Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2026, 2026:43

https://doi.org/10.28919/cmbn/9728

ISSN: 2052-2541

CONFIDENCE INTERVAL ESTIMATION IN A MULTIPREDICTOR SPLINE QUADRATIC REGRESSION MODEL FOR MODELING HBA1C LEVELS IN DIABETES MELLITUS CASES

SAMSUL ARIFIN*, DEWI ANGGRAINI

Department of Statistics, Lambung Mangkurat University, Banjarbaru 70714, Indonesia

Abstract: Modeling HbA1c levels in patients with diabetes mellitus is commonly conducted using parametric regression; however, this approach is often inadequate in capturing nonlinear relationships with metabolic predictors. This study aims to develop a multipredictor quadratic spline regression model with confidence interval estimation to model HbA1c levels flexibly. The model incorporates five predictor variables: body weight, fasting blood glucose, HDL cholesterol, LDL cholesterol, and triglycerides, and is implemented using RStudio version 2024.12.0. The results clearly demonstrate nonlinear relationships between HbA1c levels and all predictors. LDL cholesterol shows the strongest influence. Threshold effects are observed for body weight and HDL cholesterol, while glucose and triglycerides exhibit moderate nonlinear patterns. The visualization of the fitted curves and confidence bands supports a more interpretable representation of the model. Overall, the quadratic spline regression with confidence intervals provides a flexible and informative framework for modeling HbA1c levels, particularly when nonlinear associations are present.

Keywords: HbA1c; quadratic spline; confidence interval; GCV; diabetes mellitus.

2020 AMS Subject Classification: 62G08.

*Corresponding author

E-mail address: samsularrr@ulm.ac.id

Received December 03, 2025

1. INTRODUCTION

The relationship between predictor variables such as sex, age, body mass index, and glucose levels and HbA1c levels is commonly analyzed using parametric regression models, such as multiple linear regression [1]. These models are often preferred due to their ease of interpretation and simple computational procedures [2]. However, multiple linear regression has an important limitation: it assumes a specific functional form for the regression curve, such as linear or quadratic [3]. Several studies have shown that metabolic data patterns, including HbA1c, are complex and tend not to follow a linear relationship [4].

To overcome these limitations, researchers have increasingly developed nonparametric regression models. Common nonparametric approaches include kernel regression [5], local polynomial regression [6], and spline regression [7]. Among these methods, spline regression has become one of the most widely used techniques because it does not require an assumption about the functional form of the regression curve, offering greater flexibility in capturing complex relationship patterns [8].

Previous research has demonstrated that spline regression is highly effective for modeling data with nonlinear relationships [9]. For example, in a study on child growth (0–60 months), penalized splines were used to detect growth deceleration based on weight and height simultaneously, with the optimal knot points located at critical ages to show significant changes in growth rate [10]. Spline methods have also shown their effectiveness in numerical interpolation [11]. In addition, truncated splines in quantile regression have been successfully used to handle data with outliers and non-normal distributions [12].

The polynomial order used in spline regression greatly influences the shape and flexibility of the resulting curve [13]. For instance, linear splines (first order) tend to be overly rigid [14], whereas cubic splines (third order), while highly flexible, may lead to overfitting particularly with smaller sample sizes [15]. Consequently, this study applies quadratic splines (second order), which offer a balanced compromise between model complexity and estimator stability. Quadratic splines do not produce excessive fluctuations and are therefore well-suited for clinical or epidemiological data, where variation tends to be gradual [16].

Nevertheless, point estimates of the regression curve alone are insufficient to convey all the information required for statistical inference. Confidence interval estimation is recommended because it provides a range of plausible parameter values at a specified confidence level, thereby reflecting the uncertainty of the estimated parameters. For example, Setyawati et al. [17]

demonstrated that applying confidence intervals to nonparametric models allowed more reliable and accurate identification of variables affecting COVID-19 growth and mortality rates in Indonesia [17].

In light of these considerations, this study aims to develop confidence interval estimation for a multipredictor quadratic spline regression model to flexibly model HbA1c levels in patients with diabetes mellitus using predictor variables such as body weight, blood glucose levels, LDL cholesterol, HDL cholesterol, and triglycerides.

2. PRELIMINARIES

2.1 Multipredictor Nonparametric Regression Model

Suppose paired data $(x_{1i}, x_{2i}, \dots, x_{pi}, y_i)$ are given, with $i = 1,2, \dots, n$ where x_{ji} is the j-th predictor variable and y_i is the response variable [18].The nonparametric regression model can be expressed as the following equation:

$$y_i = f(x_{i1}, x_{i2}, \dots, x_{ip}) + \varepsilon_i \tag{1}$$

If the function in equation (1) is assumed to be additive and approximated by a quadratic spline function, then the regression model becomes the following equation:

$$y_i = \sum_{j=1}^{p} f_j(x_{ji}) + \varepsilon_i \tag{2}$$

For each function $f_j(x_{ji})$, the quadratic spline approximation with r knot points $k_{1j}, k_{2j}, \dots, k_{rj}$ Can be expressed in the following equation:

$$f_j(x_{ji}) = \beta_{0j} + \beta_{1j}x_{ji} + \beta_{2j}x_{ji}^2 + \sum_{h=1}^{r} \delta_{hj}\,(x_{ji} - k_{hj})_+^2 \tag{3}$$

The notation $(z)_+ = \max(0, z)$ Denotes a truncated function. Thus, for all predictor variables it can be expressed in the following equation:

$$y_i = \sum_{j=1}^{p} \left[\beta_{0j} + \beta_{1j}x_{ji} + \beta_{2j}x_{ji}^2 + \sum_{h=1}^{r} \delta_{hj}\,(x_{ji} - k_{hj})_+^2 \right] + \varepsilon_i \tag{4}$$

2.2 Matrix Notation of the Quadratic Spline Model

To express the estimation form in matrix notation, we define:

$\boldsymbol{y} = [y_1, y_2, \dots, y_n]^T$ as the vector of response variables

$\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n]^T$ as the error vector

$\boldsymbol{\beta_j} = \left[\beta_{0j}, \beta_{1j}, \beta_{2j}, \delta_{1j}, \dots, \delta_{rj}\right]^T \in \mathbb{R}^{3+r}$ as the parameter vector for the j-th predictor

Next, the matrix $\boldsymbol{X_j} \in \mathbb{R}^{n(3+r)}$ It is defined as follows:

$$\boldsymbol{X_j} = \begin{bmatrix} 1 & x_{j1} & x_{j1}^2 & \left(x_{j1} - k_{1j}\right)_+^2 & \cdots & \left(x_{j1} - k_{rj}\right)_+^2 \\ 1 & x_{j2} & x_{j2}^2 & \left(x_{j2} - k_{1j}\right)_+^2 & \cdots & \left(x_{j2} - k_{rj}\right)_+^2 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{jn} & x_{jn}^2 & \left(x_{jn} - k_{1j}\right)_+^2 & & \left(x_{jn} - k_{rj}\right)_+^2 \end{bmatrix} \tag{5}$$

The combination for all predictor variables $j = 1,2, \dots, p$ Can be written as follows:

$\boldsymbol{X} = \left[\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_p}\right] \in \mathbb{R}^{n \times p(3+r)}$ is the combined design matrix for all predictor variables

$\boldsymbol{\beta} = \left[\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \dots, \boldsymbol{\beta_p}\right]^{\boldsymbol{T}} \in \mathbb{R}^{p(3+r)}$ is the combined parameter vector.

Thus, the quadratic spline regression model can be written in matrix form as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(0, \sigma^2) \tag{6}$$

2.3 Parameter Estimation Multipredictor Quadratic Spline

The parameter estimation is carried out using the Least Squares Estimation (LSE) method. The goodness-of-fit function is defined as:

$$Q(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{X\beta}) \tag{7}$$

To obtain the estimated result, the function $Q(\boldsymbol{\beta})$ is differentiated with respect to$\boldsymbol{\beta}$, yielding the following parameter estimator:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T y} \tag{8}$$

The variance of the parameter estimator is:

$$Var\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2 (\boldsymbol{X^T X})^{-1} \tag{9}$$

The error variance estimator is given by:

$$\sigma^2 = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{10}$$

2.4 Confidence Interval (CI)

The 95% confidence interval for each coefficient $\hat{\beta}_j$ Is given by:

$$\hat{\beta}_j \pm t_{1-\alpha/2,df} \sqrt{\hat{\sigma}^2 [(\boldsymbol{X^T X})^{-1}]} \tag{11}$$

Where:

$t_{1-\alpha/2,df}$ Is the critical value from the Student's t-distribution

$df = n - p$ is the residual degrees of freedom

This confidence interval measures the level of certainty that the true coefficient lies within a specific range and is essential for assessing both the significance and stability of the parameters.

2.5 Generalized Cross Validation (GCV)

The selection of optimal knot points in the spline regression model is carried out using the GCV method:

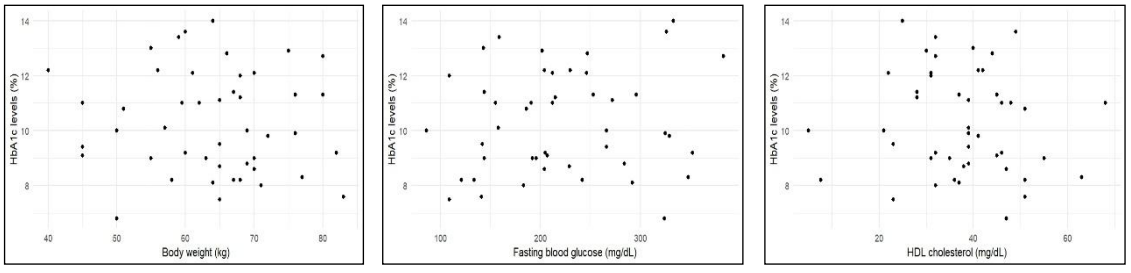
$$\mathrm{GCV}\left[k_{1j}, k_{2j}, \dots, k_{rj}\right] = \frac{MSE(k_{1j}, k_{2j}, \dots, k_{rj})}{[n^{-1} trace\left(\boldsymbol{I_n} - \boldsymbol{A}\left(k_{1j}, k_{2j}, \dots, k_{rj}\right)\right)]^2} \quad (12)$$

where $MSE\left(k_{1j}, k_{2j}, \dots, k_{rj}\right) = n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, $k_{1j}, k_{2j}, \dots, k_{rj}$ Represents the knot points, matrix $\boldsymbol{A}$ is $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and $\boldsymbol{I_n}$ Is the identity matrix [19]. The optimal knot points in the multipredictor quadratic spline model are determined based on the minimum value of GCV [20].

3. MAIN RESULTS

3.1 Exploratory Data Analysis

This study examines the relationship patterns between HbA1c levels and each predictor variable, namely body weight, fasting blood glucose, HDL cholesterol, LDL cholesterol, and triglycerides, using initial visualization through scatter plots, as presented in Figure 1. This visualization was intended to evaluate the form of the relationships between each predictor and the response variable. The scatter plots reveal that the relationships between most predictor variables and HbA1c levels do not follow a clear linear pattern and appear randomly dispersed. This observation further highlights the need for a multivariate quadratic spline regression approach, which is specifically designed to capture nonlinear variations with greater flexibility than conventional linear regression models.



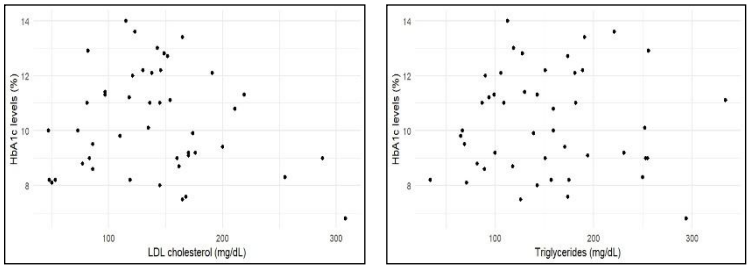


Figure 1. Scatter Plot of HbA1c with Predictor Variables

3.2 Knot Selection

Knot selection is a critical step in constructing a quadratic spline regression model, as the knot locations determine where the regression curve changes direction, directly influencing the model's flexibility. In this study, the knot points were determined based on the minimum GCV value, a common evaluation criterion in nonparametric modeling that balances goodness-of-fit and model complexity. The GCV values were calculated numerically using RStudio by testing a range of candidate values for each predictor variable, as shown in Table 1. The optimal knot points were identified at 80 for body weight, 352 for fasting glucose, 32 for HDL cholesterol, 50 for LDL cholesterol, and 34 for triglycerides.

Table 1. Selection of Optimal Knot Points

X1	GCV	X2	GCV	X3	GCV	X4	GCV	X5	GCV
80	3.678	352	3.747	55	3.800	73	3.161	65	3.748
82	3.713	348	3.749	51	3.814	77	3.169	67	3.749
77	3.786	333	3.787	49	3.821	81	3.178	69	3.750
76	3.808	329	3.797	48	3.824	82	3.181	71	3.751
75	3.822	326	3.803	47	3.827	83	3.183	82	3.760
72	3.853	325	3.804	46	3.829	86	3.191	87	3.766
51	3.859	324	3.805	32	3.830	97	3.225	89	3.768
55	3.859	296	3.830	45	3.831	110	3.275	90	3.769
71	3.860	292	3.832	44	3.832	115	3.296	94	3.775
50	3.860	121	3.837	35	3.833	118	3.310	99	3.782
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
65	3.871	230	3.858	63	3.759	288	3.565	231	3.888

3.3 HbA1c Level Model in Diabetes Mellitus Patients

Model the relationship between HbA1c levels and the predictor variables, namely body weight (x_1), blood glucose (x_2), HDL cholesterol (x_3), LDL cholesterol (x_4), and triglycerides (x_5). A quadratic spline regression model with one knot point per predictor is used, as shown in Table 2.

Table 2. Estimation of Confidence Intervals for Multipredictor Quadratic Spline

Coefficient	Estimated	Std. Error	Confidence Interval	
			Lower	Upper
Intercept	-3.032	17.039	-37.830	31.765
x_1	0.049	0.356	-0.679	0.777
x_1^2	-0.001	0.003	-0.006	0.005
$(x_1 - 80)^2$	-0.256	0.283	-0.833	0.321
x_2	0.009	0.028	-0.048	0.066
x_2^2	0.000	0.000	0.000	0.000
$(x_2 - 352)^2$	0.002	0.003	-0.004	0.007
x_3	0.095	0.142	-0.194	0.384
x_3^2	-0.001	0.002	-0.006	0.003
$(x_3 - 55)^2$	0.013	0.021	-0.029	0.055
x_4	0.203	0.431	-0.677	1.084
x_4^2	-0.001	0.003	-0.007	0.005
$(x_4 - 73)^2$	0.001	0.003	-0.005	0.007
x_5	0.031	0.324	-0.630	0.692
x_5^2	0.000	0.003	-0.005	0.005
$(x_5 - 65)^2$	0.000	0.003	-0.005	0.005

The selection of knot points is based on the minimum GCV values: 80 for body weight, 352 for glucose, 55 for HDL, 73 for LDL, and 65 for triglycerides. The multipredictor quadratic spline regression model for HbA1c level is expressed in the following equation:

$$y = -3.0323 + 0.0492x_1 - 0.0005x_1^2 - 0.2560(x_1 - 80)^2 + 0.0088x_2 + 0.0000x_2^2 + 0.0017(x_2 - 352)^2 + 0.0947x_3 - 0.0015x_3^2 + 0.0131(x_3 - 55)^2 + 0.2032x_4 - 0.0013x_4^2 + 0.0012(x_4 - 73)^2 + 0.0308x_5 - 0.0001x_5^2 + 0.0001(x_5 - 65)^2$$

The model shows that each predictor variable has a different and nonlinear influence on HbA1c levels. The body weight variable (x_1) has a complex relationship with HbA1c. The positive linear coefficient indicates that an increase in body weight is initially correlated with an increase in HbA1c. However, the negative quadratic coefficient and the negative quadratic spline at the 80 kg knot point indicate that, after surpassing this threshold, the effect decreases, suggesting a possible protective effect of higher body weight or better metabolic control in certain groups.

Blood glucose (x_2) exerts a small but consistent positive effect on HbA1c. The positive linear coefficient and the spline component at 352 mg/dL indicate that at very high glucose levels, the

effect on HbA1c tends to increase. However, overall this relationship appears flat, as reflected in the very small quadratic coefficient, which approaches zero. HDL cholesterol (x_3) shows an interesting effect, where an increase in HDL is linearly associated with higher HbA1c. Still, there is a negative quadratic effect suggesting a saturation point or upper threshold, beyond which further increases in HDL no longer contribute to HbA1c elevation. The spline effect at 55 mg/dL reinforces this pattern change, which may reflect a clinical threshold.

LDL cholesterol (x_4) is the strongest predictor, with the highest linear coefficient (0.2032). This indicates that an increase in LDL levels is directly associated with higher HbA1c levels. However, the negative quadratic coefficient and the positive spline component at 73 mg/dL indicate that the effect is also curved, with the impact of LDL diminishing at very high concentrations. Triglycerides (x_5) have a relatively small but still relevant effect. The positive linear coefficient and negative quadratic coefficient suggest a downward-curved relationship, while the spline component at 65 mg/dL provides additional flexibility in the model. This indicates that although triglycerides are not the primary predictor, they still contribute to variation in HbA1c, particularly around specific thresholds.

To understand the relationship patterns between each predictor variable and HbA1c levels in patients with diabetes mellitus, a visualization was performed using curves from the quadratic spline regression model, as shown in Figure 2. This visualization aims to display the nonlinear relationship patterns that conventional linear regression models may not adequately capture.

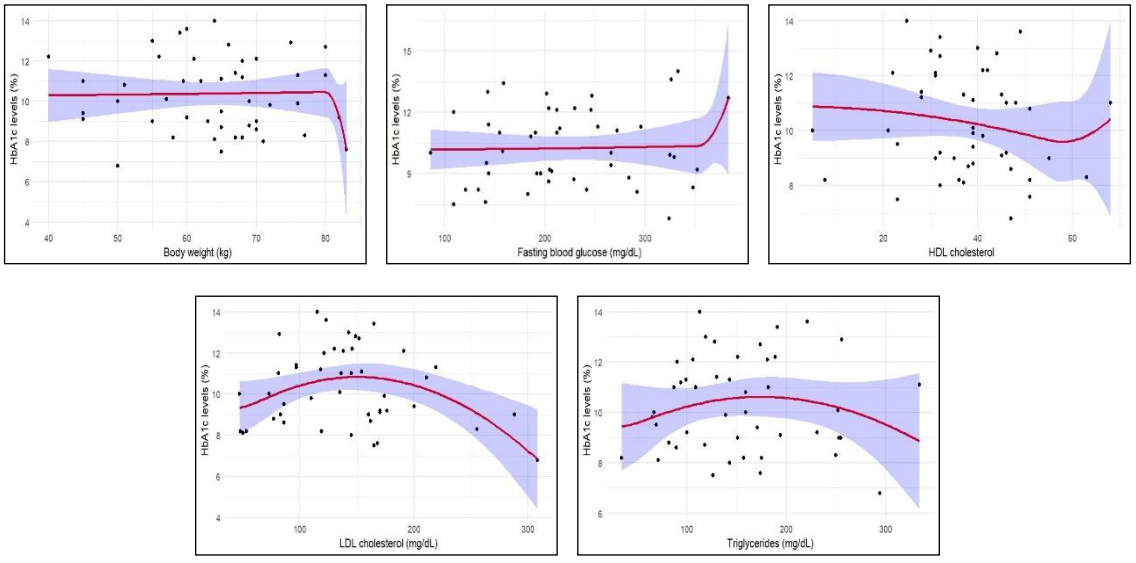


Figure 2. Estimation of Confidence Intervals for Multipredictor Quadratic Spline

Each graph displays actual observation points as black dots, the red line as the model's predicted curve, and the blue shading as the 95% confidence interval that reflects the uncertainty of model estimates across different ranges of predictor values. This visualization provides a more comprehensive depiction of how HbA1c levels change with variations in each predictor, while also highlighting the important contribution of spline approaches in capturing clinically relevant nonlinear patterns. This analysis is not only useful in statistical contexts but also has important implications for medical practice in understanding the dynamics of biochemical parameters in long-term glucose control.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] D. Tsilingiris, K. Makrilakis, A. Barmpagianni, M. Dalamaga, A. Tentolouris, et al., The Glycemic Status Determines the Direction of the Relationship Between Red Cell Distribution Width and HbA1c, J. Diabetes Complicat. 35 (2021), 108012. https://doi.org/10.1016/j.jdiacomp.2021.108012.

[2] N. Roustaei, Application and Interpretation of Linear-Regression Analysis, Med. Hypothesis Discov. Innov. Ophthalmol. 13 (2024), 151-159. https://doi.org/10.51329/mehdiophthal1506.

[3] M. Ajona, P. Vasanthi, D. Vijayan, Application of Multiple Linear and Polynomial Regression in the Sustainable Biodegradation Process of Crude Oil, Sustain. Energy Technol. Assessments 54 (2022), 102797. https://doi.org/10.1016/j.seta.2022.102797.

[4] A. Islamiyati, A. Kalondeng, N. Sunusi, M. Zakir, A.K. Amir, 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.

[5] I. Rezaei, S.H. Amirshahi, A.A. Mahbadi, Utilizing Support Vector and Kernel Ridge Regression Methods in Spectral Reconstruction, Results Opt. 11 (2023), 100405. https://doi.org/10.1016/j.rio.2023.100405.

[6] M. Sawada, T. Ishihara, D. Kurisu, Y. Matsuda, Local-Polynomial Estimation for Multivariate Regression Discontinuity Designs, arXiv:2402.08941, 2024. https://doi.org/10.48550/arXiv.2402.08941.

[7] S. Arifin, A. Islamiyati, E.T. Herdiani, Ability of Ordinal Spline Logistic Regression Model in the Classification of Nutritional Status Data, Commun. Math. Biol. Neurosci. 2023 (2023), 83. https://doi.org/10.28919/cmbn/8072.

[8] N. Chamidah, B. Lestari, H. Susilo, M.Y. Alsagaff, I.N. Budiantara, et al., Spline Estimator in Nonparametric

Ordinal Logistic Regression Model for Predicting Heart Attack Risk, Symmetry 16 (2024), 1440. https://doi.org/10.3390/sym16111440.

[9] N.A. Schuster, J.J.M. Rijnhart, J.W.R. Twisk, M.W. Heymans, Modeling Non-Linear Relationships in Epidemiological Data: The Application and Interpretation of Spline Models, Front. Epidemiol. 2 (2022), 975380. https://doi.org/10.3389/fepid.2022.975380.

[10] A. Islamiyati, A. Kalondeng, M. Zakir, S. Djibe, U. Sari, Detecting Age Prone to Growth Retardation in Children Through a Bi-Response Nonparametric Regression Model with a Penalized Spline Estimator, Iran. J. Nurs. Midwifery Res. 29 (2024), 549-554. https://doi.org/10.4103/ijnmr.ijnmr_342_22.

[11] M. Sun, L. Lan, C.G. Zhu, F. Lei, Cubic Spline Interpolation with Optimal End Conditions, J. Comput. Appl. Math. 425 (2023), 115039. https://doi.org/10.1016/j.cam.2022.115039.

[12] Anisa, A. Islamiyati, S. Sahriman, J. Massalesse, U. Sari, Truncated Spline Quantile Regression Model on Platelet Changes in Dengue Fever Patients Based on Body Temperature, Commun. Math. Biol. Neurosci. 2024 (2024), 69. https://doi.org/10.28919/cmbn/7978.

[13] F. Hamad, N. Younus, M. Jaber, Discovering the Best Choice for Spline's Knots and Intervals Using Order of Polynomial Regression Model, Open J. Stat. 14 (2024), 743-756. https://doi.org/10.4236/ojs.2024.146034.

[14] S. Arifin, D. Anggraini, N. Salam, A Comparative Study of Linear and Quadratic Spline Regression Models for Predicting HbA1c Levels in Patients with Diabetes Mellitus, Jambura J. Math. 7 (2025), 183-188. https://doi.org/10.37905/jjom.v7i2.33292.

[15] J.I. Arnes, A. Hapfelmeier, A. Horsch, T. Braaten, Greedy Knot Selection Algorithm for Restricted Cubic Spline Regression, Front. Epidemiol. 3 (2023), 1283705. https://doi.org/10.3389/fepid.2023.1283705.

[16] S. Samreen, M. Sarfraz, A. Mohamed, A Quadratic Trigonometric B-Spline as an Alternate to Cubic B-Spline, Alex. Eng. J. 61 (2022), 11433-11443. https://doi.org/10.1016/j.aej.2022.05.006.

[17] M. Setyawati, N. Chamidah, A. Kurniawan, Confidence Interval of Parameters in Multiresponse Multipredictor Semiparametric Regression Model for Longitudinal Data Based on Truncated Spline Estimator, Commun. Math. Biol. Neurosci. 2022 (2022), 107. https://doi.org/10.28919/cmbn/7672.

[18] 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. https://doi.org/10.14456/sjst-psu.2020.68.

[19] A.F. Al Barra, D.R.S. Saputro, Knot Optimization for Bi-response Spline Nonparametric Regression with Generalized Cross-Validation (GCV), BAREKENG: J. Ilmu Mat. Ter. 19 (2025), 271-280. https://doi.org/10.30598/barekengvol19iss1pp271-280.

[20] N. Chamidah, B. Lestari, H. Susilo, T.K. Dewi, T. Saifudin, et al., Modeling Coronary Heart Disease Risk Based

on Age, Fatty Food Consumption and Anxiety Factors Using Penalized Spline Nonparametric Logistic Regression, MethodsX 14 (2025), 103320. https://doi.org/10.1016/j.mex.2025.103320.