TRUNCATED SPLINE REGRESSION IN LINEAR PARTIAL MODEL FOR LONGITUDINAL DATA

Idhia Sriliana\textsuperscript{1)} I Nyoman Budiantara\textsuperscript{2)}

\textsuperscript{1)} Student of Statistics Department, Faculty of Mathematics and Natural Sciences, Institut of Technology Sepuluh Nopember, Surabaya, Indonesia (atha_muflih@yahoo.co.id)
\textsuperscript{2)} Lecture of Statistics Department, Faculty of Mathematics and Natural Sciences, Institut of Technology Sepuluh Nopember, Surabaya, Indonesia (i_nyoman.b@statistika.its.ac.id)

ABSTRACT

Theoretically and practically, Spline regression estimator in semiparametric or partially linear model is not only suitable for handling cross section data, but it can be generalized for longitudinal data. Linear partial models for longitudinal data with \( y_i = Xβ + f(t_i) + ε_i, j = 1, 2, \ldots, n, i = 1, 2, \ldots, n \), there are \( n \) subjects which have the \( i \)th observation. By using the truncated spline regression approach and the weighted least squares optimization, the estimation in partially linear model for longitudinal data was

\[
\hat{Y} = A(k_1, k_2, \ldots, k_s)Y, \text{ where } A(k_1, k_2, \ldots, k_s) = X(k_1)(x(k_2)wX(k_1))^T X(k_1)w. \text{ The application was used to predict the pattern of the relationship among the number of leukocytes in leukemia patients (y) toward the number of platelets (x) and checkup time (t). The result of model estimation showed the number of leukocytes leukemia (y) had linear pattern with the number of platelets (x) and it also had pattern which followed a linear spline two knots toward checkup time (t). This model had a GCV value of 0.4118387 with MSE value of 0.0001967 and the coefficient of determination \((R^2)\) of 0.9879.}

Keywords: Leukemia, Linear Partial Model, Longitudinal Data, Truncated Spline

I. INTRODUCTION

Regression analysis is one of the data analysis that is used to determine the model patterns of relationships between predictor variables and response variables. If the pattern of relationships between predictor variables and response variables are unknown, then the nonparametric regression is a regression approach which appropriate for the data pattern of an unknown form of the regression curve [1]. Spline is one of the nonparametric regressions that have specific statistical interpretation and an excellent visual interpretation [2]. Moreover, spline is able to handle the smooth character data and it has an excellent ability to handle data that changes its behavior in specific sub-interval [3]. Spline and partial spline estimator theoretically and practically is not only suitable for handling cross section data, but can be generalized to longitudinal data [4]. In a study of longitudinal data, in general, the observations are made of \( n \) mutually independent subjects where each subject is observed repeatedly in different time periods [5].

One of the methods that can be used for estimating parameters in linear partial models for longitudinal data is truncated spline regression approach. Truncated spline regression approach has several advantages including easier mathematically and the interpretations same as the parametric regression. Basically, the main issue in the estimation of parameters using truncated spline regression is about the selection of optimal knot points with ordinary least squares optimization. However, estimation of parameters in linear partial models for longitudinal data is conducted using weighted least squares optimization, which includes \( W \) value for the completion of the smallest squares optimization. This is caused by the observations of longitudinal data are dependent on the same subjects while for different subjects are independent, so that the correlation between observations cause ordinary least squares optimization need to be refined by a value.

There are several studies which analyze longitudinal data modeling using nonparametric regression approach. They are Wu and Chiang [6] who using kernel estimators, Zhang [7] using generalized spline regression, Rice and Wu [8] using the spline approach which based on a mixed effects model to estimate the regression curve for longitudinal data. On the other hand, there are several studies on longitudinal data based on the semiparametric model in particular linear partial model, including the Fan and Zhang [9], Guo [10], Liang and Xiao [11], and Laome [12].

In this study, it used spline regression approach with weighted least square optimization to estimate parameters in partially linear model for longitudinal data. Then, it was conducted the selection of optimal knots point to find the best spline model estimator. In the last step, the proposed method was applied to
predict the relationships pattern of the number of leukocytes in leukemia patients based on the number of platelets and checkup time. The results showed truncated spline estimators in linear partial models for longitudinal data have the characteristics of linearity. The selection of optimal knot points was conducted by GCV method. Application of the model showed the pattern of the relationship among the number of leukocytes in leukemia patients, the number of platelets, and checkup time may be formed in a linear partial model, with the number of platelets as parametric component and checkup time as nonparametric component. Based on the result of model estimation, it showed that the number of leukocytes in leukemia patients had linear pattern on the number of platelets and it also had pattern which followed a linear partial two knots in checkup time.

II. METHODS

The method used in this study was truncated spline regression with weighted least square optimization. Spline is polynomial pieces that have continuous segmented characteristics thus it can explain the local characteristics of the data function. In general, the spline function of degree $m$ with $K$ knots is any function that written as:

$$s(t) = \sum_{j=1}^{r} \alpha_j t^j + \sum_{j=1}^{m-r} \rho_j (t-k)^j$$

with, $$(t-k)^j = \begin{cases} (t-k)^j, & t \geq k \\ 0, & t < k \end{cases}$$

where $\alpha$ is polynomial parameter, $\rho$ is truncated parameter, and $k_1, k_2, \ldots, k_K$ are knot points.

In a study of longitudinal data, the observations were conducted on $n$ subjects which were mutually independent, where each subject was observed continuously in a specified period. If it is given the longitudinal data $(t_{ij}, x_{ij}, y_{ij}), j = 1, 2, \ldots, n$, then the linear partial model for longitudinal data is obtained from [14]:

$$y_i = X' \beta + f(t_i) + \varepsilon_i, \quad j = 1, 2, \ldots, n, \quad i = 1, 2, \ldots, n$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_r)'$ is $1 \times r$-dimensional vector of parametric regression coefficient, $X$, $X' = (x_{i1}, x_{i2}, \ldots, x_{in})$, $X_{ij} = (x_{ij}, x_{ij+1}, \ldots, x_{ij+m})$, $f(t_i)$ is an differentiable function and $\varepsilon_i$ is random error, $j = 1, 2, \ldots, n, \quad i = 1, 2, \ldots, n$.

The optimal curve of spline regression estimator depends on the location of the knots points $k_1, k_2, \ldots, k_K$. Therefore, it needs to choose optimal knots point to determine the best spline model. Knot points are the interface point to show the change of behavior of spline functions on different intervals. The method used in the selection of optimal knot points is generalized cross validation (GCV), which is defined as follows [13]:

$$GCV(k) = \frac{N'Y' (I - A(k))' W (I - A(k)) Y}{(N \text{ trace} (I - A(k)))}$$

Estimation of linear partial models for longitudinal data with truncated spline regression approach was performed with these following steps:

a. Approaching $f(t_i)$ with truncated spline degree $m$ and knots $k_1, k_2, \ldots, k_K$.
b. Indicating the linear partial model for longitudinal data as follows:

$$y_i = X' \beta + \sum_{j=1}^{r} \alpha_j t^j + \sum_{j=1}^{m-r} \varepsilon_j (t_i - k_j) + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

c. Writing model in step (b) as follows:

$$Y = X(k)B + \varepsilon, \quad k = (k_1, k_2, \ldots, k_K)$$

d. Determining $B$ by using weighted least squares optimization:

$$\min_{b \in \mathbb{R}^{p \times 1}} \{ (Y - X(k)B)' W (Y - X(k)B) \}$$

e. Obtaining the spline estimation in linear partial model as follows:

$$\hat{y}_i = X' \hat{\beta} + \sum_{j=1}^{r} \hat{\alpha}_j t^j + \sum_{j=1}^{m-r} \hat{\varepsilon}_j (t_i - k_j)$$

Furthermore, the results of estimates were applied to observe the pattern of relationship among the number of leukemia leukocyte toward the number of platelets and the checkup time with these following steps:

a. Creating data plot $(t_{ij}, x_{ij}, y_{ij})$.
b. Modeling the data with spline approach, 
c. Choosing the optimal knot points by the GCV method, 
d. Finding the smallest value of GCV, 
e. Estimating model patterns of relationships with leukemia leukocyte count, 
f. Calculating the value of determination coefficient and mean square error.

III. RESULTS

A. Estimation Spline Truncated with Weighted Least Square Optimization in Linear Partial Models for Longitudinal Data

If the curve regression in a linear partial model for longitudinal data was approximated by truncated spline regression,

$$s(t_i) = \sum_{j=1}^{r} \alpha_j t^j + \sum_{j=1}^{m-r} \varepsilon_j (t_i - k_j)^j$$

where $i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, n$. 


Hence, the linear partial models for longitudinal data can be expressed in matrix notation as follows:

$$Y = X(k)B + \varepsilon$$

(5)

where $k_i = (k_{1i}, k_{2i}, \ldots, k_{ni})$, $i = 1, 2, \ldots, n$.

By using weighted $W$, estimates of $B$ in (5) can be obtained by solving the optimization:

$$\min_{B} \left\{ (Y - X(k)B)^T W (Y - X(k)B) \right\}$$

(6)

with $W$ is a matrix:

$$W = \begin{bmatrix} W_1 & 0 & \ldots & 0 \\ 0 & W_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & W_n \end{bmatrix}$$

so that, we obtained:

$$\hat{B} = (X(k)^T WX(k))^{-1} X(k)^T W Y$$

(7)

Based on the estimation of $B$ in (7), then obtained:

$$\hat{Y} = X(k)\hat{B} = A(k_{1i}, k_{2i}, \ldots, k_{ni}) Y, \ i = 1, 2, \ldots, n$$

where,

$$A(k_{1i}, k_{2i}, \ldots, k_{ni}) = X(k) (X(k) WX(k))^T X(k)^T W X(k)$$

B. Applications to Data Number of Leukocytes in Leukemia Patients

The data used in this study was secondary data derived from studies conducted Oktiriani [14], that was the data from leukemia patients treated at Surabaya Hajj General Hospital in 2009. It was about the development of leukocytes numbers in leukemia patients who were influenced to the checkup time and the number of platelets. Based on data from medical records of Surabaya Hajj General Hospital in 2009, there were four patients who had mutually independent leukemia and assumed only having leukemia. In each of patients was measured the number of leukocyte cell and platelet cell in a certain period to be analyzed by using a truncated spline model. The study was focused to predict the pattern of the relationship among the number of leukocytes in leukemia patients ($y$) toward the number of platelets ($x$) and time ($t$), where the number of platelets medically as explanatory variables that was assumed as parametric component in partially linear model.

In the analysis process, it would be conducted linear partial model estimation simultaneously to obtained the optimal linear partial model. In addition, the criteria for the best model with a truncated spline approach were measured by the smallest value of GCV at the point of optimum knots [15]. However, other criteria of the goodness of the model were considered through the MSE and the coefficient of determination $R^2$. Thus, the smallest GCV value was obtained through a combination of the number of knots and the degree of the polynomial used. Repetition which were conducted for each patient caused the correlation between observations in the same subjects. The indication of a correlation could be seen in the response variable that affected to the time variable.

So the estimation of the optimal spline model was obtained from a weighting. In this case, the analysis was conducted by using weighting of variance covariance matrix.

Furthermore, the combinations of GCV with the number of knots and the degree of the polynomial spline for estimating the linear partial model are shown in Table 1.

<table>
<thead>
<tr>
<th>Knots</th>
<th>The Degree of Polynomial Spline</th>
<th>GCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.6445039</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3.1611685</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>4.0323595</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.4118387</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3.0689753</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4.1987952</td>
</tr>
</tbody>
</table>

Based on Table 1, it was obtained the smallest GCV values on a combination of two knots and one degree of the polynomial (linear spline) of 0.4118387. Knot points for each patient are shown in Table 2 below:

<table>
<thead>
<tr>
<th>Subject</th>
<th>Knots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>138</td>
</tr>
<tr>
<td>Patient 2</td>
<td>106</td>
</tr>
<tr>
<td>Patient 3</td>
<td>102</td>
</tr>
<tr>
<td>Patient 4</td>
<td>120</td>
</tr>
</tbody>
</table>

Thus, the estimation of linear partial model about the pattern of the relationship among the number of leukocytes in leukemia patients toward the number of platelets and checkup time is given:

$$\hat{y}_j = 0.24789 + 0.04287x_j + 0.00065t_j + 0.00237(t_j - 138) + 0.00945(t_j - 154), \ j = 1, 2, \ldots, 7$$

$$\hat{y}_j = 0.02549 + 0.00262x_j - 0.0004t_j + 0.00097(t_j - 106) - 0.00079(t_j - 162), \ j = 1, 2, \ldots, 8$$

$$\hat{y}_j = 0.01704 - 0.11162x_j + 0.00010t_j - 0.00008(t_j - 102) - 0.0001(t_j - 153), \ j = 1, 2, \ldots, 9$$

$$\hat{y}_j = 0.01503 + 0.00038x_j + 0.00004t_j + 0.00031(t_j - 120) - 0.00137(t_j - 174), \ j = 1, 2, \ldots, 5$$
This model has a MSE value of 0.0001967 and coefficient of determination $R^2$ of 0.9879.

Based on (9), (10), (11), and (12) show that the estimation of the number of leukocytes for each leukemia patient $(y)$ is as follows:

(i) It has linear pattern with the number of platelets $(x)$, and

(ii) It patterns which follows a linear spline with two knots with checkup time $(t)$.

Based on the model, it appears that in patient 2 and patient 4 have a tendency that the number of leukocytes keeps increasing with the length of time of checkup. It means that the longer the time the patient undergoes the checkup; therefore the patient's condition keep declining instead, although in the beginning of each patient has different patterns. In contrast, patient 1 and patient 3 have a tendency that the number of leukocytes keeps decreasing along the length of time of checkup, which means that the longer the patient undergoes the checkup of the patient's condition improved. This difference may be influenced by the circumstances and background of different patients, as well as the influence of other factors that are not included in the model. While the pattern of the relationship between the number of platelets and the numbers of leukocytes, leukocyte numbers have a tendency to increase as the numbers of platelets also increase.

VII. CONCLUSION

If given a linear partial model for longitudinal data $y = X\beta + f(t) + \varepsilon$, $j = 1, 2, ..., n$, $i = 1, 2, ..., n$ then by using a truncated spline regression approach with weighted least squares (WLS) optimization, so it is obtained $\hat{Y} = A(k_1, k_2, ..., k_L)\hat{y}$, where

$A(k_1, k_2, ..., k_L) = X(k) (X(k)^TW(k))^{-1}X(k)^TW$.

The results of model application show a pattern of relationship among the number of leukocytes in leukemia patients, the number of platelets, and checkup time in the same time. They may be formed in a linear partial model, with the number of platelets as parametric component and checkup time as nonparametric component. Model estimation results show that the number of leukocytes leukemia patients $(y)$ have linear pattern toward the number of platelets $(x)$ and pattern which follow a linear spline two knots for checkup time $(t)$. This model has a GCV value of 0.4118387 with MSE value of 0.0001967 and the coefficient of determination ($R^2$) of 0.9879.

ACKNOWLEDGEMENT

This paper mostly based on guidance from ICCBS II.

REFERENCES


