Received: 22 Oct 2024 | Revised: 15 Mar 2025 | Accepted: 18 Mar 2025 | Available online: 24 Aug 2025
https://doi.org/10.13057/ijas.v8i1.94370

E-ISSN: 2621-086X
https://jurnal.uns.ac.id/ijas

Indonesian Journal of Applied Statistics Vol. 8, No.1, pp. 13-23, 2025

Penalized Spline Semiparametric Regression for Bivariate Response in Modeling Macro Poverty Indicators

Cinta Rizki Oktarina, Idhia Sriliana*, Sigit Nugroho
Department of Statistics, Universitas Bengkulu, Bengkulu, Indonesia
* Corresponding Author. E-mail: idhiasriliana@unib.ac.id

Abstract

Semiparametric spline regression has become an increasingly popular method for modeling data due to its flexibility and objectivity, especially as a parameter estimation method. Spline functions are highly effective in semiparametric regression because they offer unique statistical interpretations by segmenting each predictor variable in relation to the response variable. Bivariate semiparametric regression can be applied to data where observations tend to have disparities between regions, making it suitable for poverty data, particularly the poverty depth index and the poverty severity index. The objective of this research is to analyze the models of the poverty depth index and poverty severity index, as well as to perform segmentation and interpretation of these models. This study utilized observations from 60 districts/cities in the southern part of Sumatra. Several predictor variables were considered, including the percentage of households with a floor area of ≤19 m^2, labor force participation rate, and life expectancy as parametric components, while the nonparametric components included the average length of schooling and the percentage of households with tap water sources. The estimation methods used were penalized least squares and penalized weighted least squares, involving a full search algorithm for selecting the number and location of knots. The results of the study indicated that the penalized weighted least squares method was the best estimator, with an MSE value of 0.3122 and two knots for each predictor, yielding GCV values of 4.3604 and 4.0794.

Keywords: semiparametric regression; bivariate response; poverty; knot; penalized weighted least square

This is an open access article under the Creative Commons Attrribution-ShareAlike 4.0 International License

How to Cite:
C. R. Oktarina, I. Sriliana, and S. Nugroho, “Penalized spline semiparametric regression for bivariate response in modeling macro poverty indicators,” *Indonesian Journal of Applied Statistics,* vol. 8, no. 1, pp. 13-23, 2025, doi: 10.13057/ijas.v8i1.94370.

1. INTRODUCTION

Regression analysis is a method used to explain how one or more response variables depend on one or more predictor variables. According to [1], there are three approaches used to estimate the regression function: parametric, nonparametric, and semiparametric. Semiparametric regression is used when there are parametric components with known relationship patterns and nonparametric components with unknown patterns, allowing the estimation curve to adjust to the data. Semiparametric regression analysis has been developed not only for univariate response analysis but also for bivariate and multivariate responses. Bivariate response analysis involves two correlated response variables, where significant correlation between response variables is a key requirement [2]. Spline has a unique statistical interpretation in explaining how segments of each predictor variable relate to the response variable. Additionally, spline can handle fluctuating data behavior in certain sub-intervals [3]. Therefore, spline is a suitable function for addressing varying data behavior, making the spline function complex. According to [4], a penalized function is required to control the complexity of the spline function and prevent overfitting.

Copyright © 2025, Author(s)

Research on splines has been widely applied. [5] conducted a study on outlier identification using penalized spline regression to model the poverty depth index as the response variable. The results of this study obtained an R-square value of 69.10%, with the optimal number of knots for each predictor variable being 1, 2, 4, 1, 5, 3, and 1, respectively. Subsequent research by [6] focused on modeling the factors affecting the poverty severity index in 2015 using truncated spline nonparametric regression. Based on the conducted research, it was found that the optimal number of knots for each predictor variable was 3, 3, 2, 3, and 3. Therefore, based on previous research, this study will focus on involving the poverty depth index and the poverty severity index, which are macro poverty indicators, as response variables with significant correlations. The predictor variables suspected to have an influence will be a combination of parametric and nonparametric components.

Poverty is a common social issue faced by developing countries, including Indonesia. Poverty in Indonesia has been a fundamental issue since independence. In the early decades of independence, Indonesia's poverty rate was estimated to be over 50% of the total population, lasting for approximately 20 years. The issue of poverty not only focuses on the percentage of poor people but also on defining poor people differently, such as the poverty depth index and poverty severity index.

2. METHODS

2.1. Regression Analysis

Regression analysis is a method for investigating and modeling the relationships between response and predictor variables [7]. It examines the dependence of one or more response variables on one or more predictor variables [8]. As science progresses, regression modeling has advanced. According to [1], there are three approaches to estimating regression curves: parametric, nonparametric, and semiparametric. Parametric regression assumes a specific form for the relationship, nonparametric regression does not assume any particular form and adapts to the data, while semiparametric regression combines both parametric and nonparametric elements, allowing for more flexibility in modeling complex relationships.

2.2. Bivariate Response Regression

Bivariate response regression involves two response variables that have a significant correlation when estimating data [2]. The estimation process considers the relationship between the response variables and a set of predictor variables, assuming that each response variable follows its own regression model. Generally, the bivariate response regression model with the OLS estimator can be described as follows [9]:

$$y_i^{(d)} = \beta_0^{(d)} + \beta_1^{(d)} x_{1i} + \cdots + \beta_p^{(d)} x_{pi} + \varepsilon_i^{(d)}, d = 1,2 \quad (1)$$

Based on the initial concept of bivariate response regression, the response variables must have a significant relationship, which can be measured using correlation analysis. Pearson correlation analysis is commonly used for this purpose. The Pearson correlation coefficient, denoted by r, ranges between -1 and 1 and can be calculated using the following equation [10]:

$$r = \frac{s_{y^{(1)}y^{(2)}}}{s_{y^{(1)}}s_{y^{(2)}}} \quad (2)$$

2.3. Bivariate Semiparametric Penalized Spline Regression

Bivariate semiparametric penalized spline regression examines the dependence of one or more response variables on one or more predictor variables using the penalized spline estimator. Given paired data $(x_1, x_2, \dots, x_p, t_1, t_2, \dots, t_S)$, the relationship between the variables x_b and $y^{(d)}$is known, while the relationship between the variables t_g and $y^{(d)}$is unknown. The relationship among the variables $x_b, t_g, y^{(d)}$is assumed to follow a semiparametric regression model.

Copyright © 2025, Author(s)
https://doi.org/10.13057/ijas.v8i1.94370
E-ISSN: 2621-086X

One of the estimators that can be used to estimate parameters in the bivariate semiparametric penalized spline method is the penalized least square. Penalized least square essentially assumes constant variance in the error, commonly referred to as homoskedasticity. Detection of homoskedasticity can be performed using several tests, one of which is the Glejser test [8]. The Glejser test is a popular method for detecting heteroskedasticity by regressing the predictor variables against the absolute values of the model's errors.

The next step is to add weights based on the results of the Glejser test. The weights will control the correlation between the responses of the resulting error model, providing more accurate and optimal estimates. The penalized weighted least square (PWLS) method is a technique used to minimize the weighted sum of squared errors, especially when there is a violation of the assumption of constant variance in the error model or the presence of heteroskedasticity. The weighting matrix in PWLS is formulated as follows [11]:

$$\boldsymbol{W} = \begin{bmatrix} s_{y^{(1)}}^2 \boldsymbol{I} & s_{y^{(1)}y^{(2)}} \boldsymbol{I} \\ s_{y^{(2)}y^{(1)}} \boldsymbol{I} & s_{y^{(2)}}^2 \boldsymbol{I} \end{bmatrix}^{-1} \tag{3}$$

The regression function $f(t_i)$ can be estimated using both PLS and PWLS. The PWLS estimator uses a smoothing parameter to control the roughness of the regression function and involves weights in estimating the parameters. PWLS includes weights in the form of the inverse of the variance-covariance matrix of the error model, denoted by $\boldsymbol{W}$, as shown in equation (3). The bivariate semiparametric regression model using the PWLS estimator can be written as follows:

$$y_i^{(d)} = \beta_0^{(d)} + \sum_{b=1}^{p} \beta_b^{(d)} x_{bi} + \delta_0^{(d)} + \sum_{g=1}^{S} \left(\delta_g t_{gi}^{(d)} + \sum_{j=1}^{k_g} \phi_{gj}^{(d)} \left(t_{gi} - \xi_{gj}\right)_+ \right) + \varepsilon_i^{(d)} \tag{4}$$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{T\delta} + \boldsymbol{\varepsilon} \tag{5}$$

The estimation of parameters in a bivariate semiparametric regression model cannot be performed simultaneously for all parameters. Therefore, it is assumed that the parameter $\boldsymbol{\beta}$ is known, allowing the focus to be on estimating the nonparametric component parameters. According to [12], this can be expressed by assuming:

$$\boldsymbol{y}^* = \boldsymbol{y} - \boldsymbol{X\beta}$$
$$\boldsymbol{y}^* = \boldsymbol{T\delta} + \boldsymbol{\varepsilon}$$

The parameter $\widehat{\boldsymbol{\delta}}$ in the spline function is obtained by minimizing the P function expressed as follow [13]:

$$P = \frac{1}{2n} \sum_{i=1}^{n} W_i (y - f(t_i))^2 + \lambda W_i \int_0^1 (f''(t_i))^2 dx \tag{6}$$

$$P = 2n^{-1} (\boldsymbol{y}^* - \boldsymbol{T\delta})' \boldsymbol{W} (\boldsymbol{y}^* - \boldsymbol{T\delta}) + \lambda \boldsymbol{\delta}' \boldsymbol{D\delta} \tag{7}$$

Based on equation (7), $\widehat{\boldsymbol{\delta}}_{\boldsymbol{PWLS}}$ is obtained by decreasing the P function against $\boldsymbol{\delta}$. The first step is to decrease the goodness of fit against $\boldsymbol{\delta}$, so that the following results are obtained:

$$\begin{aligned}
\frac{\partial (\boldsymbol{y}^* - \boldsymbol{T\delta})' \boldsymbol{W} (\boldsymbol{y}^* - \boldsymbol{T\delta})}{\partial \boldsymbol{\delta}} &= \frac{\partial (\boldsymbol{y}^{*\prime} - \boldsymbol{\delta}'\boldsymbol{T}') \boldsymbol{W} (\boldsymbol{y}^* - \boldsymbol{T\delta})}{\partial \boldsymbol{\delta}} \\
&= \frac{\partial (\boldsymbol{y}^{*\prime}\boldsymbol{W} - \boldsymbol{\delta}'\boldsymbol{T}'\boldsymbol{W})(\boldsymbol{y}^* - \boldsymbol{T\delta})}{\partial \boldsymbol{\delta}} \\
&= \frac{\partial \left(\boldsymbol{y}^{*\prime}\boldsymbol{W}\,\boldsymbol{y}^* - \boldsymbol{y}^{*\prime}\boldsymbol{WT\delta} - \boldsymbol{\delta}'\boldsymbol{T}'\boldsymbol{W}\,\boldsymbol{y}^* + \boldsymbol{\delta}'\boldsymbol{T}'\boldsymbol{WT\delta} \right)}{\partial \boldsymbol{\delta}} \\
&= \frac{\partial (\boldsymbol{y}^{*\prime}\boldsymbol{W}\,\boldsymbol{y}^* - 2\boldsymbol{\delta}'\boldsymbol{T}'\boldsymbol{W}\,\boldsymbol{y}^* + \boldsymbol{\delta}'\boldsymbol{T}'\boldsymbol{WT\delta})}{\partial \boldsymbol{\delta}} \\
&= \left(\frac{\partial}{\partial \boldsymbol{\delta}} \boldsymbol{y}^{*\prime}\boldsymbol{W}\,\boldsymbol{y}^* - \frac{\partial}{\partial \boldsymbol{\delta}} 2\boldsymbol{\delta}'\boldsymbol{T}'\boldsymbol{W}\,\boldsymbol{y}^* + \frac{\partial}{\partial \boldsymbol{\delta}} \boldsymbol{\delta}'\boldsymbol{T}'\boldsymbol{WT\delta} \right) \\
&= (0 - 2\boldsymbol{T}'\boldsymbol{W}\,\boldsymbol{y}^* + 2\boldsymbol{T}'\boldsymbol{WT\delta}) \\
&= 2\boldsymbol{T}'\boldsymbol{W}(\boldsymbol{T\delta} - \boldsymbol{y}^*) \qquad (8)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \lambda \boldsymbol{\delta}'\boldsymbol{D\delta}}{\partial \boldsymbol{\delta}} &= \lambda \frac{\partial \boldsymbol{\delta}'\boldsymbol{D\delta}}{\partial \boldsymbol{\delta}} \\
&= 2\lambda \boldsymbol{D\delta} \qquad (9)
\end{aligned}$$

Copyright © 2025, Author(s)
https://doi.org/10.13057/ijas.v8i1.94370
E-ISSN: 2621-086X

Based on equation (8) and equation (9) which will be substituted into equation (10), $\hat{\boldsymbol{\delta}}_{\boldsymbol{PWLS}}$ is obtained as follows:

$$\begin{aligned}
2n^{-1}2\boldsymbol{T}'\boldsymbol{W}\left(\boldsymbol{T}\hat{\boldsymbol{\delta}}-\boldsymbol{y}^{*}\right)+2\lambda\boldsymbol{D}\hat{\boldsymbol{\delta}} &= 0\\
n^{-1}\boldsymbol{T}'\boldsymbol{W}\left(\boldsymbol{T}\hat{\boldsymbol{\delta}}-\boldsymbol{y}^{*}\right)+\lambda\boldsymbol{D}\hat{\boldsymbol{\delta}} &= 0\\
n^{-1}\left(\boldsymbol{T}'\boldsymbol{W}\boldsymbol{T}\hat{\boldsymbol{\delta}}-\boldsymbol{T}'\boldsymbol{W}\boldsymbol{y}^{*}\right)+\lambda\boldsymbol{D}\hat{\boldsymbol{\delta}} &= 0\\
\lambda\boldsymbol{D}\hat{\boldsymbol{\delta}} &= -n^{-1}\left(\boldsymbol{T}'\boldsymbol{W}\boldsymbol{T}\hat{\boldsymbol{\delta}}-\boldsymbol{T}'\boldsymbol{W}\boldsymbol{y}^{*}\right)\\
n\lambda\boldsymbol{D}\hat{\boldsymbol{\delta}} &= \left(\boldsymbol{T}'\boldsymbol{W}\boldsymbol{y}^{*}-\boldsymbol{T}'\boldsymbol{W}\boldsymbol{T}\hat{\boldsymbol{\delta}}\right)\\
(\boldsymbol{T}'\boldsymbol{W}\boldsymbol{T}\hat{\boldsymbol{\delta}}+n\lambda\boldsymbol{D}\hat{\boldsymbol{\delta}}) &= (\boldsymbol{T}'\boldsymbol{W}\boldsymbol{y}^{*})\\
(\boldsymbol{T}'\boldsymbol{W}\boldsymbol{T}+n\lambda\boldsymbol{D})\hat{\boldsymbol{\delta}} &= (\boldsymbol{T}'\boldsymbol{W}\boldsymbol{y}^{*})\\
\hat{\boldsymbol{\delta}}_{\boldsymbol{PWLS}} &= (\boldsymbol{T}'\boldsymbol{W}\boldsymbol{T}+n\lambda\boldsymbol{D})^{-1}\boldsymbol{T}'\boldsymbol{W}\boldsymbol{y}^{*} \qquad (10)
\end{aligned}$$

Based on the $\hat{\boldsymbol{\delta}}_{\boldsymbol{PWLS}}$ estimator, the $\boldsymbol{y}$ function estimation is then obtained as follows:

$$\begin{aligned}
\hat{\boldsymbol{y}}^{*} &= \boldsymbol{T}\hat{\boldsymbol{\delta}}_{\boldsymbol{PWLS}}\\
&= \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{W}\boldsymbol{T}+n\lambda\boldsymbol{D})^{-1}\boldsymbol{T}'\boldsymbol{W}\boldsymbol{y}^{*}\\
&= \boldsymbol{A}\boldsymbol{y}^{*} \qquad (11)
\end{aligned}$$

Based on equation (5) according to [12] $\hat{\boldsymbol{\beta}}$ is obtained through decreasing the K function against $\boldsymbol{\beta}$.

$$\begin{aligned}
K &= \left(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{T}\hat{\boldsymbol{\delta}}_{\boldsymbol{PWLS}}\right)'(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{T}\hat{\boldsymbol{\delta}}_{\boldsymbol{PWLS}})\\
&= (\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{A}\boldsymbol{y}^{*})'(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{A}\boldsymbol{y}^{*})\\
&= \left(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{X\beta})\right)'\left(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{X\beta})\right)\\
&= (\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Ay}-\boldsymbol{AX\beta})'(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Ay}-\boldsymbol{AX\beta})\\
&= (\boldsymbol{y}-\boldsymbol{Ay}-\boldsymbol{X\beta}-\boldsymbol{AX\beta})'(\boldsymbol{y}-\boldsymbol{Ay}-\boldsymbol{X\beta}-\boldsymbol{AX\beta})\\
&= ((\boldsymbol{I}-\boldsymbol{A})\boldsymbol{y}-(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X\beta})'\,((\boldsymbol{I}-\boldsymbol{A})\boldsymbol{y}-(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X\beta})\\
&= (\boldsymbol{y}(\boldsymbol{I}-\boldsymbol{A})'-\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})')((\boldsymbol{I}-\boldsymbol{A})\boldsymbol{y}-(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X\beta})\\
&= \boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{y}-\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X\beta}-\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\,\boldsymbol{y}+\\
&\quad\ \boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X\beta})\\
&= \boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{y}-2(\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X\beta})+\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\,\boldsymbol{X\beta})
\end{aligned}$$

The minimum value of K is reached when $\frac{\partial K}{\partial \boldsymbol{\beta}} = 0$, thus obtained:

$$\begin{aligned}
&0-2(\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{y})+\boldsymbol{2X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X}\hat{\boldsymbol{\beta}}=0\\
&2(-(\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{y})+\boldsymbol{2X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X}\hat{\boldsymbol{\beta}}=0\\
&2(-(\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{y})+\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X}\hat{\boldsymbol{\beta}})=0\\
&-(\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{y})+\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X}\hat{\boldsymbol{\beta}}=0\\
&\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X}\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{y})\\
&\qquad\qquad\qquad\quad \hat{\boldsymbol{\beta}}=(\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{y} \qquad (12)
\end{aligned}$$

The parameter estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}_{\boldsymbol{PWLS}}$ that have been obtained in equation (10) and equation (11), so according to [12] can be substituted in equation (5), as follows:

$$\begin{aligned}
\hat{\boldsymbol{y}} &= \boldsymbol{C}_{\boldsymbol{par}}\boldsymbol{y}+\boldsymbol{C}_{\boldsymbol{nonpar}}\boldsymbol{y}\\
&= (\boldsymbol{C}_{\boldsymbol{par}}+\boldsymbol{C}_{\boldsymbol{nonpar}})\boldsymbol{y}\\
&= \boldsymbol{C}_{\boldsymbol{semipar}}\boldsymbol{y} \qquad (13)
\end{aligned}$$

$$\boldsymbol{C}=\boldsymbol{X}(\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})'(\boldsymbol{I}-\boldsymbol{A})$$

2.4. Optimal number and location of knots

A knot (ξ_j) is a point where there is a change in the behavior of a function at different intervals. Penalized spline regression applies knots located at quantile points which are unique values of the predictor variables after the data is sorted. In determining the location of knots using penalized spline regression, it can be written as follows [14]:

$$\xi_j = \frac{j}{k+1,}, j = 1,2,3,..,k \qquad (14)$$

In determining the optimal location and number of knots, a frequently used method is the full search method. This method allows to systematically search for the knot configuration that best fits the data.

Copyright © 2025, Author(s)
https://doi.org/10.13057/ijas.v8i1.94370
E-ISSN: 2621-086X

Meanwhile, in determining the smoothing parameter λ as well as the optimal number and location of knots, a commonly used method is to examine the generalized cross validation (GCV) value that reaches the minimum [15] .The advantage of the GCV method is that it is asymptotically optimal [16]. The smoothing parameter λ plays a role in controlling the roughness penalty. As the value of λ increases, the function estimate becomes smoother, while decreasing the value of λ will result in a coarser estimate. The GCV method can be defined as follows [17]:

$$GCV(\xi,\lambda) = \frac{MSE(\xi,\lambda)}{\left(1-2n^{-1}tr(\boldsymbol{A})\right)^2} \tag{15}$$

The full-search method is an algorithm that relies on the minimum GCV value to determine the most optimal number of knots by comparing the GCV values for $\xi = 1$ and $\xi = 2$. When the GCV value of $\xi = 1 < \xi = 2$ then the iteration process will stop and have the optimal number of knots which is 1. Meanwhile, if the GCV value of $\xi = 1 > \xi = 2$ then the iteration process will continue by comparing the GCV values of $\xi = 2$ and $\xi = 3$. Comparing GCV values is done in the same way until the minimum GCV value is obtained [17].

2.5. Types and Sources of Data

This study utilizes data from BPS publications covering five provinces in southern Sumatra in 2022, known as "Provinsi dalam Angka". Observations include 60 regencies/cities in these provinces. The study used R programming language for analysis. The response variables are the poverty depth index (Y_1) and the poverty severity index (Y_2). The predictor variables include the percentage of households with floor area ≤19 m^2 (X_1), labor force participation rate (X_2), life expectancy (X_3), average years of schooling (X_4), and the percentage of households with piped drinking water (X_5).

2.6. Data Analysis

The steps in this study are as follows:

1. Collect and describe data on the poverty depth index, poverty severity index, and influencing factors based on theoretical reviews and previous research.
2. Measure the correlation between the two response variables using Pearson correlation (equation (2)).
3. Visualize data with scatterplots to determine relationship patterns between response and predictor variables.
4. Validate parametric and nonparametric components using the Terasvirta linearity test.
5. Estimate the model with the PLS estimator:
   a) Determine knot locations and count using the full search method and optimize lambda based on minimum GCV (equation (15)).
   b) Estimate the PLS model using the optimal knots and lambda.
   c) Obtain the estimated functions $y^{(1)}$ and $y^{(2)}$
   d) Test for heteroscedasticity in the PLS error covariance matrix using the Glejser test.
   e) Define the weighting matrix $\boldsymbol{W}$ based on heteroscedasticity test results.
6. Estimate the model with the PWLS estimator:
   a) Determine optimal knot locations, count, and lambda using the full search method and minimum GCV.
   b) Estimate the PWLS model using the optimal parameters and weighting matrix $\boldsymbol{W}$ (equation (7)).
   c) Obtain the estimated functions $y^{(1)}$ and $y^{(2)}$.
7. Select the best model based on minimum MSE
8. Assess model fit using eta square
9. Segment and interpret the best model, create a semiparametric biresponse model, and plot observation data and estimated response variables.

Copyright © 2025, Author(s)
https://doi.org/10.13057/ijas.v8i1.94370
E-ISSN: 2621-086X

3. RESULTS AND DISCUSSION

3.1. Correlation on Response Variables

Based on Table 1, the calculated t_{value} is 37.7987, which is greater than the critical t_{table} of 2.0000. Thus, H_0 is rejected, indicating a strong and significant correlation of 0.9803 between the response variables, poverty depth index, and poverty severity index, in southern Sumatra in 2022 at a 5% significance level. This confirms the assumption of correlation, indicating that modeling these indices using a bivariate response approach is appropriate.

Table 1. Correlation output of $Y^{(1)}$ and $Y^{(2)}$

$H_0: \rho = 0$	
Statistics	Value
r	0.9803
t_{value}	37.7987
t_{table}	2.0000
p-value	<0.001

3.2. Determination of Parametric and Nonparametric Components

The next step is to test the linearity between the response and predictor variables (Table 2). This test analyzes whether the relationship can be represented linearly. If linearity is confirmed, parametric components may suffice. However, if the relationship is nonlinear, nonparametric components like penalized splines should be considered to capture the more complex relationship between the response and predictor variables.

Table 2. Results of Terasvirta linearity test

Variable	p-value		Pattern of Relationship		Component
	Y_1	Y_2	Y_1	Y_2	
X_1	0.3815	0.3169	Linear	Linear	Parametric
X_2	0.3623	0.3860	Linear	Linear	Parametric
X_3	0.1447	0.1699	Linear	Linear	Parametric
X_4	0.0316	0.0401	Nonlinear	Nonlinear	Nonparametric
X_5	0.0113	0.0223	Nonlinear	Nonlinear	Nonparametric

3.3. Penalized Least Square Model Estimation

The penalized least square estimator is a smooth estimator used to obtain a regression function that fits the data. It addresses the challenge of splines fitting every data point too closely, which can create a rough curve. Therefore, a smoothing parameter is needed to balance goodness of fit and roughness penalty. The next steps involve determining the optimal number of knots, their locations, and the smoothing parameter (Table 3 and Table 4):

Table 3. The most optimal knot results t_1

Number of Knots	Knot	Lambda	GCV Minimum
1	8.2000	100	0.9265**
2	7.8000 8.4000	100	0.9266

Copyright © 2025, Author(s)
https://doi.org/10.13057/ijas.v8i1.94370
E-ISSN: 2621-086X

Table 4. The most optimal knot results t_2

Number of Knots	Knot	Lambda	GCV Minimum
1	6.0800	6.3800	0.9149
2	2.81600 8.62600	7.5300	0.9069
3	1.76500 6.0800 10.8650	12.6500	0.9041**

Notes: ** number of knots, knot locations and optimal lambda

When estimating parameters with penalized least square, the upper and lower limits of lambda and increment values must be determined by involving the linear order and the most optimal number and location of knots for each nonparametric variable. So that the parameter estimates are obtained as follows:

$$\begin{aligned}\hat{y}_i^{(1)} = {} & 2.19 \times 10^{-14} + 0.0668x_{1i} - 0.0977x_{2i} - 0.1971x_{3i} + 13.9718 + 0.1561t_{1i} \\ & - 0.0002(t_{1i} - 8.2000)_+ + 0.0008t_{2i} - 0.0001(t_{2i} - 1.7650)_+ \\ & - 0.0017(t_{2i} - 6.0800)_+ - 0.0032(t_{2i} - 10.8650)_+\end{aligned}$$

$$\begin{aligned}\hat{y}_i^{(2)} = {} & -2.08 \times 10^{-14} + 0.0292x_{1i} - 0.2656x_{2i} + 0.0531x_{3i} + 3.8550 + 0.0433t_{1i} \\ & - 4.31 \times 10^{-5}(t_{1i} - 8.2000)_+ - 0.00046t_{2i} \\ & + -4.42 \times 10^{-5}(t_{2i} - 1.7650)_+ - 0.0004(t_{2i} - 6.0800)_+ \\ & - 0.0006(\mathrm{t}_{2\mathrm{i}} - 10.8650)_+\end{aligned}$$

3.4. Testing the Variance of the PLS Error Model

The penalized least square estimator assumes that there is a constant variance in the error, which is called homoscedasticity. Homoscedasticity can be detected by various tests, one of which is the Glejser test (Table 5):

Table 5. Testing the variance of the error model

Source	Degree Freedom	Sum of Squares	Mean Square	F_{value}	p-value
Regression	4	15.3631	3.8408	48.6519	<0.001
Error	114	8.9996	0.0789		
Sum	119	24.3628			
		df_1	4		
		df_2	114		

Based on Table 5, the p-value is <0.001, which is smaller than the significance level of 0.05. Therefore, the null hypothesis (H_0) is rejected, indicating a variance difference in residuals between responses at the 5% significance level. This suggests the presence of heteroskedasticity in the regression model, which can affect the reliability of the analysis. It is important to consider steps to correct model assumptions for more accurate and reliable results. So that the weight matrix symbolized by $\boldsymbol{W}$ is obtained as follows:

$$\boldsymbol{W} = \begin{bmatrix} 49.9252\boldsymbol{I} & | & -154.1068\boldsymbol{I} \\ --- & -- & -- \\ -154.1068\boldsymbol{I} & | & 520.58\boldsymbol{I} \end{bmatrix}$$

3.5. Penalized Weigthed Least Square Model Estimation

The penalized weighted least square (PWLS) method can be used by minimizing the weighted sum of squared errors when the assumption of constant variance in the error model is violated or

Copyright © 2025, Author(s)
https://doi.org/10.13057/ijas.v8i1.94370
E-ISSN: 2621-086X

heteroscedasticity. The next step is to determine the optimal number of knots, knot locations and smoothing parameters as in equation (7) which will involve the weight matrix (Table 6 and Table 7).

Table 6. The most optimal knot results t_1

Number of Knots	Knot	Lambda	GCV Minimum
1	8.2000	1	4.5363
2	7.8000 8.4000	1	4.4493**
3	7.6450 8.2000 8.6300	1	4.4567

Table 7. The most optimal knot results t_2

Number of Knots	Knot	Lambda	GCV Minimum
1	6.0800	1.81	4.3878
2	2.8167 8.6267	1.00	4.0794**
3	1.7650 6.0800 10.8600	1.54	4.1263

Notes: ** number of knots, knot locations and optimal lambda

So that the parameter estimates are obtained as follows:

$$\begin{aligned}\hat{y}_i^{(1)} = 4.3525 \times 10^{-13} + 0.0629x_{1i} - 0.2060x_{2i} - 0.1944x_{3i} + 13.8598 + 0.1554t_{1i} \\ - 0.0028(t_{1i} - 7.8000)_+ - 0.0009(t_{1i} - 8.4000)_+ + 0.0032t_{2i} \\ + 0.0035(t_{2i} - 2.8167)_+ - 0.0120(t_{2i} - 8.267)_+\end{aligned}$$

$$\begin{aligned}\hat{y}_i^{(2)} = 1.2458 \times 10^{-14} + 0.0305x_{1i} - 0.1737x_{2i} - 0.0543x_{3i} + 3.9045 + 0.0393t_{1i} \\ + 0.0073(t_{1i} - 7.8000)_+ + 0.0004(t_{1i} - 8.4000)_+ + 0.0082t_{2i} \\ - 0.0159(t_{2i} - 2.8167)_+ + 0.0069(t_{2i} - 8.6267)_+\end{aligned}$$

3.6. Best Model Segmentation and Interpretation

After determining the best PWLS model for the poverty index in southern Sumatra in 2022, the focus shifts to segmenting and interpreting the model. Segmenting uses model variables to divide the population into groups with similar poverty characteristics. Interpretation analyzes regression coefficients to understand the predictors' relative influence on the poverty index and explore related patterns or trends. These steps aim to deepen understanding of factors affecting poverty in the region, providing a solid foundation for effective decision-making and policy planning.

Based on the above model, the interpretation for each predictor variable on response 1 is as follows: an increase of one unit in variable x_1 will increase the poverty depth index by 0.0629, assuming other predictors remain constant; an increase of one unit in variable x_2 will decrease the poverty depth index by 0.2060, assuming other predictors remain constant; and an increase of one unit in variable x_3 will decrease the poverty depth index by 0.1944, assuming other predictors remain constant. The interpretation of the nonparametric component will be based on the penalized weighted least square estimator for each predictor variable in the form of piecewise functions. The piecewise function for the average years of schooling is as follows:

$$f^{(1)}(t_{1i}) = 0.1554t_{1i} - 0.0028(t_{1i} - 7.8000)_+ - 0.0009(t_{1i} - 8.4000)_+$$

Copyright © 2025, Author(s)
https://doi.org/10.13057/ijas.v8i1.94370
E-ISSN: 2621-086X

$$f^{(1)}(t_{1i}) = \begin{cases} 0.1554t_{1i}. & 0 < t_{1i} \leq 7.8000 \\ 0.1526t_{1i} + 0.0218. & 7.8000 < t_{1i} \leq 8.4000 \\ 0.1517t_{1i} + 0.0075. & t_{1i} > 8.4000 \end{cases}$$

Based on the piecewise function above, when other predictor variables are constant, the interpretation of t_1 on the poverty depth index is as follows: when t_1 is less than or equal to 7.8000, every one-year increase in t_1 tends to increase the poverty depth index by 0.1554. When the average years of schooling t_1 is in the interval of 7.8000 to 8.4000, every one-year increase in t_1 tends to increase the poverty depth index by 0.1526. If t_1 is more than 8.4000, every one-year increase in t_1 tends to increase the poverty depth index by 0.1517.

$$f^{(1)}(t_{2i}) = 0.0032t_{2i} + 0.0035(t_{2i} - 2.8167)_+ - 0.0120(t_{2i} - 8.6267)_+$$

$$f^{(1)}(t_{2i}) = \begin{cases} 0.0032t_{2i}. & 0 < t_{2i} \leq 2.8167 \\ 0.0067t_{2i} - 0.0098. & 2.8167 < t_{2i} \leq 8.6267 \\ -0.0053t_{2i} + 0.1035. & t_{2i} > 8.6267 \end{cases}$$

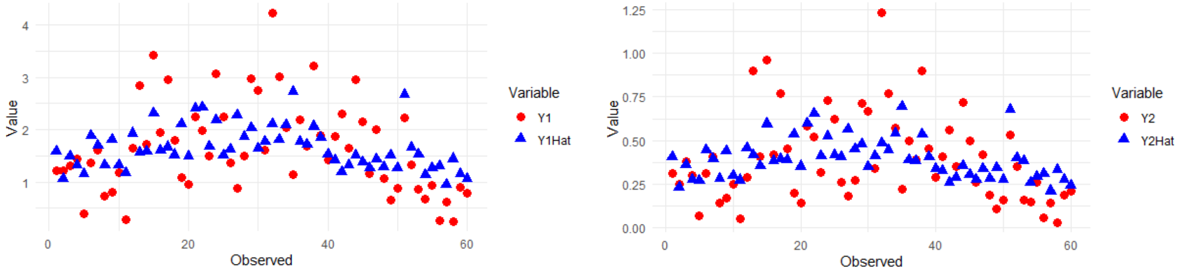
$$f^{(2)}(t_{1i}) = 0.0393t_{1i} + 0.0073(t_{1i} - 7.8000)_+ + 0.0004(t_{1i} - 8.4000)_+$$

$$f^{(2)}(t_{1i}) = \begin{cases} 0.0393t_{1i}. & 0 < t_{1i} \leq 7.8000 \\ 0.0466t_{1i} - 0.0547. & 7.8000 < t_{1i} \leq 8.4000 \\ 0.0470t_{1i} - 0.0034. & t_{1i} > 8.4000 \end{cases}$$

$$f^{(2)}(t_{2i}) = 0.0082t_{2i} - 0.0159(t_{2i} - 2.8167)_+ + 0.0069(t_{2i} - 8.6267)_+$$

$$f^{(2)}(t_{2i}) = \begin{cases} 0.0082t_{2i}. & 0 < t_{2i} \leq 2.8167 \\ -0.0077t_{2i} + 0.0448. & 2.8167 < t_{2i} \leq 8.6267 \\ -0.0008t_{2i} - 0.0595. & t_{2i} > 8.6267 \end{cases}$$

Based on the estimated function $y^{(d)}$ obtained, $y^{(d)}$ and $\hat{y}^{(d)}$can then be plotted to find out how much difference between the estimated results and the original can be seen in Figure 1 as follows:



(a) (b)

Figure 1. Plot of the estimated function $y^{(d)}$; (a) response variable $y^{(1)}$ and (b) response variable $y^{(2)}$

The scatter plots compare observed values (red circles) with predicted values (blue triangles) for two response variables, $y^{(1)}$ and $y^{(2)}$, across 60 observations. For both variables, the predicted values generally follow the observed values, indicating that the model captures the overall trends. However, there are some discrepancies between the observed and predicted values, suggesting that the model does not perfectly fit all data points.

4. CONCLUSIONS

This study applied penalized semiparametric spline regression to model the poverty depth index and poverty severity index, considering both parametric and nonparametric components. The penalized

Copyright © 2025, Author(s)
https://doi.org/10.13057/ijas.v8i1.94370
E-ISSN: 2621-086X

weighted least square (PWLS) method was found to be more accurate than penalized least square (PLS), as indicated by a lower mean squared error (MSE) of 0.3122 and optimal generalized cross-validation (GCV) values of 4.3604 and 4.0794. The findings revealed that average years of schooling and the percentage of households with piped water exhibited a nonparametric relationship, while other variables, such as the percentage of households with a floor area ≤19 m^2, labor force participation rate, and life expectancy, followed a parametric pattern. The segmentation analysis further emphasized the importance of these variables in explaining poverty levels across different districts and cities in southern Sumatra. By incorporating both parametric and nonparametric elements, this study successfully captured the complex relationships between predictor variables and poverty indicators. The results provide valuable insights for policymakers in designing targeted interventions to reduce poverty, particularly by focusing on education, housing conditions, and access to clean water. These findings highlight the potential of bivariate semiparametric regression as a robust statistical approach for analyzing socio-economic issues and guiding effective poverty alleviation strategies.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my main supervisor for his valuable guidance, my co-supervisor for his very useful input, and the examiners for their very constructive suggestions.

REFERENCES

[1] I. N. Budiantara, *Regresi Nonparametrik Spline Truncated*. Surabaya: ITS Press, 2019.

[2] P. P. Gabrela, J. D. T. Purnomo, and I. N. Budiantara, "The estimation of mixed truncated spline and fourier series estimator in bi-response nonparametric regression," *AIP Conf. Proc.*, vol. 2903, no. 1, 2023, doi: 10.1063/5.0177224.

[3] N. Y. Adrianingsih, I. N. Budiantara, and J. D. T. Purnomo, "Mixture model nonparametric regression and its application," *J. Phys. Conf. Ser.*, vol. 1842, no. 1, 2021, doi: 10.1088/1742-6596/1842/1/012044.

[4] W. Griggs, *Penalized Spline Regression and Its Application*. Whitman College, 2013.

[5] A. R. Fadilah, A. Fitrianto, and I. M. Sumertajaya, "Outlier identification on penalized spline regression modeling for poverty gap index in Java," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 16, no. 4, pp. 1231–1240, 2022, doi: 10.30598/barekengvol16iss4pp1231-1240.

[6] W. P. Aji, "Pemodelan faktor-faktor yang mempengaruhi indeks keparahan kemiskinan pada tahun 2015 menggunakan regresi spline truncated," Institut Teknologi Seputuh Nopember, 2018. [Online]. Available: https://repository.its.ac.id/51114/1/1313100034-undergraduate_theses.pdf

[7] D. C. Montgomery, E. A. Peck, and G. . Vinning, *Linear Regression Analysis*. John Wiley & Sons, Inc, 2021. doi: 10.2307/1268395.

[8] D. N. Gujarati, *Econometrics by Example*, 2nd ed. Palgrave, 2015.

[9] R. Jhonson and D. Winchern, *Applied Multivariate Statistical Analysis*. Pearson Education Limited, 2014.

[10] A. C. Rencher and W. F. Cristensen, *Methods of multivariate analysis*. A JOHN WILEY & SONS, INC., PUBLICATION, 2012. doi: 10.5860/choice.33-1586.

[11] H. Nurcahayani, I. N. Budiantara, and I. Zain, "The curve estimation of combined truncated spline and fourier series estimators for multiresponse nonparametric regression," *Mathematics*, vol. 9, no. 10, 2021, doi: 10.3390/math9101141.

Copyright © 2025, Author(s)
https://doi.org/10.13057/ijas.v8i1.94370
E-ISSN: 2621-086X

[12] Z. Azizah, “Interval Konfidensi untuk Parameter Model Regresi Semiparametrik Birespon dengan Pendekatan Spline Truncated (Aplikasi pada Data Kemiskinan dan Pengeluaran Per Kapita Makanan Provinsi Jawa Timur),” Institut Teknologi Seputuh Nopember, 2018.

[13] P. . Green and B. . Silverman, *Nonparametric Regression and Generalized Linear Models*. Springer Science Business Media, 1994.

[14] L. Yang and Y. Hong, “Adaptive penalized splines for data smoothing,” *Comput. Stat. Data Anal.*, vol. 108, pp. 70–83, 2017, doi: 10.1016/j.csda.2016.10.022.

[15] X. Li *et al.*, “A practical and effective regularized polynomial smoothing (RPS) method for high-gradient strain field measurement in digital image correlation,” *Opt. Lasers Eng.*, vol. 121, no. April, pp. 215–226, 2019, doi: 10.1016/j.optlaseng.2019.04.017.

[16] G. Wahba, *Spline Models for Observational Data.* Pennsylvania: Society For Industrial And Applied Mathematics., 1990.

[17] D. Ruppert, M. P. Wand, and R. . Carroll, *Semiparametric Regression*. Cambridge University Press, 2003. doi: 10.1201/9781420091984-c17.

Copyright © 2025, Author(s)

https://doi.org/10.13057/ijas.v8i1.94370
E-ISSN: 2621-086X