Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/methodsx

Method for modelling the number of HIV and AIDS cases using least square spline Biresponse nonparametric negative binomial regression[☆,☆☆]

Arip Ramadan[a,b], Nur Chamidah[c,d,*], I Nyoman Budiantara[e], Budi Lestari[f], Dursun Aydin[g,h]

[a] *Doctoral Study Program of Mathematics and Natural Sciences, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia*
[b] *Information System Study Program, Department of Industrial and System Engineering, Telkom University Surabaya Campus, Surabaya 60231, Indonesia*
[c] *Department of Mathematics, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia*
[d] *Research Group of Statistical Modeling in Life Science, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia*
[e] *Statistics Department, Faculty of Science and Data Analytics, Sepuluh Nopember Institute of Technology, Surabaya 60111, Indonesia*
[f] *Department of Mathematics, Faculty of Mathematics and Natural Sciences, The University of Jember, Jember 68121, Indonesia*
[g] *Department of Statistics, Faculty of Science, Muğla Sıtkı Koçman University, Muğla 48000, Turkey*
[h] *Research Scholar at Department of Mathematics, University of Wisconsin, Oshkosh Algoma Blvd, Oshkosh, WI 54901, USA*

ARTICLE INFO

Method name:
LS-Spline BNNBR

Keywords:
HIV
AIDS
Least square spline
Early-age marriage
Contraceptive user
BLSSNNBR model

ABSTRACT

We propose a new statistical modeling method to analyze functional relationship between predictor variables and correlated response variables, and apply it for modeling the number of HIV and AIDS cases influenced by early-age marriages and contraceptive users. The proposed method is LS-Spline BNNBR that is a Biresponse Nonparametric Negative Binomial Regression (BNNBR) based on least squares spline (LS-Spline). To validate the proposed method, we compare it with classical method, namely Ordinary Least Square Biresponse Parametric Negative Binomial Regression (OLS-BPNBR). The results show that the LS-Spline BNNBR method provided deviance value of 7.542334 which is smaller than that of OLS-BPNBR method, namely 7.993935. Thus, the proposed method has better prediction fit than the classical method.

- This study discusses new statistical modeling method for the number of HIV and AIDS cases influenced by early-age marriages and contraceptive users where the response variables are correlated with each other.
- The model estimation results show that the number of HIV and AIDS cases in Indonesia will decrease if the early-age marriages percentage increases and the contraceptive users percentage decreases.
- The results are useful to provide a scientific basis for making policies to control the number of HIV and AIDS cases in Indonesia.

[☆] Related research article: None.
[☆☆] For a published article: None.
[*] Corresponding author at: Department of Mathematics, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia.
E-mail address: nur-c@fst.unair.ac.id (N. Chamidah).

https://doi.org/10.1016/j.mex.2025.103336
Received 4 December 2024; Accepted 24 April 2025
Available online 25 April 2025

Specifications table

Subject area	Mathematics and Statistics
More specific subject area	Nonparametric Regression.
Name of your method	LS-Spline BNNBR.
Name and reference of original method	Not Applicable.
Resource availability	**Data Availability:** The authors attest that, upon reasonable request, the corresponding author will provide the data used in this article.

Background

Infectious diseases, such as HIV (Human Immunodeficiency Virus) and AIDS (Acquired Immune Deficiency Syndrome), have attacked many people and have received very serious attention from several countries, including Europe, South America and North America [1], Africa [2,3], China [4–7], Philippines [8], Brazil [9], Liberia [10], Ethiopia [11], and Indonesia [12], in an effort to reduce the spread and number of victims due to HIV and AIDS [13–15]. These HIV and AIDS spread from sufferers to other people through a transmission [16,17]. Some factors that influence the transmission include early-age marriages, drug users, contraceptive users, education and the number of unemployed. Several researchers have studied the influencing factors of HIV and AIDS cases, for examples, Tohari et al. [18,19] included drug users and contraceptive users factors; Ramadan et al. [20] included early-age marriages and contraceptive users; Igwe et al. [21] included education and the number of unemployed factors; and Farman et al. [22] included antiretroviral treatment compartment. A common statistical tool used to analyze the functional relationship between predictor variables and response variables is the regression model. The regression model consists of parametric regression model and nonparametric regression model. In real cases, we often also use a combination of the two regression models, which is called a semiparametric regression model. Furthermore, to estimate a regression model describing the relationship between the number of HIV and AIDS cases and influencing factors, some researchers employed statistical models of nonparametric and semiparametric regressions [18–20] and parametric regressions [21,22]. In those modeling, we often encounter the use of estimators to estimate these models which are applied to HIV and AIDS cases dataset, for examples, local linear estimator was used to estimate the number of cases of HIV and AIDS models influenced by drug users and contraceptive users percentage in Indonesia [18] and in East Java [19]; LS-Spline estimator was used to estimate a model of the number of HIV cases in Indonesia influenced by early-age marriages and contraceptive users percentage [20]. On the other hand, other estimators have also been widely used by several previous researchers to estimate nonparametric and semiparametric regression models applied to medical researches and other real cases, for prediction and interpretation purposes, for examples, smoothing spline [23–26]; penalized spline [27]; mixed estimator of smoothing spline and Fourier series [28]; and truncated spline [29,30].

In this study, we want to analysis the correlation between the number of HIV and AIDS cases for each province in Indonesia. Since there is two correlated response variables, then we use a biresponse regression model. Therefore, we develop a regression analysis method called the LS-Spline BNNBR for modeling and analyzing HIV and AIDS cases occurred in all provinces of Indonesia influenced by early-age marriages percentage and contraceptive users percentage. LS-Spline is one of spline estimators that has the ability to explain changes in data behavior at certain sub-intervals, so it is very suitable for prediction and interpretation purposes [31,32]. Hereinafter, although Tohari et al. [18,19] have discussed HIV and AIDS modeling using local linear estimator, but according to Wang [33], the local linear estimator is very well applied to data which tends to follow an uptrend or downtrend, and is less good for data that fluctuates in sub-intervals. Apart from that, the predictor variables used by Tohari et al. [18,19] is drug users and contraceptive users percentages, while here we use early-age marriages and contraceptive users percentages. Whereas, Ramadan et al. [20] discussed modeling the HIV cases only, no AIDS cases. In addition, we use a regression model with two response variables correlated with each other. Thus, our proposed method is a development of the methods discussed by Tohari et al. [18,19] and Ramadan et al. [20].

Method details

In this section we provide details of method including defining structure of dataset, determining Pearson correlation coefficient, defining Biresponse Negative Binomial Regression (BNBR) model, determining Least Square Spline Biresponse Nonparametric Regression (LS-Spline BNR) model, and parameter selection method.

Defining structure of dataset

The dataset used in this study are secondary dataset collected from the Indonesian Central Statistics Agency in Surabaya City, Indonesia. The dataset consist of the number of HIV and AIDS cases that occurred in thirty provinces in Indonesia throughout 2023 and the percentage of influencing factors, namely early-age marriages and contraceptive users. In this case, we consider a paired dataset of the number of cases of HIV and the number of cases of AIDS in each province of Indonesia that are correlated with each other. Also, we consider dataset of early-age marriages percentage and contraceptive users percentage. Here, the number cases of HIV and AIDS are as response variables where the number of HIV cases is as the first response variable (Y_1) and the number of AIDS cases is as the second response variable (Y_2), and between the first response variable (Y_1) and the second response variable (Y_2) there is a correlation. While, percentages of early-age marriages and contraceptive users are as predictor variables where percentage of

Table 1
The Number Cases of HIV and AIDS in 30 Provinces of Indonesia in 2023.

Number	Province	HIV Cases (Y_1)	AIDS Cases (Y_2)	Early-Age Marriages (X_1)	Contraceptive Users (X_2)
1	D. I. Aceh	175	50	4.60	40.54
2	Sumatera Utara	1904	326	4.82	38.27
3	Sumatera Barat	364	89	3.48	41.98
⋮	⋮	⋮	⋮	⋮	⋮
30	Papua	1790	254	13.21	19.33

Source: Indonesian Central Statistics Agency.

early-age marriages is as the first predictor variable (X_1) and percentage of contraceptive users is as the second predictor variable (X_2). The structure of the dataset is presented in Table 1.

Determining Pearson correlation coefficient

In common, a measure of linear correlation in a regression model is determined by coefficient value of the Pearson correlation (ρ). It is a value from minus one (–1) to one (1) which describes a measure of the strength and direction of a linear relationship between two variables [34]. The coefficient of Pearson correlation is determined by the covariance value of two variables that is divided by multiplication value of their standard deviations [35]. Mathematically, the coefficient value of Pearson correlation between two variables, namely Y_1 and Y_2, is given by the following equation [34,35]:

$$\rho_{Y_1,Y_2} = \frac{Cov(Y_1, Y_2)}{\sqrt{Var(Y_1)}\sqrt{Var(Y_2)}} \tag{1}$$

Defining Biresponse negative binomial regression (BNBR) model

Biresponse Negative Binomial Regression (BNBR) model is a Negative Binomial Regression (NBR) model with two variables of response that are correlated with each other [36]. Let two random variables Y_{1i} and Y_{2i}, $i = 1, 2, \dots, n$, follow a mixed distribution of Poisson-Gamma. The function of joint probability of the negative binomial response is given by the following equation [36]:

$$f(y_1, y_2) = \frac{\Gamma\left(\frac{1}{\alpha} + y_1 + y_2\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(y_1 + 1)\Gamma(y_2 + 1)} \mu_1^{y_1} \mu_2^{y_2} \alpha^{\frac{1}{\alpha}} \left(\frac{1}{\alpha} + y_1 + y_2\right)^{-\left(\frac{1}{\alpha} + y_1 + y_2\right)} \tag{2}$$

where $y_1, y_2 = 0, 1, 2, \dots$; $\alpha \geq 0$ is a dispersion parameter; $E(Y_r) = \mu_r$; $Var(Y_r) = \mu_r(1 + \alpha\mu_r)$; $r = 1, 2$; and $Corr(Y_1, Y_2) = \sqrt{\frac{\mu_1\mu_2\alpha^2}{(1+\mu_1\alpha)(1+\mu_2\alpha)}}$.

Suppose we have a paired data $(\mathbf{x}_i, y_{1i}, y_{21})$ where $\mathbf{x}_i = (x_{1i}, x_{2i}, \dots, x_{pi})^T$, y_{1i} and y_{2i} are vectors of predictor variables of the first response and the second response, respectively. Next, if y_{1i} and y_{2i} are discrete type response variables with a random sample of size n which is assumed to have a Biresponse Negative Binomial (BNB) distribution and is mutually correlated, then the function of joint probability distribution of y_{1i} and y_{2i} is given as follows [37]:

$$f(y_{1i}, y_{2i}|\mathbf{x}_i) = \frac{\Gamma\left(\frac{1}{\alpha} + y_{1i} + y_{2i}\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(y_{1i} + 1)\Gamma(y_{2i} + 1)} \mu_1^{y_{1i}} \mu_2^{y_{2i}} \alpha^{\frac{1}{\alpha}} \left(\frac{1}{\alpha} + y_{1i} + y_{2i}\right)^{-\left(\frac{1}{\alpha} + y_{1i} + y_{2i}\right)} \tag{3}$$

where $\mu_{1i} = \exp(\mathbf{x}_i^T \boldsymbol{\beta}_1)$ and $\mu_{2i} = \exp(\mathbf{x}_i^T \boldsymbol{\beta}_2)$.

Determining least square spline Biresponse nonparametric regression (LS-Spline BNR) model

Suppose that the paired dataset (x_i, y_i), $i = 1, 2, \dots, n$ where y_i is the i th value of response variable, and x_i is the i th value of predictor variable, follow the following regression model:

$$y_i = \sum_{j=1}^{p} s(x_{ji}) + \varepsilon_i \tag{4}$$

where $s(x_{ji})$ is a function of regression curve and its shape is unknown, and ε_i is the i th value of random error. Here, the value of the regression function can be determined by using a least square spline regression approximation with knots $\tau_1, \tau_2, \dots, \tau_m$. Hence, we may express it the form of the following equation [38]:

$$\sum_{j=1}^{p} s(x_{ji}) = s(x_{1i}) + s(x_{2i}) + \dots + s(x_{pi}) \tag{5}$$

where $i = 1, 2, \ldots, n$; and

$$s(x_{ji}) = \sum_{l=0}^{q} \beta_{jl} x_{ji}^{l} + \sum_{k=1}^{m} \beta_{j(q+k)} (x_{ji} - \tau_{jk})_{+}^{q} \tag{6}$$

If there is one response variable in the nonparametric least square spline regression analysis, then a single response nonparametric least square spline regression model will be obtained. Meanwhile, if there are two response variables in the nonparametric least square spline regression, a biresponse nonparametric least squared spline regression model will be obtained. Biresponse regression (BR) model can be defined as a regression model that has two response variables where scientifically (logic) and mathematically they are correlated or strongly related to each other [39]. Furthermore, the BR model with a regression curve that does not follow a certain shape, it is called a biresponse nonparametric regression (BNR) model. The BNR model with the least square spline approach can be written as follows:

$$\left.\begin{aligned} y_{1i} &= s_1(x_i) + \varepsilon_{1i} \\ y_{2i} &= s_2(x_i) + \varepsilon_{2i} \end{aligned}\right\} \tag{7}$$

where $i = 1, 2, \ldots, n$, and ε_{1i} and ε_{2i} are the i th values of random error for the first response and for the second response, respectively. Here, it is assumed that the curve shape of the regression functions $s_1(x_i)$ and $s_2(x_i)$ does not follow a certain shape. The curves of $s_1(x_i)$ and $s_2(x_i)$ change over certain sub-intervals with the general form given by Eq. (6). Also, the random errors ε_{1i} and ε_{2i} are correlated each other, and have mean of zero and variance of σ^2.

Based on models that have been discussed in previous sections, we develop a regression model, namely Biresponse Nonparametric Negative Binomial Regression (BNNBR) model. The BNNBR is a negative binomial regression model that has two correlated response variables and the curve shape of its regression function does not follow certain shape, and the regression function is assumed to be smooth only.

Parameter selection method

Parameter selection in the Least Squares Spline (LS-Spline) method involves determining the node points (knot points) and smoothing parameters to minimize the error between the data and the spline curve, taking into account the smoothness of the curve. The LS-Spline method aims to find the best spline function (with the smallest error) to model the data, taking into account the smoothness of the curve. Knot points are points that divide the data into segments, where each segment is interpolated by a polynomial function (usually a quadratic or cubic polynomial). The smoothing parameter controls how “straight” or “squiggly” the spline curve is. Higher values of the smoothing parameter will produce a smoother curve, but may not fit the data well, while smaller values will produce a curve that fits the data better, but may not be smooth. Optimal parameter selection can be done with various approach methods, such as Penalized Least Squares (PLS), an approach method that combines goodness of fit with smoothness of the curve through smoothing parameters; Cross-Validation, an approach method used to select the best smoothing parameters based on the predictive performance of the model; Bayesian Information, an approach method that uses a priori information to select smoothing parameters.

The selection of the right parameters in the LS-Spline method is very important to obtain a good model, namely a model that fits the data and also has adequate smoothness. The selection of these parameters can be done with various approaches, and the most suitable approach will depend on the characteristics of the data and the purpose of the analysis. In this study, the selecting optimal parameter and knots are performed based on the MLCV (maximum likelihood cross-validation) criterion proposed by Duin [40], and Fong and Holmes [41].

Method validation

In this section, we provide validation of method including testing the correlation between response variables, identifying the relationship between response and predictor variables, testing overdispersion, determining optimal knot points, and estimating the BPNBR and BNNBR models.

Testing the correlation between response variables

The correlation value between Y_1 and Y_2 is 0.7946. By using the significance level 0.05 and obtaining a p-value of 0.0001, it can be concluded that the correlation between the two variables is significant. Thus, it can be interpreted that these two variables influence each other.

The scatter-plot given by Fig. 1 shows that there is a positive correlation between the number of cases of HIV and the number of cases of AIDS. This fact indicates that the use of the BR model is appropriate for modelling these cases.

Identifying the relationship between response and predictor variables

Hereinafter, the following figures (see Figs. 2 and 3) present scatter plots for identifying the relationship of variables in the model, namely between response and predictor variables.

Figs. 2 and 3 show that scatter-plot results of predictor variables and response variables do not indicate a certain pattern that is a pattern that usual occurs in the case of the parametric regression pattern. This fact supports that the use of the BNR model approach is very suitable for modelling these cases.

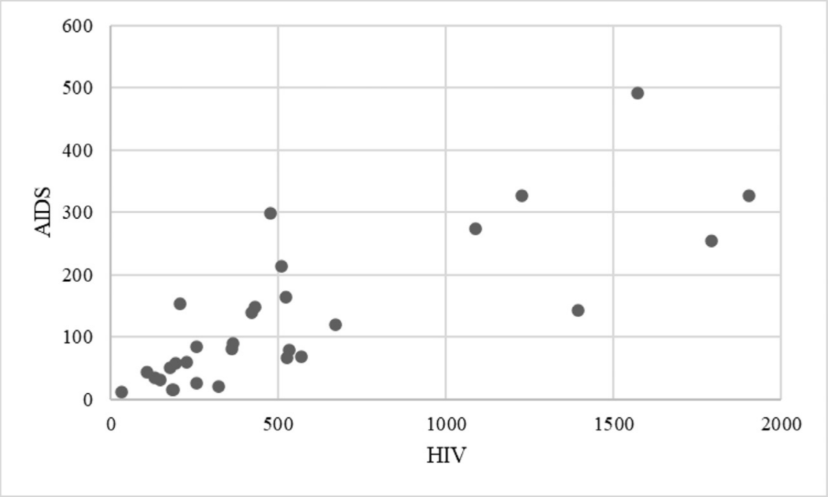

Fig. 1. Scatterplot of the Number of Cases for HIV versus AIDS.

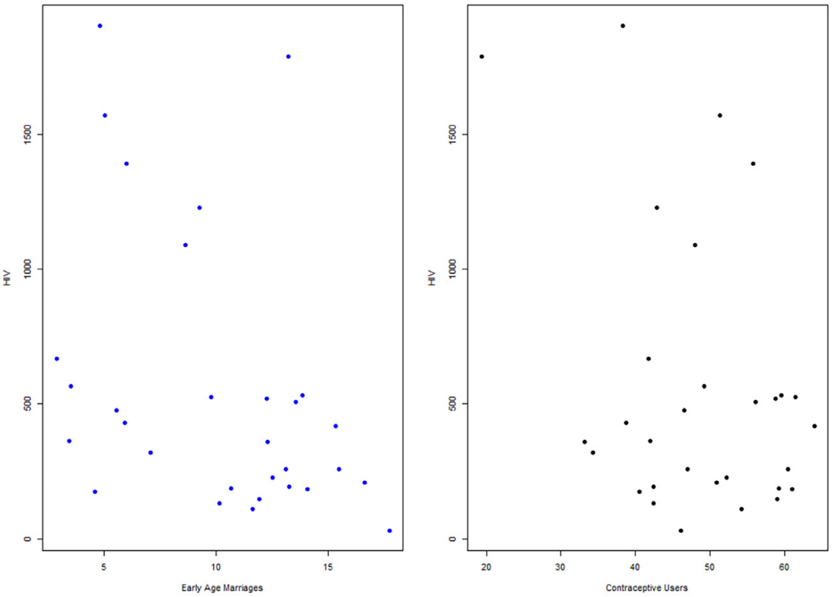

Fig. 2. Scatter-plot Early-Age Marriage and Contraceptive Users versus the Number Cases of HIV.

Testing overdispersion

The next step is to investigate the presence of overdispersion for the variables Y_1 and Y_2 using the dispersion test, the results of which are presented in Table 2.

Based on Table 2, it can be observed that the same spread (equidispersion) assumption for variables Y_1 and Y_2 is not met. Therefore, it is more appropriate to model them using distribution of Negative Binomial rather than distribution of Poisson.

Table 2
Results of Overdispersion Test.

	Y_1	Y_2
Dispersion Ratio	4.9905	4.9148
Pearson's Chi-Squared	40.1133	40.1133
p-value	< 0.0001	< 0.0001
Poisson Bivariate?	No	No

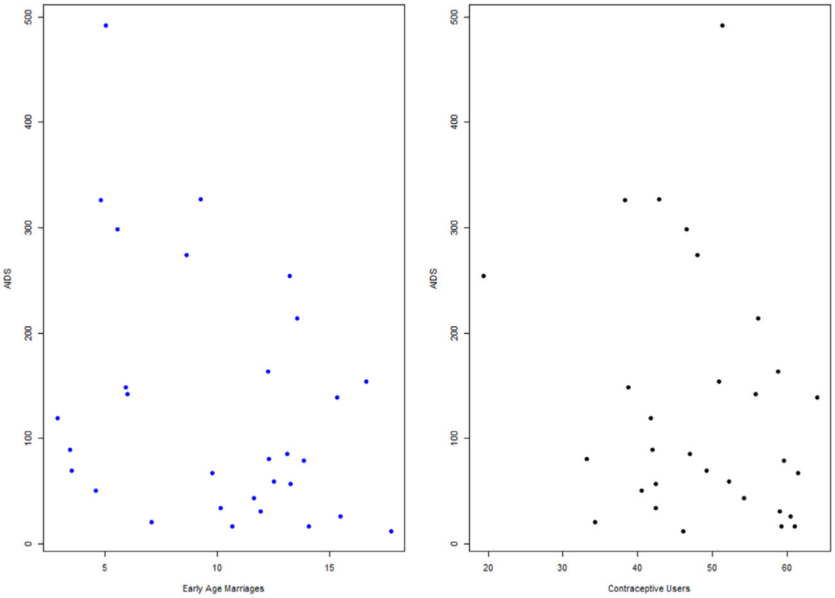

Fig. 3. Scatter-plot Early-Age Marriage and Contraceptive Users versus the Number Cases of AIDS.

Table 3
Optimal Knot Point Combination Based on the Largest MLCV.

Predictor Variables	Y_1		Y_2	
	Number of Knots	Knot Points	Number of Knots	Knot Points
X_1	3	5.9625; 11.1550; 13.2475	3	5.9625; 11.1550; 13.2475
X_2	3	42.0850; 48.6100; 58.1450	3	42.0850; 48.6100; 58.1450
MLCV	–217.1424		–213.7064	

Determining optimal knot points

The optimal combination of knot points selected is the one with the largest MLCV (maximum likelihood cross-validation) value proposed by Duin [40], and Fong and Holmes [41]. The optimal combination of knot points for each variable of predictor is presented in Table 3.

We can observe from Table 3 that the largest MLCV value for the combination in the relationship with Y_1 is −217.1424 and in the relationship with Y_2 is −213.7064.

Estimating the BPNBR and BNNBR models

Hereinafter, we determine the estimation values of parameters of the BNNBR model by using a maximum likelihood estimation method. Finally, we compare the proposed model approach, i.e., BNNBR model, with the classical model, i.e., BPNBR model to show suitability of the proposed model in predicting based on the deviance values of these models [42]. The model with a smallest deviance value is to be the best model that can be used for prediction purpose [43]. The results of estimating parameters of regression models of BPNBR (Biresponse Parametric Negative Binomial Regression) based on Ordinary Least Square (OLS) estimator and BNNBR (Biresponse Nonparametric Negative Binomial Regression) based on Least Square Spline (LS-Spline) estimator for the number of cases of HIV and AIDS, and Deviance values for those two cases are presented in Tables 4 and 5, respectively.

Based on the deviance values in Table 5, it is found that the Deviance value of BNNBR model is smaller than that of BPNBR model. Also, it can be shown that the Deviance value of BNNBR model is smaller than value of Chi-Squared distribution, i.e., $\chi^2_{n-2} = 41.33714$, and its p-value is greater than significance level value of 0.005. This means that the best model approach for modelling the number of cases of HIV and AIDS is BNNBR model. In other words, the BNNBR model based on least square spline estimator is an appropriate model approach for these cases. Thus, the proposed model approach in this research, namely BNNBR model, with the estimation results of model parameters presented in Table 4, is a better model approach for modelling these cases than classical model approach, namely BPNBR model. Next, we obtain the estimated BNNBR model based on LS-Spline estimator for the first response variable as

Table 4
Estimation Results of Models for HIV and AIDS Cases Using Least Square Spline Estimator.

HIV			AIDS		
Estimator	BPNBR Model	BNNBR Model	Estimator	BPNBR Model	BNNBR Model
$\hat{\beta}_0$	8.0416	93.5820	$\hat{\beta}_0$	6.0725	54.4770
$\hat{\beta}_1$	−0.0782	−0.0400	$\hat{\beta}_1$	−0.0647	0.0802
$\hat{\beta}_{1,1}$	–	0.0568	$\hat{\beta}_{1,1}$	–	−0.0876
$\hat{\beta}_{1,2}$	–	−0.0360	$\hat{\beta}_{1,2}$	–	0.2417
$\hat{\beta}_{1,3}$	–	−0.5246	$\hat{\beta}_{1,3}$	–	−0.6963
$\hat{\beta}_2$	−0.0210	−0.0551	$\hat{\beta}_2$	−0.0126	−0.0101
$\hat{\beta}_{2,1}$	–	−0.0585	$\hat{\beta}_{2,1}$	–	−0.0608
$\hat{\beta}_{2,2}$	–	0.0423	$\hat{\beta}_{2,2}$	–	−0.0927
$\hat{\beta}_{2,3}$	–	0.6147	$\hat{\beta}_{2,3}$	–	0.6982
$\hat{\alpha}$	0.5247	0.8525	$\hat{\alpha}$	0.5247	0.8525

Table 5
Deviance for BPNBR Model and BNNBR model.

	BPNBR model	BNNBR model
Deviance	7.993935	7.542334
p-value	0.999	0.999

follows:

$$\begin{aligned}\hat{\mu}_1 &= \exp[\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_{1,1}(x_1 - \tau_{1,1})_+ + \hat{\beta}_{1,2}(x_1 - \tau_{1,2})_+ + \hat{\beta}_{1,3}(x_1 - \tau_{1,3})_+ \\ &\quad + \hat{\beta}_2 x_2 + \hat{\beta}_{2,1}(x_2 - \tau_{2,1})_+ + \hat{\beta}_{2,2}(x_2 - \tau_{2,2})_+ + \hat{\beta}_{2,3}(x_2 - \tau_{2,3})_+] \\ &= \exp[93.582 - 0.04x_1 + 0.0568(x_1 - 5.9625)_+ - 0.036(x_1 - 11.155)_+ - 0.5246(x_1 - 13.2475)_+ \\ &\quad + - 0.0551x_2 - 0.0585(x_2 - 42.085)_+ + 0.0423(x_2 - 48.61)_+ + 0.6147(x_2 - 58.145)_+].\end{aligned}$$

Hence, we obtain the estimated Least Square Spline (LS-Spline) regression function for the first response variable as follows:

$$\hat{s}_1(x_1) = 93.582 - 0.04x_1 + 0.0568(x_1 - 5.9625)_+ - 0.036(x_1 - 11.155)_+ - 0.5246(x_1 - 13.2475)_+$$

$$= \begin{cases} 93.582 - 0.04x_1 & \text{for} \quad x_1 < 5.9625 \\ 93.2433 + 0.0168x_1 & \text{for} \quad 5.9625 \le x_1 < 11.155 \\ 93.6449 - 0.0192x_1 & \text{for} \quad 11.155 \le x_1 < 13.2475 \\ 100.5945 - 0.5438x_1 & \text{for} \quad x_1 \ge 13.2475 \end{cases}$$

$$\hat{s}_1(x_2) = 93.582 - 0.0551x_2 - 0.0585(x_2 - 42.085)_+ + 0.0423(x_2 - 48.61)_+ + 0.6147(x_2 - 58.145)_+$$

$$= \begin{cases} 93.582 - 0.0551x_2 & \text{for} \quad x_2 < 42.085 \\ 96.044 - 0.1136x_2 & \text{for} \quad 42.085 \le x_2 < 48.61 \\ 93.9878 - 0.0713x_2 & \text{for} \quad 48.61 \le x_2 < 58.145 \\ 58.2461 + 0.5434x_2 & \text{for} \quad x_2 \ge 58.145 \end{cases}.$$

Also, we obtain the estimated BNNBR model based on LS-Spline estimator for the second response variable as follows:

$$\begin{aligned}\hat{\mu}_2 &= \exp[\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_{1,1}(x_1 - \tau_{1,1})_+ + \hat{\beta}_{1,2}(x_1 - \tau_{1,2})_+ + \hat{\beta}_{1,3}(x_1 - \tau_{1,3})_+ \\ &\quad + \hat{\beta}_2 x_2 + \hat{\beta}_{2,1}(x_2 - \tau_{2,1})_+ + \hat{\beta}_{2,2}(x_2 - \tau_{2,2})_+ + \hat{\beta}_{2,3}(x_2 - \tau_{2,3})_+] \\ &= \exp[54.477 + 0.0802x_1 - 0.0876(x_1 - 5.9625)_+ + 0.2417(x_1 - 11.155)_+ - 0.6963(x_1 - 13.2475)_+ \\ &\quad + - 0.0101x_2 - 0.0608(x_2 - 42.085)_+ - 0.0927(x_2 - 48.61)_+ + 0.6982(x_2 - 58.145)_+].\end{aligned}$$

Hence, we obtain the estimated Least Square Spline (LS-Spline) regression function for the second response variable as follows:

$$\hat{s}_2(x_1) = 54.477 + 0.0802x_1 - 0.0876(x_1 - 5.9625)_+ + 0.2417(x_1 - 11.155)_+ - 0.6963(x_1 - 13.2475)_+$$

$$= \begin{cases} 54.477 + 0.0802x_1 & \text{for} \quad x_1 < 5.9625 \\ 54.9993 - 0.0074x_1 & \text{for} \quad 5.9625 \le x_1 < 11.155 \\ 52.3031 + 0.2343x_1 & \text{for} \quad 11.155 \le x_1 < 13.2475 \\ 61.5273 - 0.462x_1 & \text{for} \quad x_1 \ge 13.2475 \end{cases}$$

Table 6
Estimation Results of BPNBR and BNNBR Models for HIV Cases in Provinces of Indonesia.

Province	Actual value	BPNBR estimates	BNNBR estimates
D. I. Aceh	175	924.8305	1034.3020
Sumatera Utara	1904	953.5007	1161.7960
Sumatera Barat	364	979.3528	999.1675
Riau	476	756.3893	549.5239
Jambi	187	388.5071	421.9925
Sumatera Selatan	521	346.6077	333.4724
Bengkulu	146	353.2486	384.2741
Lampung	526	397.7340	1396.5450
Kepulauan Bangka Belitung	184	286.9720	718.1603
Kepulauan Riau	669	1030.5550	1036.0460
D. I. Yogyakarta	567	839.1410	453.9010
Banten	1392	601.9347	257.7956
Bali	1571	710.6282	365.4187
Nusa Tenggara Barat	207	291.5693	62.2019
Nusa Tenggara Timur	429	864.8434	1083.9120
Kalimantan Barat	531	301.0060	358.2154
Kalimantan Tengah	257	259.8611	244.7520
Kalimantan Selatan	419	244.6545	1798.6030
Kalimantan Timur	1089	575.9843	478.4879
Kalimantan Utara	130	575.8438	931.4379
Sulawesi Utara	508	330.6063	221.6458
Sulawesi Tengah	227	390.0507	353.2633
Sulawesi Selatan	1227	612.2225	869.6165
Sulawesi Tenggara	193	451.5318	899.7073
Gorontalo	109	399.8054	310.1686
Sulawesi Barat	31	295.1199	52.6925
Maluku	320	867.7301	1403.5600
Maluku Utara	257	416.0015	543.1579
Papua Barat	360	593.2297	1572.7500
Papua	1790	736.9071	3305.1660

$$\hat{s}_2(x_2) = 54.477 - 0.0101x_2 - 0.0608(x_2 - 42.085)_+ - 0.0927(x_2 - 48.61)_+ + 0.6982(x_2 - 58.145)_+$$

$$= \begin{cases} 54.477 - 0.0101x_2 & \text{for} \quad x_2 < 42.085 \\ 57.0358 - 0.0709x_2 & \text{for} \quad 42.085 \le x_2 < 48.61 \\ 61.5419 - 0.1636x_2 & \text{for} \quad 48.61 \le x_2 < 58.145 \\ 20.9451 + 0.5346x_2 & \text{for} \quad x_2 \ge 58.145 \end{cases}.$$

The formula $\hat{s}_1(x_1)$ explains the relationship between X_1 and Y_1, the formula $\hat{s}_1(x_2)$ explains the relationship between X_2 and Y_1. Also, the formula $\hat{s}_2(x_1)$ explains the relationship between X_1 and Y_2, and the formula $\hat{s}_2(x_2)$ explains the relationship between X_2 and Y_2. Next, when the X_1 value is between 5.9625 and 11.1550, adding the X_1 value will increase Y_1, other than in this range adding the X_1 value will decrease Y_1. When the X_2 value is >58.1450, the Y_1 value will increase, otherwise, the Y_1 value will decrease. When the X_1 value is <5.9625 and between 11.1550 and 13.2475, the Y_2 value will increase. When the X_2 value is >58.1450, the Y_2 value will increase. Based on the estimation of the BNNBR models, it can be observed that when the percentage of the number of early-age marriages is between 5.9625 and 11.1550, increasing the number of early-marriages will increase the number of cases of HIV. On the other hand, when the percentage of the number of contraceptive users is >58.1450, the number of cases of HIV will increase. Also, when the percentage of the number of early-age marriages is <5.9625 and between 11.1550 and 13.2475, the number of cases of AIDS will increase. Whereas when the percentage of the contraceptive users is >58.1450, the number of cases of AIDS will increase.

Next, based on the estimated BPNBR and BNNBR models above, the estimation results of the BPNBR model and the BNNBR model for cases of HIV and AIDS in Provinces of Indonesia are presented in Tables 6 and 7, respectively.

In the following figure (see Fig. 4), we provide plot results of estimation values of BPNBR and BNNBR models compared with the actual values for HIV cases and AIDS cases.

Fig. 4 shows that visually, the estimation curve of the number of HIV and AIDS cases using the BNNBR model approach based on the Least Square Spline (LS-Spline), which is called LS-Spline BNNBR method, is more representative than using the BPNBR model approach based on the Ordinary Least Square (OLS), which is called OLS-BPNBR method. This is in accordance with the results of the statistical model estimation calculations in the discussion section above. This fact shows that the use of proposed method, namely LS-Spline BNNBR, is more suitable to model the number of HIV and AIDS cases in Indonesia affected by percentages of early-age marriages and contraceptive users than the use of classical method, namely OLS-BPNBR.

Table 7
Estimation Results of BPNBR and BNNBR Models for AIDS Cases in Provinces of Indonesia.

Province	Actual value	BPNBR estimates	BNNBR estimates
D. I. Aceh	50	193.1880	222.6137
Sumatera Utara	326	195.9927	231.8532
Sumatera Barat	89	203.9569	200.5402
Riau	298	168.3712	172.0723
Jambi	16	103.0856	55.4942
Sumatera Selatan	163	93.6184	57.2711
Bengkulu	31	95.2129	60.8734
Lampung	67	106.2309	184.0213
Kepulauan Bangka Belitung	16	80.9490	166.6755
Kepulauan Riau	119	212.5029	191.7193
D. I. Yogyakarta	69	185.7665	115.2610
Banten	142	145.6403	47.7602
Bali	492	163.5997	91.2873
Nusa Tenggara Barat	154	78.1226	35.7430
Nusa Tenggara Timur	148	181.1759	252.7148
Kalimantan Barat	79	83.6125	82.8065
Kalimantan Tengah	26	74.3718	64.1185
Kalimantan Selatan	139	71.9458	450.4949
Kalimantan Timur	274	135.3478	157.1256
Kalimantan Utara	34	131.7264	231.7817
Sulawesi Utara	214	88.8937	60.9751
Sulawesi Tengah	59	100.0115	113.2853
Sulawesi Selatan	327	138.8830	225.6909
Sulawesi Tenggara	57	107.7482	372.5032
Gorontalo	43	103.0841	65.9548
Sulawesi Barat	12	77.1433	36.7339
Maluku	20	177.9342	262.2399
Maluku Utara	85	102.8879	261.8785
Papua Barat	80	129.1699	334.5676
Papua	254	144.6874	479.7195

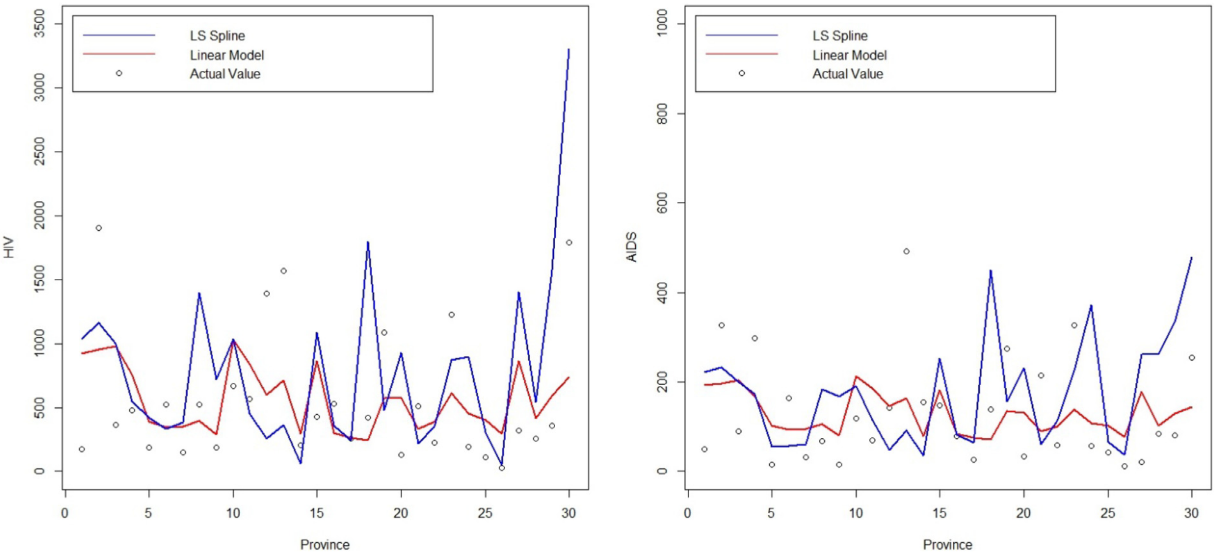

Fig. 4. Plots of actual values (plaid), estimation values of BPNBR (red) and BNNBR (blue) for HIV (left) and AIDS (right).

Conclusion

The data analysis results using the proposed method, namely LS-Spline BNNBR, show that the best combination of knots for the first predictor variable (X_1) and the second predictor variable (X_2) is combination of three knots for each response variable, with the largest Maximum Likelihood Cross-Validation (MLCV) values for the first response variable (Y_1) and for the second response variable (Y_2) are −217.1424 and −213.7064, respectively. Also, the results show that LS-Spline BNNBR method is the best method for modelling the number of cases of HIV and AIDS in Indonesia. Hereinafter, based on the estimation results using the proposed method, namely

LS-Spline BNNBR, show that when the percentage of the number of early-age marriages is between 5.9625 and 11.1550, increasing the number of early-marriages will increase the number of HIV cases in Indonesia. On the other hand, when the percentage of the number of contraceptive users is >58.1450, the number of HIV cases in Indonesia will increase. Also, when the percentages of the number of early-age marriages are <5.9625 and between 11.1550 and 13.2475, the number of AIDS cases in Indonesia will increase. Whereas when the percentage of the contraceptive users is >58.1450, the number of AIDS cases in Indonesia will increase. Therefore, to reduce the number of cases of HIV and AIDS in Indonesia, the early-age marriages percentage must be reduced, while the contraceptive users percentage must be increased.

Limitations

Not applicable.

Ethics statements

Not applicable.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Arip Ramadan: Conceptualization, Methodology, Software, Writing – original draft, Validation. **Nur Chamidah:** Conceptualization, Methodology, Supervision, Validation, Software, Writing – review & editing. **I Nyoman Budiantara:** Conceptualization, Methodology, Supervision, Validation. **Budi Lestari:** Supervision, Validation, Writing – review & editing. **Dursun Aydin:** Supervision, Validation, Writing – review & editing.

Data availability

Data will be made available on request.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Next, authors thank the Indonesian Central Statistics Agency (Badan Pusat Statistik–BPS) in Surabaya City, Indonesia for providing data used in this research.

References

[1] D. Govender, M.J. Hashim, M.A.B. Khan, H. Mustafa, G. Khan, Global Epidemiology of HIV/AIDS: a resurgence in North America and Europe, J. Epidemiol. Glob. Health 11 (3) (2021) 296–301.
[2] K.A. Risher, A. Cori, G. Reniers, M. Marston, C. Calvert, A. Crampin, et al., Age patterns of HIV incidence in Eastern and Southern Africa: a modelling analysis of observational population-based cohort studies, The Lancet HIV 8 (7e429–e439) (2021).
[3] Kusilika, The factors that have led to increased HIV/AIDS prevalence among adolescents aged 13-21 years in Hoima Regional Referral Hospital, Hoima District: a cross-sectional study, Student's J. Health Res. Africa 3 (3) (2022) 1–14 14.
[4] N. He, Research progress in the epidemiology of HIV/AIDS in China, CCDC Weekly 3 (48) (2021) 1022–1030.
[5] Z. Dou, Y. Luo, Y. Zhao, X. Zheng, M. Han, Preplanned studies: trends in mortality and prevalence of reported HIV/AIDS cases—China, 2002–2021, CCDC Weekly 5 (42) (2023) 943–947.
[6] N. Wang, G. Lan, Q. Zhu, H. Chen, J. Huang, Q. Meng, Z. Shen, S. Liang, X. Wu, L. Luo, R. Ye, J. Chen, S. Tan, H. Xing, Y. Shao, Y. Ruan, M. Lin, HIV epidemiology, care, and treatment outcomes among student and nonstudent youths living with HIV in Southwest China between 1996 and 2019: historical cohort study, JMIR Public Health Surveill 9 (e38881) (2023).
[7] L.A. Mohamud, A.M. Hassan, J.A. Nasir, Determinants of HIV/Aids knowledge among females in Somalia: findings from 2018 to 2019 SDHS data, HIV AIDS–Res. Palliative Care 15 (2023) 435–444.
[8] L.M.A. Gangcuangco, P.C. Eustaquio, The State of the HIV epidemic in the Philippines: progress and challenges in 2023, Trop. Med. Infect. Dis. 8 (5) (2023) 1–16 258.
[9] L.A. Andrade, T. de F. Amorim, W.S. da Paz, et al., Reduced HIV/AIDS diagnosis rates and increased AIDS mortality due to late diagnosis in Brazil during the COVID-19 pandemic, Sci. Rep. 13 (23003) (2023).
[10] B.L. Seifu, G. Eshun, G.A. Tesema, F. Kyei-Arthur, Comprehensive knowledge about HIV/AIDS and associated factors among reproductive age women in Liberia, BMC Public Health 24 (619) (2024) 1–10.
[11] S.A. Yilema, Y.A. Shiferaw, A.T. Belay, D.B. Belay, Mapping the spatial disparities of HIV prevalence in Ethiopian zones using the generalized additive model, Sci. Rep. 14 (6215) (2024) 1–10.
[12] F.M. Nasution Jocelyn, N.A. Nasution, M.H. Asshiddiqi, N.H. Kimura, M.H.T. Siburian, Z.Y.N. Rusdi, A.R. Munthe, I. Chairenza, M.C.F.B.G. Munthe, P. Sianipar, S.P. Gultom, D. Simamora, I.R. Uswanas, E. Salim, K. Khairunnisa, R.A. Syahputra, HIV/AIDS in Indonesia: current treatment landscape, future therapeutic horizons, and herbal approaches, Front. Public Health 12 (1298297) (2024) 1–11.
[13] R.H. Remien, M.J. Stirratt, N. Nguyen, R.N. Robbins, A.N. Pala, C.A. Mellins, Mental health and HIV/AIDS the need for an integrated response, AIDS 33 (9) (2019) 1411–1420.
[14] T.A. Schwetz, A.S. Fauci, The Extended impact of Human Immunodeficiency Virus/AIDS research, J. Infect. Dis. 219 (1) (2019) 6–9.

[15] GBD 2021 HIV Collaborators, Global, regional, and national burden of HIV/AIDS, 1990–2021, and forecasts to 2050, for 204 countries and territories: the global burden of disease study 2021, Lancet HIV 11 (e807) (2024) –22.

[16] G.T. Vu, B.X. Tran, C.L. Hoang, B.J. Hall, H.T. Phan, G.H. Ha, C.A. Latkin, C.S.H. Ho, R.C.M. Ho, Global research on quality of life of patients with HIV/AIDS: is it socio-culturally addressed? (GAPRESEARCH), Int. J. Environ. Res. Public Health 17 (2127) (2020).

[17] S. Payagala, A. Pozniak, The Global Burden of HIV, Clin. Dermatol. 42 (2) (2024) 119–127.

[18] A. Tohari, N. Chamidah, F. Fatmawati, Modelling of HIV and AIDS cases in Indonesia using Bi-response negative binomial regression approach based on local linear estimator, Ann. Biol. 36 (2) (2020) 215–219.

[19] A. Tohari, N. Chamidah, F. Fatmawati, B. Lestari, Modelling the number of HIV and AIDS cases in East Java using biresponse multipredictor negative binomial regression based on local linear estimator, Commun. Math. Biol. Neurosci. 2021 (73) (2021) 1–17.

[20] A. Ramadan, N. Chamidah, I.N. Budiantara, Modelling the number of HIV cases in Indonesia using negative binomial regression based on least square spline estimator, Commun. Math. Biol. Neurosci. 2024 (79) (2024) 1–16.

[21] M.C. Igwe, E.I. Obeagu, A.O. Ogbuabor, Analysis of the factors and predictors of adherence to healthcare of people living with HIV/AIDS in tertiary health institutions in Enugu State, Madonna Univ. J. Med. Health Sci. 2 (3) (2022) 42–57.

[22] M. Farman, A. Akgül, M.T. Tekin, M.M. Akram, A. Ahmad, Fractal fractional-order derivative for HIV/AIDS model with Mittag-Leffler kernel, Alexandria Eng. J. 61 (12) (2022) 10965–10980.

[23] B. Lestari, Fatmawati, I.N. Budiantara, Spline estimator and its asymptotic properties in multiresponse nonparametric regression model, Songklanakarin J. Sci. Technol. 42 (3) (2020) 533–548.

[24] B. Lestari, N. Chamidah, D. Aydin, E. Yilmaz, Reproducing kernel hilbert space approach to multiresponse smoothing spline regression function, Symmetry (Basel) 14 (11) (2022) 2227.

[25] B. Lestari, N. Chamidah, I.N. Budiantara, D. Aydin, Determining confidence interval and asymptotic distribution for parameters of multiresponse semiparametric regression model using smoothing spline estimator, J. King Saud Univ. - Sci. 35 (5) (2023) 102664.

[26] D. Aydın, E. Yılmaz, N. Chamidah, B. Lestari, Right-censored partially linear regression model with error in variables: application with carotid endarterectomy dataset, Int. J. Biostatistic. 20 (1) (2024) 245–278.

[27] N. Chamidah, B. Lestari, T. Saifudin, R. Rulaningtyas, P. Wardhani, I.N. Budiantara, Estimating the number of malaria parasites on blood smears microscopic images using penalized spline nonparametric poisson regression, Commun. Math. Biol. Neurosci. 2024 (2024) 60.

[28] N. Chamidah, B. Lestari, I.N. Budiantara, D. Aydin, Estimation of multiresponse multipredictor nonparametric regression model using mixed estimator, Symmetry (Basel) 16 (4) (2024) 386.

[29] N. Chamidah, B. Lestari, I.N. Budiantara, T. Saifudin, R. Rulaningtyas, A. Aryati, P. Wardani, Consistency and asymptotic normality of estimator for parameters in multiresponse multipredictor semiparametric regression model, Symmetry (Basel) 14 (2) (2022) 336.

[30] N. Chamidah, B. Lestari, H. Susilo, M.Y. Alsagaff, I.N. Budiantara, D. Aydin, Spline estimator in nonparametric ordinal logistic regression model for predicting heart attack risk, Symmetry (Basel) 16 (11) (1440) 2024.

[31] R.L. Eubank, Spline Smoothing and Nonparametric Regression, Marcel Deker, New York, NY, 1988.

[32] G. Wahba, Spline Models for Observational Data, SIAM, Philadelphia, PA, 1990.

[33] Y. Wang, Smoothing Splines, Methods and Applications, CRC Press, New York, NY, 2011.

[34] P. Schober, C. Boer, L.A. Schwarte, Correlation coefficients: appropriate use and interpretation, Anesth. Analg. 126 (5) (2018) 1763–1768.

[35] M. Zhang, W. Li, L. Zhang, H. Jin, Y. Mu, L. Wang, A pearson correlation-based adaptive variable grouping method for large-scale multi-objective optimization, Inf. Sci. (N.Y.) 639 (2023) 118737.

[36] A. Islamiyati, A. Kalondeng, N. Sunusi, M. Zakir, A.K. Amir, Biresponse nonparametric regression model in principal component analysis with truncated spline estimator, J. King Saud Univ.-Sci. 34 (3) (2022) 101892.

[37] S. Cheon, S.H. Song, B.C. Jung, Tests for independence in a bivariate negative binomial model, J. Korean Stat. Soc. 38 (2) (2009) 185–190.

[38] R.N.S. Setiawan, I.N. Budiantara, V. Ratnasari, Application of confidence intervals for parameters of nonparametric spline truncated regression on index development gender in East Java, IPTEK J. Sci. 2 (3) (2017).

[39] T. Similä, J. Tikka, Input selection and shrinkage in multiresponse linear regression, Comput. Stat. Data Anal. 52 (1) (2007) 406–422.

[40] Duin, On the choice of smoothing parameters for parzen estimators of probability density functions, IEEE Trans. Comput. 25 (11) (1976) 1175–1179 C–.

[41] E. Fong, C.C. Holmes, On the marginal likelihood and cross-validation, Biometrika 107 (2) (2020) 489–496.

[42] D.T. Ailobhio, J.A. Ikughur, A review of some goodness-of-fit tests for logistic regression model, Asian J. Prob. Stats. 26 (7) (2024) 75–85.

[43] M. Ibrahim, E. Altun, H. Goual, H.M. Yousof, Modified goodness-of-fit type test for censored validation under a new burr type XII distribution with different methods of estimation and regression modeling, Eurasian Bull. Math. 3 (3) (2021) 162–182.