Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2024, 2024:69

https://doi.org/10.28919/cmbn/7978

ISSN: 2052-2541

TRUNCATED SPLINE QUANTILE REGRESSION MODEL ON PLATELET CHANGES IN DENGUE FEVER PATIENTS BASED ON BODY TEMPERATURE

ANISA[1], ANNA ISLAMIYATI[1,*], SITTI SAHRIMAN[1], JUSMAWATI MASSALESSE[2], UMMI SARI[3]

[1]Department of Statistics, Faculty of Mathematical and Natural Sciences, Hasanuddin University, Makassar, 90245, Indonesia

[2]Department of Mathematics, Faculty of Mathematical and Natural Sciences, Hasanuddin University, Makassar, 90245, Indonesia

[3]Hasanuddin University Teaching Hospital, Makassar, 90245, Indonesia

Abstract: Quantile regression is one of the models used for data containing outliers. The data that has been analyzed by many researchers using quantile regression is dengue fever data. In this study, we propose using quantile regression with truncated spline in modeling platelet data based on the body temperature of dengue patients. We use knot points ranging from 1, 2, to 3 points and the optimal model uses 3 knot points, namely 36, 37.1, and 39.1. In the 0.25 and 0.75 spline quantile models, four patterns of platelet changes were found that tended to be the same. In contrast to the 0.50 spline quantile model, it is seen that platelets actually decrease when the temperature is below 36°C. However, the three models showed that platelets decreased when the body temperature was high, reaching 39.1°C.

Keywords: dengue fever; knots; spline quantile.

2020 AMS Subject Classification: 92D30.

[*]Corresponding author

E-mail address: annaislamiyati701@gmail.com

Received April 01, 2023

1. INTRODUCTION

Nonparametric regression is one model that is able to model the pattern of the relationship between the response variable and the predictor variable. There are several estimators in nonparametric regression that have been developed by researchers, including spline [1], kernel [2], local polynomial [3], and Fourier series [4]. For splines, it has also developed rapidly both in terms of theory and application. There are several spline estimators, including truncated splines [5], penalized splines [6]–[8], and smoothing splines [9]. Spline is used when the data is not patterned parametric or we want to segment the pattern of changes in the data, and both in quantitative [10].

In subsequent developments, nonparametric regression, in general, has been developed for data containing outliers, for example, robust spline regression [11] and quantile smoothing spline regression [12]. Its more flexible use causes the resulting estimation results to be more accurate. Therefore, we use a quantile regression approach with a truncated spline estimator in this article. The advantage of the truncated spline is that it contains knots in the goodness of fit function, allowing us to detect patterns of change that can occur in event intervals in predictor variables [13].

Quantile regression is developed on data that has outliers. Koenker and Bassett (1978) first introduced a regression with conditional quantile distribution of the response variables expressed as covariate functions. Estimation of the quantile function of the response conditional distribution is carried out at various desired quantile values. Each quantile characterizes a certain point of the conditional distribution. The advantage of quantile regression is flexibility in modeling data with heterogeneous conditional distributions, especially when tails or extreme percentiles are interesting or when the continuous outcome variable of interest does not follow a normal distribution [14]. Quantile regression has been developed in multivariable cases eg anesthesia data using quantiles 10, 25, 50, 75, and 90 [15]. Quantiles can also consider the interaction effect between 2 factors, for example, between class size and class composition on educational attainment [16].

In this study, truncated spline quantile regression was used to analyze dengue fever (DHF) data. DHF is one of the infectious diseases that can cause death in a short time and often causes outbreaks in an area. This is because dengue has a very fast spread, but no cure has been found, nor has there been a vaccine for prevention. DHF is an environmental-based infectious disease that had become a significant cause of morbidity and mortality in the community. The indicators of the severity of DHF are low platelet count (thrombocytopenia) and plasma leakage which is characterized by hemoconcentration. The decrease in the number of platelets generally occurs before there is an increase in the hematocrit and occurs before the body temperature drops [17]. For this reason, in this study, we focused on the platelet count factor of DHF patients with their body temperature factor. This article is divided into 4 parts, namely the second part describes the materials and research methods used. The third section describes the results and discussion related to the truncated spline quantile regression model on DHF data and the last section is the conclusion of our article.

2. PRELIMINARIES

This study examines changes in the number of platelets based on the body temperature of DHF patients. Data were obtained from medical records at the Hasanuddin University Teaching Hospital, Makassar, Indonesia, and data were obtained for 158 patients. The response variable in this study was the platelets of patients with DHF, namely the results of laboratory tests of patients obtained on the first day of hospitalization which was expressed as 103 platelets/µl. Furthermore, the predictor variable is body temperature (x), which is the patient's body temperature measured on the first day the patient is hospitalized expressed in ºC.

In this article, we use a truncated spline quantile regression model to analyze the data. For example, given that data $\{x_i, y_i\}, i = 1, 2, \ldots, n$, is a paired set of random variables that are independently distributed and identical to quantile $\theta \in (0,1)$. The data has a conditional probability distribution function $F(y|x_i) = P(Y \leq y|x_i)$ and an inverse function $F^{-1}(\theta) = \inf\{y: F(y) \geq \theta\}$ which is defined as $Q_y(\theta) = \inf\{y: F(y) \geq \theta\} = F^{-1}(\theta)$ which is a quantile function θ of the response variable y.

If the relationship between the response variable and the predictor is expressed in a function

f whose form is unknown, then the relationship can be expressed in the form of a regression equation as follows:

$$y_i(\theta) = f(x_i) + \varepsilon_i(\theta) \tag{1}$$

Where the function f is the truncated spline function on each quantile θ with the order q and the number of knots as much as m which can be expressed as follows:

$$f(x_i) = \sum_{j=0}^{q} \beta_j(\theta) x_i^j + \sum_{k=1}^{m} \beta_{q+k}(\theta)(x_i - \tau_k)_+^q \tag{2}$$

Where x_i is the i-th predictor variable, β_0 is the intercept, β_j is the nonparametric logistic regression coefficient of the truncated spline, $\boldsymbol{\tau}_k$ is the knot point with $k = 1,2,\ldots,m,$ and $(x_i - \tau_k)_+^q$ is a truncated polynomial function which can be expressed as following:

$$(x_i - \tau_k)_+^q = \begin{cases} (x_i - \tau_k)_+^q & ; x_i \geq \tau_k \\ 0 & ; x_i < \tau_k \end{cases}$$

Equation (2) can also be expressed in the form of a matrix, namely:

$$\boldsymbol{f} = \boldsymbol{X\beta}(\theta) + \varepsilon(\theta) \tag{3}$$

Based on Equation (3), the truncated spline quantile regression model can be written as follows:

$$\boldsymbol{y}(\theta) = \boldsymbol{X\beta}(\theta) + \varepsilon(\theta)$$

Where $y(\theta) = \left(y_1(\theta)\, y_2(\theta) \cdots y_n(\theta)\right)^T$, $\varepsilon(\theta) = \left(\varepsilon_1(\theta)\, \varepsilon_2(\theta) \cdots \varepsilon_n(\theta)\right)^T$ and the matrix $\boldsymbol{X}$ can be written as follows:

$$X = \begin{bmatrix} 1 & x_1 & \cdots & x_1^q & (x_1 - \tau_1)^q & \cdots & (x_1 - \tau_m)^q \\ 1 & x_2 & \cdots & x_2^q & (x_2 - \tau_1)^q & \cdots & (x_2 - \tau_m)^q \\ \vdots & \vdots & \ddots & \cdots & \vdots & \vdots & \vdots \\ 1 & x_n & \cdots & x_n^q & (x_n - \tau_1)^q & \cdots & (x_n - \tau_m)^q \end{bmatrix}$$

The quantile regression model with truncated splines contains the parameter β which is estimated using the least absolute deviation (LAD) method which considers the weighting θ of positive errors (upper deviation) and weighting $(1-\theta)$ for negative errors (lower deviation). Therefore, the parameter β estimation is obtained by minimizing the following estimation criteria:

$$\hat{\beta}(\theta) = \min_{\beta(\theta)} \left\{ \theta \sum_{i=1}^{n} |\varepsilon_{i1}(\theta)| + (1-\theta) \sum_{i=1}^{n} |\varepsilon_{i2}(\theta)| \right\}$$

The optimal knot point for the best spline regression model is obtained from the smallest GCV value. The GCV method can be written as follows:

$$GCV(\tau) = \frac{MSE(\tau)}{(n^{-1}trace[\boldsymbol{I} - \boldsymbol{A}(\tau)])^2}$$

Where $MSE(\tau) = n^{-1}\sum_{i=0}^{n}(y_i - \hat{y})^2$, $\boldsymbol{I}$ is the identity matrix, and $\boldsymbol{A}(\tau) = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

3. MAIN RESULTS

In this study, 158 DHF data were analyzed with the average platelets being 137.45 x 10^3 platelets/μl from the normal platelet size in the blood which is around 150-400 x 10^3 platelets/µl. The average body temperature of DHF patients is 37.53℃ from the normal body temperature ranging from 36.1℃-37.7℃. This shows that the average platelet count of DHF patients is below the normal number and the patient's body temperature is still in the normal range. Next, we make a boxplot for data on platelet counts and body temperature to show the presence of outliers in the data as shown in Figure 1.

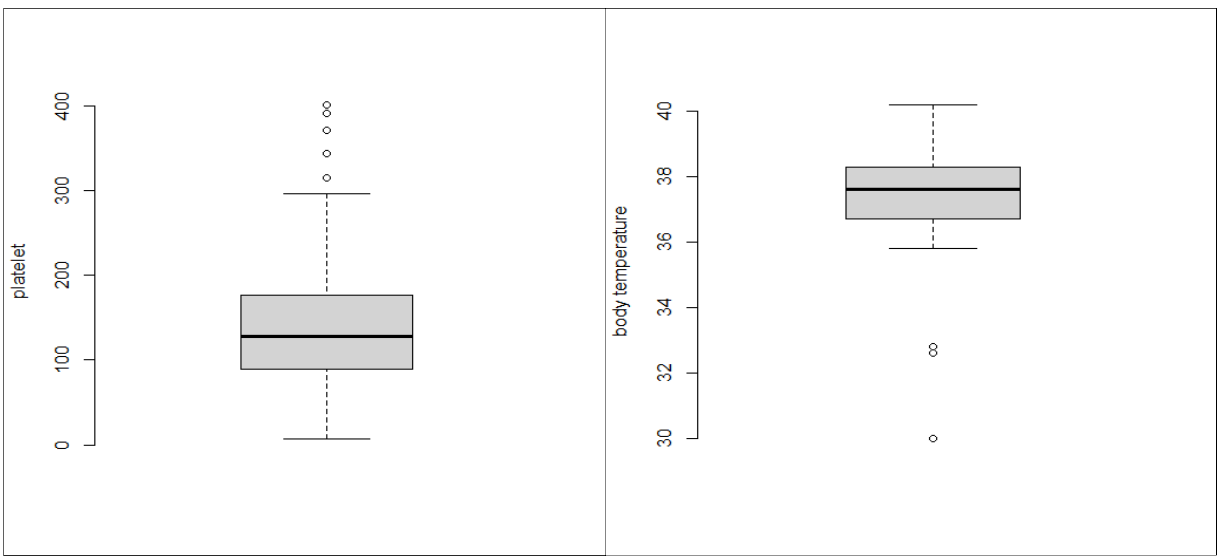

FIGURE 1. Boxplot on data on platelet count and body temperature of DHF patients

We can see in Figure 1 that there are outliers in data on platelet counts and body temperatures. The scatter plot shown in Figure 2 shows that the data plot does not follow a parametric form so the use of parametric quantiles is not appropriate for the data. Therefore, we used a nonparametric quantile regression model using a truncated spline estimator.

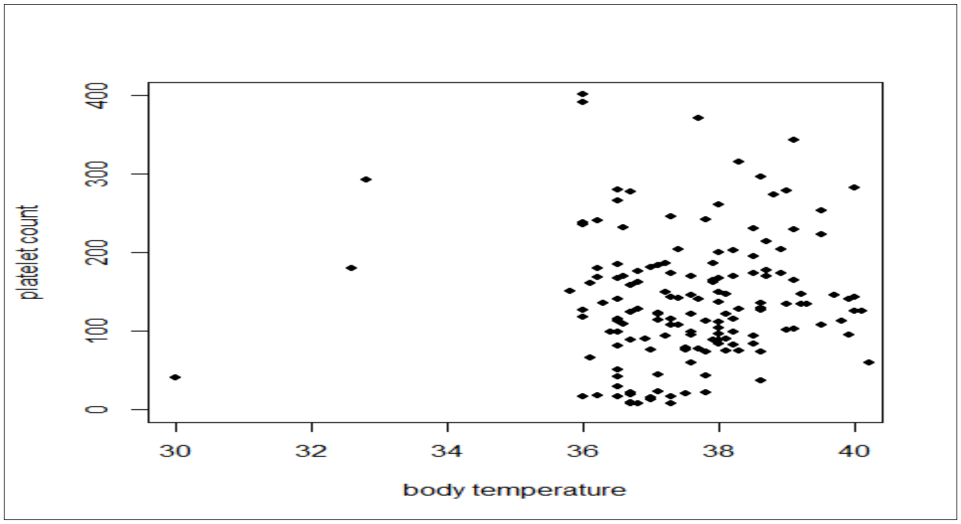

FIGURE 2. Scatter plot of patient platelet data based on body temperature

To model the data, we used a truncated spline quantile regression model of order 1 and tested it at 1-3 knot points. The selected quantiles are 0.25, 0.50, and 0.75 with the following regression model:

$$y(\theta) = \beta_0(\theta) + \beta_1(\theta)x_i + \sum_{\mathrm{k}=1}^{3} \beta_{1+k}(\theta)(x_i - \tau_k)_+, \theta = 25, 50, 75$$

The analysis stage begins with selecting the optimal knot point for each quantile. The knot point starts from 1 to 3 knots. GCV values are shown in Table 1. Based on the results of the selection of knot points, Table 1 only shows 1-knot point from the many knot points that can be formed in the data. The values in Table 1 are knot points which give GCV values at 1,2, and 3-knot points, respectively. At 1 knot point, which gives a minimum GCV value of 37.1 at 6395.62, at 2-knot points at 36 and 37.1 with a GCV value of 6278.57 and 3-knot points, namely 36, 37.1, and 39.1 at 6055.11. This value indicates that the minimum GCV value is given at the use of 3-knot points. Therefore, the data were analyzed by spline quantile regression model using 3-knot points.

TABLE 1. GCV values for 1, 2, and 3 knot points

1 Knot	GCV	2 Knot		GCV	3 Knot			GCV
τ_1	value	τ_1	τ_2	value	τ_1	τ_2	τ_3	value
37.10	6395.62	36.00	37.10	6278.57	36.00	37.10	39.10	6055.11*
37.00	6395.65	36.10	37.10	6314.16	36.00	37.10	39.20	6072.09
37.20	6402.17	37.00	39.00	6335.76	36.00	38.80	39.30	6077.61
36.90	6405.25	36.20	37.10	6344.03	36.00	37.10	39.50	6085.02
37.30	6408.25	37.00	38.80	6358.30	36.10	38.50	39.10	6098.66
36.80	6417.42	36.90	39.00	6364.66	36.10	37.10	39.20	6115.19
37.40	6419.84	36.90	38.80	6388.19	36.10	38.80	39.30	6120.45
37.50	6426.80	36.80	39.00	6395.52	36.00	37.10	39.70	6126.78
36.70	6430.81	37.00	38.50	6415.96	36.10	38.10	39.50	6127.12
37.60	6443.57	36.80	38.80	6419.43	36.20	38.50	39.10	6135.65

*) : minimum GCV value

Based on Figure 3, it can be seen that in each spline quantile regression curve there are several patterns of changes that occur in the patient's platelet count. This change corresponds to the optimal knot point obtained in Table 1. The regression curve used for the data is the regression model in Figure 3(c), namely the quantile regression curve at 3-knot points. This is because at the 3-knot points it gives the smallest GCV value. The quantile regression model with truncated splines with 3 optimal knot points $\theta = 0.25, 0.50$, dan 0.75 is stated as follows:

$$\hat{y}_i(0.25) = -95 + 4.53x_i - 26.53(x_i - 36) + 69.78(x_i - 37.1) - 103.5(x_i - 39.1)$$

$$\hat{y}_i(0.50) = 474.43 - 9.06x_i - 34.89(x_i - 36) + 76.54(x_i - 37.1) - 72.58(x_i - 39.1)$$

$$\hat{y}_i(0.75) = -357.94 + 16.47x_i - 114.97(x_i - 36) + 149.67(x_i - 37.1) - 146.73(x_i - 39.1)$$

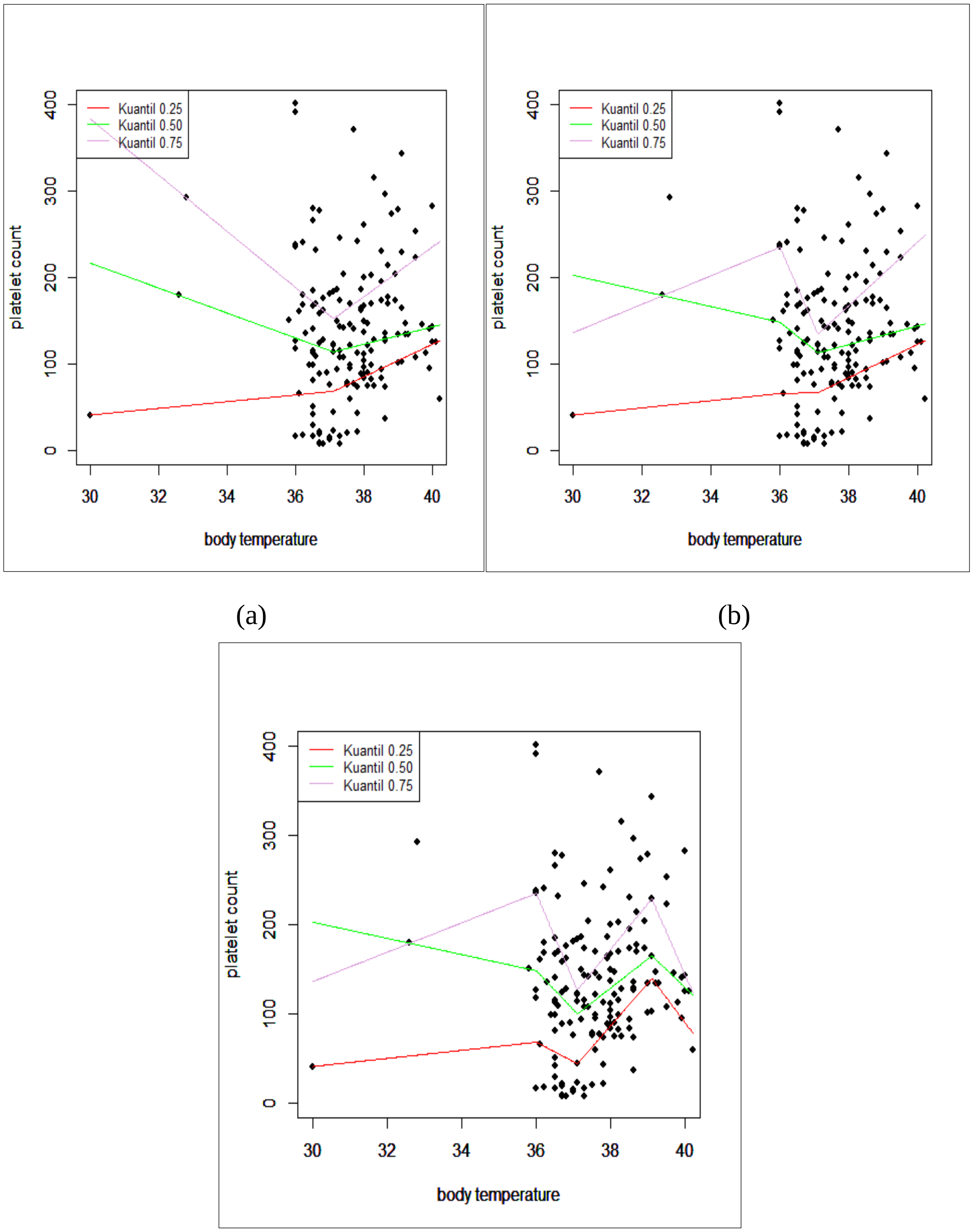

FIGURE 3. Spline quantile regression curve estimation at (a) 1 knot point, (b) 2 knot point, (c) 3 knot point

The optimal model at quantile $\theta = 0.25$ corresponds to Figure 3 (c) for the red curve. The results show that for data around the 0.25 quantile, the $\hat{y}_i(0.25)$ model should be used. From the model estimation, it can be seen that when the temperature is below 36ºC, platelets tend to increase by 4.53 x 10^3 platelets/µl, but after reaching 36ºC, platelets have decreased by 26.53 x 10^3 platelets/µl. The increase in body temperature causes immediate medical action and it can be seen that when the body temperature reaches 37.1ºC, the number of platelets has increased quite significantly around 69.78 x 103 platelets/µl. However, when the temperature got higher reaching 39.1ºC or more, the platelet count decreased very drastically, which was around 103.5 x 103 platelets/µl. Furthermore, it can be seen that the quantile $\theta = 0.50$ which corresponds to the green color has a different pattern for the 0.25 and 0.75 quantiles. It can be seen that when the temperature is below 36ºC and above, the platelet count actually decreases. This certainly needs attention because a decrease in platelets can occur due to other factors that must immediately get a doctor's treatment. The increase in platelets will occur when the body temperature reaches 37.1ºC and falls again after the temperature increases to 39.1ºC or above. For the quantile $\theta = 0.75$, the model corresponds to the purple regression curve and shows the same pattern as the 0.25 quantile, only the difference in the change in the platelet count.

The overall results of the spline quantile regression model showed that there was a change in the condition of DHF patients, especially in their platelet count based on body temperature. Of course, some segmentation of the change pattern obtained from the spline quantile regression model can be a reference in the treatment of DHF.

4. CONCLUSION

Spline quantile regression is used on data that is not parametrically patterned and has outliers. As is the case with DHF data, there tend to have outlier data due to the wide variety of patient conditions. Based on the results of the model estimation, there are 3 optimal spline quantile regression models at the use of 3-knot points. There are different patterns of change in each quantile. There are times when the change in platelets decreases and there are times when they increase again. However, the overall model shows that all models in the 3 quantiles, namely $\theta = 0.25, 0.50, \text{and } 0.75$ have a very drastic drop pattern when the body temperature is high, reaching 39.1ºC or above that temperature.

ACKNOWLEDGMENTS

Many thanks to LP2M Hasanuddin University for Internal Research Grants for Unhas Basic Research Scheme in 2022 with research contract No: 1474/UN4.22/PT.01.03/2022 dated June 9, 2022.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] A. Islamiyati, A. Kalondeng, N. Sunusi, et al. Biresponse nonparametric regression model in principal component analysis with truncated spline estimator, J. King Saud Univ. - Sci. 34 (2022), 101892. https://doi.org/10.1016/j.jksus.2022.101892.

[2] R. Hidayat, I.N. Budiantara, B.W. Otok, et al. The regression curve estimation by using mixed smoothing spline and kernel (MsS-K) model, Commun. Stat. - Theory Methods 50 (2020), 3942–3953. https://doi.org/10.1080/03610926.2019.1710201.

[3] N. Chamidah, B. Lestari, T. Saifudin, Modeling of blood pressures based on stress score using least square spline estimator in bi-response non-parametric regression, Int. J. Innov. Creat. Chang. 5 (2019), 1200-1216.

[4] M.F.F. Mardianto, E. Tjahjono, M. Rifada, Semiparametric regression based on three forms of trigonometric function in fourier series estimator, J. Phys.: Conf. Ser. 1277 (2019), 012052. https://doi.org/10.1088/1742-6596/1277/1/012052.

[5] A. Islamiyati, Raupong, A. Kalondeng, et al. Estimating the confidence interval of the regression coefficient of the blood sugar model through a multivariable linear spline with known variance, Stat. Transit. 23 (2022), 201-212.

[6] A. Islamiyati, Fatmawati, N. Chamidah, Estimation of covariance matrix on bi-response longitudinal data

analysis with penalized spline regression, J. Phys.: Conf. Ser. 979 (2018), 012093. https://doi.org/10.1088/1742-6596/979/1/012093.

[7] A. Islamiyati, N. Sunusi, A. Kalondeng, et al. Use of two smoothing parameters in penalized spline estimator for bi-variate predictor non-parametric regression model, J. Sci. Islam. Repub. Iran. 31 (2020), 175-183. https://doi.org/10.22059/jsciences.2020.286949.1007435.

[8] A. Islamiyati, Fatmawati, N. Chamidah, Penalized spline estimator with multi smoothing parameters in bi-response multi-predictor nonparametric regression model for longitudinal data, Songklanakarin J. Sci. Technol. 42 (2020), 897-909.

[9] B. Lestari, Fatmawati, I.N. Budiantara, Spline estimator and its asymptotic properties in multiresponse nonparametric regression model, Songklanakarin J. Sci. Technol. 42 (2020), 533-548.

[10] A. Islamiyati, Fatmawati, N. Chamidah, Changes in blood glucose 2 hours after meals in type 2 diabetes patients based on length of treatment at Hasanuddin University Hospital, Indonesia, Rawal Med. J. 45 (2020), 31-34.

[11] C.G. Morales, V.L. Davila, C.B. Cabral, et al. Robust quantile regression using a generalized class of skewed distributions, Stat 6 (2017), 113-130. https://doi.org/10.1002/sta4.140.

[12] M. Yuan, GACV for quantile smoothing splines, Comput. Stat. Data Anal. 50 (2006), 813-829. https://doi.org/10.1016/j.csda.2004.10.008.

[13] A. Islamiyati, Spline longitudinal multi-response model for the detection of lifestyle- based changes in blood glucose of diabetic patients, Curr. Diabetes Rev. 18 (2022), 98-104. https://doi.org/10.2174/1573399818666211117113856.

[14] R. Koenker, P. Ng, S. Portnoy, Quantile smoothing splines, Biometrika 81 (1994), 673-680. https://doi.org/10.1093/biomet/81.4.673.

[15] S.J. Staffa, D.S. Kohane, D. Zurakowski, Quantile regression and its applications: A primer for anesthesiologists, Anesth. Analg. 128 (2019), 820-830. https://doi.org/10.1213/ane.0000000000004017.

[16] M. Harding, C. Lamarche, Estimating and testing a quantile regression model with interactive effects, J. Econom. 178 (2014), 101-113. https://doi.org/10.1016/j.jeconom.2013.08.010.

[17] A.P. Spann, J.E. Campbell, S.R. Fitzgibbon, et al. The effect of hematocrit on platelet adhesion: Experiments and simulations, Biophys. J. 111 (2016), 577-588. https://doi.org/10.1016/j.bpj.2016.06.024.