Nonparametric Regression Estimator of Multivariable Truncated Spline For Categorical Data

Afiqah Saffa Suriaslan, I Nyoman Budiantara, and Vita Ratnasari

Abstract—**Recent years have witnessed significant interest in truncated spline estimators for nonparametric regression with quantitative data. However, the applicability of these estimators is limited by the frequent occurrence of categorical response variables in real-world applications. A paucity of nonparametric estimators exists for handling categorical response data. This necessitates a method capable of modeling relationships between variables exhibiting pattern shifts across sub-intervals with categorical outcomes. This article thus introduces a novel multivariate truncated spline nonparametric regression estimator for categorical data, developed through a synthesis of literature and theoretical research. The developed method was applied to Indonesia's 2023 poverty depth index data and East Java's 2020 gender development index data. A comparative analysis of the Truncated Spline nonparametric regression model and the binary logistic regression model for estimating categorical data revealed that the Truncated Spline approach yielded superior estimations. Some of the highlights of the proposed method are: 1) This study employs truncated spline nonparametric regression to model categorical response data, 2) Optimal knot placement is determined using the Akaike Information Criterion (AIC), 3) The method's overall performance is demonstrated through its application to two established datasets, and 4) This study compares truncated spline nonparametric regression and logistic regression.**

Index Terms—**Nonparametric Regression, Truncated Spline, Maximum Likelihood Estimation, Categorical Data**

I. INTRODUCTION

Modeling In statistics, the focus is on several basic principles that form the foundation, concept Simplicity, interpretability, and ease of use including mathematical simplicity operationalization. Regression analysis is a common statistical technique for modeling employed across numerous research areas [1]. One of the approaches to regression analysis is nonparametric regression. Nonparametric methods are flexible for unknown data, letting the data determine the regression curve without researcher bias [2]. Smoothness in nonparametric regression curves is mathematically formalized by the curves residing within a designated function space. Consequently, the nonparametric regression method offers considerable flexibility, employing smoothing techniques, nonparametric estimation allows for modeling based directly on the observed data. Among various nonparametric methods, researchers frequently develop and utilize Spline estimator. Spline estimators achieve this adaptability by relying on knot points [3]. This method can be effectively applied when the connections between variables change within certain sub-intervals. Various research [4]-[8] use Spline nonparametric estimator.

Research [9]-[11] applied the nonparametric regression using Truncated Spline estimators. Additionally, [12]-[14] apply semiparametric regression using Truncated Splines [15], [16] develop bi-response regression utilizing Development of a nonparametric regression mixture estimator using Truncated Splines is detailed in [17]-[21]. Although research has been conducted, it primarily centers on quantitative responses, ignoring the prevalence of qualitative responses (categorical data). Thus, the prior Truncated Spline model could not handle categorical responses. A method capable of dealing with categorical responses is logistic regression, which provides insights into the connections between variables with a categorical scale that has two (binary) categories or more using one or more predictor variables.

This research aims to study and develop theories in the statistical modeling area, particulary Truncated Spline nonparametric regression modeling. The is subsequently used on poverty depth index data from 34 provinces in Indonesia for 2023 and The 2020 Gender Development Index data from 38 East Javanese districts/cities reveals a significant challenge: Indonesia's poverty rate, at 9.36% according to the Central Bureau of Statistics, negatively impacts community welfare. Comprehensive analysis incorporating all contributing factors is therefore crucial for effective intervention. Measuring the welfare of a region's population can be done through the poverty depth index. Meanwhile, the strategy adopted from 2019 onward to accelerate economic growth, crucial for improving people's welfare, is centered on human resource development. Promoting quality and competent human resources requires gender balance. The gender development index provides a comparison of human development achievements between women and men.

The research results obtained are the development of Truncated Spline nonparametric regression model theory, as well as models based on poverty depth index data and gender development index and factors considered to impact. The theoretical development of the The performance of a

Manuscript received August 15, 2024; revised April 19, 2025.

This work is supported by the Directorate General of Science and Technology Resources and Higher Education, Kemristekdikti, through the Master's Education Towards Doctoral Program for Superior Scholars (PMDSU) scholarship, which funds this research based on decision letter number 0964/E5/AL.04/2024 and contract number 38/E5/PG.02.00.PL/2024.

Afiqah Saffa Suriaslan is a postgraduate student of the Statistics Department, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia (e-mail: afiqahfq75@gmail.com).

I Nyoman Budiantara is a Professor in Statistics Department, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia. (corresponding author to provide phone: 0815-5352-7408; e-mail: nyomanbudiantara65@gmail.com)

Vita Ratnsari is a Professor in Statistics Department, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia. (e-mail: vita_ratna@its.ac.id).

truncated spline nonparametric regression model is assessed by comparison with logistic regression This evaluation includes a comparison of prediction accuracy and model fit through several criteria. Thus, this research focuses not only on theory development, but also on validation and application of the model to real data.

II. Preliminaries

A. Truncated Spline Multivariable

If the variable function is approximated using a Spline Truncated function of degree m and knot points K_{1j}, $K_{2j,\dots,}$ K_{rj}, where j is 1, 2, ... , p. Then it can be written into the following equation [22]:

$$f(x_{1i}, x_{2i}, \dots, x_{pi}) = \beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{m}\beta_{jk}x_{ji}^{k} + \sum_{j=1}^{p}\sum_{u=1}^{r}\beta_{j(m+u)}(x_{ji} - K_{ju})_{+}^{m} \quad (1)$$

with $i = 1,2,\dots,n$

The Truncated function is given by:

$$(x_{ji} - K_{ju})_{+}^{m} = \begin{cases}(x_{ji} - K_{ju})^{m}, x_{ji} \geq K_{ju}\\ 0 \quad\quad\quad\quad , x_{ji} < K_{ju}\end{cases} \quad (2)$$

Where β_0, β_{jk}, and $\beta_{j(m+u)}$, $j = 1,2,\dots,p$, $k = 1,2,\dots,m$, $u = 1,2,\dots,r$ are the model parameters in the Truncated Spline function.

B. Regression For Categorical Data

The response variable in regression for categorical data (binary logistic) will have a probability $\pi(\boldsymbol{x_i})$ if it is 1, and has a probability value of $1 - \pi(\boldsymbol{x_i})$ if it is 0. The probability distribution of binary logistic regression can be written as follows [23]:

$$Y_i \sim B(1, \pi(\boldsymbol{x_i})) \quad (3)$$

The probability distribution function is:

$$P(Y_i = y_i) = \pi(\boldsymbol{x_i})^{y_i}(1 - \pi(\boldsymbol{x_i}))^{1-y_i}, \; y_i = 0,1 \; , 0 < \pi(\boldsymbol{x_i}) < 1 \quad (4)$$

The regression model for logistic regression is:

$$\pi(\boldsymbol{x_i}) = \frac{\exp(w)}{1 + \exp(w)} \quad (5)$$

Where w is the value that is a function of the predictor variables.

III. Results

A. Regression of Nonparametric Truncated Splie For Categorical Data

To obtain nonparametric regression estimator of multivariable Truncated Spline for categorical data, several steps are required. First, build a Truncated Spline nonparametric regression model with optimal knot points. Next, to optimize the model, first, the log-likelihood function is established, and its derivatives are calculated. Then, the Newton-Raphson method is used to find the parameter estimates via numerical iterations

Given $x_{1i}, x_{2i}, \dots, x_{pi}$; $i = 1, 2, \dots, n$, The number of predictor variables is p, with a probability distribution of

$$Y_i \sim B\left(1, \pi(x_{1i}, x_{2i}, \dots, x_{pi})\right), \; i = 1,2,\dots,n \quad (6)$$

where the probability of success is

$$P(Y_i = 1) = \pi(x_{1i}, x_{2i}, \dots, x_{pi}) = \pi(\boldsymbol{x_i}) \quad (7)$$

and the probability of unsucces:

$$P(Y_i = 0) = 1 - \pi(x_{1i}, x_{2i}, \dots, x_{pi}) = 1 - \pi(\boldsymbol{x_i}) \quad (8)$$

with the probability function:

$$P(Y_i = y_i) = \pi(\boldsymbol{x_i})^{y_i}(1 - \pi(\boldsymbol{x_i}))^{1-y_i} \quad (9)$$

Where $y_i = 0,1$; $i = 1,2,\dots,n$ and $\pi(\boldsymbol{x_i})$ is defined in the probability distribution function $P(Y_i = y_i)$ as follows:

$$P(Y_i = y_i) = \pi(\boldsymbol{x_i})^{y_i}(1 - \pi(\boldsymbol{x_i}))^{1-y_i} = \pi(\boldsymbol{x_i})^{y_i}(1 - \pi(\boldsymbol{x_i}))^{1}(1 - \pi(\boldsymbol{x_i}))^{-y_i}$$

$$P(Y_i = y_i) = \left(\frac{\pi(\boldsymbol{x_i})}{1 - \pi(\boldsymbol{x_i})}\right)^{y_i}(1 - \pi(\boldsymbol{x_i})) \quad (10)$$

Lemma 1

The link function simplifies the logit transformation, applied as in equation (10), facilitates simpler parameter estimation within the logistic regression model.

$$\ln\left(\frac{\pi(\boldsymbol{x_i})}{1 - \pi(\boldsymbol{x_i})}\right) = \beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{m}\beta_{jk}x_{ji}^{k} + \sum_{j=1}^{p}\sum_{u=1}^{r}\beta_{j(m+u)}(x_{ji} - K_{ju})_{+}^{m} \quad (11)$$

Proof of Lemma 1

Based on (10), then we made in the natural logarithm function (ln)

$$\ln P(Y_i = y_i) = y_i \ln\left(\frac{\pi(\boldsymbol{x_i})}{1 - \pi(\boldsymbol{x_i})}\right) + \ln(1 - \pi(\boldsymbol{x_i})) \quad (12)$$

If made in exponential form, (12) forms the exponential family distribution function which is written as follows

$$f(y_i, w) = \exp\left(\frac{y_i.w - b(w)}{a(\emptyset)} + c(w, \emptyset)\right)$$

Thus,

$$P(Y_i = y_i) = \exp\left(\frac{y_i \ln\left(\frac{\pi(\boldsymbol{x_i})}{1 - \pi(\boldsymbol{x_i})}\right) - (-\ln(1 - \pi(\boldsymbol{x_i})))}{1}\right)$$

where, the logit function is obtained:

$$\mathrm{w} = \ln\left(\frac{\pi(\boldsymbol{x_i})}{1 - \pi(\boldsymbol{x_i})}\right) \quad (13)$$

$$\ln(\exp(w)) = \ln\left(\frac{\pi(\boldsymbol{x_i})}{1 - \pi(\boldsymbol{x_i})}\right)$$

$$\exp(w) = \frac{\pi(\boldsymbol{x_i})}{1 - \pi(\boldsymbol{x_i})}$$

$$\exp(w) = \pi(\boldsymbol{x_i}) + \exp(w)\,\pi(\boldsymbol{x_i})$$

$$\pi(\boldsymbol{x_i}) = \frac{\exp(w)}{1 + \exp(w)} \quad (14)$$

The Logistic Regression model can be written as follows (14) and based on (13) logit transformation of $\pi(\boldsymbol{x_i})$ is defined as follows:

$$\ln\left(\frac{\pi(\boldsymbol{x_i})}{1 - \pi(\boldsymbol{x_i})}\right) = w = f(x_{1i}, \dots, x_{pi})$$

$f(x_{1i}, \dots, x_{pi})$ It is modeled using a degree-m truncated spline function with specified knot point $K_{1j}, K_{2j,\dots,}K_{rj}$, where j is 1, 2, … , p, Therefore, the logit equation is derived as:

$$\ln\left(\frac{\pi(\boldsymbol{x_i})}{1-\pi(\boldsymbol{x_i})}\right) = \beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{m}\beta_{jk}x_{ji}^{k} + \sum_{j=1}^{p}\sum_{u=1}^{r}\beta_{j(m+u)}\left(x_{ji} - K_{ju}\right)_{+}^{m}$$

where, β_0, β_{jk}, and $\beta_{j(m+u)}$, $j = 1,2,\dots,p$, $k = 1,2,\dots,m$, $u = 1,2,\dots,r$ are the model parameters in the Truncated Spline function and Truncated function is defined as follows:

$$\left(x_{ji} - K_{ju}\right)_{+}^{m} = \begin{cases}\left(x_{ji} - K_{ju}\right)^{m}, x_{ji} \geq K_{ju}\\ 0 \qquad\qquad , x_{ji} < K_{ju}\end{cases}$$

The logit function can be presented in matrix form:

$$\ln\left(\frac{\pi(\boldsymbol{x_i})}{1-\pi(\boldsymbol{x_i})}\right) = \boldsymbol{X^T\beta}$$

Where

$$\boldsymbol{X^T} = \begin{bmatrix} 1 & x_{11} & \cdots & {x_{11}}^m & \cdots & {x_{p1}}^m & \cdots & \left(x_{p1} - K_{1r}\right)_{+}^{m}\\ 1 & x_{12} & \cdots & {x_{12}}^m & \cdots & {x_{p2}}^m & \cdots & \left(x_{p2} - K_{1r}\right)_{+}^{m}\\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots\\ 1 & x_{1n} & \cdots & {x_{1n}}^m & \cdots & {x_{pn}}^m & \cdots & \left(x_{pn} - K_{1r}\right)_{+}^{m}\end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix}\beta_0\\ \beta_{11}\\ \beta_{12}\\ \vdots\\ \beta_{1m}\\ \vdots\\ \beta_{p(r+m)}\end{bmatrix}$$

Theorem 1

A nonparametric regression model employing truncated splines with multiple predictors for categorical response variables on (11) in Lemma 1, gives:

$$\pi(\boldsymbol{x_i}) = \frac{\exp(\beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{m}\beta_{jk}x_{ji}^{k} + \sum_{j=1}^{p}\sum_{u=1}^{r}\beta_{j(m+u)}\left(x_{ji} - K_{ju}\right)_{+}^{m})}{1+(\beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{m}\beta_{jk}x_{ji}^{k} + \sum_{j=1}^{p}\sum_{u=1}^{r}\beta_{j(m+u)}\left(x_{ji} - K_{ju}\right)_{+}^{m})}$$

where $i = 1,2,\dots,n$

Proof of Theorem 1

Based on Lemma 1 the logit function is obtained

$$\ln\left(\frac{\pi(\boldsymbol{x_i})}{1-\pi(\boldsymbol{x_i})}\right) = \exp(\beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{m}\beta_{jk}x_{ji}^{k} + \sum_{j=1}^{p}\sum_{u=1}^{r}\beta_{j(m+u)}\left(x_{ji} - K_{ju}\right)_{+}^{m})$$

$$\ln\left(\frac{\pi(\boldsymbol{x_i})}{1-\pi(\boldsymbol{x_i})}\right) = \ln\left(\exp\left(\beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{m}\beta_{jk}x_{ji}^{k} + \sum_{j=1}^{p}\sum_{u=1}^{r}\beta_{j(m+u)}\left(x_{ji} - K_{ju}\right)_{+}^{m}\right)\right)$$

$$\frac{\pi(\boldsymbol{x_i})}{1-\pi(\boldsymbol{x_i})} = \exp\left(\beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{m}\beta_{jk}x_{ji}^{k} + \sum_{j=1}^{p}\sum_{u=1}^{r}\beta_{j(m+u)}\left(x_{ji} - K_{ju}\right)_{+}^{m}\right)$$

$$\pi(\boldsymbol{x_i}) = \frac{\exp(\beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{m}\beta_{jk}x_{ji}^{k} + \sum_{j=1}^{p}\sum_{u=1}^{r}\beta_{j(m+u)}\left(x_{ji} - K_{ju}\right)_{+}^{m})}{1+(\beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{m}\beta_{jk}x_{ji}^{k} + \sum_{j=1}^{p}\sum_{u=1}^{r}\beta_{j(m+u)}\left(x_{ji} - K_{ju}\right)_{+}^{m})}$$

B. Regression Nonparametric Regression Estimator of Multivariable Truncated Spline For Categorical Data

Parameter estimation obtained is $\boldsymbol{\beta}$, where

$$\boldsymbol{\beta} = \begin{pmatrix}\beta_0 & \beta_{11} & \dots & \beta_{1,(\mathrm{r+m})} & \vdots & \cdots & \vdots & \cdots & \beta_{\mathrm{p},(\mathrm{r+m})}\end{pmatrix}^{\boldsymbol{T}}$$

Lemma 2

The form using Maximum Likelihood Estimation (MLE) method as follows [23] :

$$\begin{aligned} l(\boldsymbol{\beta}) &= \prod_{i=1}^{n} P(Y_i = y_i)\\ &= \prod_{i=1}^{n}\pi(\boldsymbol{x_i})^{y_i}(1-\pi(\boldsymbol{x_i}))^{1-y_i}\\ &= \pi(\boldsymbol{x_i})^{\sum_{i=1}^{n}y_i}(1-\pi(\boldsymbol{x_i}))^{n-\sum_{i=1}^{n}y_i} \end{aligned} \quad (15)$$

Maximum likelihood estimation (MLE) determines parameter values by maximizing the log-likelihood function's first derivative, a process facilitated by the function's form as shown in equation (15).

$$\begin{aligned} \ln\left[l(\boldsymbol{\beta})\right] = L(\boldsymbol{\beta}) &= \ln\ \pi(\boldsymbol{x_i})^{\sum_{i=1}^{n}y_i}(1-\pi(\boldsymbol{x_i}))^{n-\sum_{i=1}^{n}y_i}\\ &= \sum_{i=1}^{n}\left\{y_i \ln\left(\frac{\pi(\boldsymbol{x_i})}{1-\pi(\boldsymbol{x_i})}\right) + \ln[1-\pi(\boldsymbol{x_i})\,]\right\}\\ &= \sum_{i=1}^{n}\left\{y_i\left(f(x_{1i},\ \dots,\ x_{pi})\right) - \ln\left[1 + \exp(f(x_{1i},\ \dots,\ x_{pi}))\right]\right\} \end{aligned} \quad (16)$$

The estimators $\widehat{\boldsymbol{\beta}}$ is obtained by deriving the (16) partially with respect to each parameter. The same rule is applied for the derivative of $L(\boldsymbol{\beta})$ so that for the first derivative of $L(\boldsymbol{\beta})$ the general equation is given as follows:

$$\begin{aligned} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_o} &= \sum_{i=1}^{n}\{y_i - \pi(\boldsymbol{x_i})\}\\ \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{jk}} &= \sum_{i=1}^{n}\left\{y_i {x_{ji}}^k - \pi(\boldsymbol{x_i}){x_{ji}}^k\right\}\\ \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{j,(m+u)}} &= \sum_{i=1}^{n}\left\{y_i\left(x_{ji} - K_{ju}\right)_{+}^{m} - \pi(\boldsymbol{x_i})\left(x_{ji} - K_{ju}\right)_{+}^{m}\right\} \end{aligned} \quad (17)$$

Proof of Lemma 2

Derivation of $L(\beta)$ with respect to β_0

$$\begin{aligned} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} &= \sum_{i=1}^{n}\left\{y_i\frac{\partial}{\partial \beta_0}f(x_1,..,x_p) - \frac{\partial}{\partial \beta_0}\ln\left[1+\exp f(x_1,..,x_p)\right]\right\}\\ &= \sum_{i=1}^{n}\left\{y_i(1) - \frac{\exp f(x_1,..,x_p)}{1+\exp f(x_1,..,x_p)}(1)\right\} = \sum_{i=1}^{n}\{y_i - \pi(\boldsymbol{x_i})\} \end{aligned}$$

Derivation of $L(\boldsymbol{\beta})$ with respect to β_{11}

$$\begin{aligned} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{11}} &= \sum_{i=1}^{n}\left\{y_i\frac{\partial}{\partial \beta_{11}}f(x_1,..,x_p) - \frac{\partial}{\partial \beta_{11}}\ln\left[1+\exp f(x_1,..,x_p)\right]\right\}\\ &= \sum_{i=1}^{n}\left\{y_i x_{1i} - \frac{\exp f(x_1,..,x_p)}{1+\exp f(x_1,..,x_p)}x_{1i}\right\}\\ &= \sum_{i=1}^{n}\{y_i x_{1i} - \pi(\boldsymbol{x_i})x_{1i}\}\\ &\vdots \end{aligned}$$

Derivation of $L(\boldsymbol{\beta})$ with respect to $\beta_{1,(m+r)}$

$$\begin{aligned} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{1,(m+r)}} &= \sum_{i=1}^{n}\left\{y_i\frac{\partial}{\partial \beta_{1,(m+r)}}f(x_1,..,x_p) - \frac{\partial}{\partial \beta_{1,(m+r)}}\ln\left[1+\exp f(x_1,..,x_p)\right]\right\}\\ &= \sum_{i=1}^{n}\left\{y_i\left(x_{1i} - K_{1r}\right)_{+}^{m} - \frac{\exp f(x_1,..,x_p)}{1+\exp f(x_1,..,x_p)}\left(x_{1i} - K_{1r}\right)_{+}^{m}\right\}\\ &= \sum_{i=1}^{n}\{y_i\left(x_{1i} - K_{1r}\right)_{+}^{m} - \pi(\boldsymbol{x_i})\left(x_{1i} - K_{1r}\right)_{+}^{m}\}\\ &\vdots \end{aligned}$$

Derivation of $L(\boldsymbol{\beta})$ with respect to $\beta_{p,(m+r)}$

$$\begin{aligned} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{p,(m+r)}} &= \sum_{i=1}^{n}\left\{y_i\frac{\partial}{\partial \beta_{p,(m+r)}}f(x_1,..,x_p) - \frac{\partial}{\partial \beta_{p,(m+r)}}\ln\left[1+\exp f(x_1,..,x_p)\right]\right\}\\ &= \sum_{i=1}^{n}\left\{y_i\left(x_{pi} - K_{pr}\right)_{+}^{m} - \frac{\exp f(x_1,..,x_p)}{1+\exp f(x_1,..,x_p)}\left(x_{pi} - K_{pr}\right)_{+}^{m}\right\}\\ &= \sum_{i=1}^{n}\left\{y_i\left(x_{pi} - K_{pr}\right)_{+}^{m} - \pi(\boldsymbol{x_i})\left(x_{pi} - K_{pr}\right)_{+}^{m}\right\} \end{aligned}$$

The estimator $\widehat{\boldsymbol{\beta}}$ determined when the derivative equals to 0. As the derived equation cannot be solved directly, numerical iteration must be used.

Lemma 3

Numerical iteration was achieved by implementing the Newton-Raphson method, using this equation [24]:

$$\boldsymbol{\beta}_{(t+1)} = \boldsymbol{\beta}_{(t)} - \left(\boldsymbol{H}_{(t)}\right)^{-1}\boldsymbol{g}_{(t)} \quad (18)$$

Where $\boldsymbol{\beta}_{(t+1)}$ and $\boldsymbol{\beta}_{(t)}$ are t-th iteration of parameter value, $t = 1,2,\ldots,$converged. $\boldsymbol{g}_{(t)}$ is vector with first derivative of ln likelihood function and $\boldsymbol{H}_{(t)}$ is the hessian matrix, calculated using the second derivatives of the log-likelihood function (shown below), is:

$$\boldsymbol{g}_{(t)} = \left(\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0}, \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{11}}, \cdots, \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{1,(r+m)}}, \cdots, \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{p1}}, , \cdots, \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{p,(m+r)}}\right)$$

$$\boldsymbol{H}_{(t)} = \begin{bmatrix} \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_0{}^2} & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_0 \partial \beta_{11}} & \cdots & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_0 \partial \beta_{p,(m+r)}} \\ \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_{11} \partial \beta_0} & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_{11}{}^2} & \cdots & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_{11} \partial \beta_{p,(m+r)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_{p,(m+r)} \partial \beta_0} & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_{p,(m+r)} \partial \beta_{11}} & \cdots & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_{p,(m+r)}{}^2} \end{bmatrix}$$

$\widehat{\boldsymbol{\beta}}$ will obtained when [24]

$$\left|\boldsymbol{\beta}_{(t+1)} - \boldsymbol{\beta}_{(t)}\right| < \varepsilon, \varepsilon = 0.000001$$

Proof of Lemma 3

Based on (16) will be performed the second derivative of $L(\boldsymbol{\beta})$ for $\boldsymbol{H}_{(t)}$

Derivation of $L(\boldsymbol{\beta})$ with respect to β_0 and β_0

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_o{}^2} = \sum_{i=1}^{n}\left\{\frac{\partial}{\partial \beta_0} y_i - \frac{\partial}{\partial \beta_0}\frac{\exp f(x_{1,..,}x_p)}{1+\exp f(x_{1,..,}x_p)}\right\}$$
$$= \sum_{i=1}^{n}\left\{0 - \frac{[\exp f(x_{1,..,}x_p)][1+\exp f(x_{1,..,}x_p)]-[\exp f(x_{1,..,}x_p)].[\exp f(x_{1,..,}x_p)]}{[1+\exp f(x_{1,..,}x_p)]^2}\right\}$$
$$= \sum_{i=1}^{n}\left\{0 - \frac{[\exp f(x_{1,..,}x_p)]}{[1+\exp f(x_{1,..,}x_p)]}\frac{1}{[1+\exp f(x_{1,..,}x_p)]}\right\}$$
$$= -\sum_{i=1}^{n}\pi(x_i)\,(1-\pi(x_i))$$
$$\vdots$$

Derivation of $L(\boldsymbol{\beta})$ with respect to β_{11} and β_{1m}

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{1m}\partial \beta_{11}} = \sum_{i=1}^{n}\left\{\frac{\partial}{\partial \beta_{1m}} y_i x_{1i} - \frac{\partial}{\partial \beta_{1m}}\frac{\exp f(x_{1,..,}x_p)}{1+\exp f(x_{1,..,}x_p)} x_{1i}\right\}$$
$$= \sum_{i=1}^{n}\left\{0 - \frac{[\exp f(x_{1,..,}x_p)][1+\exp f(x_{1,..,}x_p)]-[\exp f(x_{1,..,}x_p)].[\exp f(x_{1,..,}x_p)]}{[1+\exp f(x_{1,..,}x_p)]^2} {x_{1i}}^m x_{1i}\right\}$$
$$= \sum_{i=1}^{n}\left\{0 - \frac{[\exp f(x_{1,..,}x_p)]}{[1+\exp f(x_{1,..,}x_p)]}\frac{1}{[1+\exp f(x_{1,..,}x_p)]} {x_{1i}}^m x_{1i}\right\}$$
$$= -\sum_{i=1}^{n}\pi(x_i)\left(1-\pi(x_i)\right){x_{1i}}^m x_{1i}$$
$$\vdots$$

Derivation of $L(\boldsymbol{\beta})$ with respect to $\beta_{p,(m+r)}$ and $\beta_{q,(m+r)}$

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{q,(m+r)}\partial \beta_{p,(m+r)}} = \sum_{i=1}^{n}\left\{\frac{\partial}{\partial \beta_{q,(m+r)}} y_i \left(x_{pi} - K_{pr}\right)_+^m - \frac{\partial}{\partial \beta_{q,(m+r)}}\frac{\exp f(x_{1,..,}x_p)}{1+\exp f(x_{1,..,}x_p)}\left(x_{pi} - K_{pr}\right)_+^m\right\}$$
$$= \sum_{i=1}^{n}\left\{0 - \frac{[\exp f(x_{1,..,}x_p)][1+\exp f(x_{1,..,}x_p)]-[\exp f(x_{1,..,}x_p)].[\exp f(x_{1,..,}x_p)]}{[1+\exp f(x_{1,..,}x_p)]^2}\left(x_{qi} - K_{qr}\right)_+^m\left(x_{pi} - K_{pr}\right)_+^m\right\}$$
$$= \sum_{i=1}^{n}\left\{0 - \frac{[\exp f(x_{1,..,}x_p)]}{[1+\exp f(x_{1,..,}x_p)]}\frac{1}{[1+\exp f(x_{1,..,}x_p)]}\left(x_{qi} - K_{qr}\right)_+^m\left(x_{pi} - K_{pr}\right)_+^m\right\}$$
$$= -\sum_{i=1}^{n}\pi(x_i)\,(1-\pi(x_i))\left(x_{qi} - K_{qr}\right)_+^m\left(x_{pi} - K_{pr}\right)_+^m$$

Thus, the estimator $\widehat{\boldsymbol{\beta}}$ obtained is

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix}\hat{\beta}_0 & \hat{\beta}_{11} & \ldots & \hat{\beta}_{1,(r+m)} & \vdots & \cdots & \vdots & \cdots & \hat{\beta}_{p,(r+m)}\end{pmatrix}^{\boldsymbol{T}}$$

IV. DATA APPLICATION

This study employed the multivariable truncated spline nonparametric regression method on two datasets: Indonesia's 2023 poverty depth index and East Java's 2020 gender development index, to analyze categorical data.

A. Case 1: Data of Poverty Depth Index in Indonesia for 2023

This study employs secondary data on Indonesia's poverty depth index, comprising one response variable and three predictor variables hypothesized to influence it. All data were obtained from publications of Indonesia's Central Bureau of Statistics (BPS). The detail of the variables are described in Table I.

TABLE I
VARIABLE DESCRIPTION

Variable	Notation	Description
Response	y	1 = High Poverty Depth Index 0 = Low Poverty Depth Index
Predictor (X_j)	X_1	Average Years of Schooling
	X_2	Open Unemployment Rate
	X_3	Labor Force Participation Rate

Fig. 1 provides detailed information about response categorized. The data is categorized based on certain criteria, namely the average poverty depth index value in Indonesia for 2023, which is 1.53. Provinces that have a poverty index value above the average are classified as having a high poverty depth index, while provinces that have a poverty index value below the average are classified as having a low poverty depth index. This categorization aims to highlight regional disparities in poverty levels and provide a basis for further analysis.

Based on the data, there are 17 provinces with a high poverty depth index that will be categorized as 1, indicating a greater severity of poverty in these regions. Conversely, there are 17 provinces with a low poverty depth index that will be categorized as 0, suggesting relatively better socioeconomic conditions.

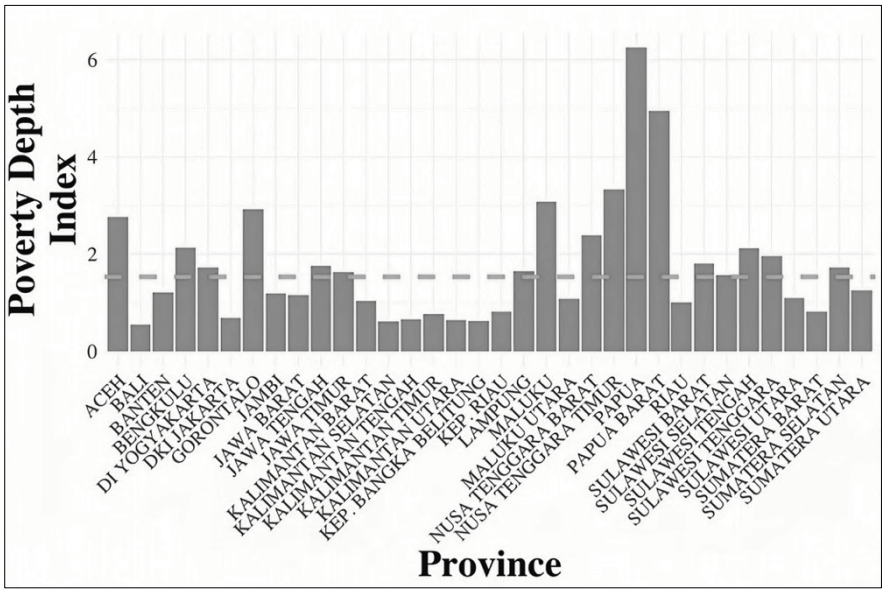

Fig. 1: Categories of Poverty Depth Index in Indonesia

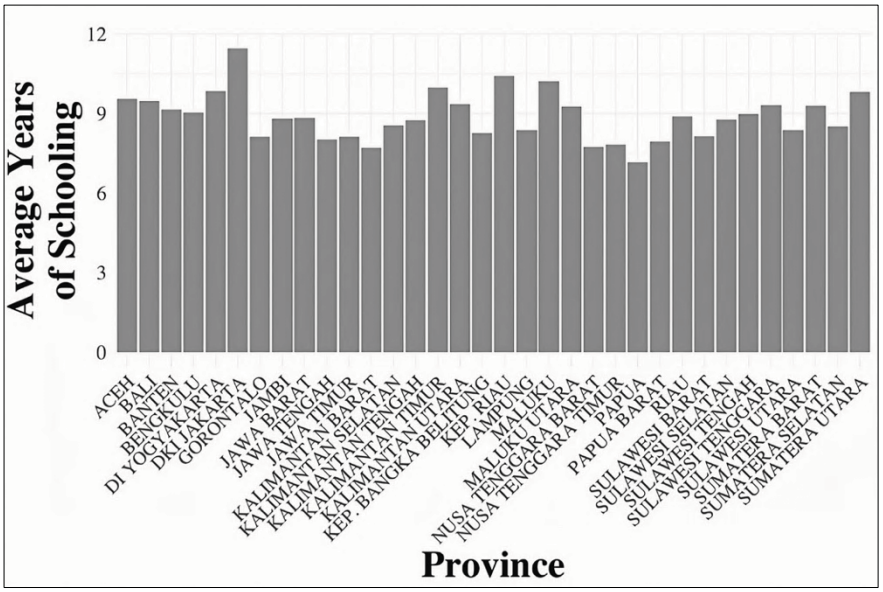

Fig. 2: Characteristics of X_1

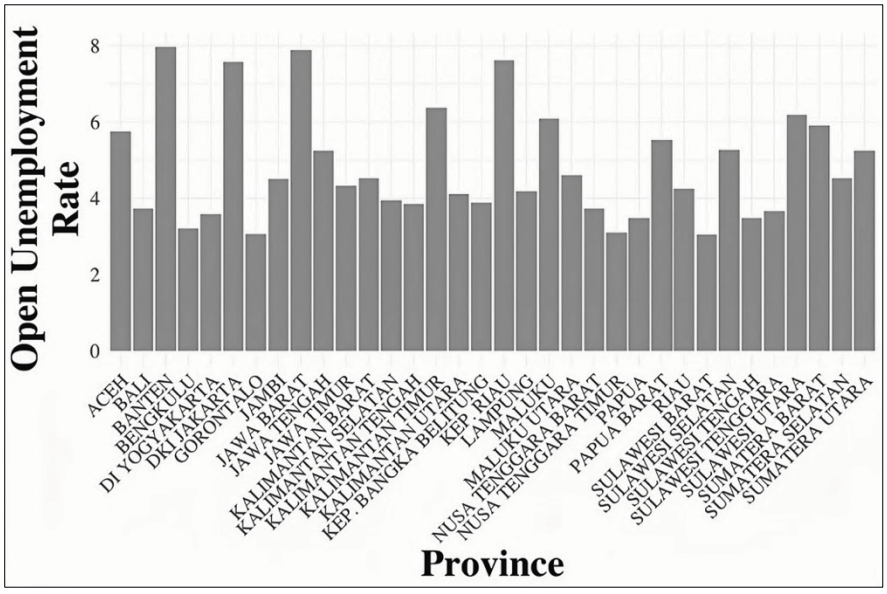

Fig. 3: Characteristics of X_2

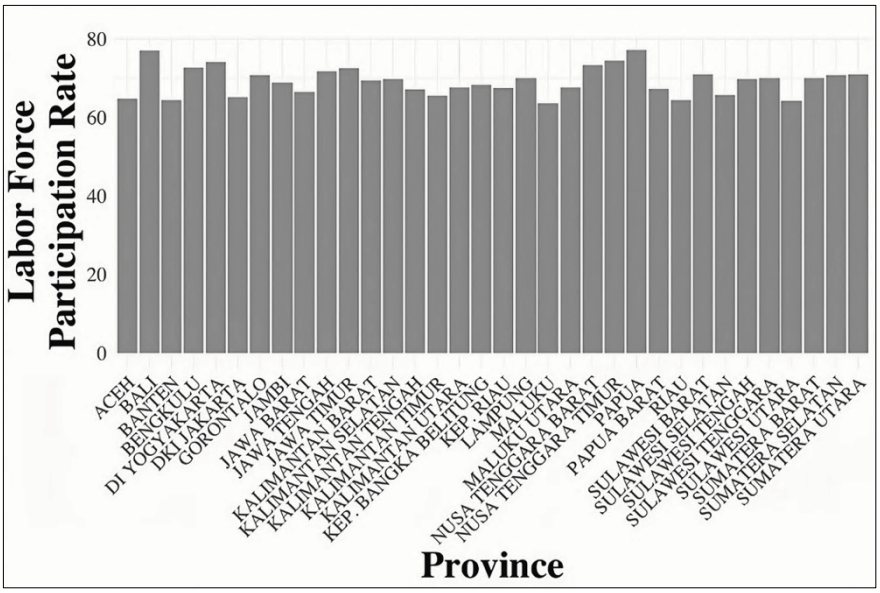

Fig. 4: Characteristics of X_3

In Fig. 2, it can be seen that DKI Jakarta is the province with the highest average years of schooling, and Papua is the province with the lowest average years of schooling. Fig. 3 illustrates that Banten Province reports the highest open unemployment rate, conversely, West Sulawesi Province reports the lowest. Then, Fig. 4 illustrates the disparity in labor force participation rates across Indonesian provinces, with Bali exhibiting the highest rate and Maluku the lowest.

Table II presents the descriptive statistics characterizing each variable.

TABLE II
DESCIPTIVE STATISTICS OF RESEARCH VARIABLES

X_j	Category Response	Mean	Median	Variance	Minimum	Maximum
X_1	1	8.558	8.360	8.360	8.360	10.20
	0	9.194	9.150	9.150	9.150	11.450
X_2	1	4.192	3.730	3.730	3.730	6.080
	0	5.419	4.60	4.60	4.60	7.970
X_3	1	70.58	70.79	70.79	70.79	77.20
	0	67.93	67.62	67.62	67.62	77.08

Table II details the variable characteristics of the variables. In addition, it is obtained that there is no high correlation among predictor variables.
Each variable is shown through the plot in below:

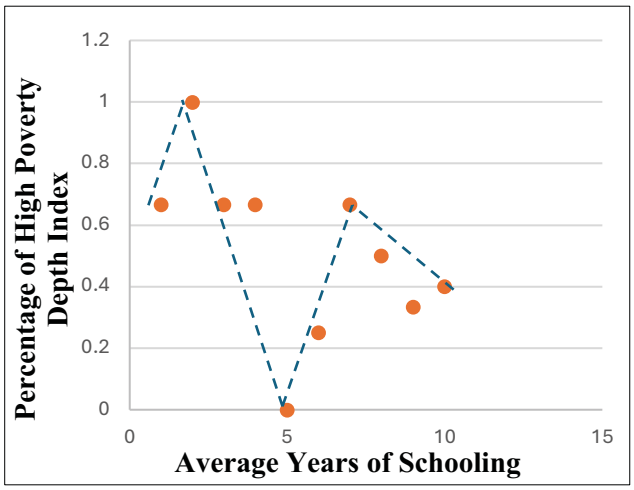

Fig. 5: Scatterplots of X_1

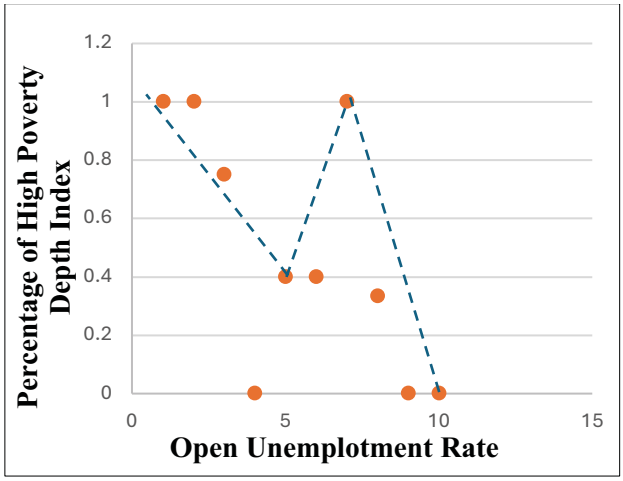

Fig. 6: Scatterplots of X_2

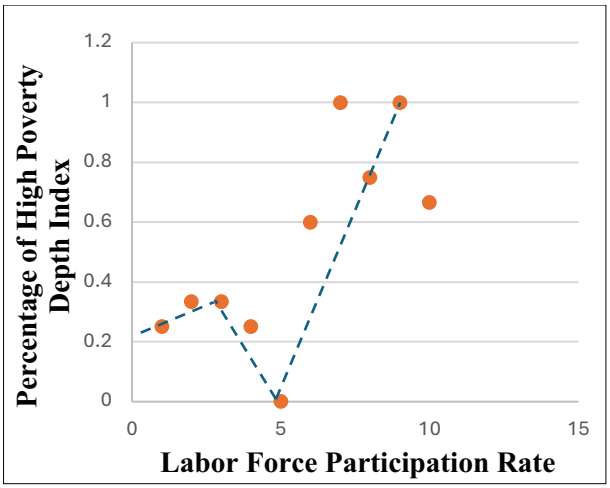

Fig. 7: Scatterplots of X_3

A scatterplot is constructed to know the relationship between variables by making them into groups of intervals [23]. Based on Fig. 5, the average years of schooling in group 2 data has an upward pattern tendency. Next, the patterns of the data in group 5 tend to decrease. Meanwhile, the patterns of the data in group 7 and above tend to increase. Based on Fig. 6, open unemployment rate in the data in group 5 has a downward pattern tendency. Then, the unemployment rate in

the data in group 5-7, tend to increase. Meanwhile, in group 10 the patterns tend to decrease again. Figure 7 indicates a rising trend in labor force participation for group 3, while overall patterns show a subsequent decline. Next, group 5 tends to decrease. Lastly, in group 9, the pattern tends to increase. The scatterplot suggests a pattern changes at certain sub-intervals, indicating the suitability of a truncated linear spline model. Optimal model selection requires, the researchers use a combination of the number of knot points limited to 3.

Equation (5) presents the following truncated spline nonparametric linear regression model for Indonesia's 2023 poverty depth index data.

$$\pi(\boldsymbol{x_i}) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{3}\beta_{j1}x_{ji} + \sum_{j=1}^{3}\sum_{u=1}^{r}\beta_{j(1+u)}\left(x_{ji} - K_{ju}\right)_+\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{3}\beta_{j1}x_{ji} + \sum_{j=1}^{3}\sum_{u=1}^{r}\beta_{j(1+u)}\left(x_{ji} - K_{ju}\right)_+\right)}$$

Table III presents the knot point locations and AIC values derived from the model. The optimal knot points identified are 10.017 for X_1, 3.862, 4.683, and 5.505 for X_2, and 68.133 and 70.400 for X_3, yielding a minimum AIC value of .20.

TABLE III
AIC VALUE BASED ON KNOT POINT CANDIDATE

Number of Knot Points	K_{ju}		Knot Point Value	AIC (K)
1,1,1	K_{1u}	K_{11}	9.300	
		K_{12}		
		K_{13}		
	K_{2u}	K_{21}	3.862	
		K_{22}		42.345
		K_{23}		
	K_{3u}	K_{31}	74.933	
		K_{32}		
		K_{33}		
			⋮	
3,2,2	K_{1u}	K_{11}	7.867	
		K_{12}	10.017	
		K_{13}	10.733	
	K_{2u}	K_{21}	3.862	
		K_{22}	5.505	22
		K_{23}		
	K_{3u}	K_{31}	68.133	
		K_{32}	70.400	
		K_{33}		
1,3,2	$\boldsymbol{K_{1u}}$	$\boldsymbol{K_{11}}$	**10.017**	
		$\boldsymbol{K_{12}}$		
		$\boldsymbol{K_{13}}$		
	$\boldsymbol{K_{2u}}$	$\boldsymbol{K_{21}}$	**3.862**	
		$\boldsymbol{K_{22}}$	**4.683**	**20***
		$\boldsymbol{K_{23}}$	**5.505**	
	$\boldsymbol{K_{3u}}$	$\boldsymbol{K_{31}}$	**68.133**	
		$\boldsymbol{K_{32}}$	**70.400**	
		$\boldsymbol{K_{33}}$		
			⋮	
3,3,3	K_{1u}	K_{11}	7.867	
		K_{12}	10.017	
		K_{13}	10.733	
	K_{2u}	K_{21}	3.862	
		K_{22}	6.327	26
		K_{23}	7.148	
	K_{3u}	K_{31}	65.867	
		K_{32}	68.133	
		K_{33}	70.400	

The following presents a nonparametric linear regression model using a truncated spline with optimally determined knots.

$$\pi(\boldsymbol{x_i}) = \frac{\exp(w)}{1 + \exp(w)}$$

Where:

$$w = \beta_0 + \sum_{j=1}^{3}\beta_{j1}x_{ji} + \beta_{12}(x_{1i} - 10.017)_+ + \beta_{22}(x_{2i} - 3.862)_+ + \beta_{23}(x_{2i} - 4.683)_+ + \beta_{24}(x_{2i} - 5.505)_+ + \beta_{32}(x_{3i} - 68.133)_+ + \beta_{33}(x_{3i} - 70.400)_+$$

Table IV presents the estimated model parameters for Indonesia's 2023 poverty depth index, obtained using truncated spline nonparametric linear regression.

TABLE IV
PARAMETER ESTIMATION RESULTS

Parameters	Estimations	Parameters	Estimations
β_0	6.525	β_{23}	8.889
β_{11}	-0.635	β_{24}	-15.276
β_{12}	9.569	β_{31}	0.157
β_{21}	-3.254	β_{32}	0.544
β_{22}	1.372	β_{33}	-0.787

Table IV presents the estimated model parameters for Indonesia's 2023 poverty depth index, obtained using truncated spline nonparametric linear regression.

$$\pi(\boldsymbol{x_i}) = \frac{\exp(\mathrm{w})}{1 + \exp(\mathrm{w})}$$

Where:

$$w = 6.525 - 0.635x_{1i} + 9.569(x_{1i} - 10.017)_+ - 3.254x_{2i} + 1.372\,(x_{2i} - 3.862)_+ + 8.889(x_{2i} - 4.683) - 15.276(x_{2i} - 5.505)_+ + 0.157x_{3i} + 0.544(x_{3i} - 68.133)_+ - 0.787(x_{3i} - 70.400)_+$$

The researchers subsequently employed two models to analyze the data: a truncated spline nonparametric regression model for categorical variables and a binary logistic regression model for the poverty depth index.

$$\pi(\boldsymbol{x_i}) = \frac{\exp\left(1.106 - 0.563x_{1i} - 0.455x_{2i} + 0.087x_{3i}\right)}{1 + \exp\left(1.106 - 0.563x_{1i} - 0.455x_{2i} + 0.087x_{3i}\right)}$$

B. Case 2: Data of Gender Development Index in East Java for 2020

This research utilizes secondary gender development index data to analyze the relationship between a single response variable and two hypothesized predictor variables, detailed in Table V.

TABLE V
VARIABLE DESCRIPTION

Variable	Notation	Description
Response	y	1 = High Gender Development Index 0 = Low Gender Development Index
Predictor (X_j)	X_1	Percentage of Population Not Consuming Tobacco (> 5 years)
	X_2	Net Enrollment Rate

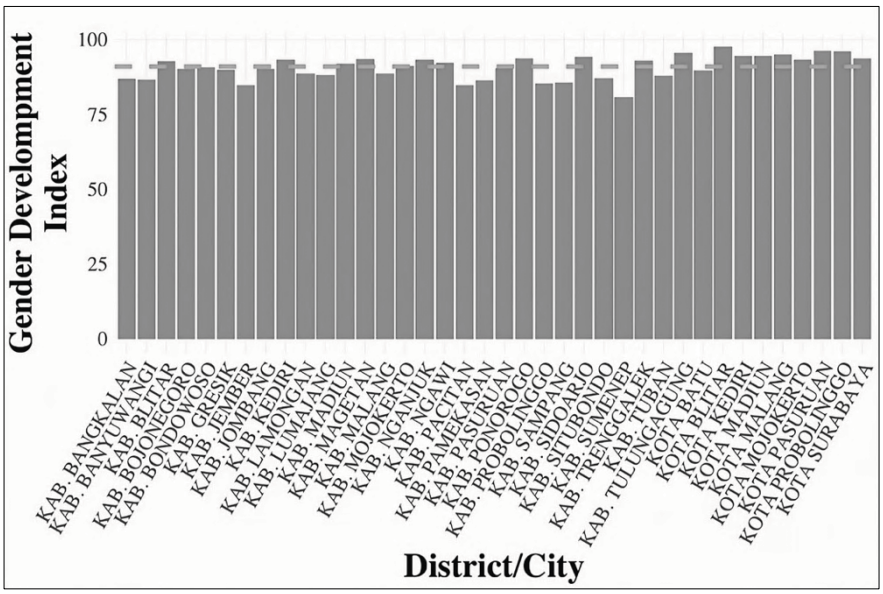

Fig. 8: Categories of Gender Development Index

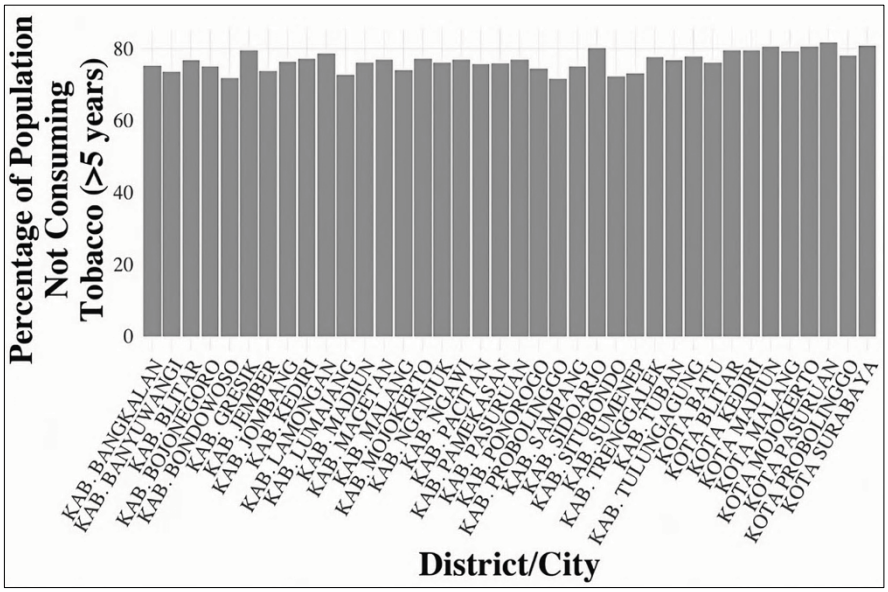

Fig. 9: Characteristics of X_1

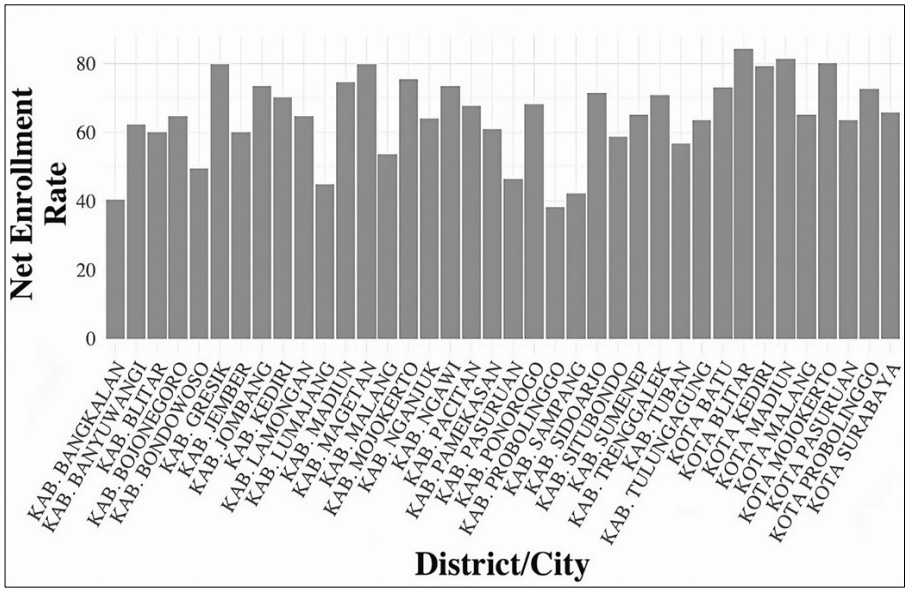

Fig. 10: Characteristics of X_2

Fig. 8, this study presents categorized response data. Categorization is based on East Java's average Gender Development Index (GDI) of 90.91, resulting in two groups: 19 districts/cities with a high poverty depth index (categorized as 1) and 19 with a low poverty depth index (categorized as 0). Descriptive statistics characterize the variables, the results of which are presented below.

TABLE VI
DESCIPTIVE STATISTICS OF RESEARCH VARIABLES

X_j	Category Response	Mean	Median	Variance	Minimum	Maximum
X_1	1	74.88	75.01	4.97	71.50	79.42
	0	78.18	77.76	3.86	74.39	81.53
X_2	1	57.98	60.01	135.37	38.31	79.71
	0	71.75	71.43	49.56	59.90	84.18

Table VI presents the variable characteristics. In addition, it is obtained that there is no multicollinearity between the predictor variables. In Fig. 9, Pasuruan City exhibits the highest percentage of residents who have not consumed tobacco for over five years, while Probolinggo Regency shows the lowest. Fig. 10 illustrates that Kota Blitar exhibits the highest net enrollment rate, while Kabupaten Probolinggo shows the lowest.

The scatterplots of the data for each variable is shown through the plot in below:

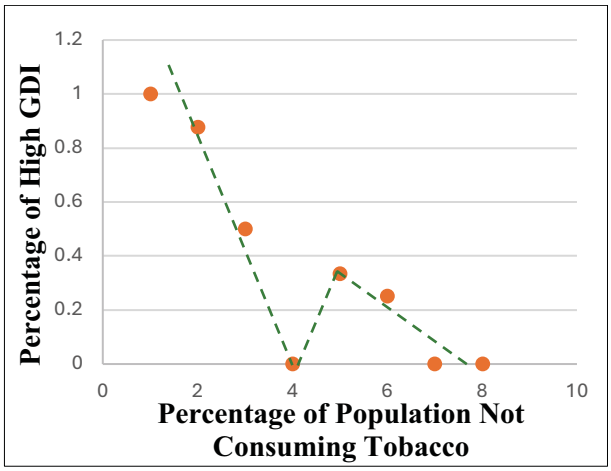

Fig. 11: Scatterplots of X_1

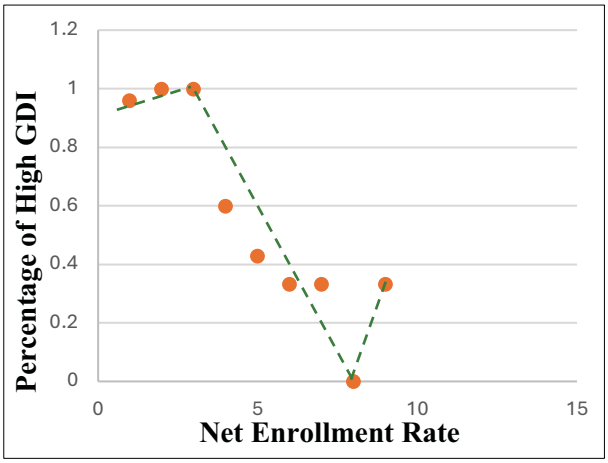

Fig. 12: Scatterplots of X_2

Based on Fig. 11, the percentage of population not consuming tobacco (> 5 years) in group 4 data has an downward pattern tendency. Next, the patterns of the data in group 5 tend to increase. Meanwhile, the patterns of the data in group 6 and above tend to decrease. Based on Fig. 12, net enrollment rate in the data in group 3 has a upward pattern tendency. Then, the net enrollment rate in the data in group 8, tend to decrease. Meanwhile, in group 9, the patterns are tend to increase again. The scatterplot suggests data pattern changes at certain sub-intervals, making a linear truncated spline an appropriate modeling approach. Optimal model selection requires, the researchers use a maximum of three knot points were used.

The Truncated Spline nonparametric linear regression model for poverty depth index data in 2023 in shown as follow:

$$\pi(\boldsymbol{x_i}) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{2}\beta_{j1}x_{ji} + \sum_{j=1}^{2}\sum_{u=1}^{r}\beta_{j(1+u)}(x_{ji} - K_{ju})_+\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{2}\beta_{j1}x_{ji} + \sum_{j=1}^{2}\sum_{u=1}^{r}\beta_{j(1+u)}(x_{ji} - K_{ju})_+\right)}$$

Optimal knot point candidates are presented in the following table:

TABLE VII
AIC VALUE BASED ON KNOT POINT CANDIDATE

Number of Knot Points	K_{ju}		Knot Point Value	AIC (K)
1,1	K_{1u}	K_{11}	74.008	
		K_{12}		37.407
		K_{13}		
	K_{2u}	K_{21}	61.245	
		K_{22}		
		K_{23}		
			⋮	
3,1	$\boldsymbol{K_{1u}}$	$\boldsymbol{K_{11}}$	**74.008**	
		$\boldsymbol{K_{12}}$	**75.261**	**34.766***
		$\boldsymbol{K_{13}}$	**76.515**	
	$\boldsymbol{K_{2u}}$	$\boldsymbol{K_{21}}$	**61.245**	
		K_{22}		
		K_{23}		
3,2	K_{1u}	K_{11}	74.008	
		K_{12}	75.261	36.442
		K_{13}	76.515	
	K_{2u}	K_{21}	61.245	
		K_{22}	78.446	
		K_{23}		
3,3	K_{1u}	K_{11}	74.008	
		K_{12}	75.261	26
		K_{13}	76.515	
	K_{2u}	K_{21}	61.245	
		K_{22}	72.712	
		K_{23}	78.446	

Table VII shows that the optimal knot points for the model minimizing the AIC (34.766) are 74.008, 75.261, and 76.515 for X_1 and 61.245 for X_2. The resulting truncated spline nonparametric linear regression model is:

$$\pi(\boldsymbol{x_i}) = \frac{\exp(w)}{1 + \exp(w)}$$

Where:

$$w = \beta_0 + \sum_{j=1}^{3}\beta_{j1}x_{ji} + \beta_{12}(x_{1i} - 74.008)_+ + \beta_{13}(x_{1i} - 75.261)_+ + \beta_{14}(x_{1i} - 76.515)_+ + \beta_{22}(x_{2i} - 61.245)_+$$

Table VIII presents the estimated model parameters for the 2020 Gender Development Index in East Java, derived using truncated spline nonparametric linear regression.

TABLE VIII
PARAMETER ESTIMATION RESULTS

Parameters	Estimations	Parameters	Estimations
β_0	2.655	β_{14}	3.803
β_{11}	0.704	β_{21}	-0.867
β_{12}	0.249	β_{22}	0.862
β_{13}	-4.759		

Table VIII's estimation results yield the following truncated spline nonparametric linear regression model:

$$\pi(\boldsymbol{x_i}) = \frac{\exp(\mathrm{w})}{1+\exp(\mathrm{w})}$$

Where:

$$w = 2.655 + 0.704x_{1i} + 0.249(x_{1i} - 74.008)_+ - 4.759(x_{2i} - 75.261)_+ + 3.803(x_{1i} - 76.515)_+ - 0.867x_{2i} + 0.862(x_{2i} - 61.245)_+$$

Subsequently, the researchers evaluate This study employed two models: a truncated spline nonparametric regression model for categorical data and a binary logistic regression model to analyze the gender development index:

$$\pi(\boldsymbol{x_i}) = \frac{\exp(50.317 - 0.566x_{1i} - 0.106x_{2i})}{1+\exp(50.317 - 0.566x_{1i} - 0.106x_{2i})}$$

C. Comparison of Truncated Spline Nonparametric Regression for Categorical Data and Binary Logistic Regression

Using the deviance statistical test, the regression model with the lowest deviance was selected. The test results are as follows:

TABLE IX
COMPARISON OF DEVIANCE VALUES

Case	Methods	Deviance Value
Case 1	Truncated Spline Nonparametric Regression	26.767
	Binary Logistic Regression	38.631
Case 2	Truncated Spline Nonparametric Regression	26.608
	Binary Logistic Regression	30.097

Table IX shows that the nonparametric regression model's deviance (26.767) for the poverty depth index data is lower than the parametric model's deviance (38.631). For data of gender development indeks using the nonparametric regression model (26.608) is smaller than the deviance value for the parametric regression model (30.097). So that the Truncated Spline Nonparametric Regression model is a better model for both cases.

In determining the best model, evaluation criteria also be used. The calculation of each criterion can be obtained from confusion matrix, the following results were obtained X. Table X presents the accuracy of Truncated Spline and binary logistic regression models in predicting poverty depth and gender development indices. For case 1, the Truncated Spline model correctly classified 13 provinces as low and 15 as high poverty depth, while the logistic regression model correctly classified 11 and 13 provinces, respectively. Similarly, for case 2, the Truncated Spline model correctly classified 16 districts/cities as low and 15 as high gender development, compared to 11 and 13 districts/cities, respectively, for the logistic regression model.

TABLE X
CONFUSION MATRIX

Case 1			
Truncated Spline nonparametric for Categorical Data		Prediction	
		0	1
Actual	0	13	4
	1	2	15
Binary Logistic Regression		Prediction	
		0	1
Actual	0	11	6
	1	4	13
Case 2			
Truncated Spline nonparametric for Categorical Data		Prediction	
		0	1
Actual	0	16	3
	1	4	15
Binary Logistic Regression		Prediction	
		0	1
Actual	0	15	4
	1	5	14

Based on the data presented in the table X, it can be concluded that the Truncated Spline nonparametric regression model for categorical data is more effective in classifying than the binary logistic regression model. This can be seen from the number of provinces and districts/cities that are correctly classified by the Truncated Spline model which is more than the logistic regression model. Based on Table X show that the Truncated Spline nonparametric regression model outperforms the binary logistic regression model in classification, with more correct predictions. A summary of accuracy, recall, specificity, and precision for both models is presented in Table XI.

TABLE XI
COMPARISON OF EVALUATION CRITERIA VALUE

Case 1		
Evaluation Criteria	Methods	
	Spline Truncated Nonparametric Regression For Categorical Data	Binary Logistic Regression
Accuracy	82.35%	70.59%
sensitivity	88.24%	76.47%
Specifitiy	76.47%	76.47%
Precision	78.95%	64.71%
Case 2		
Evaluation Criteria	Methods	
	Spline Truncated Nonparametric Regression For Categorical Data	Binary Logistic Regression
Accuracy	81.58%	76.32%
sensitivity	83.33%	77.78%
Specifitiy	80%	75%
Precision	78.95%	73.68%

Based on Table XI, it can be seen that the Truncated Spline nonparametric regression model for categorical data has a higher accuracy value for case 1 and case 2 (82.35% and 81.58%) compared to binary logistic regression (70.59% and 76.32%). This indicates that nonparametric model is more demonstrates greater overall reliability in producing accurate predictions. The truncated spline nonparametric regression model demonstrated superior sensitivity (88.24% and 83.33% for cases 1 and 2, respectively) compared to logistic regression (76.47% and 77.78%), indicating greater effectiveness in identifying high poverty depth and gender development indices. This superiority extended to specificity (76.47% and 80% vs. 76.47% and 75% for logistic regression), minimizing false negatives. Furthermore, the nonparametric model achieved higher precision (78.95% vs. 64.71% and 73.68% for logistic regression).

In addition, test the stability of the accuracy of the model classification by calculating Press'Q, with the hypothesis as follows.

H_0: Model classification results are inconsistent

H_1: Model classification results are consistent

The Press'Q values obtained based on table XII in both methods are given as follows

TABLE XII
COMPARISON OF PRESS'Q VALUES

Case	Methods	Press'Q
Case 1	Truncated Spline Nonparametric Regression	16.941
	Binary Logistic Regression	7.529
Case 2	Truncated Spline Nonparametric Regression	15.16
	Binary Logistic Regression	10.53

Based on Table XII the nonparametric Truncated Spline regression for categorical data has a larger Press'Q values for case 1 and case 2 (16.941 and 15.16) than the binary logistic regression (7.529 and 10.53), indicating that the nonparametric regression model has a greater chance of rejecting H_0 or Press's Q > Chi Square. This indicates that the nonparametric regression model provides more consistent and accurate classification results compared to binary logistic regression, strengthening the argument that this model is more appropriate for categorical data with more complex patterns.

The comparison through the plot for case 1: data of poverty depth index can be seen in Fig. 13 as follows:

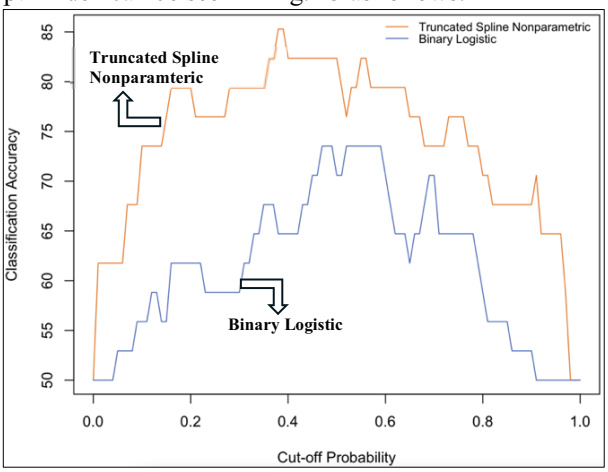

Fig. 13: Plot of Method Comparison against Classification Accuracy (Case 1)

Fig. 13 shows evidence that the Nonparametric regression modeling of categorical data using truncated splines is able to provide higher and more consistent classification accuracy compared to binary logistic regression at most cut-off probabilities.

The comparison through the plot for case 2: data of gender development index can be seen in Fig. 14 as follows:

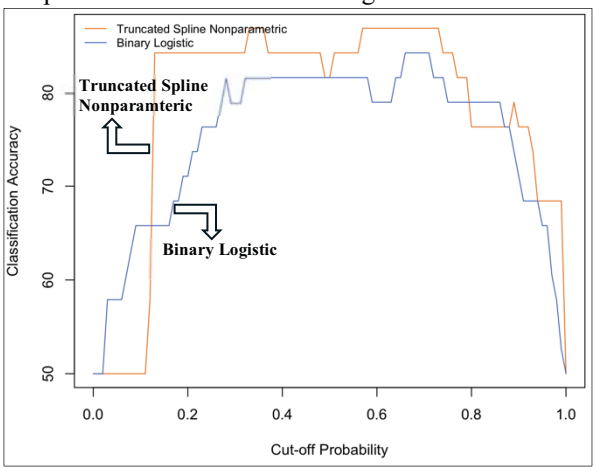

Fig. 14. Plot of Method Comparison against Classification Accuracy (Case 2)

Fig. 14 shows that although not entirely, the probability of classification accuracy in the Nonparametric truncated spline regression models for categorical data generally exhibit higher performance compared to binary logistic regression, which proves that Truncated spline nonparametric regression models can provide better classification accuracy. Therefore, nonparametric models can be a more appropriate choice in handling categorical (binary) response data, especially when the pattern of relationships between variables changes at certain sub-intervals.

V. CONCLUSION

The following presents the Truncated Spline nonparametric regression model for categorical data, as derived from the preceding discussion:

$$\pi(x_i) = \frac{\exp(\beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{m}\beta_{jk}x_{ji}^{k} + \sum_{j=1}^{p}\sum_{u=1}^{r}\beta_{j(m+u)}(x_{ji} - K_{ju})_{+}^{m})}{1 + (\beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{m}\beta_{jk}x_{ji}^{k} + \sum_{j=1}^{p}\sum_{u=1}^{r}\beta_{j(m+u)}(x_{ji} - K_{ju})_{+}^{m})}$$

With $i = 1,2,\ldots,n$ and truncated function is

$$(x_{ji} - K_{ju})_{+}^{m} = \begin{cases} (x_{ji} - K_{ju})^{m}, & x_{ji} \geq K_{ju} \\ 0, & x_{ji} < K_{ju} \end{cases}$$

Parameter estimation is constructed using the Bernoulli distribution framework, employing the Maximum Likelihood Estimation (MLE) technique, and then continued with Newton Raphson iterations. This model was applied to comparable real-world datasets, including Indonesia's 2023 Poverty Depth Index and East Java's 2020 Gender Development Index.

The study shows The truncated spline nonparametric regression model yielded superior performance to the binary logistic regression model, as indicated by lower deviance and improved evaluation metrics. A significant advantage is its capacity to model non-linear relationships without assuming a specific functional form, which results in more accurate estimates. Therefore, A truncated spline nonparametric model may be preferable for categorical data in handling

categorical (binary) response data, especially when the relationship pattern between variables changes at certain sub-intervals.

REFERENCES

[1] D. Montgomery, E. A. Peck, and G. G. Vining, Introduction to Linear Regression Analysis. United States: Wiley, 2012.

[2] G. Wahba, Spline Models for Observational Data. Madison, WI: Society for Industrial and Applied Mathematics, 1990.

[3] L. R. Eubank, Nonparametric Regression and Spline Smoothing. New York, NY: Marcel Dekker, 1999.

[4] P. Craven and G. Wahba, "Smoothing Noise Data with Spline Functions," Numer. Math., vol. 31, pp. 377–403, 1979.

[5] I. N. Budiantara, M. Ratna, I. Zain, and W. Wibowo, "A nonparametric spline regression model was used to estimate Indonesian poverty rates. (International Journal of Basic and Applied Sciences, 2012, vol. 12, no. 6, pp. 119-124).

[6] B. Lestari, Fatmawati, and I. N. Budiantara, "Asymptotic Analysis of Spline Estimators for Multiresponse Nonparametric Regression" Songklanakarin J. Sci. Technol., vol. 42, no. 3, pp. 533–548, 2020.

[7] Fatmawati, I. N. Budiantara, and B. Lestari, "Comparison of Smoothing and Truncated Spline Estimators in Estimating Blood Pressure Models," Int. J. Innov. Creat. Change, vol. 5, no. 3, pp. 1177–1199, 2019.

[8] L. Yang and Y. Hong, "Adaptive Penalized Splines for Data Smoothing," Comput. Stat. Data Anal., vol. 108, pp. 70–83, 2017.

[9] V. Ratnasari, I. N. Budiantara, I. Zain, M. Ratna, and N. P. A. M. Mariati, "A comparative analysis of truncated spline and Fourier series methods in multivariable nonparametric regression, applied to poverty data from Papua, Indonesia," Int. J. Basic Appl. Sci., vol. 15, pp. 9–12, 2015.

[10] B. Sifriyani, I. N. Budiantara, S. H. Kartiko, and Gunardi, "Spatial Heterogeneity-Influenced Hypothesis Testing for Truncated Spline Nonparametric Regression: A Novel Method and Application," Abstr. Appl. Anal., pp. 1–13, 2018.

[11] A. A. Puspitasari, A. A. R. Fernandes, A. Efendi, S. Astutik, and E. Sumarminingsih, "Development of Nonparametric Truncated Spline at Various Levels of Autocorrelation of Longitudinal Generating Data," J. Stat. Appl. Probab., vol. 12, no. 2, pp. 757–766, 2023.

[12] Y. Zhang, L. Hua, and J. Huang, "A Spline-Based Semiparametric Maximum Likelihood Estimation Method for the Cox Model with Interval-Censored Data," Scand. J. Stat., vol. 37, no. 2, pp. 338–354, 2010.

[13] M. L. Hazelton and B. A. Turlach, "Semiparametric Regression With Shape-Constrained Penalized Splines," Comput. Stat. Data Anal., vol. 55, no. 10, pp. 2871–2789, 2011.

[14] M. Setyawati, N. Chamidah, and A. Kurniawan, "Confidence Interval of Parameters in Multiresponse Multipredictor Semiparametric Regression Model for Longitudinal Data Based on Truncated Spline Estimator," Commun. Math. Biol. Neurosci., vol. 107, pp. 1–18, 2022.

[15] A. Islamiyati, A. Kalondeng, N. Sunusi, M. Zakir, and A. K. Amir, "Biresponse Nonparametric Regression Model in Principal Component Analysis with Truncated Spline Estimator," J. King Saud Univ. Sci., vol. 34, no. 3, pp. 1–9, 2011.

[16] A. Islamiyati, Fatmawati, and N. Chamidah, "Nonparametric regression modeling of longitudinal bi-response data with multiple smoothing parameters using penalized spline estimation," Songklanakarin J. Sci. Technol., vol. 42, no. 4, 2020.

[17] M. A. D. Octavanny, I. N. Budiantara, H. Kuswanto, and D. P. Rahmawati, "Nonparametric Regression Model for Longitudinal Data with Mixed Truncated Spline and Fourier Series," Abstr. Appl. Anal., Hindawi, 2020.

[18] I. S. Sriliana, I. N. Budiantara, and V. Ratnasari, "A Truncated Spline and Local Linear Mixed Estimator in Nonparametric Regression for Longitudinal Data and Its Application," Symmetry, vol. 14, no. 12, pp. 1–19, 2022.

[19] I. S. Wayan, I. N. Budiantara, S. Suhartono, and S. W. Purnami, "Combined Estimator Fourier Series and Truncated Spline in Multivariable Nonparametric Regression," Appl. Math. Sci., vol. 9, pp. 4997–5010, 2015.

[20] L. Laome, I. N. Budiantara, and V. Ratnasari, "Estimation Curve of Mixed Truncated Spline and Fourier Series Estimator for Geographically Weighted Nonparametric Regression," Mathematics, vol. 11, no. 152, pp. 1–13, 2023.

[21] B. Sifriyani, A. T. R. Dani, M. Fauziyah, M. N. Hayati, S. Wahyuningsih, and S. Prangga, "Spline and Kernel Mixed Estimators in Multivariable Nonparametric Regression for Dengue Hemorrhagic Fever Model," Commun. Math. Biol. Neurosci., vol. 11, pp. 1–15, 2023.

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

[23] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression. New York, NY: John Wiley & Sons, 2000.

[24] A. Agresti, Categorical Data Analysis. New Jersey, NJ: John Wiley & Sons, 2002.

Afiqah Saffa Suriaslan, born on November 7, 2000 in Makassar, is a postgraduate student at the Department of Statistics, Faculty of Data Science and Analysis, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. She earned her S.Stat degree at Universitas Negeri Makassar, Makassar, Indonesia. Her focuses on Nonparametric Regression.

I Nyoman Budiantara, born on June 3, 1965 in Keliki, is a Professor at the Department of Statistics, Faculty of Data Science and Analysis, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. He earned his Ph.D from Gadjah Mada University, Yogyakarta, Indonesia. His research focuses on Nonparametric Regression. Various results of his research have been published in National Journals, International Proceedings, and Scopus indexed International Journals.

Vita Ratnasari, born on September 10, 1970, is a Professor at the Department of Statistics, Faculty of Data Science and Analysis, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. She earned her Ph.D from Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. Her research focuses on Categorical Data Analysis. Various results of her research have been published in National Journals, International Proceedings, and Scopus indexed International Journals.