Eksakta : Berkala Ilmiah Bidang MIPA
VOLUME 25 NO 02 2024, pp 138-150
ISSN : Print 1411-3724 — Online 2549-7464
DOI : https://doi.org/10.24036/eksakta/vol25-iss01/491

Eksakta
Berkala Ilmiah Bidang MIPA
http://www.eksakta.ppj.unp.ac.id/index.php/eksakta

Article

Generalized Additive Models for Modeling Pneumonia Cases in Toddlers in West Java based on the Penalized Spline Estimator

Article Info

Article history :

Received January 26, 2024
Revised April 20, 2024
Accepted April 28,2024
Published June 30, 2024

Keywords :
Pneumonia, GAM, penalized spline, full search, GCV local scoring

Azkanul Wahyu[1*], Nurul Gusriani[1], Kankan Parmikanti[1]

[1]Department of Mathematics, Faculty of Mathematics and Natural Science (FMIPA), Universitas Padjadjaran, Bandung, Indonesia

Abstract. Acute Respiratory Infections (ARI) are one of the causes of high mortality in the world, such as pneumonia in toddlers. Pneumonia cases in West Java are high compared to other provinces. In this study, pneumonia cases will be modeled with Generalized Additive Models (GAM) based on penalized spline estimators. The optimal number of knots is determined using the full search algorithm and the optimal smoothing parameter is obtained based on the minimum Generalized Cross Validation (GCV) value of order one or two. Then, GAM parameter estimation is performed using the local scoring algorithm. Formed model based on the order, number of knots, and smoothing parameters of each predictor variable with order one, number of knots two, and optimal smoothing parameter one for X_1, order two, number of knots three, and optimal smoothing parameter one for X_2, and order one, number of knots two, and optimal smoothing parameter for X_3 whose parameters were estimated by local scoring resulted in a coefficient of determination of 0.679. This indicates that 67.9% of the factors from the predictor variables affect the percentage of pneumonia cases among under-fives while the remaining 32.1% is influenced by other factors outside the model.

Corresponding Author :
Azkanul Wahyu
Department of Mathematics, Faculty of Mathematics and Natural Science (FMIPA), Universitas Padjadjaran, Bandung, Indonesia
Email : azkanul20001@mail.unpad.ac.id

1. Introduction

Cardiovascular diseases, respiratory problems, and neonatal deaths account for three of the world's disease deaths. Respiratory problems such as Acute Respiratory Infection (ARI), a lower respiratory infection, are one of the diseases that cause a high number of deaths in the world each year and are ranked as the fourth highest cause of death. Based on the data of the *World Health Organization* (WHO) (2020) [1], the disease caused the deaths of more than 2.6 million people in 2019. One of the areas that the government concentrates on to raise societal wellbeing is health [2]. Pneumonia is one of the diseases included in ARI and is a contagious disease [3-4]. Cough, dyspnea, sputum production, and chest pain are the signs of pneumonia [5].

The relationship between pneumonia and its influencing factors can be seen using regression analysis. Regression analysis is a study of the relationship between one response variable and one or more predictor variables [6-7]. One type of regression that can be used is Generalized Additive Models (GAM) which is included in nonparametric regression. Some previous studies that modeled with GAM are [8-9]. According to [8], GAM parameter estimation were obtained using the local scoring algorithm and a kernel estimator. On the contrary, [9] used a thin plate spline to estimate the nonparametric function. The smoothing function is estimated by minimizing the penalized sum of squares criterion with the optimal smoothing parameters obtained using the cross-validation criterion.

In a study by [10], the incidence of pneumonia in children under-five was influenced by one factor, exclusive breastfeeding. According to [11], factors that influence the percentage of pneumonia cases are low birth weight (LBW) and immunization status. In another study by [12], it was concluded that there was an influence of the percentage of households with clean and healthy living behaviors (PHBS) on the percentage of pneumonia cases.

In this study, the percentage of pneumonia cases among under-fives in West Jawa Province in 2021 will be modeled using Generalized Additive Models (GAM) based on the influencing factors, namely PHBS, exclusive breastfeeding, and LBW. Penalized spline can overcome the problem of knot point selection because penalized spline uses a very large number of knots, with a smoothing parameter (λ) that controls the smoothness of the curve [13]. The optimal smoothing parameter can be obtained using the full search algorithm based on the Generalized Cross-Validation (GCV) criterion [14-15]. Furthermore, the regression model will be estimated with the local scoring algorithm [16].

2. Materials and Methods

2.1. Materials

The object of this study is data on the percentage of pneumonia cases among under-fives in West Java Province and its influencing factors, such as the percentage of households with clean and healthy living behaviors (PHBS), the percentage of exclusive breastfeeding among under-fives, and the percentage of low birth weight (LBW) in 2021 obtained from the official website of the West Java regional government (opendata.jabarprov.go.id). The observation (n) units used are 27 districts/cities in West Java.

2.2. Regression Analysis

Statistics, a scientific discipline, has found widespread application among the people [17]. Regression analysis is a technique to determine the pattern of relationship between response variables and predictor variables. The response variable is written as variable Y followed by predictor variables $X_1, X_2, \ldots, X_{p^*}$ with p^* stating the number of predictor variables. Regression analysis can be used to see the effect of each predictor variable on the model, analyze the effect of changes in the value of predictor variables, or predict the value of the response variable that is influenced by its predictor variables based on the equation it has [18]. Another form of regression analysis is multiple linear regression which used to predict the connection between the variables when there are numerous predictor variable [19].

According to [20], the approach to regression analysis is divided into three approaches, namely parametric, nonparametric, and semiparametric regression approaches. Nonparametric regression can be used when the relationship pattern between the response variable (Y) and one or more predictor variables (X) is unknown. In general, the form of the nonparametric regression model is as equation (1).

$$y_i = f(x_i) + \varepsilon_i \tag{1}$$

Nonparametric regression can overcome the problems in parametric regression where the form of the regression is unknown. In addition, in parametric regression, there are several assumptions that must be met, such as normality, no autocorrelation, heteroscedasticity, and multicollinearity. This is not considered in nonparametric regression because it is not bound by the assumption of the shape of the regression curve [21].

2.3 Generalized Additive Models (GAM)

GLM model has a link function that connects the mean of response variable with the predictor variable [22]. Generalized Additive Models (GAM) are extensions of GLMs by adding additive functions that replace linear functions and by replacing nonlinear predictors with smoothing functions [23]. GAM helps to maximize the model accuracy of the response variable (Y) coming from a diverse distribution by estimating a nonparametric function of the predictor variables (X) connected by a link function [24]. The general form of the GAM model can be formulated as in equation (2)

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p^*} f_j(x_{ji}), \tag{2}$$

with $Y_i \sim f_{\theta_i}(y_i)$, β_0 is the intercept coefficient, $X_j$is the jth predictor variable, p^* is the number of predictor variables and $f_j(x_{ji})$ is the smoothing function of the jth predictor variable [16].

2.4 Smoothing Function Penalized Spline

Smoothing functions are used to estimate the regression function in nonparametric regression approaches. This technique is commonly known as smoothing technique. Some smoothing techniques that can be used in nonparametric regression approaches are kernel estimators, spline estimators, Fourier series, orthogonal series, and wavelets [21]. One of these smoothing techniques, the spline estimator, has very flexible properties and is able to handle data with changing behavior well. One form of spline estimator that is often used is penalized spline [14]. The penalized spline estimator has an order, knots, number of knots, and smoothing parameter λ so that it can produce smoother smoothing results [25].

In the penalized spline estimator, the estimated form of the regression function f_j with order p_j and knots $k_1, k_2, \dots, k_{u_j}$ is obtained by an approximation as equation (3)

$$f_j(x_{ji}) = \sum_{h_j=0}^{p_j} \beta_{jh_j} x_{ji}^{h_j} + \sum_{h_j^*=1}^{u_j} \beta_{j(p_j+h_j^*)} \left(x_{ji} - k_{jh_j^*}\right)_+^{p_j} \tag{3}$$

with $\boldsymbol{\beta}_j = \left(\beta_{j0}, \beta_{j1}, \dots, \beta_{jp_j}, \beta_{j(p_j+1)}, \dots, \beta_{j(p_j+u_j)}\right)^T$ as the parameter vector, u_j is the number of knots of the jth predictor variable, $i: 1, \dots, n$, $j: 1, \dots, p^*$ and

$$\left(x_{ji} - k_{jh_j^*}\right)_+^{p_j} = \begin{cases} \left(x_{ji} - k_{jh_j^*}\right)^{p_j}, & x \geq k_{jh_j^*} \\ 0, & x < k_{jh_j^*} \end{cases}.$$

Equation (4) can be written in matrix form as equation (4).

$$\mathbf{f}_j(\mathbf{X}_j) = \mathbf{X}_j\boldsymbol{\beta}_j. \tag{4}$$

According to [14], to obtain the estimated value of $\boldsymbol{\beta}_j$ from the value of each predictor variable X_j can be obtained by minimizing the Penalized Least Square (PLS) function. The PLS function is expressed in the equation (5)

$$L_j = \sum_{i=1}^{n}\left(y_i - f_j(x_{ji})\right)^2 + \lambda_j \sum_{h_j^*=1}^{u_j} \beta^2_{j\left(p_j+h_j^*\right)}, \tag{5}$$

where λ_j is the smoothing parameter for X_j, p_j is the polynomial order for X_j, u_j is the number of knots for X_j, and

$$\sum_{h_j^*=1}^{u_j} \beta^2_{j(p_j+h_j^*)} = \beta^2_{j(p_j+1)} + \beta^2_{j(p_j+2)} + \cdots + \beta^2_{j(p_j+u_j)}. \tag{6}$$

by differentiating (5) against $\boldsymbol{\beta}_j$ can be obtained equation (7)

$$\boldsymbol{\beta_j}\left(\mathbf{X_j^T X_j} + \boldsymbol{\lambda_j \mathbf{D}_j}\right) = \mathbf{X_j^T y}. \tag{7}$$

Based on equation (7), the estimated value of $\boldsymbol{\beta}_j$ can be obtaines as equation (8)

$$\widehat{\boldsymbol{\beta}}_{j_{OLS}} = \left(\mathbf{X}_j^T\mathbf{X}_j + \lambda_j\mathbf{D}_j\right)^{-1}\mathbf{X}_j^T\boldsymbol{y}. \tag{8}$$

Substitute equation (8) into equation (4) to obtain the penalized spline function of the predictor variable X_j is equation (9)

$$\hat{\mathbf{f}}_j(\mathbf{X}_j) = \mathbf{X}_j\left(\mathbf{X}_j^T\mathbf{X}_j + \lambda_j\mathbf{D}_j\right)^{-1}\mathbf{X}_j^T\boldsymbol{y}. \tag{9}$$

The penalized spline function estimate of the predictor variable X_j in the equation can be expressed as equation (10)

$$\hat{\mathbf{f}}_j(\mathbf{X}_j) = \mathbf{A}_{\lambda_j}\boldsymbol{y}, \tag{10}$$

with

$$\mathbf{A}_{\lambda_j} = \mathbf{X}_j\left(\mathbf{X}_j^T\mathbf{X}_j + \lambda_j\mathbf{D}_j\right)^{-1}\mathbf{X}_j^T. \tag{11}$$

2.5 Optimal Smoothing Parameter

The selection of optimal smoothing parameters is very important to obtain smooth curve results. Determining the optimal smoothing parameter can be done using Generalized Cross Validation (GCV). The smoothing parameter is obtained from the minimum GCV value. According to [14], the GCV value is obtained from the formula on equation (12)

$$GCV_{\lambda_j} = \frac{MSE_{\lambda_j}}{\left[1 - \frac{tr\left(\mathbf{A}_{\lambda_j}\right)}{n}\right]^2}, \tag{12}$$

with MSE (Mean Squared Error) used to calculate the GCV value in finding the optimal λ rewritten as MSE_{λ_j}, namely

$$MSE_{\lambda_j} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{13}$$

2.6 Choosing Optimal The Number of Knot

Knot is the point that determines the change in the behavior of the data [26]. Knots are located at the quantile points of the unique (single) value of the predictor variable $\{x_i\}_{i=1}^{n}$ with each knot equidistant. The number of knot points can also be called the number of knots (u) [14].

The full search algorithm is one of the algorithms that can be used to select the number of knots. It searches for all possible knot counts and selects the knot count with the minimum GCV value as the optimal knot candidate. The number of knots (u) searched will be less than the number of observations. The following are the steps in the full search algorithm:

1. For $u = 1$ and $u = 2$, compare each smoothing parameter that minimizes the GCV value.
   a. If the GCV value at $u = 2$ is more than 0,98 times the GCV value at $u = 1$, the process will be stopped. Then, select the number of knots (u) between $u = 1$ and $u = 2$ that has the smallest GCV value.
   b. If the GCV value at $u = 2$ is equal to or less than 0,98 times the GCV value at $u = 1$, the process will continue.
2. For $u = 2$ and $u = 3$, compare each smoothing parameter that minimizes the GCV value.
   a. If the GCV value at $u = 2$ is more than 0,98 times the GCV value at $u = 1$, the process will be stopped. Then, select the number of knots (u) between $u = 1$ and $u = 2$ that has the smallest GCV value.
   b. If the GCV value at $u = 2$ is equal to or less than 0,98 times the GCV value at $u = 1$, the process will continue.
3. The process continues in the same way to compare the GCV values for $u = 3$ and $u = 4$. And so on until the optimal number of knots (u) is obtained [14].

2.7 Kolmogorov-Smirnov Test

The distribution of the response variable in GAM must belong to the exponential distribution family, such as the normal distribution. The Kolmogorov-Smirnov test can be used to test whether the data to be analyzed follows a certain distribution or not [27].

Hypothesis test:

$$H_0: F(x) = F^*(x)$$
$$H_1: F(x) \neq F^*(x)$$

Statistic test:

$$D = \max\{|F_n(x) - F^*(x)|\} \tag{14}$$

with the hypothesis set, namely

$F_n(x)$: Empirical distribution function of the sample

$F^*(x)$: Cumulative distribution function

H_0: Data is normally distributed

H_1: Data is not normally distributed

The test criterion of the Kolmogorov-Smirnov test is if the value of D resulting from the test statistic is less than the $D_{\alpha,n}$ value, then accept H_0. However, if the value of D value resulting from the test statistic is more than the $D_{\alpha,n}$ value, then reject H_0.

2.8 Coefficient of Determination

The coefficient of determination (R^2) can be interpreted as a measuring tool to measure the extent to which variations or changes in the response variable can be explained by variations in one or more predictor variables. The coefficient of determination is between 0 and 1. The results of the coefficient of determination are more informative than other measurement tools, such as Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute

Percentage Error (MAPE) which have limitations to interpret the results [28]. The coefficient of determination can be formulated as equation (15).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{15}$$

2.9 Local Scoring Algorithm

The regression function in GAM can be estimated using the local scoring algorithm. This function can estimate equation (3) faster than local likelihood estimation [13].

1. Initialize the initial values $\hat{f}_j^l(x_{ji}) = \hat{f}_1^l(x_{1i}) = \cdots = \hat{f}_{p^*}^l(x_{p^*i})$ for each of the 0th iteration predictor variables, that is, when $l = 0$ for $l = 0,1,2,\ldots, j = 1,2,\ldots,p^*$ and $i = 1,2,\ldots,n$.
2. Iterate local scoring with the following steps:
   a. Determining the adjusted dependent variable value (z) as equation (16)
   $$z_i^{(l)} = \eta_i^{(l)} + \left(y_i - \mu_i^{(l)}\right)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{(l)} \tag{16}$$
   b. Define the weight matrix $(\mathbf{W})$ as equation (17)
   $$w_{ii} = \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2_{(l)} \left(Var(y_i)^{(l)}\right)^{-1} \tag{17}$$
3. Iterate the model with weights using the following backfitting algorithm:
   a. Calculate the partial residual with equation (18)
   $$\mathbf{R}_j^{(l+1)} = \mathbf{z} - \sum_{s \neq j}^{p^*} f_s^{(l)}(X_s) \tag{18}$$
   b. Determine the smoothing function in the model with the equation (19)
   $$\mathbf{f}_j^{(l+1)} = \mathbf{A}_{\lambda_j}\mathbf{R}_j^{(l+1)} \tag{19}$$
   c. Calculate the mean value of the squared residuals with the equation (20)
   $$Avg\left(RSS^{(l+1)}\right) = \frac{1}{n}\{(\mathbf{y} - \hat{\boldsymbol{\mu}})^T(\mathbf{y} - \hat{\boldsymbol{\mu}})\} \tag{20}$$
   d. Iterate until the RSS value meets the convergence in equation (21)
   $$\left|Avg\left(RSS^{(l+1)}\right) - Avg\left(RSS^{(l)}\right)\right| < \varepsilon \tag{21}$$
4. The local scoring and backfitting steps will continue until a small average deviance value is obtained in equation (22)
   $$Avg(Dev) = \frac{1}{n}\{(\mathbf{y} - \hat{\boldsymbol{\mu}})^T\mathbf{W}_i(\mathbf{y} - \hat{\boldsymbol{\mu}})\} \tag{22}$$
   and meets the convergence criterion in equation (23)
   $$\left|Avg(Dev)^{l+1} - Avg(Dev)^{l}\right| < \varepsilon \tag{23}$$

2.10 Generalized Additive Models with Penalized Spline Estimator

According to [29], Weighted Least Square (WLS) is used to determine parameter estimates using local scoring procedures in GAM. The Penalized Least Square (PLS) function will be minimized by the WLS method to obtain the estimate $\widehat{\boldsymbol{\beta}}_j$ for each predictor variable X_j as formulated by equation (24).

$$L_j = \sum_{i=1}^{n} w_i\left(y_i - f_j(x_{ji})\right)^2 + \lambda_j \sum_{h_j^*=1}^{u_j} \beta^2_{j\left(p_j + h_j^*\right)}, \tag{24}$$

by differentiating (24) against $\boldsymbol{\beta}_j$ can be obtained equation (25)

$$\widehat{\boldsymbol{\beta}}_{j_{WLS}} = \left(\mathbf{X}_j^T\mathbf{W}\mathbf{X}_j + \lambda_j\mathbf{D}_j\right)^{-1}\mathbf{X}_j^T\mathbf{W}\mathbf{y}. \tag{25}$$

Substitute equation (25) into equation (5) to obtain the penalized spline function of the predictor variable X_j as equation (26)

$$\hat{\mathbf{f}}_j(\mathbf{X}_j) = \mathbf{X}_j \widehat{\boldsymbol{\beta}}_{j_{WLS}}. \tag{26}$$

Then, equation (26) can be expressed as equation (27)

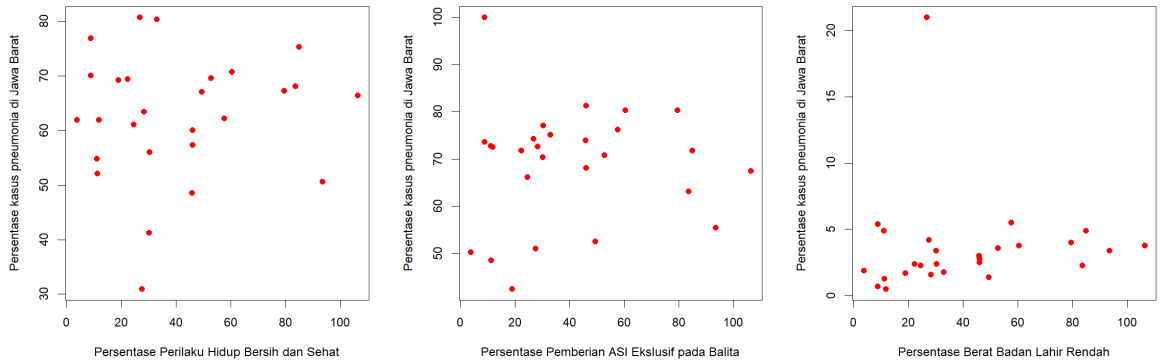
$$\hat{\mathbf{f}}_j(\mathbf{X}_j) = \mathbf{A}^*_{\lambda_j} \boldsymbol{y}, \tag{27}$$

with

$$\mathbf{A}^*_{\lambda_j} = \mathbf{X}_j \left(\mathbf{X}_j^T \mathbf{W} \mathbf{X}_j + \lambda_j \mathbf{D}_j\right)^{-1} \mathbf{X}_j^T \mathbf{W}. \tag{28}$$

3. Results and Discussion

Before modeling, it is necessary to conduct an initial analysis first so that modeling can be carried out. The first thing to do is to make a scatterplot between each predictor variable and the response variable to ensure that a nonparametric regression model can be performed in this case. The results are showns as Figure 1A, Figure 1B, and Figure 1C



A B C

Figure 1. Scatter plot Y against A) the percentage of households with clean and healthy living behaviors (PHBS), B) the percentage of exclusive breastfeeding among under-fives, and C) the percentage of low birth weight (LBW)

Based on Figure 1A, Figure 1B, and Figure 1C, the scatter plots between Y and X_j do not follow a specific function graph pattern. In this case, nonparametric regression can be used to overcome the nonlinear relationship between variables. Furthermore, the response variable is assumed to be in the exponential family. The Kolmogorov Smirnov test will be used to ensure the assumption that the response variable Y belongs to the exponential family distribution, such as normal distribution which is the chosen distribution in this study

Table 1. Kolmogorov-Smirnov Test Results Data on the Percentage of Pneumonia Cases in Toddlers in West Java in 2021

n	Significance Rate (α)	D	$D_{(\alpha,n)}$	Hypothesis Test
27	0,05	0,16352	0,254	H_0 is accepted

Based on Table 1, the D value generated from the test statistic is less than the $D_{(0.05,27)}$ value, so H_0 is accepted. Therefore, the data to be modeled is normally distributed.

3.1 Optimal Smoothing Parameter Selection

The optimal smoothing parameters will be selected using the full search algorithm by [12][30] below.

Table 2. The Results of Full Search Algorithm for Each Variable Predictor

Variable	Order	Number of Knots	Knots Location	Optimal Smoothing Parameter
X_1	1	2	60.7967, 68.53	1
X_2	2	3	64.675, 71.85, 74.73	1
X_3	1	1	1.9, 3.6	1

According to the data in Table 2, the model form that can be formed is in the following equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_{11}x_{1i}^1 + \hat{\beta}_{12}(x_{1i} - k_{11})_+^1 + \hat{\beta}_{13}(x_{1i} - k_{12})_+^1 + \hat{\beta}_{21}x_{2i}^1 + \hat{\beta}_{22}x_{2i}^2 + \hat{\beta}_{23}(x_{2i} - k_{21})_+^2 + \hat{\beta}_{23}(x_{2i} - k_{22})_+^2 + \hat{\beta}_{23}(x_{2i} - k_{23})_+^2 + \hat{\beta}_{31}x_{3i}^1 + \hat{\beta}_{32}(x_{3i} - k_{31})_+^1 + \hat{\beta}_{33}x_{3i}^2(x_{3i} - k_{32})_+^1$$

3.2 Parameter Estimation Results of Generalized Additive Models based on Penalized Spline Estimator

Calculations to get the estimated value of the Generalized Additive Models parameters are carried by using the local scoring algorithm with a tolerance value of 10^{-1}. The iteration results are as follows:

1^{st} iteration $r = 0$

1. Determining the initial value of $f_j(x_{ji})$ from the previous step 3.3 so that we get Table 3.

Table 3. Initial Value of the Estimation Function for Each Predictor

n	$f_1^0(x_{1i})$	$f_2^0(x_{2i})$	$f_3^0(x_{3i})$
1	34.15078	68.70238	37.38869
⋮	⋮	⋮	⋮
27	29.14551	48.55649	53.94487

2. Define z using equation (16).
3. Determining the initial value of η_i and μ_i based on the 1^{st} step, we get Table 4.

Table 4. Initial Value of η_i and μ_i

n	η_i	μ_i
1	140.2418	140.2418
⋮	⋮	⋮
27	138.6398	138.6398

4. Define the matrix W as a matrix of size $(n \times n)$ with the main diagonal entries being 1.
5. Determining the partial residual of iteration 0 for each predictor, we get Table 5.

Table 5. Initial Value of Partial Residual of Each Predictor

n	R_1^1	R_2^1	R_3^1
1	-81.671	35.05779	11.86968
⋮	⋮	⋮	⋮
27	-72.471	30.57564	40.003298

6. Calculating the value of $A^*(\lambda_j)$ using equation (28), we get

$$A^*(\lambda_1) = \begin{bmatrix} 0{,}1282217 & \cdots & -0{,}1503095 \\ \vdots & \ddots & \vdots \\ -0{,}01503095 & \cdots & 0{,}2300106 \end{bmatrix}_{27\times 27},$$

$$A^*(\lambda_2) = \begin{bmatrix} 0{,}2384613 & \cdots & 0{,}1338382 \\ \vdots & \ddots & \vdots \\ 0{,}1338382 & \cdots & 0{,}1111057 \end{bmatrix}_{27\times 27},$$

and

$$A^*(\lambda_3) = \begin{bmatrix} 0{,}06835633 & \cdots & 0{,}02471354 \\ \vdots & \ddots & \vdots \\ 0{,}02471354 & \cdots & 0{,}06907664 \end{bmatrix}_{27\times 27}.$$

7. Calculating the value of $f_j(x_{ji})$ using equation 19, we get
8.

Table 6. The value of $f_j(x_{ji})$

n	$f_1^1(x_{1i})$	$f_1^1(x_{1i})$	$f_1^1(x_{1i})$
1	-48.026	60.5768	36.0667
⋮	⋮	⋮	⋮
27	-54.491	44.48754	53.7261

9. Calculate η and μ

$$\widehat{\boldsymbol{\eta}} = \widehat{\boldsymbol{\mu}} = \begin{bmatrix} 48{,}61705 \\ 38{,}05128 \\ \vdots \\ 43{,}72309 \end{bmatrix}_{27\ x\ 1}$$

10. Calculate the mean of residual sum of squares so we get

$$Avg(RSS^1) = 265{,}4421$$

Based on calculation above, we obtained $\left|Avg\left(RSS^{l+1}\right) - Avg\left(RSS^{l}\right)\right| = 265{,}4421$, where the value is still greater than the tolerance value, so, the calculation continues to the next iteration. The iteration is continued until a convergent value is obtained. Iteration of backfitting stopped in the 7th iteration. Thus, the next step is going to the scoring iteration with initial value $Avg(Dev^0) = 0$. From the last result of previous iteration, we get $\left|Avg\left(Dev^{l+1}\right) - Avg\left(Dev^{l}\right)\right| = 252.1399$, where the value is still greater than the tolerance value, so the calculation continues to the next iteration which is the calculation of backfitting iteration at 8th iteration. The iteration is stopped at this step. The estimation of Generalized Additive Models parameters are using R Studio software with the results as follows:

$\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}} = [-90{,}3589 \quad 0{,}7017 \quad 1{,}9120 \quad -2{,}9161]^{\mathrm{T}}$
$\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}} = [-624{,}4972 \quad 23{,}0888 \quad -0{,}1927 \quad -0{,}0664 \quad 1{,}9506 \quad -1{,}9377]^{\mathrm{T}}$
$\widehat{\boldsymbol{\beta}}_{\boldsymbol{3}} = [9{,}9985 \quad 9{,}7375 \quad 7{,}3801 \quad -18{,}8694]^{\mathrm{T}}$

Based on the result, we can get models based on penalized spline estimator with local scoring algorithm as equation (29)

$$\begin{aligned} \hat{y}_i = {} & -704{,}8576 + 0{,}7017x_{1i}^1 + 1{,}9120(x_{1i} - 60{,}7966)_+^1 - 2{,}9161(x_{1i} - 68{,}53)_+^1 + \\ & 23{,}0888x_{2i}^1 - 0{,}1927x_{2i}^2 - 0{,}0664(x_{2i} - 64{,}675)_+^2 + 1{,}9506(x_{2i} - 71{,}85)_+^2 - \\ & 1{,}9377(x_{2i} - 74{,}73)_+^2 + 9{,}7375x_{3i}^1 + 7{,}3801(x_{3i} - 1{,}9)_+^1 - 18{,}8694(x_{3i} - 3{,}6)_+^1 \end{aligned} \quad (29)$$

With truncated function:

$$(x_{1i}-60{,}7966)_+^1=\begin{cases}(x_{1i}-60{,}7966)_+^1 & ;x_{1i}\geq 60{,}7966\\ 0 & ;x_{1i}<60{,}7966\end{cases}$$

$$(x_{1i}-68{,}53)_+^1=\begin{cases}(x_{1i}-68{,}53)_+^1 & ;x_{1i}\geq 68{,}53\\ 0 & ;x_{1i}<68{,}53\end{cases}$$

$$(x_{2i}-64{,}675)_+^2=\begin{cases}(x_{2i}-64{,}675)_+^2 & ;x_{2i}\geq 64{,}675\\ 0 & ;x_{2i}<64{,}675\end{cases}$$

$$(x_{2i}-71{,}85)_+^2=\begin{cases}(x_{2i}-71{,}85)_+^2 & ;x_{2i}\geq 71{,}85\\ 0 & ;x_{2i}<71{,}85\end{cases}$$

$$(x_{2i}-74{,}73)_+^2=\begin{cases}(x_{2i}-74{,}73)_+^2 & ;x_{2i}\geq 74{,}73\\ 0 & ;x_{2i}<74{,}73\end{cases}$$

$$(x_{3i}-1{,}9)_+^1=\begin{cases}(x_{3i}-1{,}9)_+^1 & ;x_{3i}\geq 1{,}9\\ 0 & ;x_{3i}<1{,}9\end{cases}$$

$$(x_{3i}-3{,}6)_+^1=\begin{cases}(x_{2i}-3{,}6)_+^2 & ;x_{3i}\geq 3{,}6\\ 0 & ;x_{3i}<3{,}6\end{cases}$$

thus obtained

$$\hat{y}_i=\begin{cases}-704{,}8576+0{,}7017x_{1i}^1 & ;x_{1i}<60{,}7966\\ -840{,}0966+2{,}6137x_{1i}^1 & ;\ 60{,}7966\leq x_{1i}<68{,}53\\ -640{,}2563-0{,}3024x_{1i}^1 & ;x_{1i}\geq 68{,}53\\ 23{,}0888x_{2i}^1-0{,}1927x_{2i}^2 & ;x_{2i}<64{,}675\\ -277{,}7416+31{,}6776x_{2i}^1-0{,}2591x_{2i}^2 & ;64{,}675\leq x_{2i}<71{,}85\\ 9792{,}0797-248{,}6235x_{2i}^1+1{,}6915x_{2i}^2 & ;71{,}85\leq x_{2i}<74{,}73\\ -1029{,}1471+40{,}9850x_{2i}^1-0{,}2462x_{2i}^2 & ;x_{2i}\geq 74{,}73\\ 9{,}7375x_{3i}^1 & ;\ x_{3i}<1{,}9\\ -14{,}02219+17{,}1176x_{3i}^1 & ;1{,}9\leq x_{3i}<3{,}6\\ 53{,}9076-1{,}7518x_{3i}^1 & ;x_{3i}\geq 3.6\end{cases}$$

3.3 Model Determination Coefficient of Percentage of Pneumonia Cases in Toddlers in West Java Province in 2021

The coefficient of determination (R^2) for the model of the percentage of pneumonia cases among under-fives based on the factors that influence it in equation (29) can be obtained using equation (16). The value was calculated using the R Studio program. The R^2 value obtained is 0.679, indicating that 67.9% of the factors in the predictor variables, namely the percentage of households with clean and healthy living behaviors (X_1), exclusive breastfeeding (X_2), and low birth weight (X_3), affect the percentage of pneumonia cases in children under-five, while the remaining 0.321 indicates that 32.1% is influenced by other factors outside the model. The coefficient of determination of 0.679 falls into the strong relation.

3.4 Parameter Estimation Results of the Model for the Percentage of Pneumonia Cases in Toddlers by Factors Affecting It in West Java Province in 2021

The estimated value obtained from the model formed will then be compared with the actual value to see whether the pattern of the prediction model can follow the actual data pattern. Based on Figure 2, the model of the percentage of pneumonia cases in children under five years old with Generalized Additive Models based on Penalized Spline is quite capable of fitting the actual data patter with the

optimal order, knot points, and smoothing parameters. This is reinforced with a model determination coefficient value of 0.679.

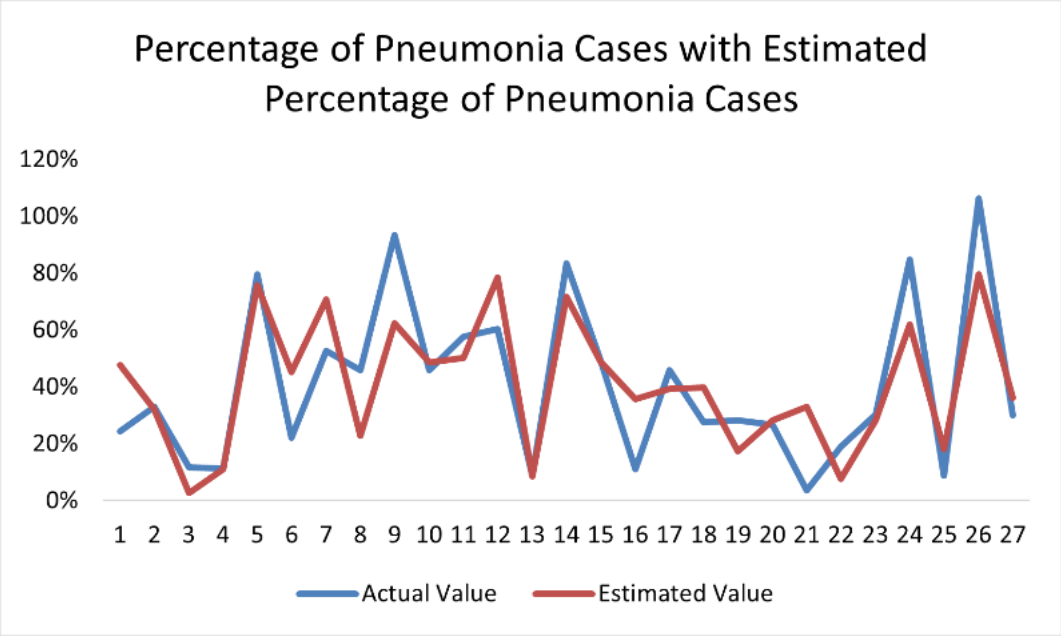


Figure 2. Comparison graph of actual and estimated values

4. Conclusion

The Generalized Additive Models based on the penalized spline estimator formed of pneumonia cases in children under five years old can be obtained with the local scoring algorithm. The model formed is the model in equation (29) with the order, number of knots, and smoothing parameters for X_1 are 1, 2, and 1, respectively, the order, number of knots, and smoothing parameters for X_2 are 2, 3, and 1, respectively, and the order, number of knots, and smoothing parameters for X_3 are 1, 2, and 1, respectively, with the coefficient of determination obtained of 0.679. Based on this, the variable of the percentage of pneumonia among under-fives in West Java Province in 2021 can be explained by the three predictor variables by 67.9%, while the remaining 32.1% is explained by other variables not included in the model.

References

[1] WHO. (2020). Retrieved From Encyclopedia, https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Diakses pada tanggal 19 Agustus 2023.

[2] Fitri, F., Sari, F. M., Gamayanti, N. F., & Utami, I. T. (2021). Infant Mortality Case: An Application of Negative Binomial Regression in order to Overcome Overdispersion in Poisson Regression. *EKSAKTA: Berkala Ilmiah Bidang MIPA*, *22*(3), 200-210.

[3] Anwar, A., & Dharmayanti, I. (2014). Pneumonia pada anak balita di Indonesia. *Kesmas: Jurnal Kesehatan Masyarakat Nasional (National Public Health Journal)*, *8*(8), 359-365.

[4] Hasan, M. M., Faruk, M. O., Biki, B. B., Riajuliislam, M., Alam, K., & Shetu, S. F. (2021, January). Prediction of Pneumonia Disease of Newborn Baby Based on Statistical Analysis of Maternal Condition Using Machine Learning Approach. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 919-924). IEEE.

[5] Nisrina, N., Handoko, B., & Andriyana, Y. (2023). The Analysis of Factors Influencing Incidence Rates of Toddler Pneumonia in Purwakarta Districts Using Panel Data Spatial Regression. *EKSAKTA: Berkala Ilmiah Bidang MIPA*, *24*(02), 122-132.

[6] Debataraja, , dan Imro'ah Nurfitri.(2020). Analisis regresi dengan metode least absolute shrinkage and selection operator (lasso) dalam mengatasi multikolinearitas. *Bimaster : Buletin Ilmiah Matematika, Statistika Dan Terapannya*. 9(1):31–38.

[7] Chen, X., & Derezinski, M. (2021). Query complexity of least absolute deviation regression via robust uniform convergence. In *Conference on Learning Theory* (pp. 1144-1179). PMLR.

[8] Rifada, M., Suliyanto, E. T., & Kesumawati, A. (2018). The logistic regression analysis with nonparametric approach based on local scoring algorithm (Case study: Diabetes mellitus type II cases in Surabaya of Indonesia). *Int. J. Adv. Soft Comput. Appl*, *10*, 167-178.

[9] Pinilla, J., & Negrín, M. (2021). Non-parametric generalized additive models as a tool for evaluating policy interventions. *Mathematics*, *9*(4), 299.

[10] Pertiwi, F. D., & Nasution, A. S. (2022). Faktor-faktor yang Berhubungan dengan Kejadian Pneumonia pada Balita di Puskesmas Semplak Kota Bogor 2020. *Promotor*, *5*(3), 273-280.

[11] Oktaviani, I., Hayati, S., & Supriatin, E. (2014). Faktor-faktor yang berhubungan dengan kejadian infeksi saluran pernafasan akut (ISPA) pada balita di Puskesmas Garuda Kota Bandung. *Jurnal Keperawatan BSI*, *2*(2).

[12] Ayu, D., Winarso, S., & Rokhmah, D. (2020). Pengaruh Indikator Perilaku Hidup Bersih dan Sehat (PHBS) Terhadap Gejala Pneumonia Pada Balita di Puskesmas Mojopanggung (perkotaan), Puskesmas Tapanrejo (pedesaan) dan Puskesmas Kedungrejo (pesisir) Banyuwangi. *Multidisciplinary Journal*, *3*(1), 1-5.

[13] Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., & Schmid, M. (2019). A review of spline function procedures in R. *BMC medical research methodology*, *19*, 1-16.

[14] Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of computational and graphical statistics*, *11*(4), 735-757.

[15] Dani, A. T. R., Ratnasari, V., & Budiantara, I. N. (2021, March). Optimal Knots Point and Bandwidth Selection in Modeling Mixed Estimator Nonparametric Regression. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1115, No. 1, p. 012020). IOP Publishing.

[16] Hastie, T., & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, *82*(398), 371-386.

[17] Sormin, C., Gusmanely, Z., & Nurhidayah, N. (2020). Generalized Poisson Regression Type-II at Jambi City Health Office. *EKSAKTA: Berkala Ilmiah Bidang MIPA*, *21*(1), 54-58.

[18] Chatterjee, S., & Hadi, A. S. (2015). *Regression analysis by example*. John Wiley & Sons.

[19] Yuinanda, S., Multahadah, C., Marisa, H., & Abdullah, M. (2023). Water Quality Model in Jambi Province using Geographically Weighted Regression. *EKSAKTA: Berkala Ilmiah Bidang MIPA*, *23*(03), 436-452.

[20] Kahar, A. M., Suparti, S., & Hakim, A. R. (2023). Analisis Indeks Harga Saham Gabungan Dan Faktor Pengaruhnya Menggunakan Pemodelan Regresi Semiparametrik Kernel Dilengkapi Gui-R. *Jurnal Gaussian*, *12*(1), 30-41.

[21] Astuti, D. A., Srinadi, I. A., & Susilawati, M. (2018). Pendekata Regresi Nonparametrik dengan Menggunakan Estimator Kernel pada Data Kurs Rupiah Terhadap Dolar Amerika Serikat. *E-Jurnal Matematika*.

[22] Dewi, K., Gusriani, N., & Parmikanti, K. (2023). Factors Affecting the Number of Infant Morality Cases in West Java for the 2019-2020 Period using Generalized Poisson Regression (GPR). *EKSAKTA: Berkala Ilmiah Bidang MIPA*, *24*(02), 259-270.

[23] Fernanda, A., & Debataraja, N. N. (2023). Pemodelan Data Produksi Kelapa Sawit Di Kalimantan Barat Menggunakan Generalized Additive Model. *Bimaster: Buletin Ilmiah Matematika, Statistika dan Terapannya*, 12(3).

[24] Ravindra, K., Rattan, P., Mor, S., & Aggarwal, A. N. (2019). Generalized additive models: Building evidence of air pollution, climate change and human health. *Environment international*, *132*, 104987.

[25] Siregar, N. R. A. A., Farida, F., Falasifah, S., El Fahmi, M. F., & Chamidah, N. (2022). Pemodelan Harga Minyak Mentah Dunia Berdasarkan Efek Pandemi Covid-19 Dengan Estimator Penalized Spline. *MUST: Journal of Mathematics Education, Science and Technology*, *7*(2), 152-166.

[26] Jao, N., Islamiyati, A., & Sunusi, N. (2022). Pemodelan Regresi Nonparametrik Spline Poisson pada Tingkat Kematian Bayi di Sulawesi Selatan. *Estimasi: Journal of Statistics and Its Application*, 14-22.

[27] Ahadi, G. D., & Zain, N. N. L. E. (2023). Pemeriksaan Uji Kenormalan dengan Kolmogorov-Smirnov, Anderson-Darling dan Shapiro-Wilk. *Eigen Mathematics Journal*, 11-19.

[28] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peerj computer science*, *7*, e623.

[29] Buja, A., Hastie, T., & Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 453-510.

[30] Jiang, W., Wu, X., Gong, Y., Yu, W., & Zhong, X. (2020). Holt–Winters smoothing enhanced by fruit fly optimization algorithm to forecast monthly electricity consumption. *Energy*, *193*, 116779.