Estimating the confidence interval of the regression coefficient of the blood sugar model through a multivariable linear spline with known variance

Anna Islamiyati[1], Raupong[2], Anisa Kalondeng[3], Ummi Sari[4]

ABSTRACT

Estimates from confidence intervals are more powerful than point estimates, because there are intervals for parameter values used to estimate populations. In relation to global conditions, involving issues such as type 2 diabetes mellitus, it is very difficult to make estimations limited to one point only. Therefore, in this article, we estimate confidence intervals in a truncated spline model for type 2 diabetes data. We use a non-parametric regression model through a multi-variable spline linear estimator. The use of the model results from the irregularity of the data, so it does not form a parametric pattern. Subsequently, we obtained the interval from beta parameter values for each predictor. Body mass index, HDL cholesterol, LDL cholesterol and triglycerides all have two regression coefficients at different intervals as the number of the found optimal knot points is one. This value is the interval for multivariable spline regression coefficients that can occur in a population of type 2 diabetes patients.

Key words: confidence interval, diabetes, known variance, spline.

1. Introduction

There are two commonly known regression coefficient estimates, namely point and interval estimation. Both are valid for all regression approaches, including nonparametric regression. Non-parametric regression is used when the assumption of the relationship between the predictor and the response is unknown, so we must estimate its function. Some estimators that have been developed include spline truncated (Aprilia, Islamiyati and Anisa, 2019), spline smoothing (Lestari, Budiantara

[1] Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Indonesia. E-mail: annaislamiyati701@gmail.com (correspondence author). ORCID: https://orcid.org/0000-0001-6441-0306.
[2] Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Indonesia. E-mail: raupong.stat.uh@gmail.com..
[3] Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Indonesia. E-mail: nkalondeng@gmail.com.
[4] Hasanuddin University Teaching Hospital, Indonesia, E-mail: ummisari.rsunhas@gmail.com.

and Chamidah, 2019), spline penalized (Islamiyati, Fatmawati and Chamidah, 2019), local polynomial (Chamidah, Gusti, Tjahjono and Lestari, 2019), Kernel (Chamidah and Saifuddin, 2013), Fourier series (Mardianto, Tjahjono and Rifada, 2020), and Gaussian process (Saegusa, 2020) , and spline principal component analysis (Islamiyati, Kalondeng, Sunusi, Zakir and Amir, 2022). For this article, we use spline truncated to estimate multi-variable non-parametric regression functions. The flexibility of the estimator by involving the knot point causes easy visual interpretation. This has become one of the main advantages of spline truncated in real applications. In addition to the knot point, the spline also considers the optimal order, which works simultaneously in the estimation model.

Confidence interval estimates are performed in cases of known or unknown variance. It depends on the available information related to population variance. We often find there is a case that was investigated by several researchers with different methods. These studies can provide information about the condition of the population variance. For example, diabetes data is a global disease that has been widely studied in various fields of study. Some studies consider the results of measurements of random blood sugar levels of patients using a penalized spline estimator of one smoothing parameter (Islamiyati, Fatmawati and Chamidah, 2020) and two smoothing parameters (Islamiyati, Sunusi, Kalondeng, Fatmawati and Chamidah, 2020). Also, some consider the patient's fasting blood sugar data (Islamiyati, Raupong and Anisa, 2019), the patient's calorie diet (Islamiyati, Fatmawati and Chamidah, 2020), and the detection of lifestyle of diabetic patients (Islamiyati, 2022). In the country of India, a Genome-Wide Association Scan study identified more than 65 common genetic variants associated with type 2 diabetes (Singh, 2015). In Indonesia, especially in Makassar City, blood sugar after meals has a variance based on the measurement results through weighted penalized spline (Islamiyati, Fatmawati and Chamidah, 2018).

However, the studies still consider a single point estimate in the non-parametric regression approach. That can cause a large difference from the estimated value of the regression coefficients obtained for research at different locations and times. Some studies on estimating confidence intervals that provide wider tolerance values in non-parametric regression coefficients, for example, a comparison of results of confidence interval estimates from spline smoothing with Bayesian (Wang and Wahba, 2003), estimation of confidence intervals with uniform distribution on non-parametric regression curves (David, Tom and Douglas, 2001) and the use of B spline estimators in polynomial spline (Mao and Zhao, 2003). Therefore, we reviewed the estimated confidence interval in diabetes data using a spline truncated. For application consideration, we use a linear spline on the dimensions of many predictor variables. The results of this article are expected to provide lower and upper limits of an interval of spline regression coefficient values from diabetes data. Diabetes data were obtained

from the Hasanuddin University Teaching Hospital in Makassar, Indonesia by taking factors in body mass index, HDL and LDL cholesterol, and triglycerides.

Furthermore, the core content of this article discusses the theoretical form of the model and proceeds to the application of diabetes data using the linear spline. The knot point being tried is limited to the use of three knots with the optimal knot point selection method through Generalized Cross-Validation (GCV) values. Based on the theory that has been widely used by non-parametric regression that a regression model that provides a minimum GCV value means that the model is the optimal model used in interpreting data conditions.

2. Methodology

2.1. Data source

Data on patients with type 2 diabetes mellitus were obtained from the Hasanuddin University Teaching Hospital, Makassar, Indonesia. The data were recorded from the medical records of diabetic patients who were hospitalized from 2014-2018. We selected 84 people as samples in this study because they had complete media records according to the factors studied (Appendix 1).

2.2. Multi-variable linear spline models in the non-parametric regression approach

The non-parametric regression function that contains one response and several predictor variables is estimated with a linear spline so it is called a multi-variable linear spline model. Multi-variable non-parametric regression models can be stated as follows:

$$y_i = f\left(t_{1i}, t_{2i}, \ldots, t_{pi}\right) + \varepsilon_i \tag{1}$$

where y_i is the response in the i-samples, and $f\left(t_{1i}, t_{2i}, \ldots, t_{pi}\right)$ is a function in the predictor $t_1, t_2, \ldots, t_p$ and ε_i is an error in the i-samples.

The function $f\left(t_{1i}, t_{2i}, \ldots, t_{pi}\right)$ in equation (1) is estimated with a spline truncated estimator. Each function in each predictor can be stated as follows:

$$\left.\begin{aligned}
f\left(t_{1i}\right) &= \alpha_{10} + \alpha_{11} t_{1i} + \sum_{u_1=1}^{s_1} \alpha_{1(1+u_1)}\left(t_{1i} - k_{1u_1}\right)_+ \\
f\left(t_{2i}\right) &= \alpha_{20} + \alpha_{21} t_{2i} + \sum_{u_2=1}^{s_2} \alpha_{2(1+u_2)}\left(t_{2i} - k_{2u_2}\right)_+ \\
&\vdots \\
f\left(t_{pi}\right) &= \alpha_{p0} + \alpha_{p1} t_{pi} + \sum_{u_p=1}^{s_p} \alpha_{p(1+u_p)}\left(t_{pi} - k_{pu_p}\right)_+
\end{aligned}\right\} \tag{2}$$

If equation (1) is expressed as a sum of the functions of each predictor, namely:

$$y_i = f\left(t_{1i}\right) + f\left(t_{2i}\right) + \ldots + f\left(t_{pi}\right) + \varepsilon_i ,$$

then the multi-variable spline regression model can be stated as follows:

$$y_i = \alpha_0 + \sum_{h=1}^{p}\left(\alpha_{h1}t_{hi} + \sum_{u_h=1}^{s_h} \alpha_{h(1+u_h)}\left(t_{hi} - k_{hu_h}\right)_+ \right) + \varepsilon_i \tag{3}$$

Equation (3) can be expressed in matrix form, namely:

$$\mathbf{y} = \mathbf{T\alpha} + \mathbf{\varepsilon} \tag{4}$$

where $\mathbf{y} = \left(y_1, y_2, \ldots, y_n\right)^T$ is the response to $i = 1, 2, \ldots, n$, $\mathbf{T} = \left(\mathbf{1}, \mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_p\right)$ is a predictor matrix containing knots, $\mathbf{\alpha} = \left(\alpha_0, \alpha_{11}, \alpha_{12}, \ldots, \alpha_{1s_1}, \alpha_{21}, \alpha_{22}, \ldots, \alpha_{2s_2}, \ldots, \alpha_{p1}, \alpha_{p2}, \ldots, \alpha_{ps_p}\right)^T$ is regression coefficient of linear multi-variable spline, and $\mathbf{\varepsilon} = \left(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\right)^T$ is an error. As for

$$\mathbf{1} = \left(1, 1, \ldots, 1\right)^T, \mathbf{T}_1 = \left(\mathbf{t}_1, \left(\mathbf{t}_1 - k_{11}\right), \left(\mathbf{t}_1 - k_{12}\right), \ldots, \left(\mathbf{t}_1 - k_{1s_1}\right)\right), \mathbf{T}_2 = \left(\mathbf{t}_2, \left(\mathbf{t}_2 - k_{21}\right), \left(\mathbf{t}_2 - k_{22}\right), \ldots, \left(\mathbf{t}_2 - k_{2s_2}\right)\right),$$

$$\ldots, \mathbf{T}_p = \left(\mathbf{t}_p, \left(\mathbf{t}_p - k_{p1}\right), \left(\mathbf{t}_p - k_{p2}\right), \ldots, \left(\mathbf{t}_p - k_{ps_p}\right)\right).$$

Estimation of multi-variable spline regression parameter $\mathbf{\alpha}$ is obtained through the least square method by minimizing the sum of the squares of the error in equation (4).

$$\mathbf{\varepsilon}^T\mathbf{\varepsilon} = \left(\mathbf{y} - \mathbf{T\alpha}\right)^T \left(\mathbf{y} - \mathbf{T\alpha}\right)$$
$$= \mathbf{y}^T\mathbf{y} - 2\mathbf{\alpha}^T\mathbf{T}^T\mathbf{y} + \mathbf{\alpha}^T\mathbf{T}^T\mathbf{T\alpha} \tag{5}$$

Next, equation (5) is derived from the vector $\mathbf{\alpha}$ and the resulting derivation is equated to zero.

$$\left.\frac{\partial\left(\mathbf{\varepsilon}^T\mathbf{\varepsilon}\right)}{\partial\mathbf{\alpha}}\right|_{\mathbf{\alpha}=\hat{\mathbf{\alpha}}} = -2\mathbf{T}^T\mathbf{y} + 2\mathbf{T}^T\mathbf{T\alpha} \tag{6}$$

If equation (6) is equated to zero, then we get:

$$\hat{\mathbf{\alpha}} = \left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{y} \tag{7}$$

From equation (7), the estimation of the multi-variable spline regression model is as follows:

$$\hat{\mathbf{y}} = \mathbf{T}\hat{\mathbf{\alpha}}$$

where $\hat{\mathbf{\alpha}}$ according to equation (7). The Generalized Cross-Validation (GCV) formula for the model is obtained as follows:

$$GCV\left(\mathbf{k}\right) = \frac{MSE\left(\mathbf{k}\right)}{n^{-1}tr\left[\mathbf{I} - \mathbf{A}\left(\mathbf{k}\right)\right]^2} \tag{8}$$

where $\mathbf{A}\left(\mathbf{k}\right) = \mathbf{T}\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T$.

2.3. Estimates of confidence intervals with variance are known

We use the pivotal quantity approach (Toulis, 2017) in estimating the confidence interval for the multi-variable spline regression coefficient. First, we determine the expected value of $\hat{\boldsymbol{\alpha}}$, namely:

$$
\begin{aligned}
E\left(\hat{\boldsymbol{\alpha}}\right) &= E\left(\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{y}\right) \\
&= \left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{T}\boldsymbol{\alpha} + E\left(\boldsymbol{\varepsilon}\right) \\
&= \boldsymbol{\alpha}
\end{aligned} \tag{9}
$$

Second, we determine the variance of $\hat{\boldsymbol{\alpha}}$, namely:

$$
\begin{aligned}
Var\left(\hat{\boldsymbol{\alpha}}\right) &= Var\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{y} \\
&= \sigma^2\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{T}\left(\mathbf{T}^T\mathbf{T}\right)^T \\
&= \sigma^2\left(\mathbf{T}^T\mathbf{T}\right)^{-1}
\end{aligned} \tag{10}
$$

Based on equations (9) and (10), the estimate of $\hat{\boldsymbol{\alpha}}$ follows the normal distribution $\hat{\boldsymbol{\alpha}} \sim N\left(\boldsymbol{\alpha}, \sigma^2\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\right)$, namely mean $\boldsymbol{\alpha}$ and variance $\sigma^2\left(\mathbf{T}^T\mathbf{T}\right)^{-1}$.

Given a form of transformation from $Z_{hu_h} = \dfrac{\hat{\alpha}_{hu_h} - \alpha_{hu_h}}{\sqrt{\sigma^2 d_{hh}}}$, $d_{hh} = \text{diag}\left(\mathbf{T}^T\mathbf{T}\right)^{-1}$ and $Z_{hu_h} \sim N(0,1)$. This means that the expectation and variance values of Z_{hu_h} are 0 and 1. The effect σ^2 is known to cause $Z_{hu_h}\left(t_1, t_2, \ldots, t_n\right)$ to be the pivotal quantity for parameter α, so the confidence interval can be stated as follows:

$$
P\left(a \le Z_{hu_h}\left(t_1, t_2, \ldots, t_n\right) \le b\right) = 1 - \gamma , \tag{11}
$$

where a and b are real number elements, $a < b$.
Equation (11) can also be stated as:

$$
P\left(a \le \frac{\hat{\alpha}_{hu_h} - \alpha_{hu_h}}{\sqrt{\sigma^2 d_{hh}}} \le b\right) = 1 - \gamma \tag{12}
$$

Equation (12) can be worked on to become:

$$
\alpha_{hu_h} = \hat{\alpha}_{hu_h} - a\sqrt{\sigma^2 d_{hh}} \text{ dan } \alpha_{hu_h} = \hat{\alpha}_{hu_h} - b\sqrt{\sigma^2 d_{hh}} . \tag{13}
$$

Based on equation (13), the confidence interval for the multi-variable linear spline regression parameters is:

$$
P\left(\hat{\alpha}_{hu_h} - b\sqrt{\sigma^2 d_{hh}} \le \alpha_{hu_h} \le \hat{\alpha}_{hu_h} - a\sqrt{\sigma^2 d_{hh}}\right) = 1 - \gamma
$$

Furthermore, the shortest confidence interval is obtained from conditional optimization by the Lagrange method, namely $\min\limits_{a,b\in R}\{a,b\}=\min\limits_{a,b\in R}\left\{(b-a)\sqrt{\sigma^2 d_{hh}}\right\}$ with the constraint function $g(b)-g(a)-(1-\gamma)=0$, where g is the cumulative probability distribution $N(0,1)$. Furthermore, the Lagrange function can be expressed as:

$$F(a,b,\lambda)=(b-a)\sqrt{\sigma^2 d_{hh}}+\lambda\big(g(b)-g(a)-(1-\gamma)\big) \qquad (14)$$

where λ is the Lagrange constant. From the results of the partial derivative of the parameter a,b,λ, we get $a=-b$, which satisfies the equation. If it is substituted into equation (14), then it is obtained:

$$P\left(\hat{\alpha}_{hu_h}-b\sqrt{\sigma^2 d_{hh}}\leq\alpha_{hu_h}\leq\hat{\alpha}_{hu_h}+b\sqrt{\sigma^2 d_{hh}}\right)=1-\gamma \qquad (15)$$

For $b=Z_{\gamma/2}$, equation (15) can also be stated as:

$P\left(\hat{\alpha}_{hu_h}-Z_{\gamma/2}\sqrt{\sigma^2 d_{hh}}\leq\alpha_{hu_h}\leq\hat{\alpha}_{hu_h}+Z_{\gamma/2}\sqrt{\sigma^2 d_{hh}}\right)=1-\gamma$, where γ is the level of significance used in research that researchers usually use 0.05.

3. Analysis and discussion

Blood sugar data of type 2 DM patients were analysed using non-parametric regression, in this case, multi-variable linear spline regression. This is because changes in a person's blood sugar can change very quickly and do not follow a certain trend. The condition is also shown in Figure 1 through scatter plots between fasting blood sugar factor (y) with body mass index/BMI (t_1), high-density lipoprotein/HDL cholesterol (t_2), low-density lipoprotein/LDL cholesterol (t_3), and triglycerides/TG (t_4). Based on Figure 1, we can see a data plot that does not follow a certain parametric pattern, for example linear, quadratic, cubic, and others. Therefore, we analysed this data using a non-parametric multi-variable linear spline regression approach.
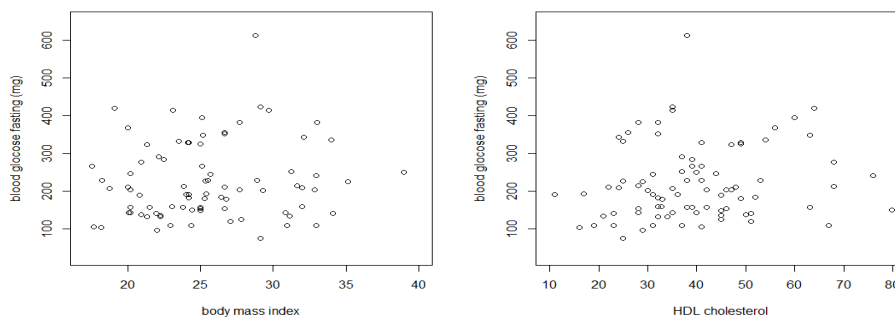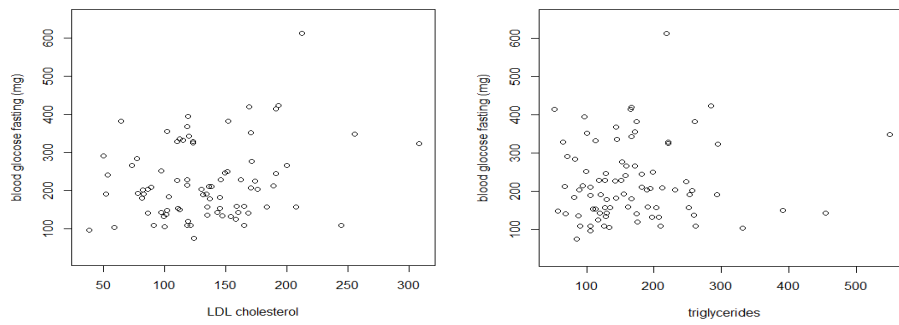
Figure 1. Scatter plot between blood sugar with body mass index, HDL cholesterol, LDL cholesterol, and triglycerides factors

Spline regression is related to the optimal knot point and the number of knots, so we need to find the values of these parameters through the minimum GCV value. Here, we show a comparison of GCV values from the use of 1, 2, and 3-knot points as in Table 1.

Table 1. GCV values at 1, 2, 3 knots

Number of knots	1	2	3
GCV value	**10,552.75**	11,033.11	10,568.35

Based on Table 1, we can see that the number of knot points that give the minimum GCV value in the multi-variable linear spline model is one knot. Therefore, the blood sugar data of type 2 DM patients were modelled with a linear spline multi-variable approach through one knot on all the predictor variables involved. Next, we obtain a multi-variable linear spline regression model as follows:
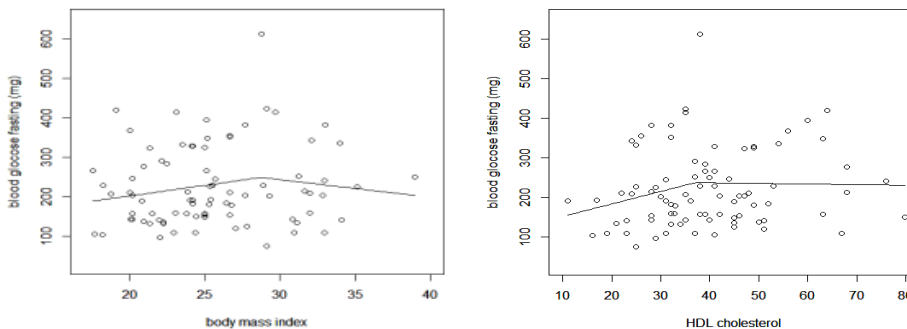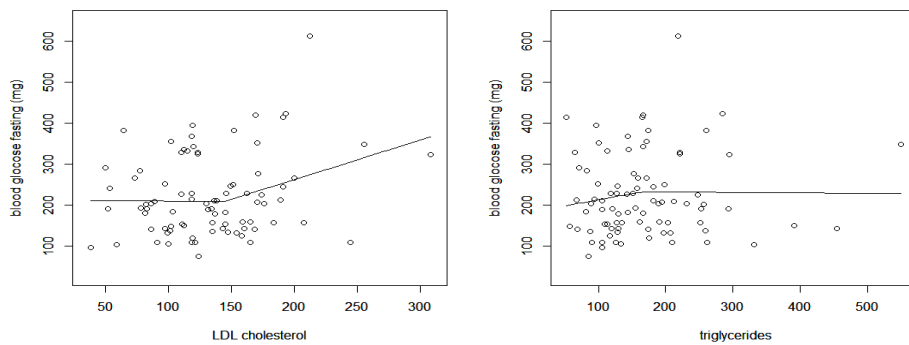
Figure 2. Estimation of multi-variable linear spline regression models with 1-knot point

The regression curve according to the model is shown in Figure 2. The optimal number of knot points is obtained at one knot for each predictor. The relationship of fasting blood glucose with Body Mass Index (BMI) is optimal at the knot point of 28.76 kg/m. These results indicate that there are two patterns of changes in fasting blood glucose based on that factor. We can see that it increases when the BMI is below the knot point, but after that, it decreases. This means that the increase based on BMI is in the pre-obesity stage. Excessive accumulation of fat in the body can cause insulin resistance which affects blood glucose levels. For variable t_2, the optimal HDL cholesterol at the one-knot point is 35 mg/dL. This shows that when HDL cholesterol reaches the point of knot, there is the accumulation of fat in certain organs. The incident triggers an increase in fasting blood glucose which can cause atherosclerosis or narrowing of the blood vessels and heart. Furthermore, fasting blood glucose has decreased slowly, which indicates an attempt to decrease after HDL increases. This certainly shows a positive trend from patients in controlling their blood sugar.

For variable t_3, the optimal LDL cholesterol at the one-knot point is 145 mg/dL. The pattern that is formed is that fasting blood glucose decreases slowly until LDL reaches that point. This means that diabetes patients keep their fasting blood glucose to prevent other diseases. In the next pattern, we see that fasting blood sugar levels increase very sharply when LDL cholesterol is more than the knot point. These results indicate the need to evaluate the patient's efforts to lower blood sugar when LDL is high. Furthermore, the optimal model for the variable t_4, namely triglycerides, is obtained at the knot point 167 mg/dL. These results indicate fasting blood glucose increases when the triglycerides are below the knot point, and after that, it decreases slowly. This means that for high triglycerides, there are patients' efforts to keep their blood sugar from rising.

Table 2. Estimated confidence intervals from spline regression coefficients for blood sugar data for type 2 DM patients at a confidence level of 95%

Variable	Parameter	Estimation of Parameter	Lower limit	Upper limit
	β_0	-76.58	-375.79	222.62
t_1	β_{11}	6.95	-1.67	15.57
t_1	β_{12}	-12.60	-34.70	9.50
t_2	β_{21}	3.91	-1.54	9.36
t_2	β_{22}	-4.27	-11.22	2.68
t_3	β_{31}	-0.38	-1.39	0.62
t_3	β_{32}	1.47	-0.26	3.20
t_4	β_{41}	0.36	-0.44	1.16
t_4	β_{42}	-0.59	-1.64	0.47

Point estimation results that have been obtained, proceed to the estimation of the confidence interval so that estimation results can be more accurate to explain the condition of the population. However, we first test the residual assumptions, especially the normal distribution. The normality assumption test is based on the Kolmogorov Smirnov test (Steinskog, Tjostheim and Kvamsto, 2007). We get the p-value, $p = 0.06 > \gamma = 0.05$, which means the residual is normally distributed. Point estimation results from the best multi-variable spline model for fasting blood glucose at the Hasanuddin University Teaching Hospital were used to construct confidence intervals for regression parameters. The results of the estimated upper and lower limit obtained are in Table 2. The table shows the model regression coefficient interval that contains a lower limit and an upper limit. These interval values provide an overview of the condition of the diabetic patient population, especially for the patient's blood sugar model based on body mass index, HDL cholesterol, LDL cholesterol, and triglycerides.

4. Conclusions

The estimation results of the regression coefficient interval of the multi-variable spline model provide interval values from the lower limit to the upper limit at the 95% confidence level. Confidence intervals represent the possible values in a diabetic patient population related to fasting blood glucose, BMI, LDL cholesterol, HDL cholesterol, and triglycerides. For type 2 diabetes, the confidence interval of the regression coefficient in Table 2 can be used as reference material for further research. The influence of these four factors of changes in blood sugar is shown by the pattern of changes in each regression curve obtained from a multi-variable simultaneous model.

Body mass index, cholesterol, and triglyceride factors can all cause an increase in blood sugar due to the accumulation of fat and cholesterol. We can pay attention to the results of the analysis, which show not an upward trend that occurs in the regression curve, but there are variations in the pattern that occurs. Segmentation of pattern changes can be demonstrated by multi-variable spline due to the presence of optimal knot parameters obtained from the minimum GCV value. The up and down patterns that occur in one regression curve indicate that many factors do indeed affect blood sugar in DM type 2 patients. Therefore, we suggest that further research adds a higher dimension to the variable by taking into account assumptions that can be violated for the data dimensions the big one.

Acknowledgment

References

Aprilia, B., Islamiyati, A., Anisa, (2019). Platelet modeling based on hematocrit in DHF patients with spline quantile regression, *International Journal of Academic and Applied Research*, Vol. 3(12)**,** pp. 51–54.

Chamidah, N., Gusti, K. H., Tjahjono, E., Lestari, B., (2019). Improving of classification accuracy of cyst and tumor using local polynomial estimator, *Telkomnika*, Vol. 17, pp. 1492–1500.

Chamidah, N., Saifudin, T., (2013). Estimation of children growth curve based on kernel smoothing in multi-response nonparametric regression, *Applied Mathematical Sciences*, Vol. 7(37), pp. 1839–1847.

David, J.C., Tom, G. F., Douglas, N., (2001). Confidence intervals for nonparametric curve estimates toward more uniform pointwise coverage, *Journal of the American Statistical Association*, Vol. 96(453), pp. 233–246.

Islamiyati, A., (2022). Spline longitudinal multi-response model for the detection of lifestyle-based changes in blood glucose of diabetic patients. *Current Diabetes Reviews*, Published on: 14 January.

Islamiyati, A., Fatmawati, Chamidah, N., (2018). Estimation of covariance matrix on bi-response longitudinal data analysis with penalized spline regression, *Journal of Physics: Conference Series*, Vol. 979**,** 012093.

Islamiyati, A., Fatmawati, Chamidah, N., (2020). Changes in blood glucose 2 hours after meals in Type 2 diabetes patients based on length of treatment at Hasanuddin University Hospital, Indonesia, *Rawal Medical Journal*, Vol. 45(1), pp. 31–34.

Islamiyati, A., Kalondeng, A., Sunusi, N., Zakir, M., Amir, A. K., (2022). Biresponse nonparametric regression model in principal component analysis with truncated spline estimator. *Journal of King Saud University-Science*, Vol. 34, 101892, pp. 1–9.

Islamiyati, A., Sunusi, N., Kalondeng, A., Fatmawati, F., Chamidah, N., (2020). Use of two smoothing parameters in penalized spline estimator for bi-variate predictor non-parametric regression model, *Journal of Sciences: Islamic Republic of Iran*, Vol. 31(2), pp. 175–183.

Islamiyati, A., Raupong, Anisa, (2019). Use of penalized spline linear to identify change in pattern of blood sugar based on the weight of diabetes patients, *International Journal of Academic and Applied Research*, Vol. 3(12)**,** pp. 75–78.

Islamiyati, A., Fatmawati, Chamidah, N., (2019). Ability of covariance matrix in bi-response multi-predictor penalized spline model through longitudinal data simulation, *International Journal of Academic and Applied Research*, Vol. 3(3), pp. 8–11.

Islamiyati, A, Fatmawati, Chamidah, N., (2020). Penalized spline estimator with multi smoothing parameters in bi-response multi-predictor regression model for longitudinal data, *Songklanakarin Journal of Science and Technology*, Vol. 42(4), pp. 897–909.

Lestari, B., Budiantara, I. N., Chamidah, N., (2019). Smoothing parameter selection method for multiresponse nonparametric regression model using smoothing spline and kernel estimators approaches. *Journal of Physics: Conference Series*, Vol. 1397, 012064.

Mao, W., Zhao, L. H., (2003). Free-knot polynomial splines with confidence intervals, *Journal of the Royal Statistical Society Series B,* Vol. 65(4), pp. 901–919.

Mardianto, M. F. F., Tjahjono, E., Rifada, M., (2020). Statistical modelling for prediction of rice production in Indonesia using semiparametric regression based on three forms of Fourier series estimator, *ARPN Journal of Engineering and Applied Sciences*, Vol. 14(15), pp. 2763–2770.

Toulis, P., (2017). A useful pivotal quantity, *The American Statistician*, Vol. 71(3), pp. 272–274.

Saegusa, T., (2020). Confidence band for a distribution function with merged data from multiple source, *Statistic in Transition*, Vol. 21(4), pp. 144–158.

Singh, S., (2015). Genetics of type 2 diabetes: advances and future prospect, *Journal of Diabetes and Metabolism*, Vol. 6(4), pp. 1–8.

Steinskog, D. J., Tjostheim, D. B., Kvamsto, N. G. A., (2007). A cautionary note on the use of the Kolmogorov–Smirnov Test for normality, *American Meteorological Society,* Vol. 135(3), pp. 1151–1157.

Wang, Y., Wahba, G., (2003). *Bootstrap confidence interval for smoothing spline and their comparison to Bayesian confidence intervals,* Technical Report University of Wisconsin-Madison USA, No. 193.

APPENDIX

Data on patients with type 2 diabetes mellitus at the Hasanuddin University Teaching Hospital, Makassar, Indonesia, 2014-2018.

Patient Number	Blood Sugar	BMI	HDL	LDL	Triglycerides
1	230	18.26	41	146	151
2	414	23.11	35	191	165
3	212	20	48	136	182
4	229	28.89	38	162	118
5	212	26.67	22	138	106
6	352	26.67	32	170	100
7	137	22.22	45	135	88
8	150	24.44	79.8	112	391
9	132	22.22	32	99	197
10	420	19.11	64	169	166
11	135	31.11	21	147	128
12	368	20	56	118	143
13	154	26.67	28	111	113
14	225	35.11	29	174	247
15	356	26.67	26	102	172
16	132	21.33	34	154	207
17	184	26.40	52	103	81
18	157	23.80	38	207	203
19	251	38.95	40	151	198
20	324	21.35	47	308	294
⋮	⋮	⋮	⋮	⋮	⋮
83	109	32.97	23	165	126
84	336	34	54	112	145

Data sharing information statement.

This study does not involve direct human interaction. Data were collected from patient medical records with the approval of hospital management. For the purposes of developing this study, the authors may share data through personal contacts.