BAREKENG: Journal of Mathematics and Its Application
December 2022 Volume 16 Issue 4 Page 1231–1240
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss4pp1231-1240

OUTLIER IDENTIFICATION ON PENALIZED SPLINE REGRESSION MODELING FOR POVERTY GAP INDEX IN JAVA

Anggita Rizky Fadilah[1], Anwar Fitrianto[2*], I Made Sumertajaya[3]

[1,2,3]Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University
Dramaga Campus, Bogor 6680, West Java, Indonesia

Corresponding author's e-mail: [2] anwarstat@gmail.com*

Abstract. *Java is one of the islands in Indonesia which has good establishment acceleration. Even though economic growth was good, poverty is still a serious problem. Three of six provinces, including DI Yogyakarta, Central Java, and East Java still have poverty rates above national rates in March 2020. This problem indicates that an imbalance in poverty happens between those regions. Several regions have extreme conditions or known as outliers. Besides that, poverty gap data have a complex pattern so modeling using a non-parametric approach is suitable. This study aims to build an appropriate model to support the success of poverty alleviation in Java and the identification of outliers was carried out using an adjusted boxplot. The best-penalized regression spline model for Poverty Gap Index in Java Island was obtained by Generalized minimum Cross-Validation (GCV) using optimum smoothing parameter (λ) 0,12 and knot combination (1, 2, 4, 1, 5, 3, and 1) for seven predictor variables. The result shows that penalized spline regression model has a higher R^2 than the OLS regression. The R^2 is obtained 69,10%, so the model is feasible to explain the variability of the poverty gap in Java. Moreover, based on the outliers' identification shows a dependency between outlier in data and residual because some districts/cities are identified as outliers in both.*

Keywords: *adjusted boxplot, outlier, penalized spline regression, Poverty Gap Index.*

Article info:

Submitted: 22nd June 2022 *Accepted: 14th October 2022*

How to cite this article:
A. R. Fadilah, A. Fitrianto, and I M. Sumertajaya, "OUTLIER IDENTIFICATION ON PENALIZED SPLINE REGRESSION MODELING FOR POVERTY GAP INDEX IN JAVA", *BAREKENG: J. Math. & App.*, vol. 16, iss. 4, pp. 1231-1240, Dec., 2022.

1. INTRODUCTION

Poverty is one of the serious problems faced by various countries every year. Poverty is not only about financial incapacity but also lack of empowerment and knowledge as an opportunity to increase income [1]. Until now, the government is still working on poverty alleviation in the national development program as a continuation of the Sustainable Development Goals (SDGs) program [2]. As the island with the largest contribution to Indonesia's Gross Domestic Product (GDP) in 2020, the economic progress on Java Island is also in line with the increasing inequality in the level of welfare among provinces. Three of the six provinces on Java Island still have poverty rates above the national rate, namely the provinces of DI Yogyakarta, Central Java, and East Java as of March 2020. In addition to differences in resources and development processes in each region, Indonesia's population growth is increasing. Developing countries are also at risk of low quality of life problems so inequality is still common [3][4]. This is in line with the research results conducted by [5] that poverty is a multidimensional problem related to social indicators and quality of life.

In general, poverty data patterns do not show a certain relationship or are difficult to determine. One of the regression modeling methods suitable for this data type is non-parametric regression [6]. Non-parametric regression has high flexibility in data patterns so that the regression curve estimation can adjust the behavior of the data without being influenced by the researcher's subjectivity [7]. The principle of non-parametric regression is to estimate the regression function by estimating the function at each point or it is called a local estimation [8]. Penalized spline regression is one type of non-parametric regression that utilizes knot points and smoothing components so that it can adjust the shape of the data pattern. The estimation of this regression model is obtained by minimizing the Penalized Least Square (PLS), an estimation criterion function that combines the Least Square function and the smoothing component [9]. However, the PLS function is not robust against outlier disturbances [10][11].

Economic inequality causes certain districts/cities to have poverty levels that are too small or too large compared to other districts/cities. In addition, most poverty data have highly skewed distributions [12]. These conditions can cause the appearance of outliers. The existence of outliers in the data leads to a non-robust model which causes errors in parameter estimation [13]. The impact is that the resulting interpretation becomes inaccurate, one of which is districts/cities that should have a high poverty rate but become low and vice versa. Based on this background, this study aims to model the Poverty Depth Index in districts/cities of Java Island in 2020 using a penalized spline regression model, followed by identifying outliers. The outlier identification step is needed to find out which districts/cities have the most poverty conditions compared to other districts/cities.

2. RESEARCH METHODS

2.1 Outlier Identification

Outliers are extreme observation values that deviate far from other observation sets, while the extreme value contained in the predictor variables is called a high leverage point [14]. While the vertical outlier is an outlier found in the residual but not in the predictor variable [15]. According to [16], the presence of outliers affects the results of estimating the Ordinary Least Square model, especially on the values and signs of the regression coefficient. This result causes the prediction as far from the actual observation. Such misrepresentations can lead to incorrect conclusions and findings [13]. The first step before the data analysis phase is identifying outliers using the diagnostic method.

One of the graphical methods that can be used in outlier detection is a line box diagram (Tukey boxplot). Tukey boxplot shows the shape of the data distribution visually based on quartile values (Q) and interquartile range (IQR). If the data have a skewed distribution, using the Tukey boxplot will cause misinterpretation. Many data will pass through the value of the upper and lower fences so that it will cause misclassification as outliers [17]. Therefore, the Adjusted Boxplot proposed by [18] emerged a new method. The Adjusted Boxplot method corrects the Tukey boxplot by using a skewness measure that is robust to outliers.

2.2 Spline Penalized Regression

Penalized spline regression contains points representing the changes in curve behavior or knots set by the smoothing parameter (λ). The smoothing parameter (λ) has a role as the controller of the balance of suitability of the curve to the data and the smoothness of the curve. According to [9], suppose $\boldsymbol{Y}$ is response variable and $\boldsymbol{X}$ is vector of predictor variables, the penalized spline regression model is expressed with

$$f(x) = \beta_0 + \sum_{j=1}^{d}\left[\sum_{l=1}^{p}\beta_j x_j^l + \sum_{m=1}^{t}\beta_{mj}\left(x_j - k_{mj}\right)_+^l\right] \quad (1)$$

with $\left(x_j - k_{mj}\right)_+^l$ as segmented function

$$\left(x_j - k_{mj}\right)_+ = \begin{cases}\left(x_j - k_{mj}\right)_+ & , x_j \geq k_{mj} \\ 0 & , x_j < k_{mj}\end{cases} \quad (2)$$

where d is the number of predictor variables, p is the polynomial ordo, t is the number of knots, and k_{mj} is the position knots of j[th] predictor variable. The estimation of penalized spline regression was obtained using Penalized Least Squares (PLS) function, which is presented as follows

$$L = \sum_{i=1}^{n}(y_i - f(\boldsymbol{X};\boldsymbol{\beta}))^2 + \sum_{j=1}^{d}\left(\lambda_j \sum_{m=1}^{t}\beta_{(p+m)j}{}^2\right) \quad (3)$$

The estimate of $\boldsymbol{\beta}$ is obtained by minimizing the residual function above. The $\boldsymbol{\beta}$ has the following expression

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X} + \boldsymbol{D_\lambda})^{-1}\boldsymbol{X^T y} \quad (4)$$

and $\boldsymbol{D_\lambda}$ is a diagonal matrix used to indicate the penalized coefficient, which can be written as follows

$$\boldsymbol{D_\lambda} = \begin{bmatrix}\boldsymbol{D_{\lambda_1}} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D_{\lambda_d}}\end{bmatrix}, \boldsymbol{D_{\lambda_d}} = \begin{bmatrix}0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \lambda_d & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \lambda_d\end{bmatrix} = diag\left(\boldsymbol{0}_{((p+1)\times 1)}, \boldsymbol{\lambda_{d_{(k\times 1)}}}\right) \quad (5)$$

2.3 Determination of Optimum Knots and Smoothing Parameter (λ)

According to [19], one of the numerical approaches used to determine the position of the knot in each predictor variable is by placing the knot (k) evenly so that the distance between knots is the same based on the following function

$$k_i = \frac{(b-a)\times i}{(t+1)} + a, \quad i = 1,2,\dots,t \quad (6)$$

where k_i is knot position, a and b are the smallest and largest observations, respectively. In addition to getting a combination of knot points, it is also necessary to determine the optimum smoothing parameter (λ) based on minimum Generalized Cross-Validation (GCV). According to [20], GCV has the capability and efficiency of computational calculations. GCV is obtained from the formula as follows

$$GCV(\lambda) = \frac{n^{-1}\left(\boldsymbol{y}-\hat{\boldsymbol{f}}(\boldsymbol{x})\right)^T\left(\boldsymbol{y}-\hat{\boldsymbol{f}}(\boldsymbol{x})\right)}{(n^{-1}\,tr[\boldsymbol{I}-\boldsymbol{H}(\lambda)])^2} \quad (7)$$

and $\boldsymbol{H}(\lambda)$ is hat matrix can be expressed as follows $\boldsymbol{H}(\lambda) = \boldsymbol{X}(\boldsymbol{X^T X} + \boldsymbol{D_\lambda})^{-1}\boldsymbol{X^T}$.

2.4 Data Description

The data used in this research were secondary data obtained from the Central Bureau of Statistics (BPS). The number of observations is 119 regencies/municipalities, in which the Poverty Gap Index (P1) roles as the response variable (Y) and the other seven variables a role as predictor variables (X). Table 1 gives a brief detail for each predictor.

Tabel 1. The Description of Predictor Variables

Variable	Description	Unit
X_1	Life expectancy (LIFE)	Year
X_2	Mean Years School (EDUC)	Year
X_3	Gross Regional Domestic Product growth rate (GRDP)	Percentage (%)
X_4	Poor people with the highest education level is senior high school (GRADUATE)	Percentage (%)
X_5	Open Unemployment Rate (OPEN)	Percentage (%)
X_6	Population growth rate (POPULATION)	Percentage (%)
X_7	Recipient household of Non-Cash Food Assistance (HOUSEHOLD)	Percentage (%)

2.5 Data Analysis Procedures

The penalized spline regression model for Poverty Gap Index data is P1 = β_0 + β_1 LIFE + β_2 EDUC + β_3 GRDP + β_4 GRADUATE + β_5 OPEN + β_6 POPULATION + β_7 HOUSEHOLD + ε. The R software is used to assist in the computation. The following steps were used to analyze the Poverty Gap Index data. First, exploration data is performed to determine the characteristics of each province in Java Island and detect outliers in data using the adjusted boxplot. Then, OLS regression and penalized spline regression is performed on the data. On penalized spline regression modeling, the optimum knots and smoothing parameter (λ) are determined based on minimum GCV. Afterward, detection of outlier in residual model is conducted to confirm the presence of outlier.

3. RESULTS AND DISCUSSION

3.1 Data Exploration

A review of the Poverty Gap Index in Java is needed to know the poverty depth in the area. Based on the summary statistics in Table 2, the highest average Poverty Gap Index is in DI Yogyakarta, while the lowest is in Banten. Based on the standard deviation value, East Java and DI Yogyakarta are higher than other provinces. This shows that the poverty gap in East Java and DI Yogyakarta are quite diverse and need to be considered for equitable economic development. Otherwise, the smallest standard deviation of the index is in Banten. It means the expenditure gap between regencies/municipalities in Banten is smaller than in other provinces.

Tabel 2. Summary of Statistics of Poverty Gap Index by the province in Java at 2020

Statistics	Province					
	Banten	DKI Jakarta	West Java	Central Java	DI Yogyakarta	East Java
Minimum	0,310	0,35	0,290	0,530	1,190	0,590
Mean	0,687	0,853	1,161	1,527	2,062	1,692
Median	0,655	0,605	1,080	1,370	1,850	1,460
Maximum	1,140	2,100	2,410	3,010	3,220	4,330
Standard Deviation	0,317	0,648	0,490	0,665	0,867	0,913

Figure 2 shows outlier detection results in predictor variables (X) and a response variable (Y) using the adjusted boxplot approach. Based on this figure, most variables have outliers. Only variable of Open Unemployment Rate (X_5) and recipient household of non-cash food assistance (X_7) does not have an outlier. Those outliers are classified into 2 groups, upper outliers (extremely high) and lower outliers (extremely low).

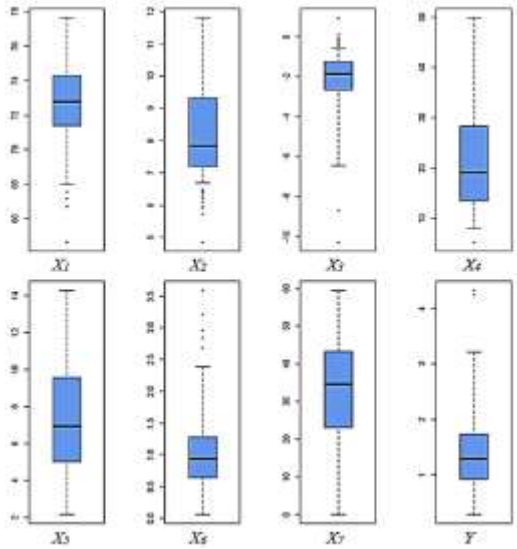

Figure 1. Adjusted boxplot of each explanatory variable (X) and response variable (Y)

Table 3 shows regencies/municipalities included as outliers in each variable. Based on Table 3, most regencies/municipalities that were classified as lower outliers in Life Expectancy (X_1) are located in Banten province. On variable Mean Years School (X_2), some lower outliers were regencies/municipalities are located in East Java province. On the Poverty Gap Index (Y) variable, Bangkalan and Sumenep district have quite severe poverty conditions compared to other regencies/cities. Those two districts were also identified as lower outliers on the Mean Years School (X_2) variable.

Table 3 Outlier Identification of Predictor Variable and Response Variable

Variable	Upper Outliers		Lower Outliers	
	Frequency	District/City	Frequency	District/City
X_1	0	-	7	Serang, Pandeglang, Cilegon city, Bondowoso, Probolinggo, Lebak, and Pamekasan.
X_2	0	-	12	Sampang, Sumenep, Bondowoso, Bangkalan, Probolinggo, Brebes, Indramayu, Lebak, Lumajang, Pemalang, Situbondo, and Jember.
X_3	11	Brebes, Jakarta Selatan city, Bogor, Bojonegoro, Sampang, Demak, Ciamis, Pangandaran, Kuningan, Majalengka, and Banjar city.	2	Cilacap and Tangerang city.
X_4	0	-	1	Pangandaran
X_5	0	-	0	-
X_6	0	-	5	Bekasi, Depok city, Tangerang Selatan city, Tangerang, and Kepulauan Seribu.
X_7	0	-	0	-
Y	2	Bangkalan and Sumenep	0	-

3.2 Penalized Spline Modelling

The formation of a penalized regression spline model involves knot (k), smoothing parameter (λ), and order of polynomial (p). In this study, the polynomial degree (p) used is 1 (linear). The knot points (k) were determined by combining up to 5 knots of each explanatory variable and optimum smoothing parameter (λ) using an iterative process from 0.01 to 5. The optimum knot points (k) and smoothing parameter (λ) were

gained with minimum GCV. Table 4 contains the GCV value based on the combination of the number of knot points along with the smoothing parameter (λ).

Table 4. GCV Value Based on Smoothing Parameter (λ) and Knots (k)

Ordo	λ	The number of knots (k) in each predictor variable							GCV
		X_1	X_2	X_3	X_4	X_5	X_6	X_7	
1	0.01	1	1	1	1	1	1	1	0.3491
1	0.01	2	1	1	1	1	1	1	0.3574
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	0.12	5	1	4	1	5	3	1	0.3240
1	**0.12**	**1**	**2**	**4**	**1**	**5**	**3**	**1**	**0.2832**
1	0.12	2	2	4	1	5	3	1	0.2925
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	5	4	5	5	5	5	5	5	0.3418
1	5	5	5	5	5	5	5	5	0.3475

Based on Table 4, the optimum of smoothing parameter (λ) is obtained 0,12 and knot points (k) were obtained 1, 2, 4, 1, 5, 3, and 1 knots for 7 explanatory variables. The minimum GCV is obtained 0,2832 from iterative numerical calculations. Tables 5 shows knot values in each predictor variable based optimum combination knot previously obtained.

Table 5. Optimal Knots Value (k) for Each Predictor Variable

Variable	Number of Knots	Knot position (k_i)				
		1	2	3	4	5
X_1	1	71,145 years				
X_2	2	7,17 years	9,49 years			
X_3	4	-8,036%	-5,792%	-3,548%	-1,304%	
X_4	1	27,58%				
X_5	5	4,182%	6,203%	8,225%	10,247%	12,268%
X_6	3	0,943%	1,825%	2,708%		
X_7	1	29,825%				

The penalized spline regression model for Poverty Gap Index in Java Island in 2020 expressed as follows:

$$\widehat{P1} = -0{,}114 + 0{,}133\ LIFE - 0{,}161(LIFE - 71{,}145)_+ - 0{,}949\ EDUC + 0{,}701(EDUC - 7{,}170)_+ - 0{,}374(EDUC - 9{,}490)_+ + 0{,}165\ GRDP + 0{,}573(GRDP + 8.036)_+ - 0{,}988(GRDP + 5.792)_+ + 0{,}274(GRDP + 3.548)_+ + \cdots + 0.005\ HOUSEHOLD + 0{,}011(HOUSEHOLD - 29.825)_+$$

This model above can be expressed for each predictor variable. Function for variable of Life expectancy (X_1) presented as follows.

$$\hat{f}_1(x_1) = 0{,}133\ LIFE - 0{,}161(LIFE - 71{,}145)_+ = \begin{cases} 0{,}133\ LIFE & LIFE < 71{,}145 \text{ years} \\ 11{,}432 - 0{,}028\ LIFE & LIFE \geq 71{,}145 \text{ years} \end{cases}$$

Based on model above, districts/cities with a life expectancy less than 71,45 years, when their life expectancy increases by 1 year, the poverty gap index will increase by 0,133. Meanwhile, districts/cities with a life expectancy of more than 71,45 years, when their life expectancy increases by 1 year, the poverty gap index will decrease by 0,028. For variable of Mean Years School (X_2) presented as follows

$$\hat{f}_2(x_2) = -0{,}949\ EDUC + 0{,}701(EDUC - 7{,}170)_+ - 0{,}374(EDUC - 9{,}490)_+ = \begin{cases} -0{,}949\ EDUC & EDUC < 7{,}17 \text{ years} \\ -5{,}025 - 0.248\ EDUC & 7{,}17 \leq EDUC < 9{,}49 \text{ years} \\ -1{,}476 - 0.622\ EDUC & EDUC \geq 9{,}49 \text{ years} \end{cases}$$

Based on the model above, districts/cities with average length of time studied less than 7,17 years, when their average length of time studied increases by 1 year, the poverty gap index will decrease 0,949. If districts/cities with average length of time studied between 7,17 years to 9,49 years, when their average length of time studied increases by 1 year, the poverty gap index will decrease by 0,248. Meanwhile, districts/cities with average length of time studied more than 9,49 years, when their average length of time studied increases by 1 year, the poverty gap index will decrease by 0,622. For variable of growth rate of Gross Regional Domestic Product (X_3) presented as follows

$$\hat{f}_3(x_3) = 0{,}165\ GRDP + 0{,}573(GRDP + 8{,}036)_+ - 0{,}988(GRDP + 5{,}792)_+ + 0{,}274(GRDP + 3{,}548)_+ 0{,}042(GRDP + 1{,}304)_+$$

$$= \begin{cases} 0{,}165\ GRDP & GRDP < -8{,}036\% \\ 4{,}606 + 0{,}738\ GRDP & -8{,}036\% \leq GRDP < -5{,}792\% \\ -1{,}115 - 0{,}250\ GRDP & -5{,}792\% \leq GRDP < -3{,}548\% \\ -0{,}142 + 0{,}025\ GRDP & -3{,}548\% \leq GRDP < -1{,}304\% \\ -0{,}087 + 0{,}066\ GRDP & GRDP \geq -1{,}304\% \end{cases}$$

Based on model above, districts/cities with Gross Regional Domestic Product growth rate less than -8,036%, when their Gross Regional Domestic Product growth rate increases by 1%, the poverty gap index will increase by 0,165. If districts/cities with Gross Regional Domestic Product growth rate between -8,036% to -5,792%, when their Gross Regional Domestic Product growth rate increases by 1%, the poverty gap index will increase 0,738. If districts/cities with Gross Regional Domestic Product growth rate between -5,792% to -3,548%, when their Gross Regional Domestic Product growth rate increases by 1%, the poverty gap index will decrease 0,250. If districts/cities with Gross Regional Domestic Product growth rate between -3,548% to -1,304%, when their Gross Regional Domestic Product growth rate increases by 1%, the poverty gap index will increase 0,025. Meanwhile, districts/cities with Gross Regional Domestic Product growth rate more than equal -1,304%, when their Gross Regional Product growth rate increases by 1%, the poverty gap index will increase 0,066. For variable of percentage of poor people with highest education is senior high school (X_4) presented as follows

$$\hat{f}_4(x_4) = 0{,}034\ GRADUATE + 0{,}006(GRADUATE - 27{,}58)_+$$

$$= \begin{cases} 0{,}034\ GRADUATE & GRADUATE < 27{,}58\% \\ -0{,}173 + 0{,}041\ GRADUATE & GRADUATE \geq 27{,}58\% \end{cases}$$

Based on model above, districts/cities with a percentage of poor people with the highest education is senior high school less than 27,58%, when their percentage of poor people with the highest education is senior high school increases by 1%, the poverty gap index will increase 0,034. Meanwhile, districts/cities that have percentage of poor people with highest education is senior high school more than equal 27,58%, when their percentage of poor people with highest education is senior high school increases by 1%, the poverty gap index will increase 0,041. For variable of Open Unemployment Rate (X_5) presented as follows

$$\hat{f}_5(x_5) = -0{,}437\ OPEN + 0{,}422(OPEN - 4{,}182)_+ - 0{,}080(OPEN - 6{,}203)_+ - 0{,}175(OPEN - 8{,}225)_+ + 0{,}262(OPEN - 10{,}247)_+ - 0{,}652(OPEN - 12{,}268)_+$$

$$= \begin{cases} -0{,}090\ OPEN & OPEN < 4{,}182\% \\ 1{,}137 + 0{,}353\ OPEN & 4{,}182\% \leq OPEN < 6{,}203\% \\ -0{,}649 - 0{,}704\ OPEN & 6{,}203\% \leq OPEN < 8{,}225\% \\ -0{,}707 - 0{,}084\ OPEN & 8{,}225\% \leq OPEN < 10{,}247\% \\ -0{,}017 + 0{,}009\ OPEN & 10{,}247\% \leq OPEN < 12{,}268\% \\ -2{,}673 - 0{,}285\ OPEN & OPEN \geq 12{,}268\% \end{cases}$$

Based on the model, districts/cities with open unemployment rate is less than 4,182%, when their open unemployment rate increases by 1%, the poverty gap index will increase by 0,353. If districts/cities with open unemployment rate between 4,182% to 6,203%, when their open unemployment rate increase by 1%, the poverty gap index will decrease by 0,704. If districts/cities with open unemployment rate between 8,225% to

10,247%, when their open unemployment rate increases by 1%, the poverty gap index will decrease 0,084. If districts/cities with open unemployment rate between 10,247% to 12,268%, when their open unemployment rate increase by 1%, the poverty gap index will increase 0,009. Meanwhile, in districts/cities with more than 12,268%, when their open unemployment rate increases by 1%, the poverty gap index will decrease by 0,285. For variable of population growth rate (X_6) presented as follows

$$\hat{f}_6(x_6) = -0{,}129\ POPULATION + 1{,}145(POPULATION - 0{,}943)_+ - 1{,}507(POPULATION - 1{,}825)_+ - 0{,}526(POPULATION - 2{,}708)_+$$

$$= \begin{cases} -0{,}129\ POPULATION & POPULATION < 0{,}943\% \\ -1{,}079 + 1{,}016\ POPULATION & 0{,}943 \leq POPULATION < 1{,}825\% \\ 1{,}671 - 0{,}491\ POPULATION & 1{,}825 \leq POPULATION < 2{,}708\% \\ 3{,}095 - 1{,}017\ POPULATION & POPULATION \geq 2{,}708\% \end{cases}$$

Based on the model above, districts/cities with a population growth rate of less than 0,943%, when their population growth rate increases by 1 year, the poverty gap index will decrease by 0,129. If districts/cities with population growth rates are between 0,943% to 1,825%, when their population growth rate increases by 1%, the poverty gap index will increase to 0,016. If districts/cities have population growth rates between 1,825% to 2,708%, when their population growth rate increases by 1%, the poverty gap index will decrease by 0,491. Meanwhile, in districts/cities with a population growth rate of more than or equal to 2,708%, when their population growth rate increases by 1%, the poverty gap index will decrease 1,017. For the variable of the percentage of Non-Cash Food Assistance's recipient household (X_7) presented as follows

$$\hat{f}_7(x_7) = 0{,}005\ HOUSEHOLD + 0{,}011(HOUSEHOLD - 29{,}825)_+$$

$$= \begin{cases} 0{,}005\ HOUSEHOLD & HOUSEHOLD < 29{,}825\% \\ -0{,}320 + 0{,}016\ HOUSEHOLD & HOUSEHOLD \geq 29{,}825\% \end{cases}$$

Based on the model above, districts/cities with a percentage of Non-Cash Food Assistance's recipient households less than 29,825%, when their percentage of Non-Cash Food Assistance's recipient household increases by 1%, the poverty gap index will increase 0,005. Meanwhile, districts/cities with a percentage of Non-Cash Food Assistance's recipient household of more than 29,825%, when their percentage of Non-Cash Food Assistance's recipient household increases by 1%, the poverty gap index will increase 0,016.

Table 5 summarizes the results of OLS regression and penalized spline regression in Poverty Gap Index data. In general, penalized spline regression as a non-parametric approach has a better result than OLS regression based on minimum R^2 and MSE. It indicates that penalized spline regression model can fit the pattern of the poverty gap curve in Java Island quite well. Based on penalized spline regression model above, the value of R^2 is obtained 0,6910. It means that life expectancy (X_1), mean years school (X_2), the growth rate of Gross Regional Domestic Product (X_3), the percentage of poor people with the highest education is senior high school (X_4), Open Unemployment Rate (X_5), population growth rate (X_6), and percentage of Non-Cash Food Assistance's Recipient household (X_7) can explain the variety of the Poverty Gap Index in Java is 69,10%.

Table 5. Summary of OLS Regression and Penalized Spline Regression

Model	R^2	*MSE*
OLS regression	0,4745	0,3147
Penalized Spline Regression	0,6910	0,1851

3.3 Identification of Outlier

Detection of outliers on residual of the penalized spline regression model using an adjusted boxplot. Based on Figure 2, the median value of residual is obtained 0.007. The figure also shows that 7 districts/cities are the upper outliers in the residual in the model. Those regencies/cities include Kuningan district in West Java, Rembang district in Central Java, Kulon Progo district in DI Yogyakarta, and other 4 districts in East Java (Bangkalan, Probolinggo, Ngawi, and Lamongan). Some of these regencies/cities also identified as outliers in predictor variables (*X*) and response variables (*Y*). Kuningan district was previously identified as

a lower outlier in the growth rate of Gross Regional Domestic Product (X_3). Then, the Bangkalan district also identified as upper outlier in Poverty Gap Index (Y) and as lower outlier in mean years school (X_2). Last, the Probolinggo district were also found as lower outlier in life expectancy (X_1) and mean years school (X_2).

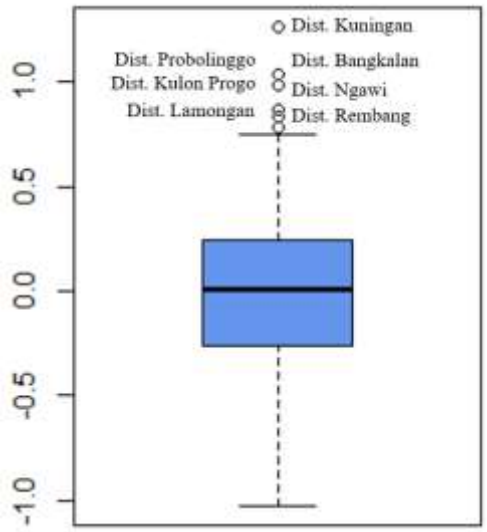

Figure 2. The Adjusted Boxplot of Residual of Penalized Spline Regression

4. CONCLUSION

The best penalized spline regression model was obtained with optimum smoothing parameter (λ) is 0,12 and knot combination 1, 2, 4, 5, 3, and 1 knots for 7 predictor variables. The penalized spline regression also has smallest R^2 and MSE than OLS regression. It shows that penalized spline regression is more flexible with poverty curve behavior than OLS regression. Moreover, it was found that some outliers were in predictor variables (X) and response variable (Y). In residual of penalized spline regression model also found 7 districts/cities as the upper outliers, where three of them also found in predictor variables (X) and response variable (Y). Those outliers were also inseparable from the outliers found in both predictor variables (X) and response variable (Y).

REFERENCES

[1] R. Gouunder, and Z. Xing, "Impact of education and health on poverty reduction: Monetary and non-monetary evidence from Fiji," *Economic Modelling*, vol. 29, no. 3, pp. 787–794, 2020, doi: 10.1016/j.econmod.2012.01.018.

[2] A. M. Arsani, B. Ario, and A. F. Ramadhan, "Impact of Education on Poverty and Health : Evidence from Indonesia," *Economics Development Analysis Journal*, vol. 9, no. 1, pp. 87–96, 2020, doi: 10.15294/edaj.v9i1.34921.

[3] A. M. Ginting, "Pengaruh Ketimpangan Pembangunan Antarwilayah Terhadap Kemiskinan Di Indonesia 2004-2013," *Pusat Penelitian - Badan Keahlian DPR RI*, vol. 20, no. 1, pp. 45–58, 2015, [Online]. Available: http://news.bisnis.com/read/20140721/15/244928/

[4] H. Hill, "What's happened to poverty and inequality in indonesia over half a century?," *Asian Development Review*, vol. 38, no. 1, pp. 68–97, 2021, doi: 10.1162/adev_a_00158.

[5] B. Žmuk, "Quality of Life Indicators in Selected European Countries: Hierarchical Cluster Analysis Approach," *Croatian Review of Economic, Business and Social Statistics*, vol. 1, no. 1–2, pp. 42–54, 2015, doi: 10.1515/crebss-2016-0004.

[6] B. Lestari, Fatmawati, I. N. Budiantara, and N. Chamidah, "Estimation of Regression Function in Multi-Response Nonparametric Regression Model Using Smoothing Spline and Kernel Estimators," *Journal of Physics: Conference Series*, vol. 1097, no. 1, 2018, doi: 10.1088/1742-6596/1097/1/012091.

[7] D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric Regression*. New York: Cambridge University Press, 2003.

[8] D. J. Henderson and A. C. Souto, "An Introduction to Nonparametric Regression for Labor Economists," *Journal of Labor Research*, vol. 39, no. Nov, pp. 355–382, 2018, doi: 10.1007/s12122-018-9279-6.

[9] A. Islamiyati, N. Sunusi, A. Kalondeng, F. Fatmawati, and N. Chamidah, "Use of two smoothing parameters in penalized spline estimator for bi-variate predictor non-parametric regression model," *Journal of Sciences, Islamic Republic of Iran*, vol. 31, no. 2, pp. 175–183, 2020, doi: 10.22059/JSCIENCES.2020.286949.1007435.

[10] B. Wang, W. Shi, and Z. Miao, "Comparative Analysis for Robust Penalized Spline Smoothing Methods," *Mathematical Problems in Engineering*, vol. 2014, no. July, 2014, doi: 10.1155/2014/642475.
[11] I. Kalogridis and S. V. Aelst, "M-type Penalized Splines With Auxiliary Scale Estimation," *Journal of Statistical Planning and Inference*, vol. 212, no. May, pp. 97–113, 2021, doi: 10.1016/j.jspi.2020.09.004.
[12] E. Alvarez, R. M. G. Fernandez, F. J. B. Encomienda, and J. F. Munoz, "The Effect of Outliers on the Economic and Social Survey on Income and Living Conditions," *International Journal of Social, Management, Economics and Business Engineering*, vol. 08, no. 10, pp. 3051–3055, 2014, doi: 10.4236/oalib.1106619.
[13] C. O. Arimie, E. O. Biu, and M. A. Ijomah, "Outlier Detection and Effects on Modeling," *Open Access Library Journal*, vol. 07, no. 09, pp. 1–30, 2020, doi: 10.4236/oalib.1106619.
[14] A. Fitrianto and S. H. Xin, "Comparisons Between Robust Regression Approaches in the Presence of Outliers and High Leverage Points," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 16, no. 1, pp. 243–252, 2022, doi: 10.30598/barekengvol16iss1pp241-250.
[15] A. M. Gad and M. E. Qura, "Regression Estimation in the Presence of Outliers: A Comparative Study," *International Journal of Probability and Statistics*, vol. 5, no. 3, pp. 65–72, 2016, doi: 10.5923/j.ijps.20160503.01.
[16] C. P. Dhakal, "Dealing With Outliers and Influential Points While Fitting Regression," *Journal of Institute of Science and Technology*, vol. 22, no. 1, pp. 61–65, 2017, doi: 10.3126/jist.v22i1.17741.
[17] B. I. Babura, M. B. Adam, A. Rahim, A. Samad, A. Fitrianto, and B. Yusif, "Analysis and Assessment of Boxplot Characters for Analysis and Assessment of Boxplot Characters for," in Journal of Physics, 2018, pp. 1–9. doi: 10.1088/1742-6596/1132/1/012078.
[18] M. Hubert and E. Vandervieren, "An adjusted boxplot for skewed distributions," *Computational Statistics and Data Analysis*, vol. 52, no. 12, pp. 5186–5201, 2008, doi: 10.1016/j.csda.2007.11.008.
[19] L. Ngo and M. P. Wand, "Smoothing with mixed model software," *Journal of Statistical Software*, vol. 9, no. 1978, pp. 1–54, 2004, doi: 10.18637/jss.v009.i01.
[20] R. L. Eubank, *Nonparametric Regression and Spline Smoothing*, *Second Edition*. New York: Marcel Dekker Inc, 1999.