Semiparametric Penalized Spline Regression for Bi-Response Longitudinal Data with An Application in Crime Study

Kosmaryati [a], Mujiati Dwi Kartikasari [b*]

Department of Statistics, Universitas Islam Indonesia
Jalan Kaliurang Km 14.5 Sleman, Yogyakarta, Indonesia 55584

[a] *kosmaryati3103@gmail.com*
[b*] *mujiatikartikasari@uii.ac.id*

Abstract: Semiparametric regression is a combination of parametric regression and nonparametric regression. Parametric regression analysis is used if the shape of curve or function is known, whereas nonparametric regression analysis is used if the curve form or the regression function is unknown. One of the short descriptions of nonparametric regression analysis is the penalized spline. The penalized spline is a segmented polynomial piece where the data characteristic is explained by knots. The advantage of the penalized approach is flexible and able to describe changes in the behavior patterns of functions within a specific subinterval. In addition, the penalized spline approach can be used to cope with or reduce data patterns that experience a sharp increase. This paper explores semiparametric regression method of the penalized spline by using the longitudinal data of bi-response. The advantage in longitudinal data usage is that it can reduce intervariable collinearity so as to produce an efficient estimate. For case study, we use criminal case data in Indonesia. Based on the results of research, the estimation of penalized spline regression model for bi-response longitudinal data is obtained. Then, the estimation of the penalized regression model is applied in case of criminality and obtained regression model with R-square value of 83.18%.

Keywords: semiparametric, penalized spline, longitudinal, bi-response, criminality.

Introduction

Crime is all forms of actions that have an economic and psychological impact and violate applicable regulations as well as social and religious norms, thus causing community arousal and punishments [1]. According to BPS data 2017, the number of criminality incidents in Indonesia tends to increase during the period of 2014-2016 [2]. News of criminality can be easily found in various media. According to Kansil (1994), it is essentially the actions of each individual in committing crimes (criminality) influenced by several factors both internally (determination of fulfillment of economic and labor needs) and external (education and environmental influences) [3]. One of the methods used in identifying factors that affect criminality is regression analysis. Regression analysis is one of the statistical methods used to determine the shape of a statistical model or the relationship between one or more predictor variables with one or more response variables [4].

The study discussing the criminality factor of one of them is Marina and Budiantara (2013) which is a model factor that affects the percentage of criminality in East Java with a semiparametric regression-spline approach [5]. In the regression analysis of several options spline and location of knots is important. Thus, it needs to be counted as many combinations of the many knots from the amount of data to determine the optimal node, and then select the optimal model based on the specific criteria. This takes a long time and if done using software requires great memory. Therefore, it takes an alternative to solve this problem, i.e. with a regression in the spline being punished where the knots are located at a unique quantile point (single) variable predictor value [6]. A regression related study based on the condemned gauges of the spline on bi-response longitudinal data between other studies conducted by Fernandes and Solimun (2016) and Islamiyati et al (2018) [7,8].

Longitudinal data is a combination of cross-section data with time series data, where data obtained from observations of n subjects that are mutually independent with each subject are observed repeatedly in T periods and between observations in the same subject correlate with each other [9]. According to Johan Harlan (2018) using the study of longitudinal data can determine individual changes, it takes a lot of subjects because the observations are repeated, and the estimation is more efficient because it is done every observation [10]. Semiparametric regression is a combination of parametric regression and nonparametric regression, where semiparametric regression includes parametric regression models and nonparametric regression [11].

Based on the explanation described, the author is interested to discuss more about Semiparametric regression model of the spline which is penalized with bi-response longitudinal data in case of criminality in Indonesia in the year 2014-2016.

Materials and Methods

Materials

We have a criminal case data in all provinces in Indonesia from 2014 to 2016. The data is obtained from BPS publication series (Statistical Central Agency) and KEMENDIKBUD (Ministry of Education and Culture). The variables used in the study consist of two types, namely response variable and predictor variable. We explain these variables in Table 1.

Tabel 1 Research Variables

Variable	Variable Name	Definition
Response	The number of crimes against the rights/property of violence (KHMK)	One way of grouping the types of crimes based on criteria of how evil is done, that is by using violence.
	The amount of crime against rights/property without use of violence (KH)	One way of grouping types of crimes based on criteria of how evil is done, that is without the use of violence.
Predictor	Percentage of vulnerable family (PKRB)	Percentage of divorce cases live in each province.
	Minimum wage value by province (UMP)	Standardization or minimum wage threshold provided by a company or an industry against its workers.
	The amount of unemployment (JP)	Population residing in the province that does not have a job (idle).
	The number of dropouts (JPS)	The number students who do not have their studies in high school education.

Methods

The stage of analysis conducted in the study is divided into two parts, namely the stage in obtaining the estimation of Semiparametric regression model of the penalized spline bi-response longitudinal data and the stage in modelling the number of crimes against the rights/property using violence and without the use of violence in Indonesia with the variables that affect it.

The stage in obtaining estimation semiparametric regression model of the penalized spline bi-response longitudinal data is describe as follows.

1. Define Semiparametric regression model of bi-response longitudinal data, as follows:

$$y_{it}^{(r)} = \beta_0^{(r)} + \sum_{v=1}^{p} \beta_v^{(r)} (x_{vit}) + \sum_{w=1}^{q} f_w^{(r)} (z_{wit}) + \varepsilon_{it}^{(r)}$$

2. Use a regression approach of spline on nonparametric component of the d-order with knots k, as follows:

$$f_w^{(r)} (z_{wit}) = \sum_{j=0}^{d_{rw}} \alpha_{wj}^{(r)} z_{wit}^{j} + \sum_{h=1}^{m_w} \theta_{wh}^{(r)} (z_{wit} - k_{wh})_+^{d_{rw}}$$

then the model form semiparametric regression of bi-response longitudinal data, can be expressed as follows:

$$y_{it}^{(r)} = \beta_0^{(r)} + \sum_{v=0}^{p} \beta_v^{(r)} x_{vit} + \sum_{w=1}^{q} \left(\sum_{j=0}^{d_{rw}} \alpha_{wj}^{(r)} z_{wit}^{j} + \sum_{h=1}^{m_w} \theta_{wh}^{(r)} (z_{wit} - k_{wh})_+^{d_{rw}} \right)$$

3. Declare semiparametric regression model of bi-response longitudinal data in matrix form, as follows:

$$\underset{\sim}{y} = X \underset{\sim}{\beta} + Z \underset{\sim}{\Phi} + \underset{\sim}{\varepsilon}$$

4. Declare a penalized estimate that minimizes the Penalized Least Square (PLS) function for the response variable, where the response variable $\underset{\sim}{y}^*$ is obtained from the following equation:

$$\underset{\sim}{y} - X\underset{\sim}{\beta} = Z\underset{\sim}{\Phi}$$

$$\underset{\sim}{y}^* = Z\underset{\sim}{\Phi}$$

5. Estimate the model by minimizing the Penalized Weighted Least Square (PWLS) criteria as follows:

$$L = (2nT)^{-1}\left(\underset{\sim}{y}^* - Z\underset{\sim}{\Phi}\right)' W\left(\underset{\sim}{y}^* - Z\underset{\sim}{\Phi}\right) + \lambda\underset{\sim}{\Phi}' D\underset{\sim}{\Phi}$$

   with **W** is a weighted matrix which is the inverse of the matrix variances Covariansi error for response 1 and Response 2. Estimating $\underset{\sim}{\Phi}$ is done by defferentiation of **L** against $\underset{\sim}{\Phi}$.

6. Estimate $\underset{\sim}{\beta}$ via the K function debiting using the WLS method which minimizes the following functions:

$$K = \left(\underset{\sim}{y} - X\underset{\sim}{\beta} - A\left(\underset{\sim}{y} - X\underset{\sim}{\beta}\right)\right)'\left(\underset{\sim}{y} - X\underset{\sim}{\beta} - A\left(\underset{\sim}{y} - X\underset{\sim}{\beta}\right)\right)$$

7. Getting matrix hat for parametric components ($A_{\text{parametric}}$) and for Nonparametric ($A_{\text{nonparametric}}$) components.
8. Declare an $A_{\text{semiparametric}}$ matrix hat = $A_{\text{parametric}}$ + $A_{\text{nonparametric}}$ to be able to calculate the generalized Cross Validation (GCV) value.

The stage in modelling the number of crimes against the rights/property using violence and without the use of violence in Indonesia with the variables that affect it is described as follows.

1. Input and describe the characteristics of the number of crimes against the rights/property using violence and without the use of violence in Indonesia and the factors that influence it.
2. Create scatterplot between response variables with each predictor variable to find out the pattern data form.Menguji korelasi antara variabel respon pertama dengan respon kedua.
3. Specify the order combination and the number point knots.
4. Do a heteroskedastisity test of the variance of the residual covariance by using Glesjer test, in case of homoskedastisity then calculate the matrix variances Covariansi from error in response 1 and error in response 2, so it is obtained, $\sigma_{11}, \sigma_{12}, \sigma_{21}$ and σ_{22} with $\sigma_{12} = \sigma_{21}$. Whereas, in case of hetereskedastisity cases count matrix variances covariansi from errors on each subject to-i; i= 1, 2, …, n at the t time; t = 1, 2, ..., T response 1 and error in response 2, so obtained, $\sigma_{11(it)}, \sigma_{12(it)}, \sigma_{21(it)}$ and $\sigma_{22(it)}$ with $\sigma_{12(it)} = \sigma_{21(it)}$.
5. Form diagonal matrix of $\sigma_{11}, \sigma_{12}, \sigma_{21}$ and σ_{22} for the case of homoskedastisity, while for the case of heteroskedastisity form the diagnonal matrix of $\sigma_{11(it)}, \sigma_{12(it)}, \sigma_{21(it)}$ and $\sigma_{22(it)}$.
6. Combine the diagonal matrix of the matrix variances Kovariansi so that the weight of the **W** weighting is obtained.
7. Calculate the estimated value of y.
8. Create the observation data plot and the estimated response variable results.
9. Calculate the R-square value.

Results and Discussion

The modelling of criminality of each province in Indonesia in 2014-2016 is carried out using semiparametric regression method of penalized spline. Semiparametric regression is a regression model that contains parametric components and nonparametric components. Determination of parametric and nonparametric components carried out using scatterplot, the scatterplot will provide information about the pattern shape of a regression curve that will be used in modelling. The scatterplot results for the response variables against each predictor variable are shown in Figure 1.

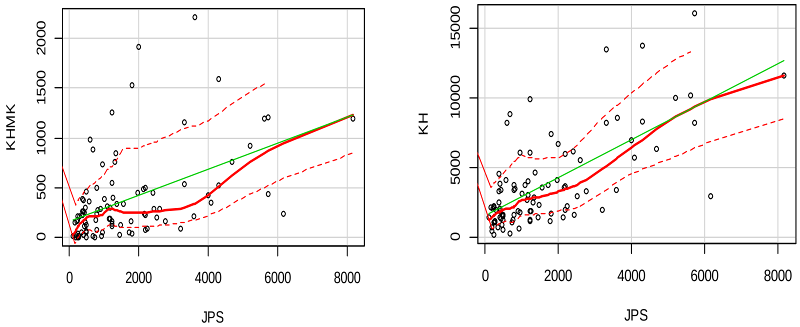

Fig. 1. Scatterplot response variables with JPS predictor variables

Figure 1 illustrates the relationship pattern between the response variables with the JPS predictor variable. Relationship patterns that are formed between the KHMK variables with JPS as well as between the KH variables with the JPS form a positive linear relationship. This is evident from the general plot movements that the higher the case of dropouts then the amount of rights/property crimes is more likely to increase violence.

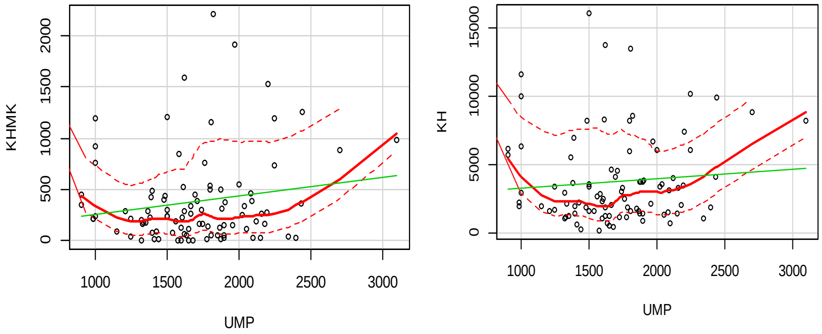

Fig. 2. Scatterplot response variables with UMP predictor variables

Figure 2 illustrates the relationship pattern between the response variables and the UMP predictor. Relationship patterns that are formed between the KHMK variables with the UMP as well as between the KH variables with the UMP tend not to follow a particular pattern. The pattern of the visible relationship also tends to change at certain sub-sub intervals. From the scatterplot, the relationship between the KHMK variable and the UMP and between the KH variables and the UMP will be approached with a nonparametric approach.

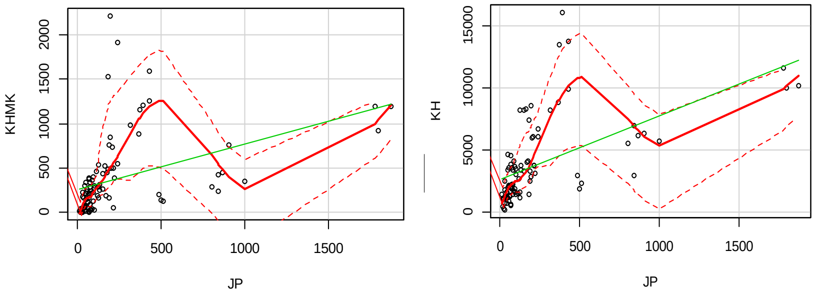

Fig. 3. Scatterplot response variables with JP predictor variables

Figure 3 illustrates the pattern of relationship between response variables with JP predictor Varibell. Relationship patterns that are formed between a variable KHMK with JP and between the variables KH with the JP are likely to undergo a change in behavior or fluctuative occurs at some intervals, it appears that the data patterns at intervals before 500 tends to Increases but at intervals 500 tends to fall and at intervals 1000 tend to increase so as to be approached with a nonparametric approach.

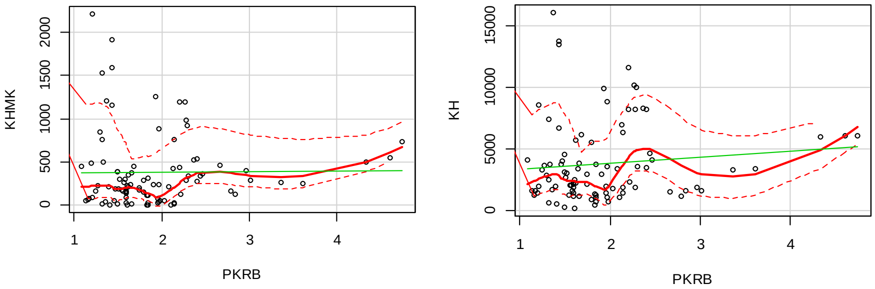

Fig. 4. Scatterplot response variables with PKRB predictor variables

Figure 4 illustrates the relationship pattern between the response variables with the PKRB predictor Varibell. Relationship patterns that are formed between the KHMK variables with the PKRB and between the KH variables and the PKRB are less likely to follow a particular pattern. The pattern of the visible relationship also tends to change at certain sub-sub intervals. From the scatterplot, the relationship between the KHMK variable and the PKRB and between the KH variable and the PKRB will be approached with a nonparametric approach.

The variables approximated by a parametric approach are symbolized by x and the variables approximated with a nonparametric approach are symbolized by Z. The full description list of each predictor variable can be seen in the Table 2.

Table 2. Results of Determining Parametric and Nonparametric Components

Variable	Approaches	Variable symbols
JP	Parametrik	x
PKRB	Nonparametrik	z_1
JPS		z_2
UMP		z_3

The next step is to know the correlation between the first response variable KHMK with the second response variable is KH using the following hypothesis:

$H_0 : \rho = 0$ (There is no correlation between KHMK and KH)

$H_1 : \rho \neq 0$ (There is a correlation between KHMK and KH)

Based on the results of the Pearson Correlation test, a correlation value of 0.796 and P-value of 0.000 < α (= 0.05) is obtained. Then it can be taken the decision to reject H_0 so that there is a correlation between the first response variable KHMK with the second response variable is KH. After the second known response KHMK and KH fulfilled the assumption of each other, the next step is to determine the combination of order and the optimal number of point knots using GCV criteria.

Determination of the combination of the optimal polynomial order in the first response and the second response as well as the optimal number of points of knots by obtaining the optimal Lambda value on each nonparametric component predictor variable based on Generalized Cross criteria Validation (GCV) minimum. Combination of polynomial order on first response and second response, number of knots and knots, as well as optimal lambda values based on minimum GCV criteria in the first nonparametric Predictor namely UMP are shown in Table 3.

Table 3. Combination of order and nonparametric points of the UMP predictor

Number of knots	Point knot	Order combinations		GCV Minimum	Lambda
		Respon 1	Respon 2		
1	1670	**1**	**1**	**11582085**	**1000**
		1	2	11624465	1000
		2	1	11865922	1000
		2	2	11912885	587
2	1530,35; 1856,67	1	1	11947287	1000
		1	2	11976653	1000
		2	1	12320084	1000
		2	2	12356283	609

At the number of points knots 2 the minimum GCV value is greater than the number of points knots 1 then the increase in the number of knots stopped. So get the smallest minimum GCV value of 11582085 lies in the number of points knots 1 at the point of knots 1670 with the optimal Lambda value 1000 contained in the first response order combination of 1 and the second response order is 1.

The order combination and the number of optimal knot points based on the minimum GCV criteria in the second nonparametric Predictor is JP displayed in Table 4.

Table 4. Combination of order and nonparametric points of the JP predictor

Number of knots	Point knot	Order combinations		GCV Minimum	Lambda
		Respon 1	Respon 2		
1	97,75	1	1	7114863	82
		1	2	7288581	1000
		2	1	7303924	39
		2	2	7568947	999
2	70,55; 177,89	1	1	6441461	405
		1	2	5747333	1000
		2	1	6734309	248
		2	2	7213571	1000
3	57,04; 97,75; 207,78	1	1	6392932	611
		1	2	6752422	1000
		2	1	6814677	390
		2	2	7245007	1000
4	49,51; 76,02; 132,02; 240,34	1	1	6690038	1000
		1	2	6690038	1000
		2	1	6804927	673
		2	2	7292732	1000
5	46,86; 70,55; 97,75; 177,89; 370,52	1	1	6117426	1000
		1	2	5892843	1000
		2	1	6694048	1000
		2	2	6474515	1000
6		1	1	6361804	1000
		1	**2**	**5741019**	**1000**

Number of knots	Point knot	Order combinations		GCV Minimum	Lambda
		Respon 1	Respon 2		
	44,46; 63,16; 79,87; 126,93; 192,47; 423,35	2	1	7037982	1000
		2	2	6382153	1000
7	40,29; 57,04; 73,18; 97,75; 151,7; 207,78; 456,58	1	1	6512780	1000
		1	2	5766912	1000
		2	1	7288560	1000
		2	2	6493437	1000

According to Table 4 is found that at the number of points knots 7 minimum GCV value is greater than the number of point knots 6 The addition of the number of knots is stopped. So get the smallest minimum GCV value of 5741019 lies in the number of Knots 6 point at a point of knots 44.46; 63.16; 79.87; 126.93; 192.47; 423.35 with the optimal Lambda value 1000 found in the first response order combination is 1 and the second response order is 2.

The order combination and the number of optimal knot points based on the minimum GCV criteria in the second nonparametric Predictor is PKRB displayed in Table 5.

Table 5. Combination of order and nonparametric points of the PKRB predictor

Number of knots	Point knot	Order combinations		GCV Minimum	Lambda
		Respon 1	Respon 2		
1	1,74	**1**	**1**	**11698024**	**1000**
		1	2	11952288	1000
		2	1	11961252	1000
		2	2	12224389	1000
2	1,58; 1,97	1	1	11698039	1000
		1	2	11952296	1000
		2	1	11961262	1000
		2	2	12224392	1000

Based on Table 5 it is found that at the number of points knots 2 the minimum GCV value is greater than the number of points knots 1 then the number of knots is stopped. So get the smallest minimum GCV value of 11698024 lies in the number of points knots 1 at the point of knots 1.74 with the optimal Lambda value 1000 contained in the first response order combination of 1 and the second response order is 1. After all the combination of order and number of knots is known each nonparametric predictor so that the residual value from the first response and the second response will be done in the next Test heteroskedastisity/ homoskedastisity between the two Residual.

Test analytic used to test heteroskedastisity in the matrix of variances covariance is the test of Glesjer with the following hypotheses:

$H_0 = \sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2 = \sigma^2$ (Homoskedastisitas)

H_0 = minimal $\sigma_i^2 \neq \sigma^2, i = 1,2,...,n$ (Heteroskedastisitas)

Based on the results of Glesjer test obtained P-value value of 0,062 > α (0.05), then the decision taken is failed to decline H_0 means that the case of homoskedastisity occurs.

The next step is to estimate the penalized semiparametric regression model of longitudinal data, which was applied to data criminality in Indonesia in 2014-2016. The estimation of the model is recognized by using a combination of order and the optimal point number of knots, the result of the estimate that is placed as follows:

$$\hat{y}^{(1)} = (-2,0813\times10^{-15}) - 650,5573 + 0,1602x_1 + 0,5044z_1 - 0,8317\times10^{-4}(z_1 - 1670)_+ + 0,1992z_2 -$$

$$0,0009(z_2 - 44,46)_+ - 0,0024(z_2 - 63,16)_+ - 0,0041(z_2 - 79,87)_+ - 0,0110(z_2 - 126,93)_+ -$$

$$1,0219(z_2 - 192,47)_+ - 0,0274(z_2 - 423,35)_+ - 69,2968z_3 + 0,5953\times10^{-4}(z_3 - 1,74)_+$$

Based on the equation result, the interpretation of each variable is as follows:

1. The relationship between JPS and the KHMK variable assuming another constant variable is that every JPS case increase by one unit will increase the case of KHMK by 0.1182 one unit.
2. The deduction function for the first nonparametric Predictor, the UMP, is expressed as follows:

$$f^{(1)}(z_1) = 0{,}5044 z_1 - 0{,}8317 \times 10^{-4}(z_1 - 1670)_+$$

$$f^{(1)}(z_1)\begin{cases} 0{,}5044 z_1 & ; z_1 < 1670 \\ 0{,}1389 + 0{,}5043 z_1 + & ; z_1 \geq 1670 \end{cases}$$

   Based on the equation of the model when the province in Indonesia with the value of UMP less than 1670 thousand increase by one unit, it will increase the case of the KHMK of 0.5044 one unit. Whereas, if the province in Indonesia with a value of UMP more than 1670 thousand increase by one unit, then the case of KHMK will increase by 0.3760 one unit.
3. The deduction function for second nonparametric predictor i.e. JP is expressed as follows:

$$f^{(1)}(z_2) = 0{,}1992 z_2 - 0{,}0009(z_2 - 44{,}46)_+ - 0{,}0024(z_2 - 63{,}16)_+ - 0{,}0041(z_2 - 79{,}87)_+ -$$
$$0{,}0110(z_2 - 126{,}93)_+ - 0{,}0219(z_2 - 192{,}47)_+ - 0{,}0274(z_2 - 423{,}35)_+$$

$$f^{(1)}(z_2) = \begin{cases} 0{,}1992 z_2 & ; \quad 0 \leq z_2 < 44{,}46 \\ 0{,}0400 + 0{,}1902 z_2 & ; \; 44{,}46 \leq z_2 < 63{,}16 \\ 0{,}1916 + 0{,}1878 z_2 & ; \; 63{,}16 \leq z_2 < 79{,}87 \\ 0{,}5191 + 0{,}1837 z_2 & ; \; 79{,}87 \leq z_2 < 126{,}93 \\ 1{,}9153 + 0{,}1727 z_2 & ; 126{,}93 \leq z_2 < 192{,}47 \\ 6{,}1304 + 0{,}1508 z_2 & ; 192{,}47 \leq z_2 < 423{,}35 \\ 17{,}7302 + 0{,}1234 z_2 & ; \quad z_2 \geq 423{,}35 \end{cases}$$

   Based on the equation of the model when the province in Indonesia with the amount of unemployment less than 44.46 thousand increases one unit, it will increase the case of the KHMK of 0.1992 one unit. If the province with unemployment between the value of 44.46 thousand to 63.16 thousand increases one unit, then the case of KHMK will increase by 0.1902 one unit. If the province with the amount of unemployment between the value of 63.16 thousand to 79.87 thousand increased one unit, then the case of KHMK will increase by 0.1878 one unit. If the province with unemployment between the value of 79.87 thousand to 126.93 thousand increases one unit, then the case of KHMK will increase by 0.1837 one unit. If the province with unemployment between the value of 126.93 to 192.47 thousand increases one unit, then the case of KHMK will increase by 0.1727 one unit. If the province with unemployment between the value of 192.47 thousand to 423.35 thousand increases one unit, then the case of KHMK will increase by 0.1508 one unit. However, if the province with unemployment of more than 423.35 thousand Elevgkat one unit, then the case of KHMK will increase by 0.1234 one unit.
4. The deduction function for the third nonparametric predictor i.e. PKRB is stated as follows:

$$f^{(1)}(z_3) = -69{,}2968 z_3 + 0{,}5953 \times 10^{-4}(z_3 - 1{,}74)_+$$

$$f^{(1)}(z_3) = \begin{cases} -69{,}2968 z_3 & ; z_3 < 1{,}74 \\ -0{,}0001 - 69{,}2967 z_3 & ; z_3 \geq 1{,}74 \end{cases}$$

   Based on the equation of the model when the province in Indonesia with the value of PKRB less than 1.74 percent increase by one unit, it will decrease the case of the KHMK of 69.2968 one unit. Whereas, if the province in Indonesia with the value of PKRB more than 1.74 percent increase by one unit, then the case of KHMK will decline by 69.2967 one unit.

The results of the models obtained for the second response can be written as follows:

$$\hat{y}^{(2)} = (1{,}7856 \times 10^{-12}) - 3528{,}518 + 1{,}3368 x_1 + 1{,}9312 z_1 + 0{,}0158(z_1 - 1670)_+ + 25{,}4053 z_2 - 0{,}1871 z_2^2 +$$
$$0{,}0123(z_2 - 44{,}46)_+{}^2 + 0{,}0730(z_2 - 63{,}16)_+{}^2 + 0{,}1439(z_2 - 79{,}87)_+{}^2 + 0{,}2048(z_2 - 126{,}93)_+{}^2 -$$
$$0{,}3480(z_2 - 192{,}47)_+{}^2 + 0{,}1081(z_2 - 423{,}35)_+{}^2 + 128{,}4694 z_3 + 0{,}2334 \times 10^{-4}(z_3 - 1{,}74)_+$$

Based on the result, the interpretation of each variable as follows:

1. The relationship between JPS and the variable KH with the assumption of another constant variable is that every JPS case increase by one unit will increase the case of KH by 1.3368 one unit.
2. The deduction function for the first nonparametric predictor i.e. UMP is expressed as follows:

$$f^{(2)}(z_1) = 1{,}9312 z_1 + 0{,}0158(z_1 - 1670)_+$$

$$f^{(2)}(z_1)\begin{cases} 1{,}9312 z_1 & ; z_1 < 1670 \\ -26{,}3860 + 1{,}9470 z_1 & ; z_1 \geq 1670 \end{cases}$$

   Based on the equation of the model if the province in Indonesia with a value of UMP less than 1670 thousand increase by one unit, it will increase the case of KH by 1.9312 one unit. Whereas, if the province in Indonesia with a value of UMP more than 1670 thousand increase by one unit, then the case of KH will increase by 1.9470 one unit.
3. The deduction function for the third nonparametric predictor i.e. JP is expressed as follows:

$$f^{(2)}(z_2) = +25{,}4053 z_2 - 0{,}1871 z_2^2 + 0{,}0123(z_2 - 44{,}46)_+{}^2 + 0{,}0730(z_2 - 63{,}16)_+{}^2 + 0{,}1439(z_2 - 79{,}87)_+{}^2 + 0{,}2048(z_2 - 126{,}93)_+{}^2 - 0{,}3480(z_2 - 192{,}47)_+{}^2 + 0{,}1081(z_2 - 423{,}35)_+{}^2$$

$$f^{(2)}(z_2) = \begin{cases} 25{,}4053 z_2 - 0{,}1871 z_2^2 & ; \quad 0 \leq z_2 < 44{,}46 \\ 24{,}3116 z_2 - 0{,}1748 z_2^2 + 24{,}3133 & ; \; 44{,}46 \leq z_2 < 63{,}16 \\ 15{,}0902 z_2 - 0{,}1018 z_2^2 + 315{,}5238 & ; \; 63{,}16 \leq z_2 < 79{,}87 \\ -7{,}8964 z_2 + 0{,}0421 z_2^2 + 1233{,}4931 & ; \; 79{,}87 \leq z_2 < 126{,}93 \\ -59{,}8869 z_2 + 0{,}2469 z_2^2 + 4533{,}0720 & ; 126{,}93 \leq z_2 < 192{,}47 \\ 74{,}0722 z_2 - 0{,}1011 z_2^2 - 8358{,}4840 & ; 192{,}47 \leq z_2 < 423{,}35 \\ -17{,}4561 z_2 + 0{,}0070 z_2^2 + 11015{,}7625 & ; \qquad z_2 \geq 423{,}35 \end{cases}$$

   Based on the equation the model is known that if the value of JP a province in Indonesia between 44.46 thousand to less than 79.87 thousand increased one unit, it will increase the case of KH. Meanwhile, if the value of JP a province in Indonesia between 79.87 to less than 192.47 is increasing, the decrease in the case of KH. If the value of JP a province in Indonesia between 192.47 thousand to less than 423.35 thousand increased one unit, it will increase the case of KH. If the value of JP a province in Indonesia between 192.47 thousand to less than 423.35 thousand increased one unit, it will increase the case of KH. Meanwhile, if the value of JP a province in Indonesia more than 423.35 thousand increases one unit, it will decrease the case of KH.
4. The deduction function for the third nonparametric predictor i.e. PKRB is stated as follows:

$$f^{(2)}(z_3) = 128{,}4694 z_3 + 0{,}2334 \times 10^{-4}(z_3 - 1{,}74)_+$$

$$f^{(2)}(z_3) = \begin{cases} 128{,}4694 z_3 & ; z_3 < 1{,}74 \\ -0{,}0004 + 128{,}4694 z_3 & ; z_3 \geq 1{,}74 \end{cases}$$

   Based on the equation of the model when the province in Indonesia with the value of PKRB less than 1.74 percent increase by one unit, it will increase the case of KH by 128.4694 one unit. Whereas, if the province in Indonesia with the value of PKRB more than 1.74 percent increase by one unit, then the case of KH will increase by 128.4694 one unit.

After obtaining the best penalized spline semiparametric regression model with an R 2 value of 0.8318, it means that the KHMK and KH variables are able to be explained by the JPS, JP, UMP, and PKRB variables by 83.18% and the remaining 16.82% is explained by other variables outside the model. The original data estimation curve with predictive data is explained in Figure 5.

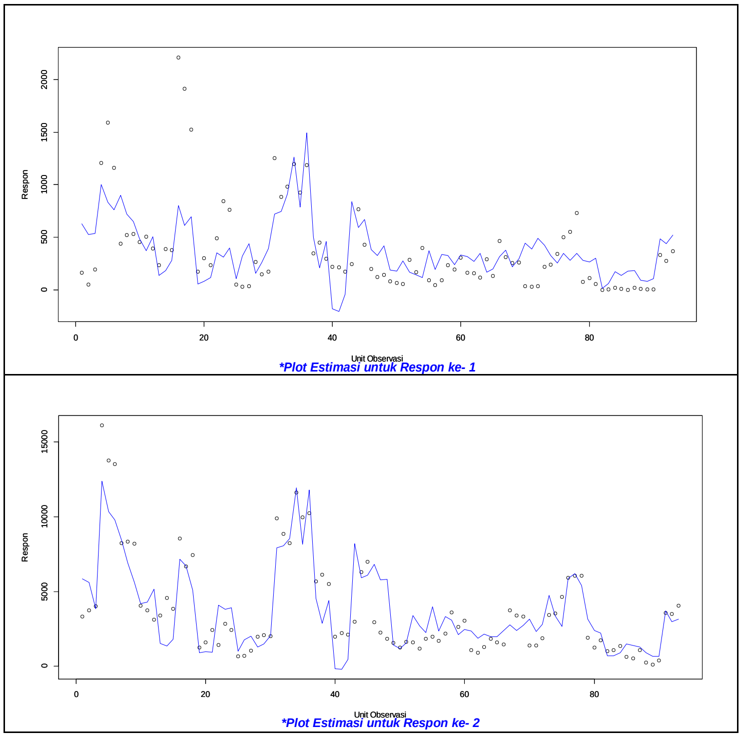

Fig. 5. Plot of Observation Data and Response Variable Estimation Data

Based on Figure 5, it can be seen that the prediction data generated tends to follow the movement of the original data, although the curve shows that the prediction value is not exactly the actual value. This means that the penalized spline semiparametric regression model obtained has the ability to adapt more effectively in overcoming sharp up or down data patterns with the help of knot points.

Conclusions

The penalized spline semiparametric regression model of crime data for each province in Indonesia in 2014-2016 obtained with an R2 of 83.18% is as follows:

$$\hat{y}^{(1)} = (-2{,}0813\times10^{-15}) - 650{,}5573 + 0{,}1602x_1 + 0{,}5044z_1 - 0{,}8317\times10^{-4}(z_1 - 1670)_+ + 0{,}1992z_2 -$$

$$0{,}0009(z_2 - 44{,}46)_+ - 0{,}0024(z_2 - 63{,}16)_+ - 0{,}0041(z_2 - 79{,}87)_+ - 0{,}0110(z_2 - 126{,}93)_+ -$$

$$1{,}0219(z_2 - 192{,}47)_+ - 0{,}0274(z_2 - 423{,}35)_+ - 69{,}2968z_3 + 0{,}5953\times10^{-4}(z_3 - 1{,}74)_+$$

$$\hat{y}^{(2)} = (1{,}7856\times10^{-12}) - 3528{,}518 + 1{,}3368x_1 + 1{,}9312z_1 + 0{,}0158(z_1 - 1670)_+ + 25{,}4053z_2 - 0{,}1871z_2^2 +$$

$$0{,}0123(z_2 - 44{,}46)_+{}^2 + 0{,}0730(z_2 - 63{,}16)_+{}^2 + 0{,}1439(z_2 - 79{,}87)_+{}^2 + 0{,}2048(z_2 - 126{,}93)_+{}^2 -$$

$$0{,}3480(z_2 - 192{,}47)_+{}^2 + 0{,}1081(z_2 - 423{,}35)_+{}^2 + 128{,}4694z_3 + 0{,}2334\times10^{-4}(z_3 - 1{,}74)_+$$

The interpretation of the model is described as follows.

1. Based on the model in the first response it is known that if the JPS, UMP, JP, and PKRB variables increase, the KHMK variable will increase, where the increase in the KHMK value with a UMP value <1670 thousand is smaller than the UMP value of 1670 thousand.
2. Based on the model in the second response it is known that if the JPS, UMP, and JP variables increase, the KH variable will increase, where the increase in KH value with a UMP value <1670 thousand is smaller than the UMP value of 1670 thousand. However, the JP variable changes significantly every piece of knots where if the value of JP <79.87 thousand and if the value of 192.47 thousand ≤ JP <423.35 thousand, the KH value

will increase. Meanwhile, if the JP value is 423.35 thousand and if 79.87 thousand ≤ JP <192.47 thousand, the KH value will increase.

References

1. Fitri, Wanda (2017). Perempuan dan Perilaku Kriminalitas: Studi Kritis Peran Stigma Sosial Pada Kasus Residivis Perempuan. Journal of *Kafa'ah*, 7 (1), 67-78.
2. BPS. *Statistik kriminal* (2017). Jakarta: Badan Pusat Statistika.
3. Dermawanti., Abdul Hoyyi., and Agus Rusgiyono (2015). Faktor-Faktor Yang Mempengaruhi Kriminalitas Di Kabupaten Batang Tahun 2013 Dengan Analisis Jalur. Journal of *Gaussian*, 4 (2), 247 – 256.
4. Astiti, Desak Ayu Wiri., I Wayan Sumarjaya., dan Made Susilawati (2016). Analisis Regresi Nonparametrik *Spline* Multivariat untuk Pemodelan Indikator Kemiskinan Di Indonesia. Joural of *E-Jurnal Matematika, 5* (3), 111 - 116.
5. Marina, Sherly Mega dan I Nyoman Budiantara (2013). Pemodelan Faktor-Faktor yang Mempengaruhi Persentase Kriminalitas di Jawa Timur dengan Pendekatan Regresi Semiparametrik *Spline*. Journal of *Sains Dan Seni Pomits*, 2 (2), 147 - 152.
6. Agustina, Novia., Suparti., dan Moch. Abdul Mukid (2015). Pemodelan Data Indeks Harga Saham Gabungan Menggunakan Regresi *Penalized Spline*. Journal of *Gaussian*, 4 (3), 603 - 612.
7. Fernandes, Adji Achmad Ronaldo dan Solimun (2016). The Effect of Correlation Between Responses In Bi-Response Nonparametric Regression Using Smoothing *Spline* For Longitudinal Data. Journal of *Communications in Applied Analysis*, 20, 335-354.
8. Islamiyati, Anna., Fatmawati., dan N. Chamidah (2018). Estimation of Covariance Matrix on Bi-Response Longitudinal Data Analysis with *Penalized Spline* Regression.*The 2nd International Conference on Science (ICOS),* IOP Conf. Series: Journal of Physics: Conf. Series 979 (2018) 012093.
9. Putri, Ristanti Febriana., Septiadi Padmadisastra., and Sri Winarni (2017). Analisis Data Longitudinal Dalam Desain Faktorial Menggunakan Linear Mixed Model. *Seminar Statistika FMIPA UNPAD 2017 (SNS VI),* ISSN: 2087-2590.
10. Harlan, Johan (2018). *Analisis Data Longitudinal*. Depok: Gunadarma.
11. Nurdiani, Nunung., Nar Herrhyanto., dan Dadan Dasari (2017). Estimasi Model Regresi Nonparametrik Menggunakan Radial Smoothing Berdasarkan *Estimator Penalized Spline*. Journal of *Eureka Matika, 5* (1), 106 - 121.