Parameter Interval Estimation of Semiparametric Spline Truncated Regression Model for Longitudinal Data

IOP Conf. Series: Materials Science and Engineering **546** (2019) 052053 doi:10.1088/1757-899X/546/5/052053

Parameter Interval Estimation of Semiparametric Spline Truncated Regression Model for Longitudinal Data

Dasty Dewi Prawanti[1*], I Nyoman Budiantara[1], Jerry D.T. Purnomo[1]

[1]Institut Teknologi Sepuluh Nopember (ITS), Surabaya 60111, Indonesia

*Corresponding author : nyomanbudiantara@gmail.com

Abstract. Regression analysis is one method in statistics that is used to determine the pattern of functional relationships between response variables with predictor variables. Semiparametric regression approach is a combination of parametric regression and nonparametric regression. The most popular estimator for nonparametric regression or semiparametric regression is spline truncated estimator. Spline is the estimation method that is most often used because it has excellent statistical interpretation and visual interpretation compared to other methods. Regression modelling using longitudinal data is often found in everyday life, where observations are carried out for each subject over a period of time. Interval estimation is often examined by nonparametric regression and semiparametric regression; this estimation aims to determine predictor variables that have a significant influence on the response variable. One indicator used in poverty analysis is the poverty line. Based on Indonesia's macro poverty analysis calculations, in the period March 2016 to March 2017, the poverty line increased by 5.67 percent, with increases in urban and rural areas at 5.79 percent and 5.19 percent respectively. Modelling using semiparametric spline truncated regression for longitudinal data on data on the percentage of poor people in Indonesia produces the best model using $\mathbf{W}_1$ weighting and one point knot. Based on the results of the study with a significance level of 0.05, it was found that the percentage of poor people was influenced by the human development index (HDI) and the unemployment rate. This semiparametric regression model has a minimum GCV value of 1.677, MSE of 5.477×10^{-2} and R^2 value of 98.67%.

1. Introduction

Regression analysis is one method in statistics that is used to determine the pattern of the relationship between the response variable and the predictor variable described in a function called the regression curve [1]. There are three approaches to estimating the regression curve, namely the parametric regression approach, nonparametric regression and semiparametric regression. The parametric approach is carried out if the functional form between the response variable and predictor variable is assumed to have a certain form such as linear, quadratic, exponential and so on, while the nonparametric approach does not assume a particular form of regression function so that the regression function is estimated using smoothing techniques. Therefore, it is expected that the data searches for its own estimation form without being influenced by the researchers' subjectivity factor [2]. Meanwhile, if the regression curve consists of parametric and nonparametric components, the appropriate regression approach used is semiparametric regression. The method often used to estimate the regression curves in nonparametric regression and semiparametric regression, namely spline, because this method has excellent statistical interpretation and visual interpretation compared to other methods. Special bases that are often used in research use spline estimations, namely truncated bases. Spline truncated is one type of polynomial piecewise which is a polynomial that is segmented or fragmented. The segmented polynomial model

causes spline truncated to have more flexibility than ordinary polynomial models, so it is more effective to explain the local characteristics of the data function [3].

Longitudinal data is a combination of cross section data and time series data where in longitudinal data between subjects are mutually independent of each other, but between observations in the subject are interdependent so there is a correlation between observations [4]. The advantage of longitudinal data is that it can know changes in individuals, does not require many subjects, and also estimates more efficiently because it is carried out every observation. The spline estimator approach for longitudinal data can accommodate correlations between observations in the same subject, which are not found in cross section data, so the problem of assuming autocorrelation can be solved [5].

Inference statistics are methods in statistics that are considered important, where statistical inference is divided into two, namely parameter estimation and hypothesis testing. In parameter estimation research, interval estimation is often discussed with nonparametric regression and semiparametric regression approaches. This estimation aims to find out predictor variables that have a significant effect on the response variable. There are several studies on interval estimation such as those carried out by [6], [7], [8] and [9], but studies examining the estimated interval parameters of a semiparametric regression model with a spline truncated approach for longitudinal data have not been done, therefore researchers are interested in studying this. The application of interval parameter estimates of the semiparametric spline truncated regression model for longitudinal data will be applied to the percentage data of the poor in Indonesia in 2011-2017. In this study, it is expected that the pattern of the relationship between the percentage of the population below the poverty line and the factors that are thought to significantly influence it is known.

2. Theoretical Review

In this section we will review some of the theories used.

2.1 Semiparametric Regression

This semiparametric regression model is more flexible than the linear model because of the presence of parametric and nonparametric components, this will accommodate the relationship between the response variable and the predictors that are linear and nonlinear in nature [10]. Suppose that data is given as follows (x_i, z_i, y), the relationship between (x_i, z_i) is assumed to follow the following semiparametric regression model:

$$y_i = f(x_i) + g(z_i) + \varepsilon_i \quad (1)$$

with y_i is the response variable in the i^{th} observation, $f(x_i)$ is a parametric component, while $g(z_i)$ is a nonparametric component and ε_i is a random error with a normal distribution with zero mean and variant σ^2. Explicitly, the semiparametric regression model in equation (1) can be expressed in the form of a matrix:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} g(z_1) \\ g(z_2) \\ \vdots \\ g(z_n) \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (2)$$

with the equation as follows,

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{g}(\mathbf{z_i}) + \boldsymbol{\varepsilon} \quad (3)$$

2.2 Semiparametric Spline Truncated Regression for Longitudinal Data

Some popular semiparametric models are spline, kernel, local polynomial, Fourier series, wavelets, MARS (Multivariate Adaptive Regression Spline) etc. [11] with the model that is most often used because of its spline flexibility. Suppose given paired data $(x_i, z_i), i = 1,2,...,n$. The relationship between the response variable and the predictor variable x_i, z_i follows the regression model so that the equation is obtained:

$$y_i = h(x_i, z_i) + \varepsilon_i \tag{4}$$

where, h is a regression curve and ε_i is a random error with normal distribution with zero mean and variance σ^2. The regression model is assumed to be an additive, so the regression curve h can be decomposed into:

$$h(x_i, z_i) = f(x_i) + g(z_i) \tag{5}$$

if the regression curve f is assumed to be a parametric component or becomes a linear function such as the parametric regression equation, while the function *g* is assumed to be a nonparametric component which is approached with a spline truncated function with degrees m and points of knots $K_1, K_2, ..., K_r$, so the *g* function can be called a regression model nonparametric spline truncated.

In semiparametric regression with longitudinal data, it is generally performed on n mutually independent objects, where each object is observed repeatedly (repeated measurement). In general, the model of semiparametric spline truncated regression with degrees 1 (*m* = 1) for longitudinal data can be written with the following equation:

$$y_{il} = \beta_i x_{il} + \gamma_i z_{il} + \sum_{u=1}^{r} \gamma_{(1+u)i} \left(z_{il} - K_{ui}\right)_+^1 + \varepsilon_{il} \tag{6}$$

where $i = 1,2,...,n$ is the subject of observation as many as *n* and $l = 1,2,...,t$ is a repetition of observations made until the *t*-period. In equation (6), $\gamma_{ik} z_{il}$ is a polynomial component with the truncated function as follows:

$$\left(z_{il} - K_{ui}\right)_+^1 = \begin{cases} \left(z_{il} - K_{ui}\right)^1 & , z_{il} \geq K_{ui} \\ 0 & , z_{il} < K_{ui} \end{cases}$$

equation (6) can be written in the form of a matrix as follows,

$$\mathbf{y} = \mathbf{Db} + \boldsymbol{\varepsilon} \tag{7}$$

based on equation (7), the estimator **b** can be obtained with Weighted Least Square (WLS) optimization as follows:

$$\min_{b \in R^{n[p+q(1+r)]}} \left\{ (\mathbf{y} - \mathbf{Db})^{\mathbf{T}} \mathbf{W} (\mathbf{y} - \mathbf{Db}) \right\} \tag{8}$$

response **y** is a vector of size $nt \times 1$, $\mathbf{D} = [\mathbf{X} : \mathbf{Z}]$ is a matrix of size $nt \times n\left[p + q(1+r)\right]$ which is a matrix that contains predictors of parametric components and nonparametric components. Vector parameter **b** has size $n\left[p + q(1+r)\right] \times 1$ consisting of parameter vectors $\boldsymbol{\beta}$ and parameters $\boldsymbol{\gamma}$, while $\boldsymbol{\varepsilon}$ is an error vector with size $nt \times 1$. Weighting matrix (**W**) which is a covariance-sized variant matrix $nt \times nt$.

2.3 Weighted Least Square (WLS)

In the application using longitudinal data in the regression model both parametric regression and nonparametric regression, there are two fundamental assumptions. The first assumption is that the variance of the random error in the model is assumed to be homogeneous for each repeated measurement in the subject and the second assumption is that the random variance-covariance error matrix in the

model is assumed to be known [12]. Based on this, the most appropriate estimation method used is the Weighted Least Square (WLS) method. The WLS method estimates parameters by minimizing the number of squares between observations and a model called the sum of squared errors. In general, in the WLS method the function that is minimized to estimate parameters is formulated as follows:

$$Q = \left(\mathbf{y} - \mathbf{D\beta}\right)^{\mathbf{T}} \mathbf{W}\left(\mathbf{y} - \mathbf{D\beta}\right) \tag{9}$$

by minimizing the number of squared errors in equation (9), the results will be obtained:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{D}^{\mathbf{T}}\mathbf{W}\,\mathbf{D}\right)^{-1}\mathbf{D}^{\mathbf{T}}\mathbf{W}\,\mathbf{y} \tag{10}$$

The result of equation (10) is used to estimate the semiparametric spline truncated regression parameter with the **W** matrix which is a diagonal matrix containing weighting for parameter estimation.

$$\mathbf{W} = \begin{bmatrix} \mathbf{W_1} & 0 & \cdots & 0 \\ 0 & \mathbf{W}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{W}_n \end{bmatrix} \text{ where } \mathbf{W}_1 = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

according to [13], there are several methods in determining weighting, including:

1. $\mathbf{W_i} = N^{-1}\mathbf{I},\ i = 1,2,...,n$, this weight gives the same treatment at each observation.
2. $\mathbf{W_i} = n^{-1}\mathbf{I},\ i = 1,2,...,n$, this weight gives the same treatment for each observation in the subject.
3. $\mathbf{W_i} = \mathbf{V_i^{-1}}$, where $\mathbf{V_i} = \mathrm{cov}(yi),\ i = 1,2,...,n$,this weight takes into account the correlation in the subject of observation.

2.4 Optimal Knot Point Selection

The spline estimator is very dependent on the choice of smoothing parameters to determine the optimal knot point. Knot point is the point when the pattern of data changes at different intervals [14]. A good method for selecting the optimal smoothing parameter in the spline estimator is generalized cross validation (GCV) [7]. The most optimum value of knots is the value of knots with the minimum GCV value. The GCV method is generally defined as follows:

$$GCV\left(\mathbf{k}\right) = \frac{MSE\left(\mathbf{k}\right)}{\left[n^{-1}trace\left(\mathbf{I} - \mathbf{A}\left(\mathbf{k}\right)\right)\right]^2} \tag{11}$$

with,

$$MSE\left(\mathbf{k}\right) = n^{-1}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \text{ dan } \mathbf{A}\left(\mathbf{k}\right) = \mathbf{D}\left(\mathbf{k}\right)\left(\mathbf{D}^{\mathbf{T}}\left(\mathbf{k}\right)\mathbf{D}\left(\mathbf{k}\right)\right)^{-1}\mathbf{D}^{\mathbf{T}}\left(\mathbf{k}\right)\mathbf{W}$$

2.5 Confidence Interval for Regression Parameters

The basis of the confidence interval approach is the concept of interval estimation. Interval estimation is an interval or distance that has a probability that has been formed, the probability includes the limit of the value of the actual unknown parameter [1]. Let $X_1, X_2, ..., X_n$ be a random sample that is mutually independent with a probability density function $f\left(x_i;\theta\right), \theta \in R$. Let $L\left(X_1, X_2, ..., X_n\right)$ and $U\left(X_1, X_2, ..., X_n\right)$ be two statistics, $L\left(X_1, X_2, ..., X_n\right) \le U\left(X_1, X_2, ..., X_n\right)$. The random interval $\left[L\left(X_1, X_2, ..., X_n\right), U\left(X_1, X_2, ..., X_n\right)\right]$ is the confidence interval for parameter θ with a confidence interval of $1-\alpha$ $\left(0 < \alpha < 1\right)$ which can be written in the form of equations as follows:

$$P\left(L\left(X_1, X_2, ..., X_n\right) \le \theta \le U\left(X_1, X_2, ..., X_n\right)\right) = 1 - \alpha \tag{12}$$

$L(X_1, X_2, ..., X_n)$ is the lower limit of the confidence interval, $U(X_1, X_2, ..., X_n)$ is the upper limit of the confidence interval and $1-\alpha$ is confidence level

3. Result and Discussion

This study uses secondary data obtained from Badan Pusat Statistik (BPS) publication with an observation unit of 33 provinces in Indonesia. The publication used in this study is the publication of Data and Information on Poverty from 2011-2017 [15]. The response variables used in this study are data on the percentage of poor people in each province in Indonesia for 7 years from 2011 to 2017 with predictor variables for the percentage of poor people in Indonesia. The predictor variables are described in the following table:

Table 1. Research Variable

Variable	Description
y_{il}	Percentage of Poor People in the i^{st} Province in l year
x_{il}	Human Development Index (HDI), Population in the 1^{st} Province in l year
z_{il}	The unemployment rate of the population in the i^{st} Province in l year

3.1 Interval Estimation of Parameters Regression Semiparametric Spline Truncated for Longitudinal Data

The semiparametric spline truncated multivariable regression model for longitudinal data can be determined in the form of equations as follows:

$$y_{il} = \beta_i x_{il} + \gamma_i z_{il} + \sum_{u=1}^{r} \gamma_{(1+u)i} \left(z_{il} - K_{ui}\right)_+^1 + \varepsilon_{il} \tag{13}$$

with,

$$\left(z_{il} - K_{ui}\right)_+^1 = \begin{cases} \left(z_{il} - K_{ui}\right)^1 & , z_{il} \geq K_{ui} \\ 0 & , z_{il} < K_{ui} \end{cases}$$

where $i = 1,2,...,n$ is the subject observed as much as n with repeated observations until the t-period, so equation (14) can be broken down into:

$$y_{il} = \beta_i x_{il} + \left\{\gamma_i z_{il} + \gamma_{2i}\left(z_{il} - K_{1i}\right)_+^1 + ... + \gamma_{(1+r)i}\left(z_{il} - K_{ri}\right)_+^1\right\} + \varepsilon_{il}$$

each observation $i = 1,2,...,n$ and $l = 1,2,...,t$ then the equation is obtained as follows:

$$y_{11} = \beta_1 x_{11} + \left\{\gamma_1 z_{11} + \gamma_{21}\left(z_{11} - K_{11}\right)_+^1 + ... + \gamma_{(1+r)1}\left(z_{11} - K_{r1}\right)_+^1\right\} + \varepsilon_{11}$$
$$\vdots \qquad\qquad \vdots$$
$$y_{1t} = \beta_1 x_{1t} + \left\{\gamma_1 z_{1t} + \gamma_{21}\left(z_{1t} - K_{11}\right)_+^1 + ... + \gamma_{(1+r)1}\left(z_{1t} - K_{r1}\right)_+^1\right\} + \varepsilon_{1t}$$
$$\vdots \qquad\qquad \vdots$$
$$y_{n1} = \beta_n x_{n1} + \left\{\gamma_n z_{n1} + \gamma_{2n}\left(z_{n1} - K_{1n}\right)_+^1 + ... + \gamma_{(1+r)n}\left(z_{n1} - K_{rn}\right)_+^1\right\} + \varepsilon_{n1}$$
$$\vdots \qquad\qquad \vdots$$
$$y_{nt} = \beta_{n1} x_{nt} + \left\{\gamma_n z_{nt} + \gamma_{2n}\left(z_{nt} - K_{1n}\right)_+^1 + ... + \gamma_{(1+r)n}\left(z_{nt} - K_{rn}\right)_+^1\right\} + \varepsilon_{nt}.$$

So the regression model semiparametric spline truncated for longitudinal data can be expressed in matrix notation:

$$\mathbf{y} = \mathbf{Db} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2\mathbf{W}\right) \tag{14}$$

with,

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix},\ \mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{D}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{D}_n \end{bmatrix},\ \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n \end{bmatrix},\ \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix}.$$

The response $\mathbf{y}$ is a vector measuring $\mathbf{W}_1, \mathbf{W}_2, \text{dan}\, \mathbf{W}_3$, matrix $\mathbf{D}$ with a size of $nt \times n(\text{p+ q}(1+\text{r}))$ is a matrix that contains parametric components that are approached with linear functions and nonparametric components that are approximated by spline truncated. Vector $\mathbf{b}$ is a parameter vector of size $n(\text{p+ q}(1+\text{r}))$ and vector $\boldsymbol{\varepsilon}$ is an error vector of size $nt \times 1$. The next step is to get the estimate $\mathbf{b}$ by completing the Weighted Least Square (WLS) optimization until the confidence interval with the unknown variance is obtained as follows [16]:

$$P\left(\hat{\mathbf{b}}_s - t_{\left(\frac{\alpha}{2}, nt-n(\text{p+q}(1+\text{r}))\right)}\sqrt{\frac{\mathbf{y}^T\mathbf{Cy}}{nt-n(\text{p+ q}(1+\text{r}))}\mathbf{V}} \le \mathbf{b}_s \le \hat{\mathbf{b}}_s + t_{\left(\frac{\alpha}{2}, nt-n(\text{p+q}(1+\text{r}))\right)}\sqrt{\frac{\mathbf{y}^T\mathbf{Cy}}{nt-n(\text{p+ q}(1+\text{r}))}\mathbf{V}}\right) = 1-\alpha. \tag{15}$$

3.2 Confidence Intervals Application of Parameters Regression Semipaametric Spline Truncated for Longitudinal Data in Indonesia

In semiparametric regression, the determination of parametric and nonparametric components can be seen by looking at the pattern of the data. The following is a pattern of the relationship between the percentage of poor people with suspected influential factors:

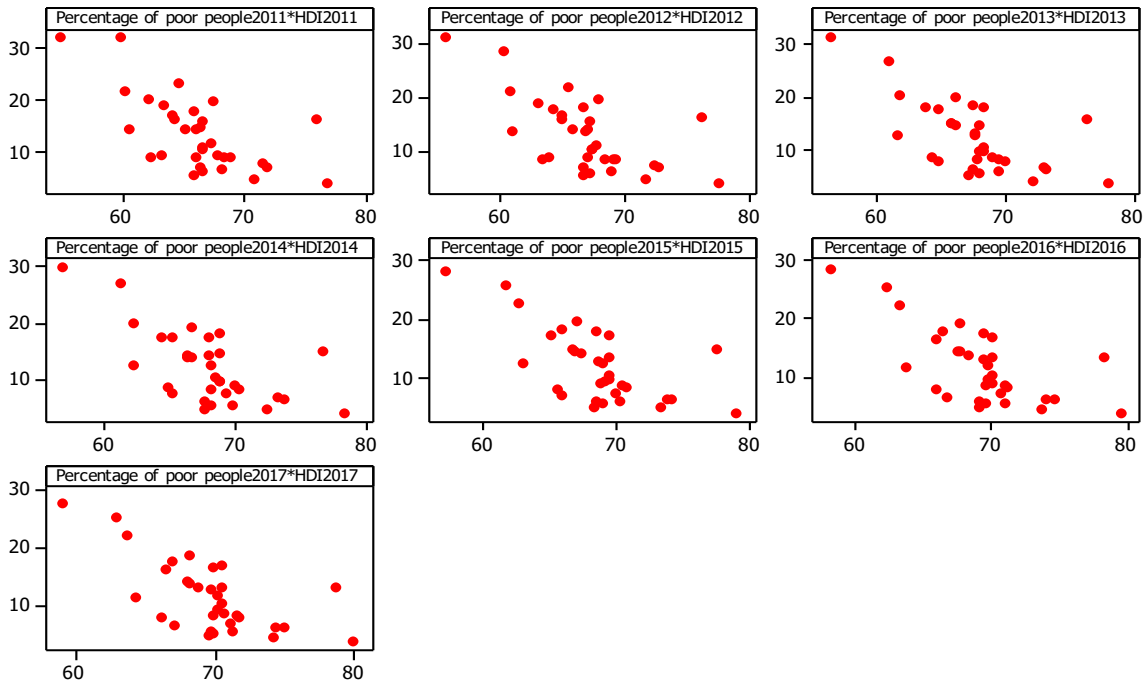

Figure 1. Scatterplot percentage of poor people with HDI in Indonesia during 2011-2017.

In figure 1, it can be seen that the scatterplot the relationship between the percentage of poor population and human development index (HDI) shows a data pattern that approaches linear patterns. Based on this, it can be concluded that HDI as a parametric component. While in figure 2 below, there is a distribution of plot relationships between the percentage of poor people and the unemployment rate showing a spread pattern of data so that the pattern is difficult to know. Therefore, the unemployment component can be assumed as a nonparametric component. The following is a scatterplot between the percentage of poor people and the unemployment rate.

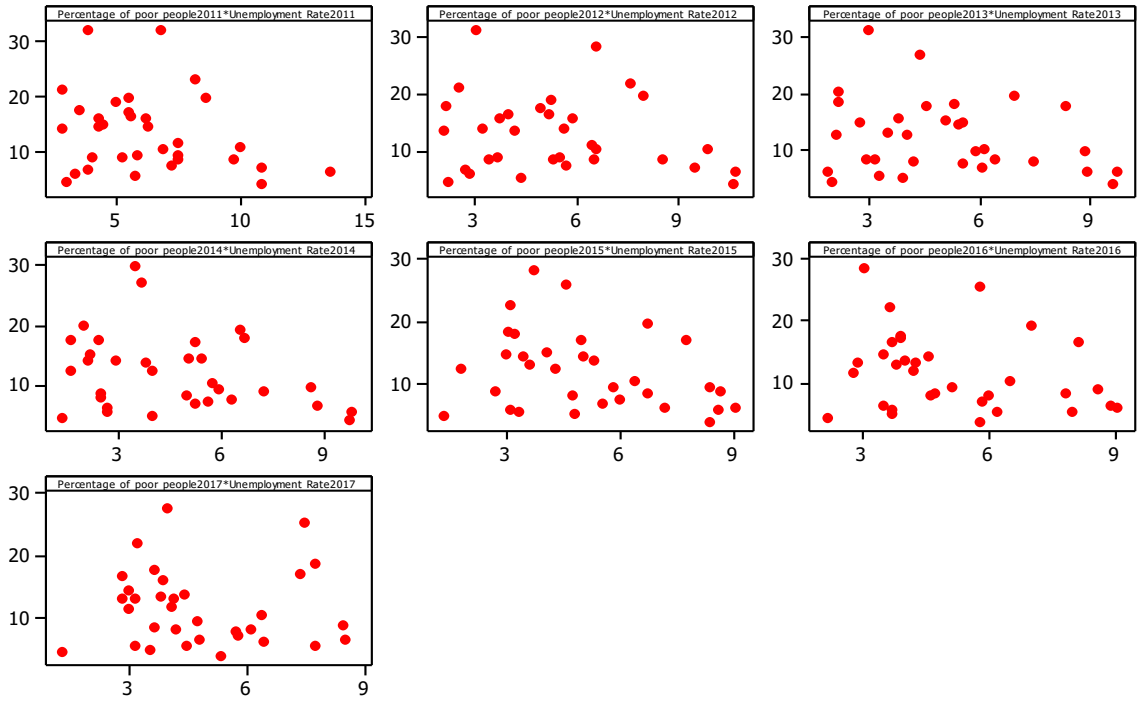

Figure 2. Scatterplot percentage of poor people with the unemployment rate in Indonesia during 2007-2017.

3.3 The Best Model Selection

The first step to choose the best model is to compare the minimum GCV value selected using weighting $\mathbf{W}_1, \mathbf{W}_2,$ dan $\mathbf{W}_3$. The following is a comparison table of GCV values in semiparametric spline truncated longitudinal data modeling with the case of the percentage of poor people in Indonesia:

Table 2. Comparison of GCV Model Values Using Weighting $\mathbf{W}_1, \mathbf{W}_2,$ dan $\mathbf{W}_1$

Weighting	GCV	R^2	MSE
$\mathbf{W}_1$	1.67724843836064	98.6693	0.54767
$\mathbf{W}_2$	1.67724843836068	98.6693	0.54767
$\mathbf{W}_3$	1.87780259488047	98.5687	0.61316

Based on Table 2, the best weighting is the first weighting ($\mathbf{W}_1$) because it has the smallest GCV value compared to the GCV value that uses the first weighting ($\mathbf{W}_2$) and the third weighting ($\mathbf{W}_3$). After determining the best weighting, the next step is to determine the optimal knot point. In this study, the criteria used in selecting the optimal knot point also use the smallest GCV value. The point of knots used is one knot with an increment of 14 increments. The smallest GCV value using one point knot will be

applied to the data on the percentage of poor people in Indonesia with the human development index as parametric variables and the unemployment rate as nonparametric variables as follows:

Table 3. Knot Points and GCV Values with W_1

Increment	Knot $(K_{ui}, u=1; i=1,2,...,33)$				GCV
	$K_{1,1}$	$K_{1,2}$	...	$K_{1,33}$	
1	8.49	7.36	...	3.89	1.6772484
2	8.35	7.25	...	3.81	1.6859862
3	8.22	7.14	...	3.74	1.7036725
4	8.09	7.04	...	3.66	1.7483046
5	7.96	6.93	...	3.59	1.7681480
6	7.82	6.82	...	3.51	1.7774347
7	7.69	6.71	...	3.44	1.7832915
8	7.55	6.60	...	3.36	1.7866767
9	7.42	6.49	...	3.29	1.7883768
10	7.28	6.38	...	3.21	1.7893804
11	7.15	6.28	...	3.14	1.7899272
12	7.02	6.17	...	3.06	1.7905828
13	6.88	6.06	...	2.99	1.7912310
14	6.75	5.95	...	2.91	1.7916209.

The calculation table in table 3, the minimum GCV value is 1.6772484, which means the optimal knot point for one point knot using weighting $\mathbf{W_1}$ is in the first increment. The optimal point of knots in each province uses weighting $\mathbf{W_1}$. So that the best model of semiparametric regression spline truncated longitudinal data on the percentage of poor people in Indonesia with one point knots and using $\mathbf{W_1}$ weighting is generally written as follows:

$$y_{il} = \beta_i x_{il} + \gamma_i z_{il} + \gamma_{2i}\left(z_{il} - K_{ui}\right)_+^1 + \varepsilon_{il}$$

where, $i = 1,2,...,33$ and $l = 1,2,...,7$.

Modeling the percentage of poor people in Indonesia in 2011-2017 using a semiparametric spline truncated regression with one knot point using weighting $\boldsymbol{W_1}$ produces a coefficient of determination (R^2) of 98.669 percent and MSE value of 5.476 × 10-1. The predictor variables that are thought to be influential are the human development index (HDI) and the unemployment rate.

3.4 Interpretation of Semiparametric Spline Truncated Regression Model for Longitudinal Data

Semiparametric spline truncated regression modeling has a good interpretation by looking at changes in the predictor variable data pattern which is characterized by the point of knots. As an illustration, the following will explain the interpretation of the model in the province with the highest average poor population in Indonesia during 2011-2017, namely the Papua Province. Using an example of a predictor variable for open unemployment and assuming another predictor variable is constant, then the semiparametric spline truncated regression model for the percentage of poor people in Papua Province is as follows:

$$y_{33l} = -1.89579x_{33l} + 0.610382z_{33l} + 17.25293(z_{33l} - 3.89)_+$$

then the interpretation of the model above is:

$$y = 0.610382z \qquad ; z < 3.89$$

$$y = 17.2529z + 67.0276 \qquad ; z \geq 3.89.$$

The relationship between the unemployment rate and the percentage of poor people assuming another predictor variable is constant in the Papua Province can be illustrated in the figure below:

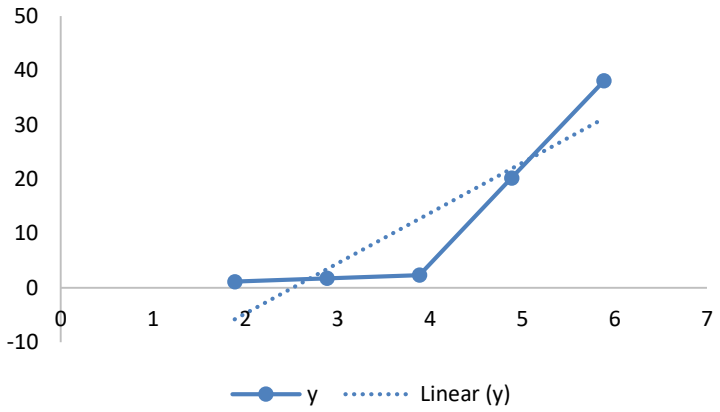

Figure 3. Relationship between the unemployment rate and the percentage of poor people in Papua Province.

The model in Papua province illustrates when the unemployment rate is less than 3.89 percent, if the unemployment rate rises by one percent, then the percentage of poor people will increase by 0.610382. Then when the unemployment rate is more than 3.89 percent, if it rises by one percent, the percentage of the poor will increase by 17.2529.

After obtaining the best semiparametric spline truncated regression model for longitudinal data, we can find out the significance of the parameters to the model using the confidence interval. Taking this conclusion is done by looking at whether the parameter confidence interval contains a value of zero or not. If the confidence interval contains a value of zero, then the parameter does not significantly affect the model. The following is a confidence interval with a 95 percent confidence level presented in table 4 bellow:

Table 4. Confidence Interval of Parameters Regression Semiparametric Spline Truncated for Longitudinal Data

Province	Variable	Parameter	Parameter Estimation	Lower Limit	Upper Limit
Aceh	x_{1l}	β_1	1.01739	0.32642	1.70838*
	z_{1l}	γ_1	0.13522	0.05583	0.21462*
		γ_{21}	5.40638	-0.05855	10.87131
Sumatera Utara	x_{2l}	β_2	1.07596	0.16790	1.98403*
	z_{2l}	γ_2	0.05343	-0.03396	0.14081
		γ_{22}	-1.82085	-5.42722	1.8554
⋮	⋮	⋮	⋮	⋮	⋮
Papua	x_{nl}	β_n	-1.89580	-3.10304	-0.68856*
	z_{nl}	γ_n	0.61038	0.54047	0.68029*
		γ_{2n}	17.25293	11.31880	23.18706*

*) Significant Level 0.05

4. Summary

Application of the semiparametric spline regression model truncated for longitudinal data with data on the percentage of poor people in Indonesia in 2011-2017 using three types of weighting, including $\mathbf{W}_1, \mathbf{W}_2,$ and $\mathbf{W}_3$ and one point knots. The selection of the best model uses the criteria for the smallest GCV value where the model is the best, using weighting $\mathbf{W}_1$. The selected model has a coefficient of determination (R^2) of 98.669 percent and the MSE value of 5.476×10^{-1}. This shows that the model is feasible to use and by using the confidence interval as one of the statistical inferences we can find out the significance of the parameters to the model.

References.

[1] Gujarati, D. N., & Porter, D. C. (2015). Basics Econometrics 2nd Edition. Jakarta: Penerbit Salemba Empat.
[2] Chamidah, N., I. N. Budiantara, Sony, S., Ismaini Z. (2012). National Mathematics XVI Conf. Bandung: Padjajaran University.
[3] Eubank, R. L. (1999). Spline Smoothing and Nonparametric Regression. New York: Marcel Dekker.
[4] Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). Analysis of Longitudinal Data 2nd Edition. Oxford Statistical Science 25. New York.
[5] Fernandes, A. A. R. (2016). Spline Estimator in Birespon Nonparametric Regression for Longitudinal Data. Thesis of Institut Teknologi Sepuluh Nopember Surabaya.
[6] Wahba, G. (1983). Bayesia "Confidence Intervals" for The Cross-validated Smoothing Spline. J. R. Statist. Soc. B., 45, 133-150.
[7] Wahba, G. (1990). Smoothing Spline ANOVA with Component-Wise Bayesian "Confidence Intervals". Journal of Computational and Graphical Statistics, 2, 9-117.
[8] Wang, Y. (1998). Smoothing Spline Models with Correlated Random Errors. J. Am. Stat. Assoc, 15, 341-348.
[9] Mao, W. and Zhao, L. H. (2003). Free-knot polynomial splines with confidence intervals. J. R. Statist. Soc. B.,65, 901-919.
[10] Egle, R. F., Grager, C. W. J., Rice, J. and Weiss, A. (1986). Semiparametric Estimates of The Relation Between Weather and Electricity Sales. J. Am. Stat Assoc, 81, 310-320.
[11] Budiantara, I.N. (2009). Spline in Nonparametric and Semiparametric Regression: A Present and Future Model of Statistics, Inagural Speech fot Professor's Position at the Department of Statistics. Surabaya: ITS Press.
[12] Budiantara, I. N. and Jerry, D. T. P. (2010). Mathematics XV National Conf. Manado: Manado University.
[13] Wu, H., and Zhang, J. T. (2006). Nonparametric Regression Methods for Longitudinal Data Analysis. New Jersey: Jihn Willey and Sons. Inc.
[14] Budiantara, I. N. (2006). Spline Model with Optimal Knots. Journal. Basic Science of Jember University, 7, 77-85.
[15] Badan Pusat Statistik. (2017). Poverty Data and Information 2017. Jakarta: BPS.
[16] Prawanti, D. D. (2019). Parameter Interval Estimation of Semiparametric Spline Truncated Regression Model for Longitudinal Data. Thesis of Institut Teknologi Sepuluh Nopember Surabaya.