Research Article

A New Mixed Estimator in Nonparametric Regression for Longitudinal Data

Made Ayu Dwi Octavanny [iD]**, I Nyoman Budiantara** [iD]**, Heri Kuswanto** [iD]**, and Dyah Putri Rahmawati** [iD]

Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia

Correspondence should be addressed to I Nyoman Budiantara; nyomanbudiantara65@gmail.com

We introduce a new method for estimating the nonparametric regression curve for longitudinal data. This method combines two estimators: truncated spline and Fourier series. This estimation is completed by minimizing the penalized weighted least squares and weighted least squares. This paper also provides the properties of the new mixed estimator, which are biased and linear in the observations. The best model is selected using the smallest value of generalized cross-validation. The performance of the new method is demonstrated by a simulation study with a variety of time points. Then, the proposed approach is applied to a stroke patient dataset. The results show that simulated data and real data yield consistent findings.

1. Introduction

Nonparametric regression is a statistical method used if the data show an unknown regression curve. The strength of this method is its great flexibility since the data are used to find the form of its estimated regression curve without being influenced by subjective judgements [1]. Some estimators used are spline, Fourier series, kernel, and polynomial.

The truncated spline is a function where there is a change in the behaviour of the curve in certain subintervals. The spline is one of the popular estimators in nonparametric regression because it has an excellent visual interpretation. Montoya et al. [2] conducted a simulation study to compare knot selection methods in a penalized regression spline model. Next, the Fourier series is much used to describe curves that present sine and cosine waves. This estimator is commonly used when the data have the characteristics of a periodicity. Bilodeau [3] estimated additive components with functions consisting of truncated Fourier cosine series, using penalized least squares to obtain the coefficients. Furthermore, Tripena [4] developed a Fourier series estimator for bi-response nonparametric regression. Gong and

Gao [5] discussed a nonparametric kernel estimation of tax policy's impact on the demand for private health insurance in Australia.

A single estimator in nonparametric regression is commonly used but does not limit the possibility of developing into a mixed estimator. There are many cases where each predictor variable has a different pattern. Therefore, applying only a single estimator can make the regression model's estimation incorrect and produce a large error. Some previous studies that have been mentioned have limitations in that they can only be used for cross-sectional data. To overcome this limitation, a longitudinal data model has been developed. The use of longitudinal data has increased in recent years because it can be applied in various fields. Longitudinal data are data obtained in observations of independent subjects, where each subject is observed repeatedly at a certain period. This has the advantage of being able to observe changes based on time [6].

In [7], the mixed estimator is limited in the sample size it can treat. The present study extends the use of the mixed truncated spline and Fourier series (MTSFS) model to larger sample sizes and various time point designs. Some properties

of the new mixed estimator will also be provided. We use generalized cross-validation (GCV) to determine the best model from various knots, oscillations, and smoothing parameters. A simulation study and real data are employed to demonstrate the performance of the proposed method. The case study includes the factors that affect the Glasgow Coma Scale (GCS) on stroke patients, i.e., body temperature (BT) and pulse rate (PR).

The rest of this paper is divided into several main topics. Section 2 introduces the materials and methods used in this study. Section 3 consists of five subsections: the theory of the new mixed estimator, its properties, the selection of the best model, the simulation study, and a case study. The details of the new mixed estimator are presented in Section 3.1, followed by its properties in Section 3.2. Section 3.3 presents how to select the optimum knot point, oscillation parameter, and smoothing parameter to obtain the best model. The results of a simulation study of the proposed method are presented in Section 3.4, followed by an application to real data in Section 3.5. Section 4 concludes the paper.

2. Materials and Methods

Suppose y_{it} is the response variable and x_{it} and z_{it} are the corresponding predictor variables with sample size of n subjects $(i = 1, 2, \ldots, n)$, with each subject having T observations $(t = 1, 2, \ldots, T)$. The relations between the response and predictors for the nonparametric regression model for longitudinal data are

$$y_{it} = \mu\left(x_{1it}, \ldots, x_{pit}, z_{1it}, \ldots, z_{qit}\right) + \varepsilon_{it}, \tag{1}$$

where μ is the regression curve and ε_{it} is a random error. Assume that the form of the regression curve μ is unknown and additive, so that

$$\mu\left(x_{1it}, \ldots, x_{pit}, z_{1it}, \ldots, z_{qit}\right) = \sum_{j=1}^{p} f_{ji}\left(x_{jit}\right) + \sum_{k=1}^{q} g_{ki}\left(z_{kit}\right), \tag{2}$$

where $\sum_{j=1}^{p} f_{ji}\left(x_{jit}\right)$ is the truncated spline component and $\sum_{k=1}^{q} g_{ki}\left(z_{kit}\right)$ is the Fourier series component.

The function f_{ji}, $j = 1, 2, \ldots p$, is an approximation using truncated spline functions and g_{ki}, $k = 1, 2, \ldots, q$, is that using Fourier series. The estimator μ is obtained through a two-step optimization, i.e., penalized weighted least squares (PWLS) and weighted least squares (WLS).

3. Results and Discussion

3.1. MTSFS Model for Longitudinal Data. Some lemmas and theorems are provided to obtain an MTSFS model for longitudinal data. Lemma 1 presents the goodness of fit of the Fourier series component, Lemma 2 presents the penalty component, and Lemma 3 gives the solution for the truncated spline component. Theorem 1 presents the PWLS optimization, and Theorem 2 presents WLS optimization. The results are summarized as follows.

Lemma 1. *If the Fourier series component in equation (2) is given by $\sum_{k=1}^{q} g_{ki}\left(z_{kit}\right)$, then the goodness of fit can be formulated as follows:*

$$N^{-1}\left(\mathbf{y}^* - \mathbf{Zc}\right)' \mathbf{W}\left(\mathbf{y}^* - \mathbf{Zc}\right), \tag{3}$$

where $N = n \times T$, $\mathbf{W}$ is the $nT \times nT$ weighting matrix, $\mathbf{y}^* = \mathbf{y} - \mathbf{f}$, and

$$\mathbf{y}^* = \left[\mathbf{y}_1^{*\prime} \ \mathbf{y}_2^{*\prime} \ \ldots \ \mathbf{y}_n^{*\prime} \right]',$$

$$\mathbf{y}_i^* = \left[y_{i1}^* \ y_{i2}^* \ \cdots \ y_{iT}^* \right]',$$

$$\mathbf{f} = \left[\mathbf{f}_1' \ \mathbf{f}_2' \ \ldots \ \mathbf{f}_n' \right]',$$

$$\mathbf{f}_i = \left(\sum_{j=1}^{p} f_{ji}\left(x_{ji1}\right), \sum_{j=1}^{p} f_{ji}\left(x_{ji2}\right), \ldots, \sum_{j=1}^{p} f_{ji}\left(x_{jiT}\right) \right)',$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{Z}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{Z}_n \end{bmatrix},$$

$$
\mathbf{c} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_n \end{bmatrix},
$$

$$
\mathbf{Z}_i = \begin{bmatrix}
z_{1i1} & \frac{1}{2} & \cos z_{1i1} & \cos 2z_{1i1} & \dots & \cos Hz_{1i1} & \dots & z_{qi1} & \frac{1}{2} & \cos z_{qi1} & \cos 2z_{qi1} & \dots & \cos Hz_{qi1} \\
z_{1i2} & \frac{1}{2} & \cos z_{1i2} & \cos 2z_{1i2} & \dots & \cos Hz_{1i2} & \dots & z_{qi2} & \frac{1}{2} & \cos z_{qi2} & \cos 2z_{qi2} & \dots & \cos Hz_{qi2} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
z_{1iT} & \frac{1}{2} & \cos z_{1iT} & \cos 2z_{1iT} & \dots & \cos Hz_{1iT} & \dots & z_{qiT} & \frac{1}{2} & \cos z_{qiT} & \cos 2z_{qiT} & \dots & \cos Hz_{qiT}
\end{bmatrix},
$$

$$
\mathbf{c}_i = \begin{bmatrix} d_{1i} & c_{01i} & c_{11i} & c_{21i} & \dots & c_{H1i} & \dots & d_{qi} & c_{0qi} & c_{1qi} & c_{2qi} & \dots & c_{Hqi} \end{bmatrix}'.
$$

$$(4)$$

Proof. The function g_{ki} is assumed to be unknown and contained in the space $C(0, \pi)$. The function is approximated using Fourier series with a trend, modified from Bilodeau [3]:

$$
g_{ki}(z_{kit}) = d_{ki} z_{kit} + \frac{1}{2} c_{0ki} + \sum_{h=1}^{H} c_{hki} \cos h z_{kit}. \quad (5)
$$

Equation (5) can be written in matrix form

$$
\mathbf{g} = \begin{bmatrix} \mathbf{g}_1' \\ \mathbf{g}_2' \\ \vdots \\ \mathbf{g}_n' \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1 \mathbf{c}_1 \\ \mathbf{Z}_2 \mathbf{c}_2 \\ \vdots \\ \mathbf{Z}_n \mathbf{c}_n \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1 & 0 & \dots & 0 \\ 0 & \mathbf{Z}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{Z}_n \end{bmatrix} \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_n \end{bmatrix} = \mathbf{Z}\mathbf{c}, \quad (6)
$$

where $\mathbf{Z}$ is a $nT \times (2 + H)nq$ matrix and $\mathbf{c}$ is a $(2 + H)nq$ vector.

The nonparametric model in equation (2) can be rewritten as

$$
y_{it} - \sum_{j=1}^{p} f_{ji}(x_{jit}) = \sum_{k=1}^{q} g_{ki}(z_{kit}) + \varepsilon_{it},
$$

$$
y_{it}^* = \sum_{k=1}^{q} g_{ki}(z_{kit}) + \varepsilon_{it},
$$

$$
i = 1, 2, \dots, n, t = 1, 2, \dots, T.
$$

$$(7)$$

Then, the goodness of fit for equation (7) can be written as

$$
N^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} w_{it} \left(y_{it}^* - \sum_{k=1}^{q} \sum_{t=1}^{T} g_{ki}(z_{kit}) \right)^2
$$

$$
= N^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} w_{it} \left(y_{it}^* - \sum_{k=1}^{q} \sum_{t=1}^{T} \left(d_{ki} z_{kit} + \frac{1}{2} c_{0ki} + \sum_{h=1}^{H} c_{hki} \cos h z_{kit} \right) \right)^2.
$$

$$(8)$$

The model in equation (8) can be written in matrix form:

$$
N^{-1} \mathbf{W} (\mathbf{y}^* - \mathbf{Z}\mathbf{c})^2 = N^{-1} (\mathbf{y}^* - \mathbf{Z}\mathbf{c})' \mathbf{W} (\mathbf{y}^* - \mathbf{Z}\mathbf{c}). \quad (9)
$$
□

Lemma 2. *If the penalty component is given,*

$$
\sum_{k=1}^{q} \lambda_k \int_0^\pi \frac{2}{\pi} (g_{ki}''(z_k))^2 \, dz_k, \quad (10)
$$

then

$$
\sum_{k=1}^{q} \lambda_k \int_0^\pi \frac{2}{\pi} (g_{ki}''(z_k))^2 \, dz_k = c' \mathbf{D}(\lambda) \mathbf{c}, \quad (11)
$$

where

$$\mathbf{c} = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \dots & \mathbf{c}_n \end{bmatrix}',$$

$$\mathbf{c}_i = \begin{bmatrix} d_{1i} & c_{01i} & c_{11i} & c_{21i} & \dots & c_{H1i} & \dots & d_{qi} & c_{0qi} & c_{1qi} & c_{2qi} & \dots & c_{Hqi} \end{bmatrix}',$$

$$\mathbf{D}(\lambda) = \begin{bmatrix} \mathbf{D}_1(\lambda) & 0 & \dots & 0 \\ 0 & \mathbf{D}_2(\lambda) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{D}_n(\lambda) \end{bmatrix},$$

$$\mathbf{D}_i(\lambda) = \begin{bmatrix} \mathbf{d}_1 & 0 & \dots & 0 \\ 0 & \mathbf{d}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{d}_q \end{bmatrix},$$

$$\mathbf{I}_k = \begin{bmatrix} \lambda_k 1^4 & \lambda_k 2^4 & \dots & \lambda_k H^4 \end{bmatrix},$$

$$\mathbf{d}_k = \operatorname{diag}(0, 0, \mathbf{I}_k), \quad k = 1, 2, \dots, q. \tag{12}$$

Proof. Regarding equation (5), we define

$$g_{ki}''(z_k) = \frac{d}{dz_k}\left[\frac{d}{dz_k}\left(d_{ki}z_{kit} + \frac{1}{2}c_{0ki} + \sum_{h=1}^{H} c_{hki}\cos h z_{kit}\right)\right]$$

$$= -\sum_{h=1}^{H} h^2 c_{hki}\cos h z_{kit}. \tag{13}$$

Consequently,

$$P_k(g_k) = \int_0^\pi \frac{2}{\pi}\left(\sum_{h=1}^{H} h^2 c_{hki}\cos h z_{kit}\right)^2 dz_k$$

$$= \frac{2}{\pi}\int_0^\pi \left[\left(\sum_{h=1}^{H} h^2 c_{hki}\cos h z_{kit}\right)^2\right.$$

$$\left.+2\sum_{h<m}^{H}\left(h^2 c_{hki}\cos h z_{kit}\right)\left(m^2 c_{mki}\cos m z_{kit}\right)\right]dz_k. \tag{14}$$

Let

$$A = \frac{2}{\pi}\sum_{h=1}^{H}\int_0^\pi \left(h^2 c_{hki}\cos h z_{kit}\right)^2 dz_k, \tag{15}$$

$$B = \frac{2}{\pi}2\sum_{h<l}^{H}\int_0^\pi \left(h^2 c_{hki}\cos h z_{kit}\right)\left(l^2 c_{lki}\cos l z_{kit}\right)dz_k.$$

Thus,

$$P_k(g_k) = A + B = \sum_{h=1}^{H} h^4 c_{hki}^2. \tag{16}$$

The penalty in equation (16) can be rewritten in matrix form

$$c'\mathbf{D}(\lambda)\mathbf{c}. \tag{17}$$

Regarding the goodness of fit in Lemma 1 and the penalty component in Lemma 2, we obtain the PWLS optimization as

$$N^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T} w_{it}\left(y_{it}^* - \sum_{k=1}^{q} g_{ki}(z_{kit})\right)^2$$

$$+ \sum_{k=1}^{q}\lambda_k \int_0^\pi \frac{2}{\pi}\left(g_{ki}''(z_k)\right)^2 dz_k, \quad 0 < \lambda_k < \infty. \tag{18}$$

Equation (18) can be rewritten in the form

$$N^{-1}(\mathbf{y}^* - \mathbf{Zc})'\mathbf{W}(\mathbf{y}^* - \mathbf{Zc}) + c'\mathbf{D}(\lambda)\mathbf{c}. \tag{19}$$

Theorem 1. *If the goodness of fit is as in Lemma 1 and the penalty component is as in Lemma 2, then the Fourier series component obtained by minimizing the PWLS in equation (18) is*

$$\widehat{g}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z}) = \mathbf{L}\mathbf{y}^*, \tag{20}$$

where $\mathbf{y}^* = \mathbf{y} - \mathbf{f}$ *and* $\mathbf{L} = \mathbf{Z}[Z'\mathbf{W}\mathbf{Z} + N\mathbf{D}(\lambda)]^{-1}Z'\mathbf{W}.$

Proof. The optimization in equation (18) can be written as

$$\underset{g_k \in C(0,\pi)}{\text{Min}} \left\{ N^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} w_{it} \left(y_{it}^* - \sum_{k=1}^{q} \sum_{t=1}^{T} g_{ki}(z_{kit}) \right)^2 + \sum_{k=1}^{q} \lambda_k \int_{0}^{\pi} \frac{2}{\pi} (g_k''(z_k))^2 dz_k \right\}. \tag{21}$$

Equation (21) can be rewritten in the form

$$\underset{\mathbf{c} \in C(0,\pi)}{\text{Min}} \left\{ N^{-1} (\mathbf{y}^* - \mathbf{Zc})' \mathbf{W} (\mathbf{y}^* - \mathbf{Zc}) + c' \mathbf{D}(\lambda) \mathbf{c} \right\}$$
$$= \underset{\mathbf{c} \in C(0,\pi)}{\text{Min}} \{Q(\mathbf{c})\}. \tag{22}$$

We obtain

$$Q(\mathbf{c}) = N^{-1} \mathbf{y}^{*'} \mathbf{W} \mathbf{y}^* - 2N^{-1} c' Z' \mathbf{W} \mathbf{y}^*$$
$$+ N^{-1} c' Z' \mathbf{W} \mathbf{Z} \mathbf{c} + c' \mathbf{D}(\lambda) \mathbf{c}. \tag{23}$$

The completion of the optimization in (23) is obtained by taking the partial derivative of $Q(\mathbf{c})$ with respect to $\mathbf{c}$ and setting it equal to zero, i.e.,

$$\frac{\partial Q(\mathbf{c})}{\partial \mathbf{c}} = 0, \tag{24}$$

giving the result

$$\hat{c} = [Z' \mathbf{W} \mathbf{Z} + N \mathbf{D}(\lambda)]^{-1} Z' \mathbf{W} \mathbf{y}^*. \tag{25}$$

By substituting (25) into (6), we obtain

$$\hat{g}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z}) = \mathbf{Z}\hat{c} = \mathbf{Z}[Z' \mathbf{W} \mathbf{Z} + N \mathbf{D}(\lambda)]^{-1} Z' \mathbf{W} \mathbf{y}^* = \mathbf{L} \mathbf{y}^*. \tag{26}$$

The nonparametric regression model in equation (7) can be written as

$$\mathbf{y}^* = \hat{g}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z}) = \mathbf{L} \mathbf{y}^*, \tag{27}$$

where $\mathbf{y}^* = \mathbf{y} - \mathbf{f}$ and $\mathbf{L} = \mathbf{Z}[Z' \mathbf{W} \mathbf{Z} + N \mathbf{D}(\lambda)]^{-1} Z' \mathbf{W}$. $\square$

Lemma 3. *If the truncated spline component in equation (2) is $\sum_{j=1}^{p} f_{ji}(x_{jit})$, then the WLS is*

$$[(\mathbf{I} - \mathbf{L})\mathbf{y} - (\mathbf{I} - \mathbf{L})\mathbf{M}\boldsymbol{\gamma}]' \mathbf{W} [(\mathbf{I} - \mathbf{L})\mathbf{y} - (\mathbf{I} - \mathbf{L})\mathbf{M}\boldsymbol{\gamma}], \tag{28}$$

where $\mathbf{L} = \mathbf{Z}[Z' \mathbf{W} \mathbf{Z} + N \mathbf{D}(\lambda)]^{-1} Z' \mathbf{W}$.

Proof. The function f_{ji} is a linear truncated spline function with s knot for each x_j, $j = 1, 2, \ldots, p$:

$$f_{ji}(x_{jit}) = \alpha_{ji} x_{jit} + \sum_{u=1}^{s} \beta_{uji}(x_{jit} - K_{uji})_+, \tag{29}$$

where $(x_{jit} - K_{uji})_+ = \begin{cases} (x_{jit} - K_{uji}), & x_{jit} \geq K_{uji}, \\ 0, & x_{jit} < K_{uji} \end{cases}$.

Equation (29) can be rewritten in matrix form

$$\mathbf{f} = [\mathbf{X}|\mathbf{S}] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} = \mathbf{M}\boldsymbol{\gamma}, \tag{30}$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{X}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{X}_n \end{bmatrix},$$

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{S}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{S}_n \end{bmatrix},$$

$$\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_n \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_n \end{bmatrix},$$

$$
\mathbf{X}_i = \begin{bmatrix} x_{1i1} & x_{2i1} & \cdots & x_{pi1} \\ x_{1i2} & x_{2i2} & \cdots & x_{pi2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1iT} & x_{2iT} & \cdots & x_{piT} \end{bmatrix},
$$

$$
\boldsymbol{\alpha}_i = \begin{bmatrix} \alpha_{1i} \\ \alpha_{2i} \\ \vdots \\ \alpha_{pi} \end{bmatrix},
$$

$$
\mathbf{S}_i = \begin{bmatrix} \left(x_{1i1} - K_{11i}\right)^1_+ & \cdots & \left(x_{1i1} - K_{s1i}\right)^1_+ & \cdots & \left(x_{pi1} - K_{1pi}\right)^1_+ & \cdots & \left(x_{pi1} - K_{spi}\right)^1_+ \\ \left(x_{1i2} - K_{11i}\right)^1_+ & \cdots & \left(x_{1i2} - K_{s1i}\right)^1_+ & \cdots & \left(x_{pi2} - K_{1pi}\right)^1_+ & \cdots & \left(x_{pi2} - K_{spi}\right)^1_+ \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \left(x_{1iT} - K_{11i}\right)^1_+ & \cdots & \left(x_{1iT} - K_{s1i}\right)^1_+ & \cdots & \left(x_{piT} - K_{1pi}\right)^1_+ & \cdots & \left(x_{piT} - K_{spi}\right)^1_+ \end{bmatrix},
$$

$$
\boldsymbol{\beta}_i = \begin{bmatrix} \beta_{11i} & \cdots & \beta_{s1i} & \cdots & \beta_{1pi} & \cdots & \beta_{spi} \end{bmatrix}^T.
$$

$$(31)$$

The MTSFS model for longitudinal data in equation (1) can be written in the form

$$
\mathbf{y} = \mathbf{f} + \mathbf{g} + \boldsymbol{\varepsilon}. \tag{32}
$$

By substituting (26) into (32), we obtain

$$
\mathbf{y} = \mathbf{f} + \mathbf{L}\mathbf{y}^* + \boldsymbol{\varepsilon}. \tag{33}
$$

To obtain the estimator of truncated spline component, equation (33) can be written as

$$
\mathbf{y} - \mathbf{L}\mathbf{y}^* = \mathbf{f} + \boldsymbol{\varepsilon},
$$
$$
\mathbf{y} - \mathbf{L}(\mathbf{y} - \mathbf{f}) = \mathbf{f} + \boldsymbol{\varepsilon}, \tag{34}
$$
$$
(\mathbf{I} - \mathbf{L})\mathbf{y} = (\mathbf{I} - \mathbf{L})\mathbf{f} + \boldsymbol{\varepsilon}.
$$

Equation (34) can be rewritten as

$$
(\mathbf{I} - \mathbf{L})\mathbf{y} = (\mathbf{I} - \mathbf{L})\mathbf{M}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{35}
$$

Thus,

$$
\boldsymbol{\varepsilon} = (\mathbf{I} - \mathbf{L})\mathbf{y} - (\mathbf{I} - \mathbf{L})\mathbf{M}\boldsymbol{\gamma}. \tag{36}
$$

As a consequence, the WLS is given by

$$
\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = [(\mathbf{I} - \mathbf{L})\mathbf{y} - (\mathbf{I} - \mathbf{L})\mathbf{M}\boldsymbol{\gamma}]'\mathbf{W}[(\mathbf{I} - \mathbf{L})\mathbf{y} - (\mathbf{I} - \mathbf{L})\mathbf{M}\boldsymbol{\gamma}]. \tag{37}
$$

□

Theorem 2. *Suppose the WLS is as in Lemma 3. Then, the mixed estimator obtained by minimizing WLS is*

$$
\widehat{f}_{(\mathbf{k},\mathbf{h},\boldsymbol{\lambda})}(\mathbf{x}, \mathbf{z}) = \mathbf{M}\mathbf{J}^{-1}\mathbf{K}\mathbf{y}, \tag{38}
$$

where $\mathbf{J} = (2 - \mathbf{L}')\mathbf{W}\mathbf{L}\mathbf{M} - \mathbf{W}\mathbf{M}$ *and* $\mathbf{K} = [(\mathbf{L}' - \mathbf{I})\mathbf{W}(\mathbf{I} - \mathbf{L})]$.

Proof. The WLS optimization in Lemma 3 can be written as

$$
\underset{\boldsymbol{\gamma}}{\text{Min}}\{[(\mathbf{I} - \mathbf{L})\mathbf{y} - (\mathbf{I} - \mathbf{L})\mathbf{M}\boldsymbol{\gamma}]'\mathbf{W}[(\mathbf{I} - \mathbf{L})\mathbf{y} - (\mathbf{I} - \mathbf{L})\mathbf{M}\boldsymbol{\gamma}]\}
$$
$$
= \underset{\boldsymbol{\gamma}}{\text{Min}}\{Q(\boldsymbol{\gamma})\}. \tag{39}
$$

We obtain

$$
Q(\boldsymbol{\gamma}) = \{y'\mathbf{W}\mathbf{y} + y'L'\mathbf{W}\mathbf{L}\mathbf{y} + \gamma'M'L'\mathbf{W}\mathbf{L}\mathbf{M}\boldsymbol{\gamma} + \gamma'M'\mathbf{W}\mathbf{M}\boldsymbol{\gamma} - 2y'L'\mathbf{W}\mathbf{y}
$$
$$
+ 2\gamma'M'L'\mathbf{W}\mathbf{y} + -2N^{-1}\gamma'M'\mathbf{W}\mathbf{y} - 2N^{-1}\gamma'M'L'\mathbf{W}\mathbf{L}\mathbf{y} + 2N^{-1}\gamma'M'\mathbf{W}\mathbf{L}\mathbf{y} - 2N^{-1}\gamma'M'\mathbf{W}\mathbf{L}\mathbf{M}\boldsymbol{\gamma}\}. \tag{40}
$$

The complete optimization is obtained by setting equal to zero the partial derivative of $Q(\gamma)$ with respect to γ, that is,

$$\frac{\partial Q\boldsymbol{\gamma}}{\partial \boldsymbol{\gamma}} = 0, \tag{41}$$

giving the result

$$\widehat{\gamma} = [(2 - \mathbf{L}')\mathbf{WLM} - \mathbf{WM}]^{-1}[(\mathbf{L}' - \mathbf{I})\mathbf{W}(\mathbf{I} - \mathbf{L})]\mathbf{y} = \mathbf{J}^{-1}\mathbf{Ky}. \tag{42}$$

By substituting $\widehat{\gamma}$ into equation (30), we obtain

$$\widehat{f}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z}) = \mathbf{M}\widehat{\gamma} = \mathbf{MJ}^{-1}\mathbf{Ky} = \mathbf{A}(\mathbf{k}, \mathbf{h}, \lambda)\mathbf{y}. \tag{43}$$

We obtain $\widehat{c}$ by substituting (43) into (25):

$$
\begin{aligned}
\widehat{c} &= [Z'\mathbf{WZ} + N\mathbf{D}(\lambda)]^{-1}Z'\mathbf{Wy}^* = [Z'\mathbf{WZ} + N\mathbf{D}(\lambda)]^{-1}Z'\mathbf{W}(\mathbf{y} - \widehat{f}) \\
&= [Z'\mathbf{WZ} + N\mathbf{D}(\lambda)]^{-1}Z'\mathbf{W}(\mathbf{y} - \mathbf{M}\widehat{\gamma}) = [Z'\mathbf{WZ} + N\mathbf{D}(\lambda)]^{-1}Z'\mathbf{W}(\mathbf{y} - \mathbf{MJ}^{-1}\mathbf{Ky}) \\
&= [Z'\mathbf{WZ} + N\mathbf{D}(\lambda)]^{-1}Z'\mathbf{W}(\mathbf{I} - \mathbf{MJ}^{-1}\mathbf{K})\mathbf{y}.
\end{aligned}
\tag{44}
$$

As a consequence,

$$
\begin{aligned}
\widehat{g}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z}) &= \mathbf{Z}\widehat{c} \\
&= \mathbf{Z}[Z'\mathbf{WZ} + N\mathbf{D}(\lambda)]^{-1}Z'\mathbf{W}(\mathbf{I} - \mathbf{MJ}^{-1}\mathbf{K})\mathbf{y} \\
&= \mathbf{L}(\mathbf{I} - \mathbf{MJ}^{-1}\mathbf{K})\mathbf{y} = \mathbf{B}(\mathbf{k}, \mathbf{h}, \lambda)\mathbf{y}.
\end{aligned}
\tag{45}
$$

By substituting $\widehat{\gamma}$ and $\widehat{c}$ into the MTSFS model for longitudinal data, we obtain

$$
\begin{aligned}
\widehat{y} = \widehat{\mu}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z}) &= \widehat{f}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z}) + \widehat{g}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z}) = \mathbf{M}\widehat{\gamma} + \mathbf{Z}\widehat{c} \\
&= \mathbf{MJ}^{-1}\mathbf{Ky} + \mathbf{L}(\mathbf{I} - \mathbf{MJ}^{-1}\mathbf{K})\mathbf{y} \\
&= [\mathbf{MJ}^{-1}\mathbf{K} + \mathbf{L}(\mathbf{I} - \mathbf{MJ}^{-1}\mathbf{K})]\mathbf{y} \\
&= \mathbf{C}(\mathbf{k}, \mathbf{h}, \lambda)\mathbf{y},
\end{aligned}
\tag{46}
$$

where $\mathbf{J} = (2 - L')\mathbf{WLM} - \mathbf{WM}$, $\mathbf{K} = [(L' - \mathbf{I})\mathbf{W}(\mathbf{I} - \mathbf{L})]$, and $\mathbf{L} = \mathbf{Z}[Z'\mathbf{WZ} + N\mathbf{D}(\lambda)]^{-1}Z'\mathbf{W}$.

3.2. The Properties of MTSFS Model for Longitudinal Data.
This section provides the MTSFS model's properties for longitudinal data, i.e., it is biased and linear in observations. The mixed estimator is biased, as proved by

$$
\begin{aligned}
E\left[\widehat{\mu}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z})\right] &= E\left[\widehat{f}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z}) + \widehat{g}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z})\right] \\
&= E[\mathbf{A}(\mathbf{k}, \mathbf{h}, \lambda)\mathbf{y} + \mathbf{B}(\mathbf{k}, \mathbf{h}, \lambda)\mathbf{y}] \\
&= E[\mathbf{C}(\mathbf{k}, \mathbf{h}, \lambda)\mathbf{y}] = \mathbf{C}(\mathbf{k}, \mathbf{h}, \lambda)E(\mathbf{y}).
\end{aligned}
\tag{47}
$$

Assuming that $\mathbf{C}(\mathbf{k}, \mathbf{h}, \lambda) \neq \mathbf{I}$,

$$
\begin{aligned}
E\left[\widehat{\mu}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z})\right] &= \mathbf{C}(\mathbf{k}, \mathbf{h}, \lambda)E(\mathbf{y}) \\
&= \mathbf{C}(\mathbf{k}, \mathbf{h}, \lambda)\left(\mathbf{f}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z}) + \mathbf{g}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z})\right) \\
&= \mathbf{C}(\mathbf{k}, \mathbf{h}, \lambda)\boldsymbol{\mu}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z}).
\end{aligned}
\tag{48}
$$

The result in equation (48) showed that the mixed estimator is biased because $E[\widehat{\mu}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z})] \neq \mu_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z})$. Even though the mixed estimator is biased, the mixed estimator is linear in observations proved by equation (49) below.

$$
\begin{aligned}
\widehat{\mu}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z}) &= \widehat{f}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z}) + \widehat{g}_{(\mathbf{k},\mathbf{h},\lambda)}(\mathbf{x}, \mathbf{z}) \\
&= \mathbf{A}(\mathbf{k}, \mathbf{h}, \lambda)\mathbf{y} + \mathbf{B}(\mathbf{k}, \mathbf{h}, \lambda)\mathbf{y} \\
&= \mathbf{C}(\mathbf{k}, \mathbf{h}, \lambda)\mathbf{y}.
\end{aligned}
\tag{49}
$$

3.3. The Selection of the Optimal Number of Knot Points, Oscillation Parameter, and Smoothing Parameter.
The MTSFS model is very dependent on the number of the knot points, oscillation parameter, and smoothing parameter. The best model is obtained by using the optimal values of these parameters. In semiparametric and nonparametric regression, there are several methods to obtain the best regression model. One of the popular methods is generalized cross-validation (GCV).

Craven and Wahba [8] introduced a modified cross-validation (CV) method called GCV. This method was developed to overcome the shortcomings of complex CV calculations. GCV has several advantages: it is simple and efficient in calculation, invariant to transformation, and does not require information about variant. In addition, GCV has better asymptotic properties than other methods [9, 10].

In this study, the value of GCV is a criterion that can be used to determine the best model from variety of knots, oscillations, and smoothing parameters. The criterion for selecting the best model includes taking the model with the lowest GCV value. The modified GCV method of the MTSFS model for longitudinal data is stated as follows.

$$\text{GCV}(\mathbf{k}, \mathbf{h}, \lambda) = \frac{N^{-1}\|(\mathbf{I} - \mathbf{C}(\mathbf{k}, \mathbf{h}, \lambda))\mathbf{y}\|^2}{\left[N^{-1}\text{trace}(\mathbf{I} - \mathbf{C}(\mathbf{k}, \mathbf{h}, \lambda))\right]^2}. \tag{50}$$

The optimal number of knot points, oscillation, and smoothing parameters is obtained by minimizing GCV $(\mathbf{k}, \mathbf{h}, \lambda)$.

3.4. Simulation Study. This section presents the use of the MTSFS model on simulation data to see the performance of the estimators obtained. The simulation was done with two predictor variables, sample size $n = 20$, and varied number of time points, $T = 5, 10, 15$. We considered 20 models for each subject generated from a formula that contains two different functions representing the truncated spline and Fourier series pattern. A polynomial function is used to present the truncated spline, while trigonometric functions are used to present the Fourier series. The predictors are generated from $U(0, 1)$, and the random errors ε_{it} are generated from a multivariate normal distribution.

The weight matrix is user-specified [6], and in this paper, we use $\mathbf{W} = N^{-1}\mathbf{I}$ so that each of the measurements is treated equally. In this study, we use two numbers of knots ($K = 1$ and $K = 2$) and several oscillation parameters ($H = 1, 2, 3, 4$). These simulation studies based on GCV are presented in Table 1.

Based on Table 1, for the model with $T = 5$, it appears that the smallest GCV value occurs when the oscillation $H = 3$ for $K = 1$ and $K = 2$. The same is also seen at $T = 10$ and $T = 15$. The smallest GCV occurs when the oscillation $H = 3$ for both knot points.

In general, the larger the number of time points, the smaller the resulting value of the GCV. Furthermore, the greater the number of knot points, the larger the GCV. Other results show that a larger oscillation is not guaranteed to produce large or small values of the GCV. So, it is necessary to choose the optimum oscillation that produces the smallest GCV.

3.5. Application to Real Data. The MTSFS model for longitudinal data obtained has been applied to the stroke patient dataset. These data were taken after an initial study of GCS on stroke patients and the factors that influenced it. The pattern of the relations between the predictors and response followed the characteristics of a truncated spline and a Fourier series. There is a predictor with the form of a truncated spline, which is changing in certain subintervals. The other predictor has the form of a Fourier series, with a repeating pattern.

Stroke is a noncommunicable disease. The number of those who suffer from it continues to increase in the world. In 2013, stroke was the second leading cause of death globally (11.8% of all deaths) after ischemic heart disease (14.8% of all deaths). Besides, stroke is the third cause of disability, namely, 4.5% of all causes of disability [11]. Based on the Global Burden of Disease (GBD) Study 2016, the estimated global lifetime risk of stroke for those aged 25 years or older almost reached 25% [12]. The global prevalence of stroke in 2017 was 104.2 million people. In addition, age-standardized stroke prevalence rates were highest in Eastern Europe, North Africa, the Middle East,

and Central and East Asia. Several countries in Europe, North Africa, and Central Asia have the highest rates of stroke mortality. Indonesia is one of the countries with the highest death rate due to stroke. According to the Indonesia Basic Health Survey 2007, stroke was the highest cause of death (15.4%) [13]. The prevalence of stroke in Indonesia in 2013 was 7% and increased to 10.9%, according to the Indonesia Basic Health Survey 2018. Furthermore, the stroke prevalence for those aged 15 years or older was 10.85%.

Stroke patients often experience head injuries due to falls. Trauma or head injury requires supervision to ensure further medical treatment. GCS was initially used to assess consciousness level after head injury and is now used in the medical field of both acute and trauma patients. According to Champion [14], trauma severity assessment is used to measure the severity of the injury and describe the patient's case's severity. An injury to the head will activate the immune response and release of interleukins by activating white blood cells and increasing the temperature. Injury to the hypothalamus also increases interference with body temperature regulation [15]. Fever is generally defined as an increase in body temperature above average and has been identified as one of the causes of worsening head injury [16]. Based on previous studies, the pulse was found to be dysfunctional in patients with severe brain injury. Changes in pulse rate are associated with increased mortality in brain injury patients. Therefore, changes in patients' pulse rates with severe brain injury should be carefully observed [17].

This article uses GCS as a response variable with 20 stroke patients ($n = 20$) with 14-day ($T = 14$) measurements for each subject. Simultaneously, the predictor variables are body temperature (BT) and pulse rate (PR). The partial relationship between GCS and each predictor variable is illustrated in Figure 1. The plot in Figure 1 shows a different pattern for each predictor. For this reason, we propose the MTSFS model for longitudinal data. Body temperature will be approximated by a truncated spline estimator, while the pulse rate will be approximated by a Fourier series estimator.

Based on Table 2, some scenarios were performed to compare the effectiveness of the proposed model, i.e., using a single estimator and the mixed estimator. By using the GCV formula in Section 3.3, we selected the best model according to the minimum GCV criterion. As shown in Table 2, modeling GCS in stroke patients produces the smallest GCV value of 45.3517, achieved at when the number of knots is $K = 1$ and number of oscillation is $H = 3$ with $\lambda = 10$. This study revealed that the MTSFS model is better than using a single estimator. This model yields a RMSE = 2.43. The important result was that the number of knots and the number of the oscillation parameters of the best model in the simulation study are in line with those for the case study. Similar to the simulation study, the best model had one knot and the oscillation parameter H equal to three.

Figure 2 presents a comparison between the response variable (red line) and fitted values (blue line) using the

TABLE 1: Comparison of the values of GCV based on variations in number of time points.

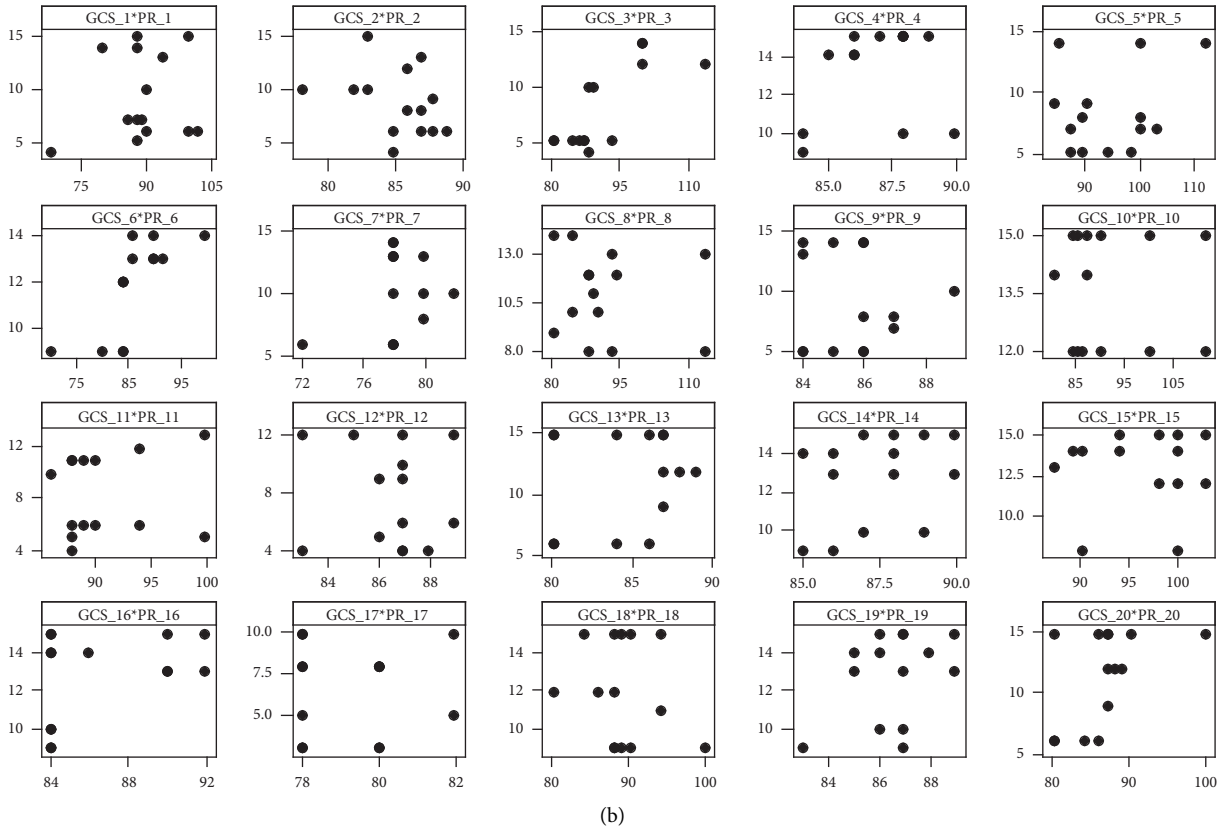| Number of time points	Number of knots	Number of oscillations	λ	GCV
5	1	1	10.0000	4.9326
		2	8.7755	4.9349
		3	0.8163	4.0640
		4	10.0000	4.9211
	2	1	10.0000	5.9051
		2	0.2041	5.9128
		3	0.8163	4.7601
		4	10.0000	5.9013
10	1	1	0.2041	3.8454
		2	10.0000	3.8051
		3	0.8163	2.7655
		4	10.0000	3.8482
	2	1	0.2041	4.1491
		2	10.0000	4.1060
		3	0.8163	2.9306
		4	10.0000	4.1519
15	1	1	0.2041	3.8133
		2	7.9592	3.8207
		3	0.8163	2.5196
		4	10.0000	3.8124
	2	1	0.2041	3.9663
		2	5.1020	3.9717
		3	0.8163	2.5693
		4	10.0000	3.9660
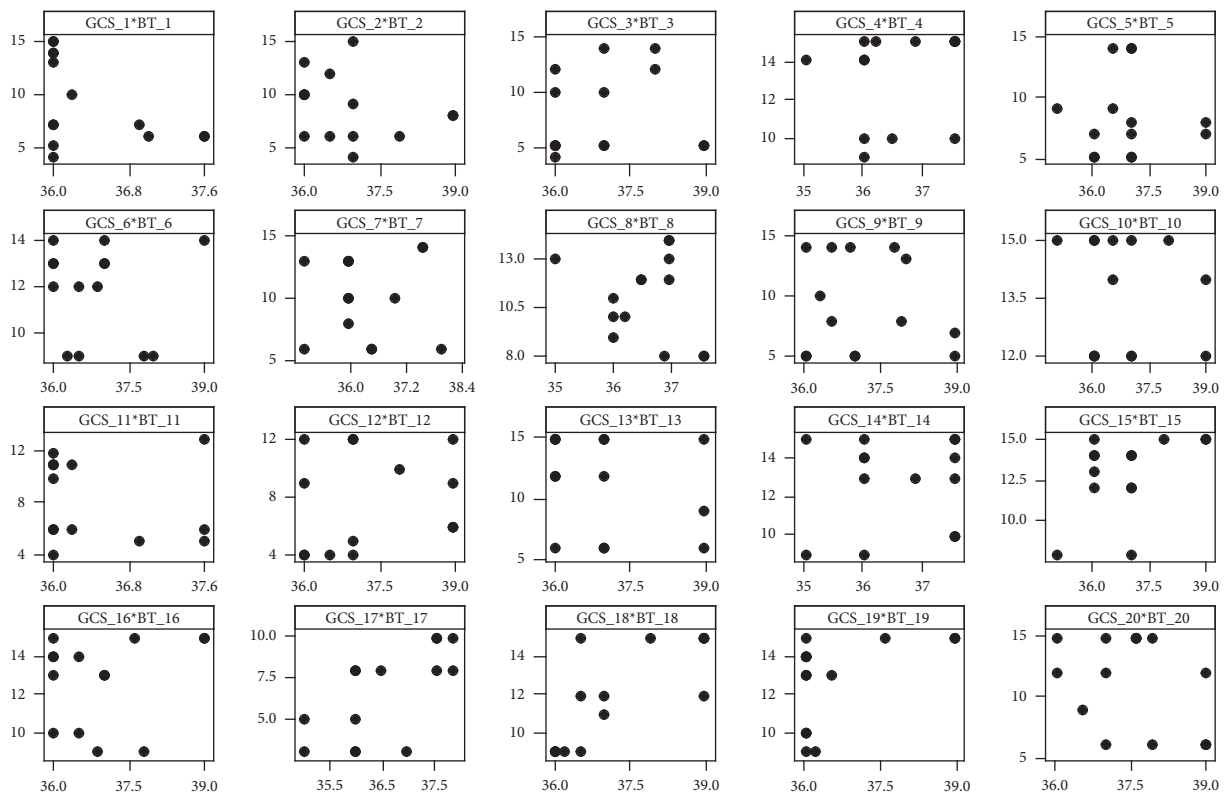
(a)

FIGURE 1: Continued.

(b)

Figure 1: Scatterplot of 20 stroke patients. (a) GCS and body temperature. (b) GCS and pulse rate.

Table 2: Summary of case study results.

Model	Number of knots	Number of oscillations	λ	GCV
Truncated spline estimator	1	—	—	391858.1
Fourier series estimator	—	1	—	323174.4
		2		161621.0
		3		101785.9
		4		75480.3
Mixed truncated spline and Fourier series	1	1	0.2041	48.5434
		2	10.0000	47.4535
		3	10.0000	**45.3517**
		4	9.3878	45.3519
	2	1	0.2041	51.6948
		2	10.0000	50.6872
		3	9.3878	48.7345
		4	7.1429	48.7347

MTSFS model. From the graph, we can see that some of the fitted values resemble the pattern of the real data; although some do not, there are deviations. Despite its limitations, the present study certainly adds to our understanding of the theory of the new mixed estimator in nonparametric regression for longitudinal data.
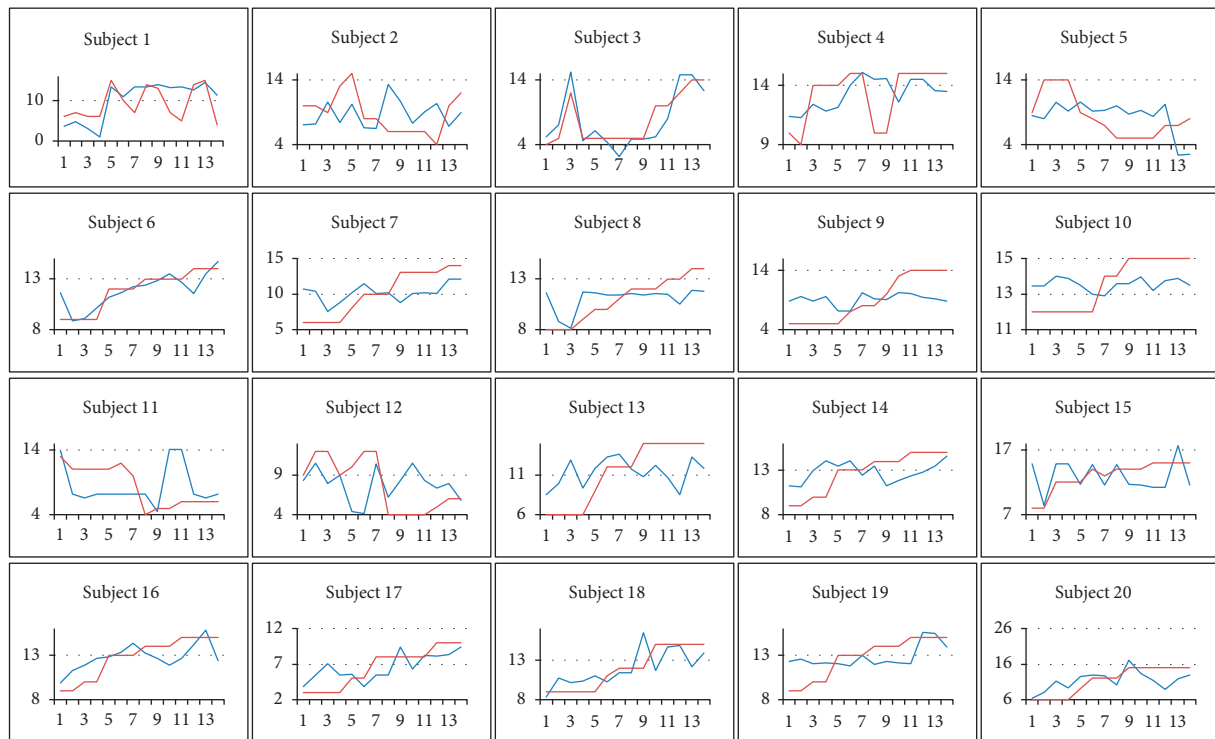
Figure 2: Comparison between actual and fitted values for each subject.

4. Conclusions

This article presented a new mixed estimator for non-parametric regression models for longitudinal data, combining the truncated spline estimator and Fourier series to obtain better estimation results if the predictors have different data patterns. A new two-step method for estimating the parameters used penalized weighted least squares and weighted least squares.

In both the simulation study and case study, the best model was selected by using the minimum generalized cross-validation (GCV). A higher number of knots or oscillation parameter does not produce a high GCV. Therefore, several combinations of knots and oscillations had to be tried to determine the best model. Interestingly, the number of knots and oscillations of the best model is the same for the simulation study as for the case study with real data: one knot and three oscillations.

The proposed model estimator is a useful alternative for estimating nonparametric regression curves for longitudinal data. The presented results suggest that the next study can use different weighting matrices and more knots, so that researchers can compare the results for improving the model performance. Another possible line of further research could be simulation studies using another function to evaluate the performance of the proposed model.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

References

[1] M.-Y. Cheng, R. L. Paige, S. Sun, and K. Yan, "Variance reduction for kernel estimators in clustered/longitudinal data analysis," *Journal of Statistical Planning and Inference*, vol. 140, no. 6, pp. 1389–1397, 2010.

[2] E. L. Montoya, N. Ulloa, and V. Miller, "A simulation study comparing knot selection methods with equally spaced knots in a penalized regression spline," *International Journal of Statistics and Probability*, vol. 3, pp. 96–110, 2014.

[3] M. Bilodeau, "Fourier smoother and additive models," *Canadian Journal of Statistics*, vol. 20, no. 3, pp. 257–269, 1992.

[4] A. Tripena, "Estimator deret fourier untuk estimasi kurva regresi nonparametrik birespon," *Magistra*, vol. 84, pp. 6–15, 2013.

[5] X. Gong and J. Gao, "Nonparametric kernel estimation of the impact of tax policy on the demand for private health insurance in Australia," *Australian & New Zealand Journal of Statistics*, vol. 60, no. 3, pp. 374–393, 2018.

[6] H. Wu and J. T. Zhang, *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches*, Wiley Series in Probability and Statistics, 2006.

[7] M. A. D. Octavanny, I. N. Budiantara, H. Kuswanto, and D. P. Rahmawati, "Nonparametric regression model for longitudinal data with mixed truncated spline and fourier series," *Abstract and Applied Analysis*, vol. 2020, Article ID 4710745, 11 pages, 2020.

[8] P. Craven and G. Wahba, "Smoothing noisy data with spline functions," *Numerische Mathematik*, vol. 31, no. 4, pp. 377–403, 1978.

[9] G. Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia, 1990.

[10] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.

[11] V. L. Feigin, B. Norrving, and G. A. Mensah, "Global burden of stroke," *Circulation Research*, vol. 120, no. 3, pp. 439–448, 2017.

[12] D. Roth, "GBD 2016 lifetime risk of stroke collaborators et al., "global, regional, and country-specific lifetime risks of stroke, 1990 and 2016," *New England Journal of Medicine*, vol. 379, pp. 2429–2437, 2018.

[13] National Institute of Health Research and Development, *Basic Health Survey 2007 Report*, National Institute of Health Research and Development, Indonesian Ministry of Health, Jakarta, 2008.

[14] H. R. Champion, "Trauma scoring," *Scandinavian Journal of Surgery*, vol. 91, no. 1, pp. 12–22, 2002.

[15] L. K. Madden and H. A. Devon, "A systematic review of the effects of body temperature on outcome after adult traumatic brain injury," *Journal of Neuroscience Nursing*, vol. 47, no. 4, pp. 190–203, 2015.

[16] C. Childs, A. Vail, P. Leach, T. Rainey, R. Protheroe, and A. King, "Brain temperature and outcome after severe traumatic brain injury," *Neurocritical Care*, vol. 5, no. 1, pp. 10–14, 2006.

[17] A. Inoue, T. Hifumi, Y. Kuroda et al., "Mild decrease in heart rate during early phase of targeted temperature management following tachycardia on admission is associated with unfavorable neurological outcomes after severe traumatic brain injury: a post hoc analysis of a multicenter randomized," *Critical Care*, vol. 22, pp. 1–9, 2018.