Semiparametric regression curve estimation for longitudinal data using mixed spline truncated and fourier series estimator

A W Wening\(^1\), I N Budiantara\(^1\), and I Zain\(^1\)

\(^1\)Department of Statistics, Faculty of Mathematics, Computation, and Data Science, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia

Email: nyomanbudiantara65@gmail.com

Abstract. Semiparametric regression is a regression approach that is used when the form of regression curve is assumed to be partly known and partly unknown. In the semiparametric regression model, the same type of estimation method is generally used for some or all of the predictor variables. There are many examples of data cases that have different patterns on each predictor variable, so if the data is forced to use only one form of estimator to estimate the regression curve, it will produce estimation that will not be appropriate to the data pattern. Therefore, a mixed estimator needs to be developed. However, mixed estimator in previous studies was only able to handle cross-sectional data. So this study uses longitudinal data that has advantages, one of which is estimation of each individual characteristic and time (period) characteristic are obtained separately in longitudinal data. Therefore, this study develops a mixed Spline Truncated and Fourier Series estimator in semiparametric regression for longitudinal data using the Weighted Least Square method. Estimation of the semiparametric regression model for longitudinal data using mixed estimator of Spline Truncated and Fourier Series is \( \hat{y} = D(K,m)y \) where \( D(K,m) = XA(K,m) + TB(K,m) + ZC(K,m) \). The selection of the best model is based on knot points and optimum oscillation parameters that should be selected optimally using minimum Generalized Cross Validation (GCV) value.

1. Introduction

Investigating the pattern of functional relationships between response variables and predictor variables through curve estimation can be done by regression analysis [1]. There are three approaches to estimate the regression curve, such as: parametric regression, nonparametric regression, and semiparametric regression [2,3]. If the form of the regression curve is known, the parametric regression approach is used. If the relationship between the response variable and the predictor variable is not known, the nonparametric regression approach is used. Whereas if the regression curve form is partly assumed to be known and partly unknown, then the semiparametric regression approach is used [4].

In the nonparametric or semiparametric regression model, the same type of estimation method is generally used for some or all of the predictor variables. Data of many cases in the real world have different patterns on their predictor variables, so if the data is forced to use only one form of estimator to estimate the regression curve, it will produce estimation that will not be appropriate to the data pattern. Besides that, it tends to produce large errors. Therefore, to overcome this problem, a mixed
regression curve estimator is used where each data pattern in the regression model is approached with the appropriate curve estimator. Research involving mixed estimators has been conducted before; regarding mixed Spline Truncated and Kernel estimators [5], mixed Kernel and Fourier Series estimators [6,7], mixed Fourier Series and Spline Truncated estimators [8].

Spline has excellent statistical and visual interpretation and has high flexibility [9]. Whereas Fourier Series are trigonometric polynomials that have flexibility, so they can adapt effectively to the local nature of the data [10]. This Fourier series estimator is usually used if the data pattern is unknown and has a recurring tendency [11], where the repetition of the dependent variable values for different independent variables.

However, a mixed estimator still has several weaknesses. One of the weaknesses is only being able to handle cross-sectional data. Longitudinal data is the data that obtained from repeated observations of each subject at different time intervals [12, 13]. Longitudinal data is better able to identify and measure effects that cannot be detected in time-series or cross-sectional data [14], it can simultaneously estimate individual characteristics by observing the dynamics between time and each variable in the study. Research on the longitudinal data is still conducted with a single estimator and has not been done with a mixed estimator.

Based on the description given above, we will examine a mixed model of Spline Truncated and Fourier Series in semiparametric regression for longitudinal data. This semiparametric regression arises because of modeling cases where the relationships between the variables are not only linear but also unknown pattern.

2. Material and Method

This chapter will discuss the literature review and procedure used in this research.

2.1 Spline Truncated on Semiparametric Regression

Semiparametric regression is a combination of parametric components and nonparametric components [4]. Spline has a high flexibility characteristic and has a very good ability to handle data whose behavior changes at certain sub-intervals [15]. This advantage can be seen in the truncated function (pieces) which are called knots. Knot points are joint fusion points where the function changes in different sub-intervals. If a large number of knots is used, it will produce a very smooth regression curve. Conversely, for a small number of knots, it will provide a very rough regression curve shape [16].

For example, given the following data \( \left( y_i, x_i, t_i \right) \) with \( i = 1, 2, \ldots, n \) and relationship between \( y_i, x_i \) and \( t_i \) is assumed to follow the semiparametric regression model as follows:

\[
y_i = f(x_i) + g(t_i) + \epsilon_i
\]  

(1)

Based on Equation (1), \( y_i \) is response variable on the \( i \)th observation, \( f(x_i) \) is parametric regression function, \( g(t_i) \) is nonparametric regression function and \( \epsilon_i \) is random errors that are assumed to be independent and are normally distributed with zero mean and variance \( \sigma^2 \). Parametric regression function \( f(x_i) \) has \( p \) predictor variables, so that it can be stated in the form of a matrix below for \( i = 1, 2, \ldots, n \).

\[
\begin{bmatrix}
f(x_1) \\
f(x_2) \\
\vdots \\
f(x_n)
\end{bmatrix} =
\begin{bmatrix}
1 & x_{11} & x_{21} & \cdots & x_{1p} \\
1 & x_{12} & x_{22} & \cdots & x_{2p} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{1n} & x_{2n} & \cdots & x_{pn}
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_p
\end{bmatrix}
\]

\[
f = X\beta
\]  

(2)
\( \tilde{f} \) is vector of function \( f \) sized \( n\times 1 \), \( X \) is a matrix sized \( n\times (p+1) \) and \( \tilde{\beta} \) is parameter vector to be estimated sized \((p+1)\times 1\). The regression curve \( g(t_i) \) in the equation (1) is approached by Spline Truncated linear function with knots \( K_1,K_2,...,K_r \). In general, the function of Spline Truncated with \( q \) is the number of Spline Truncated predictor variables can be presented in the form:

\[
g(t_1,t_2,\ldots,t_q) = \sum_{u=1}^{q} (\gamma_{u} f_{u} + \sum_{a=1}^{a_{u}} (t_{a} - K_{a_{u}}))
\]

\( \gamma_{u} f_{u} \) is polynomial component and \( \sum_{a=1}^{a_{u}} (t_{a} - K_{a_{u}}) \) is truncated component with truncated function as follows:

\[
(t_{a} - K_{a_{u}}) = \begin{cases} 
(t_{a} - K_{a_{u}}) & , t_i \geq K_u \\
0 & , t_i < K_u 
\end{cases}
\]

where \( \gamma_{u} \) and \( \gamma_{(i\_u)} \) with \( u=1,2,\ldots,r \) are unknown parameters, as well \( K_1,K_2,...,K_r \) are knot points, so for \( i=1,2,\ldots,n \) can be converted into a matrix as follows:

\[
\tilde{g} = T \tilde{\gamma}
\]

with \( \tilde{\gamma} = (\gamma_{11},\gamma_{21},\ldots,\gamma_{(i\_1)},\ldots,\gamma_{1q},\gamma_{2q},\ldots,\gamma_{(i\_q)}) \) and

\[
T = \begin{bmatrix}
t_{i1} & (t_{i1} - K_{i1}) & \ldots & (t_{iq} - K_{i1}) \\
t_{i2} & (t_{i2} - K_{i1}) & \ldots & (t_{iq} - K_{i2}) \\
\vdots & \vdots & \ddots & \vdots \\
t_{in} & (t_{in} - K_{i1}) & \ldots & (t_{in} - K_{in})
\end{bmatrix}
\]

2.2 Fourier Series on Semiparametric Regression

Fourier series estimators in nonparametric and semiparametric regression are generally used if the data has unknown patterns and there is a tendency for repetitive patterns [17]. Suppose given the following data \( (y_i, x_i, z_i) \) with \( i=1,2,\ldots,n \) and relationship between \( y_i, x_i, \) and \( z_i \) can be stated in the semiparametric regression model as follows:

\[
y_i = f(x_i) + h(z_i) + \epsilon_i
\]

\( y_i \) is response variable in the \( i^{th} \) observation, \( f(x_i) \) is parametric regression function which can be stated like equation (2), and \( \epsilon_i \) is random errors that are assumed to be independent and are normally distributed with zero mean and variance \( \sigma^2 \). While \( h(z_i) \) is nonparametric regression function which can be approached by the Fourier Series function, with \( s \) is the number of Fourier Series predictor variables.

\[
h(z_i) = \sum_{d=1}^{s} b_d z_{id} + \frac{1}{2} a_0 + \sum_{m=1}^{M} a_m \cos mz_{id}
\]

Where \( b, a_0 \) and \( a_m \) with \( m=1,2,\ldots,M \) are model parameters. Equation (6) above can be stated in matrix notation as follows:

\[
\tilde{h} = Z \tilde{a}
\]
with \( \tilde{a} = (b_1, a_0, a_1, \ldots, a_m, b_s, a_0, a_s, \ldots, a_M)^T \) and
\[
Z = \begin{pmatrix}
z_{11} \cos \frac{1}{2} \cos \frac{1}{2} \cos m z_{11} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & z_{1s} \frac{1}{2} \cos \frac{1}{2} \cos M z_{1s} \\
z_{21} \cos \frac{1}{2} \cos \frac{1}{2} \cos m z_{21} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & z_{2s} \frac{1}{2} \cos \frac{1}{2} \cos M z_{2s} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
z_{n1} \cos \frac{1}{2} \cos \frac{1}{2} \cos m z_{n1} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & z_{ns} \frac{1}{2} \cos \frac{1}{2} \cos M z_{ns}
\end{pmatrix}
\]

3. Main Results

Given the data \((x_{1ij}, x_{2ij}, \ldots, x_{qj}, t_{1ij}, t_{2ij}, \ldots, t_{qj}, z_{i1j}, z_{i2j}, \ldots, z_{iqj}, y_{ij})\) which follows a multivariable semiparametric regression model for longitudinal data as follows:
\[
y_{ij} = \mu(x_{1ij}, x_{2ij}, \ldots, x_{qj}, t_{1ij}, t_{2ij}, \ldots, t_{qj}, z_{i1j}, z_{i2j}, \ldots, z_{iqj}) + \epsilon_{ij} \\
= \sum_{h=1}^{p} f_{ih}(x_{ah}) + \sum_{l=1}^{q} g_{il}(t_{al}) + \sum_{d=1}^{k} R_{dij}(z_{ijd}) + \epsilon_{ij} \quad \text{with} \quad i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, J
\]

The relationship pattern between response variable and predictor variable can be seen by scatterplot [1]. Based on the scatterplot, the pattern between response variable and each of predictor variables in the data can be identified. The pattern between response variables \(y_{ij}\) and predictor variables \((x_{1ij}, x_{2ij}, \ldots, x_{qj})\) will follow a linear pattern. Then the pattern between response variables \(y_{ij}\) and predictor variables \((t_{1ij}, t_{2ij}, \ldots, t_{qj})\) will change at certain sub-intervals. And the pattern between response variables \(y_{ij}\) with predictor variables \((z_{i1j}, z_{i2j}, \ldots, z_{iqj})\) will have a repetitive pattern.

The model on equation (8) as a whole can be approached by semiparametric regression for longitudinal data using a mixed estimator of Spline Truncated and Fourier Series. Equation (8) contains \(n\) subjects with the \(i\) subject having \(J\) observations, which can be described as follows:
\[
y_{ij} = \beta_{0i} + \sum_{h=1}^{p} \beta_{ih} x_{ih} + \sum_{l=1}^{q} \gamma_{il} t_{il} + \sum_{u=1}^{n} \gamma_{i(u1w)} (t_{ij} - K_{uw}) + \sum_{d=1}^{k} \left( b_{id} z_{ijd} + \frac{1}{2} a_{id} + \sum_{m=1}^{M} a_{mid} \cos m z_{ijd} \right) + \epsilon_{ij}
\]

3.1 Procedure

To complete the research objectives, steps are arranged that follow the following stages:
- Get the model of semiparametric regression using mixed Spline Truncated and Fourier Series for longitudinal data as equation (8) and (9).
- Get the parameter estimates by Weighted Least Square optimization.
- Complete Weighted Least Square optimization using partial derivatives
  \[
  \frac{\partial Q(\beta, \gamma, \tilde{a})}{\beta} = 0, \quad \frac{\partial Q(\beta, \gamma, \tilde{a})}{\gamma} = 0 \quad \text{and} \quad \frac{\partial Q(\beta, \gamma, \tilde{a})}{\tilde{a}} = 0
  \]
- Get the parameter estimation of semiparametric regression using mixed Spline Truncated and Fourier Series for longitudinal data.

In the following analysis, several lemmas are given to obtain a semiparametric regression model using mixed Spline Truncated and Fourier Series in longitudinal data.
Lemma 1.
If a semiparametric regression model with mixed spline truncated and fourier series estimator was given as described by equation (9) for \( i = 1,2,\ldots,n \), then the parametric component \( \hat{f} \) can be written in the form of the following matrix:

\[
\hat{f} = X \beta
\]  

(10)

with \( \beta = [\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_p]^T \) where \( \beta \) is vector with \( n(p+1) \times 1 \) size and \( \tilde{\beta}_i = [\beta_{0i}, \beta_{1i}, \beta_{2i}, \ldots, \beta_{pi}]^T \)

and matrix predictors \( X = \text{diag}(X_1, X_2, \ldots, X_n) \) where \( X \) is matrix with \( nJ \times n(p+1) \) size and

\[
X_i = \begin{bmatrix}
1 & x_{i1} & x_{i2} & \ldots & x_{ip}
1 & x_{i2} & x_{i2} & \ldots & x_{ip}
\vdots & \vdots & \vdots & \ddots & \vdots 
1 & x_{iJ} & x_{i2} & \ldots & x_{ip}
\end{bmatrix}_{J \times (p+1)}
\]

Proof:
In equation (9), the function contains \( n \) subject, with the \( i \) subject having \( J \) observations. It can be described as follows:

For \( i = 1 \) and \( j = 1,2,\ldots,J \)

\[
\begin{align*}
\hat{f}_1 &= \beta_{01} + \beta_{11} x_{i11} + \ldots + \beta_{1p} x_{ip1} \\
\hat{f}_2 &= \beta_{02} + \beta_{11} x_{i12} + \ldots + \beta_{1p} x_{ip2} \\
&\vdots \\
\hat{f}_J &= \beta_{0J} + \beta_{11} x_{i1J} + \ldots + \beta_{1p} x_{ipJ}
\end{align*}
\]

The equation above can be expressed in the form of a matrix as follows

\[
\begin{bmatrix}
\hat{f}_1 \\
\hat{f}_2 \\
\vdots \\
\hat{f}_J
\end{bmatrix} =
\begin{bmatrix}
1 & x_{i11} & \ldots & x_{ip1} \\
1 & x_{i12} & \ldots & x_{ip2} \\
\vdots \\
1 & x_{i1J} & \ldots & x_{ipJ}
\end{bmatrix}
\begin{bmatrix}
\beta_{01} \\
\beta_{11} \\
\vdots \\
\beta_{1p}
\end{bmatrix}

\]

\( \hat{f}_i = X_i \tilde{\beta}_i \)

Do the same thing as described above for \( i = 2,\ldots,n \) so subsequently obtained

\[
\begin{bmatrix}
\hat{f}_1 \\
\hat{f}_2 \\
\vdots \\
\hat{f}_n
\end{bmatrix} =
\begin{bmatrix}
X_1 & 0 & \ldots & 0 \\
0 & X_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & X_n
\end{bmatrix}
\begin{bmatrix}
\tilde{\beta}_1 \\
\tilde{\beta}_2 \\
\vdots \\
\tilde{\beta}_n
\end{bmatrix}

\]

\( \hat{f} = X \tilde{\beta} \)

Lemma 2.
If a semiparametric regression model with mixed spline truncated and fourier series estimator was given as described by equation (9) for \( i = 1,2,\ldots,n \), then the nonparametric component \( \hat{g} \) approached by spline truncated linear with knot points \( K_{ul} : u = 1,2,\ldots,r \) and \( l = 1,2,\ldots,q \) that can be written in the form of the following matrix.

\[
\hat{g} = T \tilde{\gamma}
\]  

(11)
with \( \tilde{\beta} = [\tilde{\beta}_1 \ \tilde{\beta}_2 \ \ldots \ \tilde{\beta}_n]^T \) where \( \tilde{\beta} \) is vector with \( nq(1+r) \times 1 \) size and \( \tilde{\beta}_i = [\gamma_{i11} \ \gamma_{i12} \ \ldots \ \gamma_{i|qr|}]^T \) and matrix \( T = \text{diag}(T_1, T_2, \ldots, T_r) \) where \( T \) is matrix with \( nJ \times nq(1+r) \) size and

\[
T_i = \begin{bmatrix}
t_{i11} & (t_{i11} - K_{11})_+ & (t_{i11} - K_{21})_+ & \cdots & (t_{i|q|} - K_{1q})_+
t_{i12} & (t_{i12} - K_{11})_+ & (t_{i12} - K_{21})_+ & \cdots & (t_{i|q|} - K_{2q})_+
\vdots & \vdots & \vdots & \ddots & \vdots
t_{i1J} & (t_{i1J} - K_{11})_+ & (t_{i1J} - K_{21})_+ & \cdots & (t_{i|q|} - K_{Jq})_+
\end{bmatrix}
\]

Proof:
In equation (9), the function contains \( n \) subject, with the \( i \) subject having \( J \) observations. It can be described as follows:

For \( i = 1 \) and \( j = 1, 2, \ldots, J \)

\[
g_{i1} = \gamma_{i11}t_{i11} + \gamma_{i12}(t_{i12} - K_{11})_+ + \cdots + \gamma_{i|qr|}(t_{i|q|} - K_{1q})_+
g_{i2} = \gamma_{i11}t_{i12} + \gamma_{i12}(t_{i12} - K_{11})_+ + \cdots + \gamma_{i|qr|}(t_{i|q|} - K_{2q})_+
\]

The truncated function is defined as:

\[
(t_{ij} - K_{ij})_+ = \begin{cases}
(t_{ij} - K_{ij})_+, & t_{ij} \geq K_{ij} \\
0, & t_{ij} < K_{ij}
\end{cases}
\]

The equation above can be expressed in the form of a matrix as follows

\[
\begin{bmatrix}
g_{i1} \\
g_{i2} \\
\vdots \\
g_{iJ}
\end{bmatrix} =
\begin{bmatrix}
t_{i11} & (t_{i11} - K_{11})_+ & (t_{i11} - K_{21})_+ & \cdots & (t_{i|q|} - K_{1q})_+
t_{i12} & (t_{i12} - K_{11})_+ & (t_{i12} - K_{21})_+ & \cdots & (t_{i|q|} - K_{2q})_+
\vdots & \vdots & \vdots & \ddots & \vdots
t_{i1J} & (t_{i1J} - K_{11})_+ & (t_{i1J} - K_{21})_+ & \cdots & (t_{i|q|} - K_{Jq})_+
\end{bmatrix}
\begin{bmatrix}
\gamma_{i11} \\
\gamma_{i12} \\
\vdots \\
\gamma_{i|qr|}
\end{bmatrix}
\]

\[
\begin{bmatrix}
\tilde{\beta}_1 \\
\tilde{\beta}_2 \\
\vdots \\
\tilde{\beta}_n
\end{bmatrix} = \begin{bmatrix}
T_1 & 0 & \cdots & 0 \\
0 & T_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & T_n
\end{bmatrix}
\begin{bmatrix}
\tilde{\beta}_1 \\
\tilde{\beta}_2 \\
\vdots \\
\tilde{\beta}_n
\end{bmatrix}
\]

Do the same thing as described above for \( i = 2, \ldots, i = n \) so subsequently obtained

\[
\tilde{\beta} = T\tilde{\beta}
\]

**Lemma 3.**
If a semiparametric regression model with mixed spline truncated and fourier series estimator was given as described by equation (9) for \( i = 1, 2, \ldots, n \), then the nonparametric component \( \tilde{h} \) approached by fourier series function that can be written in the form of the following matrix.

\[
\tilde{h} = Z\tilde{a}
\]

with \( \tilde{a} = [\tilde{a}_1 \ \tilde{a}_2 \ \ldots \ \tilde{a}_n]^T \) where \( \tilde{a} \) is vector with \( ns(M+2) \times 1 \) size and \( \tilde{a}_i = [b_{i1} \ a_{i1} \ \ldots \ \ a_{i|m|}]^T \) and matrix \( Z = \text{diag}(Z_1, Z_2, \ldots, Z_n) \) where \( Z \) is matrix with \( nJ \times ns(M+2) \) size and
\[
Z_i = \begin{bmatrix}
\frac{1}{2} \cos z_{i1} & \cdots & \cos Mz_{i1} & \cdots & \frac{1}{2} \cos z_{i1} & \cdots & \cos Mz_{i1} \\
\frac{1}{2} \cos z_{i2} & \cdots & \cos Mz_{i2} & \cdots & \frac{1}{2} \cos z_{i2} & \cdots & \cos Mz_{i2} \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\frac{1}{2} \cos z_{ij} & \cdots & \cos Mz_{ij} & \cdots & \frac{1}{2} \cos z_{ij} & \cdots & \cos Mz_{ij}
\end{bmatrix}
\]

**Proof:**

In equation (9), the function contains \(n\) subject, with the \(i\) subject having \(J\) observations. It can be described as follows:

For \(i = 1,2,\ldots,J\)

\[
h_i = h_{i1} z_{i1} + \frac{1}{2} a_{i01} + \sum_{m=1}^{M} a_{mi1} \cos mz_{i1} + \frac{1}{2} a_{i01} + \sum_{m=1}^{M} a_{mi1} \cos mz_{i1}
\]

\[
h_i = h_{i2} z_{i2} + \frac{1}{2} a_{i02} + \sum_{m=1}^{M} a_{mi2} \cos mz_{i2} + \frac{1}{2} a_{i02} + \sum_{m=1}^{M} a_{mi2} \cos mz_{i2}
\]

\[
h_i = h_{ij} z_{ij} + \frac{1}{2} a_{i0j} + \sum_{m=1}^{M} a_{mi2} \cos mz_{ij} + \frac{1}{2} a_{i0j} + \sum_{m=1}^{M} a_{mi2} \cos mz_{ij}
\]

The equation above can be expressed in the form of a matrix as follows

\[
\begin{bmatrix}
\tilde{h}_1 \\
\tilde{h}_2 \\
\vdots \\
\tilde{h}_j \\
\vdots \\
\tilde{h}_n
\end{bmatrix} = Z_i \tilde{a}_i
\]

Do the same thing as described above for \(i = 2,\ldots,n\) so subsequently obtained

\[
\begin{bmatrix}
\tilde{h}_1 \\
\tilde{h}_2 \\
\vdots \\
\tilde{h}_j \\
\vdots \\
\tilde{h}_n
\end{bmatrix} = Z_1 \tilde{a}_1
\]

Based on Lemma 1, Lemma 2, and Lemma 3 subsequently obtained

\[
\tilde{y} = X \tilde{\beta} + T \tilde{\gamma} + Z \tilde{a} + \tilde{\epsilon} \quad \text{where} \quad \tilde{\epsilon} = \tilde{y} - X \tilde{\beta} - T \tilde{\gamma} - Z \tilde{a}
\]

where \(\tilde{y} = [\tilde{y}_1 \ \tilde{y}_2 \ \cdots \ \tilde{y}_n]^T\) and the error vector \(\tilde{\epsilon} = [\tilde{\epsilon}_1 \ \tilde{\epsilon}_2 \ \cdots \ \tilde{\epsilon}_n]^T\).

**Theorem 1.**

If a semiparametric regression model with mixed spline truncated and fourier series estimator was given as described by equation (9) for \(i = 1,2,\ldots,n\), then estimation for \(\hat{\beta}, \hat{\gamma}, \hat{a}\) and \(\hat{y}\) is given by

\[
\hat{\beta} = A(K,m) \tilde{y} \quad ; \quad \hat{\gamma} = B(K,m) \tilde{y} \quad ; \quad \hat{a} = C(K,m) \tilde{y} \quad \text{and} \quad \hat{y} = D(K,m) \tilde{y}
\]
where

\[ A(K,m) = \left( (I - PZUX) - (PT - PZUT)(I - RZUT)^{-1}(RX - RZUX) \right)^{-1} \]

\[ B(K,m) = (PT - PZUT)^{-1}\left[ (P - PZU) - (I - PZUX)A \right] \]

\[ C(K,m) = \left[ U - UXA - UT \left( (PT - PZUT)^{-1}\left[ (P - PZU) - (I - PZUX)A \right] \right) \right] \]

\[ D(K,m) = XA(K,m) + TB(K,m) + ZC(K,m) \]

**Proof:**

Estimation of semiparametric regression model using mixed Spline Truncated and Fourier Series for longitudinal data can be obtained using Weighted Least Square (WLS) optimization shown below:

\[ Q(\hat{\beta}, \hat{\gamma}, \hat{a}) = \text{Min}_{\beta, \gamma, a} \{ \hat{\varepsilon}^T \hat{\varepsilon} \} = \text{Min}_{\beta, \gamma, a} \left\{ \hat{y} - X\hat{\beta} - T\hat{\gamma} - Z\hat{a} \right\}^T W \left( \hat{y} - X\hat{\beta} - T\hat{\gamma} - Z\hat{a} \right) \]

Matrix \( W \) is a weighting matrix (variance-covariance matrix). Matrix \( W = \text{diag}(W_1, W_2, \ldots, W_p) \) where \( W \) is sized \( N \times N \) where \( N = nJ \). According [12], there are some methods for determining weighting matrix including:

1. \( W_i = N^{-1}I, \quad i = 1, 2, \ldots, n \), this weight gives the same treatment at each observation.
2. \( W_i = n^{-1}I, \quad i = 1, 2, \ldots, n \), this weight gives the same treatment for each observation in the subject.

Parameter estimation in equation (9) can be obtained by completing the Weighted Least Square

\[ Q(\hat{\beta}, \hat{\gamma}, \hat{a}) = \hat{y}^T W\hat{y} - 2\hat{\beta}^T X^T \hat{W}\hat{y} - 2\hat{\gamma}^T T^T \hat{W}\hat{y} - 2\hat{a}^T Z^T \hat{W}\hat{y} + 2\hat{\beta}^T X^T W\hat{a} + 2\hat{\gamma}^T T^T W\hat{a} + \hat{\beta}^T X^T \hat{W} + \hat{\gamma}^T T^T \hat{W} + \hat{a}^T Z^T \hat{W} \]

Partial derivatives obtained

\[ \frac{\partial Q(\hat{\beta}, \hat{\gamma}, \hat{a})}{\partial \hat{\beta}} = -2X^T W\hat{y} + 2X^T W\hat{a} + 2X^T \hat{W}\hat{\beta} \]

\[ \frac{\partial Q(\hat{\beta}, \hat{\gamma}, \hat{a})}{\partial \hat{\gamma}} = -2T^T W\hat{y} + 2T^T W\hat{a} + 2T^T \hat{W}\hat{\gamma} \]

\[ \frac{\partial Q(\hat{\beta}, \hat{\gamma}, \hat{a})}{\partial \hat{a}} = -2Z^T W\hat{y} + 2Z^T W\hat{\beta} + 2Z^T \hat{W}\hat{\gamma} \]

The results of derivatives obtained in the above equation are given zero, so that:

\[ \hat{\beta} = (X^T \hat{W}X)^{-1} X^T \hat{W}\hat{y} \]

\[ \hat{\gamma} = (T^T \hat{W}T)^{-1} T^T \hat{W}\hat{y} \]

\[ \hat{a} = (Z^T \hat{W}Z)^{-1} Z^T \hat{W}\hat{y} \]

To simplify the calculation, equations above can be written as follows:

\[ \hat{\beta} = P\hat{y} - PT\hat{\gamma} - P\hat{Z}\hat{a} \quad ; \quad \hat{\gamma} = R\hat{y} - RX\hat{\beta} - R\hat{Z}\hat{a} \quad ; \quad \hat{a} = U\hat{y} - UX\hat{\beta} - UT\hat{\gamma} \]

And \( P, R, U \) are as follows:

\[ P = (X^T \hat{W}X)^{-1} X^T \hat{W} \quad ; \quad R = (T^T \hat{W}T)^{-1} T^T \hat{W} \quad ; \quad U = (Z^T \hat{W}Z)^{-1} Z^T \hat{W} \]
Based on the results above, it can be seen that the parameters still contain parameters, so substitution elimination methods are used. From the steps of elimination and substitution that have been done, it is obtained $$\hat{\beta} = A(K,m)\hat{\gamma}$$ with

$$A(K,m) = \left( (I - PZUX) - (PT - PZUT)(I - RZUT)^{-1}(RX - RZUX) \right)^{-1}$$

$$\left( (P - PZU) - (PT - PZUT)(I - RZUT)^{-1}(R - RZU) \right)$$

Just like the explanation on Proof above, elimination and substitution are done to get parameters $$\hat{\beta}, \hat{\gamma}, \hat{a}$$, and $$\alpha$$, then it is obtained $$\hat{\gamma} = B(K,m)\hat{\gamma}$$ and $$\hat{a} = C(K,m)\hat{\gamma}$$ with

$$B(K,m) = (PT - PZU)^{-1}\left( (P - PZU) - (I - PZUX)A \right)$$ and

$$C(K,m) = U - UXA - UT \left( (PT - PZU)^{-1}\left( (P - PZU) - (I - PZUX)A \right) \right)$$

Parameters $$\hat{\beta}, \hat{\gamma}, \hat{a}$$ are obtained, so the estimation of semiparametric regression model for longitudinal data using mixed estimator of Spline Truncated and Fourier Series as follows:

$$\hat{\gamma} = X\hat{\beta}T + T\hat{\gamma} + Z\hat{a}$$

$$= XA(K,m)\hat{\gamma} + TB(K,m)\hat{\gamma} + ZC(K,m)\hat{\gamma}$$

$$= D(K,m)\hat{\gamma}$$

with $$D(K,m) = XA(K,m) + TB(K,m) + ZC(K,m)$$

Estimator of semiparametric regression model for longitudinal data using mixed Spline Truncated and Fourier Series contains oscillation parameters and knot points. Generalized Cross Validation (GCV) method given by:

$$GCV(K,m) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{J} (y_{ij} - \hat{y}_{ij})^2}{(nJ^{-1}tr(1 - D(K,m)))^2}$$

is used to get the optimal oscillation parameters and knot points.

4. **Conclusion**

Based on result, it can be concluded that the estimation of regression model using mixed Spline Truncated and Fourier Series estimator for longitudinal data is

$$\hat{y} = X\hat{\beta}T + T\hat{\gamma} + Z\hat{a} = D(K,m)\hat{\gamma}$$

where $$D(K,m) = XA(K,m) + TB(K,m) + ZC(K,m)$$

Estimation of the unknown parameters in semiparametric regression model $$\hat{\beta}, \hat{\gamma}, \hat{a}$$ are obtained by completing Weighted Least Square (WLS) optimization. Mixed Spline Truncated and Fourier Series estimator for longitudinal data is dependent on the number of knots, knots location, and optimal oscillation parameters that can be selected optimally using minimum Generalized Cross Validation (GCV) value.

**Acknowledgments**

The authors are grateful to Ministry of Research, Technology and Higher Education Indonesia and Department of Statistics, Faculty of Mathematics, Computing and Data Science, Institut Teknologi Sepuluh Nopember which allow this research to be published.
References

[1] Hidayat R, Budiantara I N, Otok B W and Ratnasari V 2019 Kernel-spline estimation of additive nonparametric regression model IOP Conf. Ser.: Mater. Sci. Eng. 546 052028

[2] Mariati N P A M, Budiantara I N and Ratnasari V 2019 Modeling poverty percentages in the Papua islands using fourier series in nonparametric regression multivariable J. Phys.: Conf. Ser. 1397 012071

[3] Ratnasari V, Budiantara I N, Zain I, Ratna M and Mariati N P A M 2015 Comparison truncated spline and fourier series in multivariable nonparametric regression models (applications : data of poverty in Papua, Indonesia) International Journal of Basic and Applied Sciences 15 pp 9-12

[4] Prawanti D D, Budiantara I N and Purnomo J D T 2019 Parameter interval estimation of semiparametric spline truncated regression model for longitudinal data IOP Conf. Ser.: Mater. Sci. Eng. 546 052053

[5] Ratnasari V, Budiantara I N, Ratna M and Zain I 2016 Estimation of nonparametric regression curve using mixed estimator of multivariable truncated spline and multivariable kernel Global Journal of Pure and Applied Mathematics 12 pp 5047-5057.

[6] Afifah N, Budiantara I N and Latra I N 2017 Estimator of kernel and fourier series in semiparametric regression J. Phys.: Conf. Ser. 855 012002

[7] Nisa K, Budiantara I N and Rumiati A T 2017 Multivariable semiparametric regression model with combined estimator of fourier series and kernel. IOP Conf. Ser.: Earth Environ. Sci. 58 012028

[8] Sudiarsa I W, Budiantara I N, Suhartono and Purnami S W 2015 Combined estimator fourier series and spline truncated in multivariable nonparametric regression Applied Mathematical Sciences 9 pp 4997-5010.

[9] Wahba G 1990 Smoothing spline ANOVA with component-wise bayesian confidence intervals Journal of Computational and Graphical Statistics 2 pp 9-117.

[10] Bilodeau M 1992 Fourier smoother and additive models The Canadian Journal of Statistics 20 pp 257-269

[11] Asrini L J and Budiantara I N 2014 Fourier series semiparametric regression models (case study: the production of lowland rice irrigation in central java) ARPN Journal of Engineering and Applied Sciences 9 (9) 1501-1506.