0.248; 0.565)	(−0.122; −0.013)
Length of 95% CI Boot	0.491	0.459	0.316	0.108
$\tau = 0.40$				
Quantile Method	0.186 (0.144)	0.492 (0.153)**	0.401 (0.061)**	−0.077 (0.025)*
Bootstrap Method	0.269 (0.188)**	0.527 (0.211)**	0.468 (0.143)**	−076 (0.046)
95% CI Boot	(0.047; 0.604)	(0.170; 0.838)	(0.263; 0.659)	(−0.131; 0.006)
Length of 95% CI Boot	0.557	0.667	0.395	0.138
$\tau = 0.50$				
Quantile Method	0.121 (0.143)	0.407 (0.152)**	0.556 (0.060)**	−0.092 (0.024)**
Bootstrap Method	0.233 (0.207)**	0.549 (0.216)**	0.548 (0.138)**	−0.091 (0.045)
95% CI Boot	(−0.036; 0.616)	(0.278; 0.931)	(0.328; 0.748)	(−0.155; −0.009)
Length of 95% CI Boot	0.652	0.653	0.420	0.145

* Significant at 10% level
** Significant at 5% level

Table 6. Bias of quantile estimation for τ = 0.10, 0.20, 0.30, 0.40, and 0.50

Parameter	τ = 0.10		τ = 0.20		τ = 0.30		τ = 0.40		τ = 0.50	
	SD^*_{Boot}	Bias	SD^*_{Boot}	Bias	SD^*_{Boot}	Bias	SD^*_{Boot}	Bias	SD^*_{Boot}	Bias
β_2 (Middle)	0.056	0.010	0.085	0.023	0.148	0.058	0.188	0.083	0.207	0.112
β_3 (High)	0.066	0.016	0.107	0.057	0.149	0.009	0.211	0.035	0.216	0.142
β_4 (Parity)	0.043	0.037	0.056	0.046	0.105	0.055	0.143	0.067	0.138	0.007
β_5 (Last birth interval)	0.025	0.011	0.026	0.010	0.034	0.011	0.046	0.000	0.045	0.000
SD (Bias)	2.01×10^{-8}		$0.0002.21 \times 10^{-8}$		$0.0002.21 \times 10^{-8}$		$0.0002.21 \times 10^{-8}$		$0.0002.21 \times 10^{-8}$	

Next, we measured goodness of fit of the proposed models. Several studies have reported the use of the Pseudo-R^2 to indicate goodness of fit for each selected quantile.[6,17] Table 4 shows the corresponding Pseudo-R^2 values for each selected quantile using birth weight data as the response variable. The results shown in Table 4 indicate that the 0.40th quantile is the best among all five nested models, as indicated by the highest Pseudo-R^2 value.

Although the Pseudo-R^2 values for all five lower quantiles were within an acceptable range (> 79%), this study also investigated the performance of quantile regression and its associated algorithm in recovering the true parameter. A simulation study was subsequently performed using a bootstrap approach. All 50 model fits were used to measure standard errors to calculate the 95% confidence interval of all parameters in this simulation study. The result of the bootstrap estimation method and the 95% bootstrap percentile intervals are shown in Table 5, which reveals that the quantile regression and bootstrap models yield almost identical parameter estimates. Additionally, all parameter estimates from the quantile regression method were within the 95% bootstrap percentile intervals, indicating that the parameters estimated for all selected quantiles in the proposed model were acceptable. Thus, we can conclude that the power of this study's quantile regression method yields the best fit for the proposed model.[18]

We also examined the accuracy of the quantile estimation method to determine that it is unbiased. Table 6 presents the bias estimation results between the quantile and bootstrap estimation methods for each low quantile. Bias was calculated as the difference between quantile estimation and bootstrap estimation. The quantile estimation method was unbiased if the standard deviation of bias was less than the standard deviation of bootstrap distribution.

Discussion

This present study reports on a low birth weight statistical model constructed using a quantile regression approach. Although many studies have reported on models to determine low birth weight, few studies have used the quantile approach, particularly considering the mother's education level, the number of parities, last birth interval, mother's weight gain, Hb level, and the number of pregnancy problems.

These results reveal that the mother's education level, the number of parities, and the last birth interval significantly affected low birth weight. Furthermore, a validity test used the bootstrap resampling method, with results indicating acceptability of the proposed model since it yielded identical parameter estimates. All parameter estimates of quantile regression were within 95% bootstrap percentile intervals. Next, the accuracy of the quantile regression method was tested and determined to be unbiased. This study revealed that the standard deviation of bias (i.e., the difference between quantile estimation and bootstrap estimation) was less than the standard deviation of bootstrap, which means that the parameters estimated for all selected quantiles in this study are statistically acceptable.

More research using additional data samples (> 250) is necessary to achieve a better model since quantile regression itself requires a large data sample.[6] In addition, the Bayesian approach to quantile regression also could be implemented to overcome the need for a larger data sample since more data indicate more time and more money. The Bayesian method has the ability to estimate model parameters even using small data.[19]

Conclusions

The quantile regression approach is based on its ability to enhance the understanding of the low birth weight model, where data with outliers are available. Quantile regression has the ability to overcome this problem since it can assess the association between independent variables and outcome in each conditional quantile, hence it is applicable for all data with low moderate, and high outlier values. The present study demonstrated that low, birth weight is significantly affected by mother's education level, the number of parities, and the last birth interval. This proposed model could be accepted based on validity tests using the bootstrap resampling method. All significant parameter estimates from the quantile regression were within 95% bootstrap percentile intervals.

Acknowledgments

Funding

Conflict of Interest Statement

None of the authors have a conflict of interest in relation to this study.

References

1. Hulmán A, Witte DR, Kerényi Z. Heterogeneous effect of gestational weight gain on birth weight: Quantile regression analysis from a population-based screening. *Ann Epidemiol*. 2015;25:133−7.
2. Abrevaya J, Dahl CM. The effects of birth inputs on birthweight: Evidence from quantile estimation on panel data. *J Bus Econ Stat*. 2008;26:379−97.
3. Burgette LF, Reiter JP, Science S, Carolina N. Modeling adverse birth outcomes via confirmatory factor quantile regression. *Biometrics*. 2012;68:92−100.
4. Saputri OD, Yanuar F, Devianto D. Simulation study of the implementation of quantile bootstrap method on. *Cauchy - J Mat Murni dan Apl*. 2018;5:95−101.
5. Muharisa C, Yanuar F, Devianto D. Simulation study of the using of bayesian quantile regression in non- normal error. *Cauchy - J Mat Murni dan Apl*. 2018;5:121−6.
6. Davino C, Furno M, Vistocco D. *Quantile Regression Theory and Applications*. Wiley Series; 2014.
7. Wei L, Wang D, Hutson AD. an investigation of quantile function estimators relative to quantile confidence interval coverage. *Commun Stat Theory Methods*. 2015;44:2107−35.
8. Galvao A, Montes-Rojas G. On bootstrap inference for quantile regression panel data: A monte carlo study. *Econometrics*. 2015;3:654−66.
9. Olsen CS, Clark AE, Thomas AM, Cook LJ. Comparing least-squares and quantile regression approaches to analyzing median hospital charges. *Acad Emerg Med*. 2012;19:866−75.
10. Fallah R, Kazemnejad A, Zayeri F, Shoghli A. Birthweight related factors in Northwestern Iran: Using quantile regression method. *Glob J Health Sci*. 2016;8:116−25.
11. Wei Y, Carroll RJ. Quantile regression with measurement error. *J Am Stat Assoc*. 2009;104:1129−43.
12. Feng X, Zhu L. Estimation and testing of varying coefficients in quantile regression. *J Am Stat Assoc*. 2016;111:266−74.
13. Fosdal S. *The use of logistic regression and quantile regression in medical statistics*. 2017.
14. Spyroglou I, Gunter S, AC CE, AG R, Paraskakis E. A bayesian logistic regression approach in asthma persistence prediction. *Epidemiol Biostat Public Health*. 2018;15:1−14.
15. Yanuar F. The estimation process in bayesian structural equation modeling approach. *J Phys Conf Ser*. 2014;
16. Feng X, He X, Hu J. Wild bootstrap for quantile regression. *Biometrika*. 2011;98:995−9.
17. He X, Zhu L-X. A lack-of-fit test for quantile regression. *J Am Stat Assoc*. 2003;98:1013−22.
18. Yanuar F, Ibrahim K, Jemain AA. Bayesian structural equation modeling for the health index. *J Appl Stat*. 2013;40:1254−69.
19. Rahmadita A, Yanuar F, Devianto D. The construction of patient loyalty model using bayesian structural equation modeling approach. *Cauchy - J Mat Murni dan Apl*. 2018;5:73−9.