Advances in Social Science, Education and Humanities Research, volume 474
Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019)

Analysis of Penalized Semiparametric Regression Model on Bi-Response Longitudinal Data

Kosmaryati[1*], Mujiati Dwi Kartikasari [1]

[1]*Department of Statistics, Universitas Islam Indonesia*
Jalan Kaliurang Km 14.5 Sleman, Yogyakarta, Indonesia 55584
**Corresponding author. Email: kosmaryati3103@gmail.com*

ABSTRACT
Semiparametric regression is a combination of parametric regression and nonparametric regression. Parametric regression analysis is used if a regression 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.*

1. INTRODUCTION

Regression analysis is one of the statistical methods used to determine the form of a statistical model or the relationship between one or more predictor variables with one or more response variables [1]. Estimation of curve or function of regression can be done with three kinds of approaches, i.e. parametric approach, nonparametric approach, and semiparametric approach. The parametric approach is used when the curve or function of regression is known based on theory, previous information, or other sources, while the nonparametric approach is an approach of regression model that assumes the curve or function of regression is unknown, and it is hoped that the data itself will look for the regression curve. In some cases, it is often known the curve or function of regression between the response variable and some predictor variables, but not some other predictor variables whose curves are not known. The solution to find out the function model is a semiparametric approach. The semiparametric regression approach is used if the curve or function of regression between several predictor variables on the response variable is known and some are unknown [2].

The problem that occurs in semiparametric regression analysis is in the regression model estimation, this occurs because of the nonparametric component whose curve or function is unknown. A popular nonparametric regression approach is the spline. Spline is a segmented polynomial that is joined by knots that can explain the characteristics of the data. The advantage of the spline is that it provides better flexibility in the characteristics of a functions or data smoothly, able to explain changes in functional behavior patterns within certain sub-intervals and can be used to overcome or reduce data patterns that have experienced a sharp increase [2].

In spline regression analysis, the selection of the number and location of the knots is important. Thus, it is necessary to calculate the combination of the number of knots from the amount of data to determine the optimal knots, then choose the optimal model based on certain criteria. This takes a long time and if done using software requires a large amount of memory. Therefore, an alternative is needed to overcome this problem, namely the penalized spline regression where the knots are located at the quantile points of the unique predictor variable values [3].

In this study, we implemented the semiparametric penalized spline regression method on bi-response longitudinal data. Longitudinal data is data consisting of cross-section data and time series data. The application is applied to crime data in Indonesia for each province from 2014 to 2016. This is because crime in Indonesia has increased from the previous year, news about crime can easily be found in various media. According to the Badan Pusat Statistik (BPS), during the period 2014 to 2016, the number of crimes in Indonesia tended to increase [4]. Criminality is any act that has an economic or psychological impact and violates applicable regulations and social and religious norms, so

that it can be subject to punishment based on the Undang-Undang Hukum Pidana [5].

2. Materials and Methods

2.1. Data

The data used in this study are crime data in Indonesia for each province from 2014 to 2016 obtained from BPS and Kementerian Pendidikan dan Kebudayaan (KEMENDIKBUD). The variables used in this study consisted of two types, i.e. response variables and predictor variables. These variables are presented in Table 1.

Tabel 1 Research Variables

Variable	Variable Name	Definition
Response	Number of Crimes against Rights/Property Using Violence (KHMK)	It is one way of classifying types of crimes based on the criteria for how crimes are committed, by using violence.
	Number of Crimes against Rights/Property without Using Violence (KH)	It is one way of classifying types of crimes based on the criteria for how crimes are committed, without using violence
Predictor	Percentage of Vulnerable Family (PKRB)	This is the percentage of the number of cases of living divorce in each province.
	Minimum Wage by Province (UMP)	It is a standard or minimum wage given by a company or an industry to its workers.
	Number of Unemployment (JP)	It is the number of people residing in the province who are unemployed.
	Number of Dropouts (JPS)	It is the number of students who did not complete their studies in high school education.

2.2. Semiparametric Regression Model

Semiparametric regression is a regression approach that combines a parametric component with a nonparametric component, where semiparametric regression includes a parametric regression model and a nonparametric regression model [6]. Parametric regression is a statistical method that can describe the pattern of the relationship between the response variable and the predictor variable, where the curve or function of a regression is known. According to Draper and Smith (1992) [7] the form of linear parametric regression can be stated in the following statement,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \quad i = 1,2,\cdots,n, \tag{1}$$

where y_i is the response variable, $x_{1i}, x_{2i}, \cdots, x_{pi}$ is the predictor variable, $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ is the regression parameter, and ε_i is a residual with $\varepsilon \sim \text{IIDN}(0, \sigma^2)$.

In some cases, a different relationship pattern was found between one predictor and another, where the relationship between the response variable and one of the predictor variables was known to be the pattern of the relationship. However, the relationship pattern of the response variable with other predictor variables is unknown. Variables that have a known curve or function of regression or there is previous information are classified into the parametric component. Suppose there is paired data x, y, z with a model of the relationship between variables x_i, y_i, z_i, and it is assumed to follow a semiparametric regression model, then the form of the equation is as follows,

$$y_i = x_i\beta + f(z_i) + \varepsilon_i, \quad i = 1,2,\cdots,n, \tag{2}$$

where y_i is the response variable for the observation i, x_i is a parametric component, $f(z_i)$ is a nonparametric component, and ε_i is the residual value [8].

2.3. Semiparametric Regression Model for Bi-Response Longitudinal Data

In the regression analysis, there are three types of data used, which are cross-section, time series, and longitudinal. Longitudinal data is a combination of cross-section data with time series data, where data obtained from observations of n subjects who are independent with each subject are observed repeatedly over T periods of time and between observations in the same subject are correlated [9]. According to Wu and Zhang (2006) [10] using longitudinal data studies can determine individual changes, it takes less subjects because the observations are repeated, and the estimation is more efficient because each observation is made.

In addition to being differentiated based on the use of data, regression models are also differentiated based on the number of responses involved, namely single response, bi-response and multi response regression models. Suppose that the relationship between predictors and bi-response for

longitudinal $(y_{rit}, x_{it}, z_{rit})$ data is known following the semiparametric regression model as follows,

$$y_{rit} = \beta_r(x_{it}) + f_r(z_{it}) + \varepsilon_{rit}, \quad r = 1,2; i = 1,2,\cdots,n; t = 1,2,\cdots,T. \tag{3}$$

2.4. Penalized Spline

Penalized spline is one estimator smooth and can be used to generate a regression function corresponding to the data. The following is a model with a penalized spline estimator on the bi-response longitudinal data,

$$y_{it}^{(r)} = \sum_{w=1}^{q} f_w^{(r)}(z_{wit}) + \varepsilon_{it}^{(r)}, \quad r = 1,2; i = 1,2,\cdots,n; t = 1,2,\cdots,T; w = 1,2,\cdots,q, \tag{4}$$

where $f_w^{(r)}$ is the regression function for the predictor w and the response r of unknown curve, $y_{it}^{(r)}$ is the response variable r for the observation i and at the time t, z_{wit} is the predictor variable w for the observation i and at the time t, and $\varepsilon_{it}^{(r)}$ is a random error with mean 0 and variance σ^2. Hence, Ruppert et.al. (2003) [8] states the penalized estimator function is as follows,

$$f_w^{(r)}(z_{wit}) = \sum_{h=1}^{d_w+m_w} \beta_{wh}\phi_h(z_{wit}), \tag{5}$$

where $\beta_w = (\beta_{w0}, \beta_{w1}, \cdots, \beta_{w(d_w+m_w)})^T$ denotes a parameter vector and $\phi_h(z_{wit})$ is a function defined as follows,

$$\phi_h(z_{wit}) = \begin{cases} z_{wit}^h, & 0 \le h \le d_w \\ (z_{wit} - k_{w(h-d_w)})_+^{d_w}, & d_w + 1 \le h \le m_w + 1 \end{cases} \tag{6}$$

where d_w is the order of the polynomial, m_w is the number of knots for the predictor w, and h is the index of the base function of a positive integer, and

$$(z_{wit} - k_{w(h-d_w)})_+^{d_w} = \begin{cases} (z_{wit} - k_{w(h-d_w)})_+^{d_w}, & z \ge k_{w(h-d_w)} \\ 0, & z \le k_{w(h-d_w)} \end{cases} \tag{7}$$

3. RESULTS AND DISCUSSION

Crime modeling for each province in Indonesia from 2014 to 2016 is performed using the semiparametric penalized spline regression method. Semiparametric regression is a regression model that contains a parametric component and a nonparametric component. The determination of the parametric and nonparametric components is done using a scatterplot. The scatterplot will provide information about the pattern of the regression curve that will be used in modeling. The results of the scatterplot for the response variable for each predictor variable are presented in Figure 1.

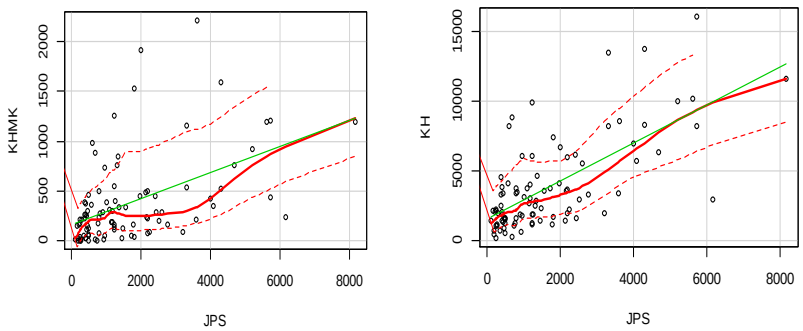

Figure 1. Scatterplot response variables with JPS predictor variable

Figure 1 illustrates the pattern of the relationship between the response variable and JPS predictor variable. The relationship pattern that is formed between KHMK and JPS and between KH and JPS formed a positive linear relationship. This can be seen from the general plot movement, which shows that the higher the number of cases of school dropouts, the more rights/property crimes that use violence and without violence tend to increase.

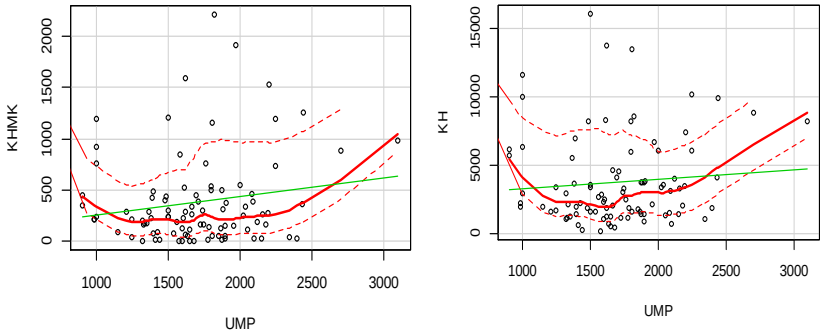

Figure 2. Scatterplot response variables with UMP predictor variable

Figure 2 illustrates the pattern of the relationship between the response variable and UMP predictor variable. The relationship pattern that is formed between KHMK and UMP and between KH and UMP tended not to follow a certain pattern. The visible pattern of relationships also tends to fluctuate in certain sub-intervals. From the results of the scatterplot, the relationship between KHMK and UMP and between KH and UMP will be approached with a nonparametric approach.

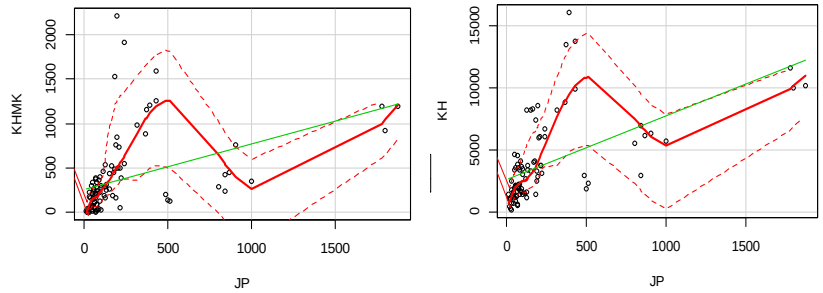

Figure 3. Scatterplot response variables with JP predictor variable

Figure 3 illustrates the pattern of the relationship between the response variable and JP predictor variable. The relationship pattern that is formed between KHMK and JP or between KH and JP tends to experience changes in behavior or fluctuates at several intervals. It can be seen that the data pattern at the interval before 500 tends to increase, but at the interval 500 tends to decrease, and at interval 1000 tends to increase. Therefore, it can be approached with a nonparametric approach.

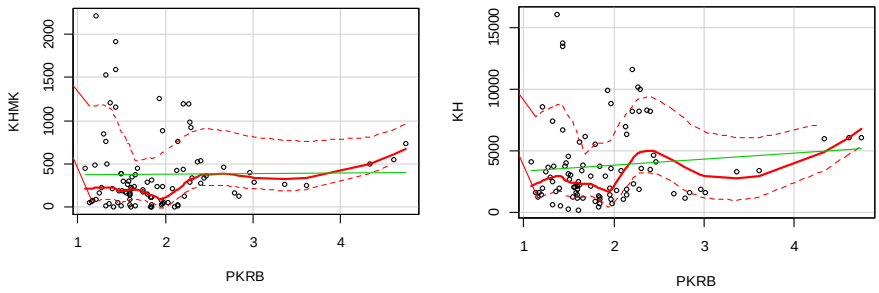

Figure 4. Scatterplot response variables with PKRB predictor variable

Figure 4 illustrates the pattern of the relationship between the response variable and PKRB predictor variable. The relationship pattern that is formed between KHMK and PKRB and between KH and PKRB tended not to follow a certain pattern. The visible pattern of relationships also tends to fluctuate in certain sub-intervals. From the results of the scatterplot, the relationship between KHMK and PKRB variables and between KH and PKRB variables will be approached with a nonparametric approach.

The variable which is approximated by the parametric approach is symbolized by x and the variable which is approximated by the nonparametric approach is symbolized by z. A complete list of information about each predictor variable can be seen in Table 2.

Table 2. Results of Determining Parametric and Nonparametric Components

Variable	Approaches	Variable symbols
JP	Parametric	x
PKRB	Nonparametric	z_1
JPS		z_2
UMP		z_3

After obtaining information from each predictor variable as shown in Table 2, then we estimate the semiparametric penalized spline regression model for bi-response longitudinal data. The model applied to crime data in Indonesia from 2014 to 2016 for the first response is

$$\hat{y}^{(1)} = (-2.0813 \times 10^{-15}) - 650.5573 + 0.1602x_1 + 0.5044z_1 - 0.8317 \times 10^{-4}(z_1 - 1670)_+ + 0.1992z_2 - 0.0009(z_2 - 44.46)_+ - 0.0024(z_2 - 63.16)_+ - 0.0041(z_2 - 79.87)_+ - 0.0110(z_2 - 126.93)_+ - 1.0219(z_2 - 192.47)_+ - 0.0274(z_2 - 423.35)_+ - 69.2968z_3 + 0.5953 \times 10^{-4}(z_3 - 1.74)_+.$$

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

1. The relationship between JPS and KHMK with the assumption that the other variables are constant is that each increase in the JPS case by one unit will increase the KHMK case by 0.1182 unit.
2. The deduction function for the first nonparametric predictor, UMP, is expressed as

$$f^{(1)}(z_1) = 0.5044z_1 - 0.8317 \times 10^{-4}(z_1 - 1670)_+$$

$$f^{(1)}(z_1) = \begin{cases} 0.5044z_1, & z_1 < 1670 \\ 0.1389 + 0.5043z_1, & z_1 \geq 1670. \end{cases}$$

Based on the model equation, if provinces in Indonesia with UMP less than 1670 thousand increase by one unit, it will increase KHMK case by 0.5044 unit. Meanwhile, if provinces in Indonesia with UMP more than 1670 thousand increase by one unit, KHMK case will increase by 0.3760 unit.

3. The deduction function for the second nonparametric predictor, JP, is expressed as

$$f^{(1)}(z_2) = 0.1992z_2 - 0.0009(z_2 - 44.46)_+ - 0.0024(z_2 - 63.16)_+ - 0.0041(z_2 - 79.87)_+ - 0.0110(z_2 - 126.93)_+ - 0.0219(z_2 - 192.47)_+ - 0.0274(z_2 - 423.35)$$

$$f^{(1)}(z_2) = \begin{cases} 0.1992z_2, & 0 \leq z_2 < 44.46 \\ 0.0400 + 0.1902z_2, & 44.46 \leq z_2 < 63.16 \\ 0.1916 + 0.1878z_2, & 63.16 \leq z_2 < 79.87 \\ 0.5191 + 0.1837z_2, & 79.87 \leq z_2 < 126.93 \\ 1.9153 + 0.1727z_2, & 126.93 \leq z_2 < 192.47 \\ 6.1304 + 0.1508z_2, & 192.47 \leq z_2 < 423.35 \\ 17.7302 + 0.1234z_2, & z_2 \geq 423.35. \end{cases}$$

Based on the model equation, if provinces in Indonesia with JP less than 44.46 thousand increase by one unit, it will increase KHMK case by 0.1992 unit. If the province with JP between the values of 44.46 thousand to 63.16 thousand increases by one unit, then KHMK case will increase by 0.102 unit. If the province with JP is between 63.16 thousand and 79.87 thousand per unit, then KHMK case will increase by 0.1878 unit. If the province with JP between 79.87 thousand and 126.93 thousand increased by one unit, KHMK case will increase by 0.1837 unit. If the province with JP between the values of 126.93 to 192.47 thousand increases by one unit, then KHMK case will increase by 0.1727 unit. If the province JP between the values of 192.47 thousand to 423.35 thousand increased by one unit, KHMK case will increase by 0.1508 unit. In addition, if a province with JP is more than 423.35 thousand by one unit, KHMK case will increase by 0.1234 unit.

4. The deduction function for the third nonparametric predictor, PKRB, is expressed as

$$f^{(1)}(z_3) = -69.2968z_3 + 0.5953 \times 10^{-4}(z_3 - 1.74)_+$$

$$f^{(1)}(z_3) = \begin{cases} -69.2968z_3, & z_3 < 1.74 \\ -0.0001 - 69.2967z_3, & z_3 \geq 1.74. \end{cases}$$

Based on the model equation, if provinces in Indonesia with PKRB less than 1.74 percent increase by one unit, it will reduce the cases of KHMK by 69.2968 unit. Meanwhile, if provinces in Indonesia with PKRB more than 1.74 percent increased by one unit, then KHMK case will decrease by 69.2967 unit.

The result of the semiparametric penalized spline regression model obtained for the second response in crime data can be written as

$$\hat{y}^{(2)} = (-1.7856 \times 10^{-12}) - 3528.518 + 1.3368x_1 + 1.9312z_1 + 0.0158(z_1 - 1670)_+ + 25.4053z_2 - 0.1871z_2^2 + 0.0123(z_2 - 44.46)_+^2 + 0.0730(z_2 - 63.16)_+^2 + 0.1439(z_2 - 79.87)_+^2 + 0.2048(z_2 - 126.93)_+^2 - 0.3480(z_2 - 192.47)_+^2 + 0.1081(z_2 - 423.35)_+^2 + 128.4694z_3 + 0.2334 \times 10^{-4}(z_3 - 1.74)_+.$$

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

1. The relationship between JPS and KH with the assumption that the other variables are constant is that every JPS case increase by one unit will increase the case of KH by 1.3368 unit.
2. The deduction function for the first nonparametric predictor, UMP, is expressed as

$$f^{(2)}(z_1) = 1.9312z_1 + 0.0158(z_1 - 1670)_+$$

$$f^{(2)}(z_1) = \begin{cases} 1.9312z_1, & z_1 < 1670 \\ -26.3860 + 1.9470z_1, & z_1 \geq 1670. \end{cases}$$

Based on the model equation, if provinces in Indonesia with UMP less than 1670 thousand increase by one unit, it will increase KH case by 1.9312 unit. Meanwhile, if provinces in Indonesia with UMP more than 1670 thousand increase by one unit, then KH case will increase by 1.9470 unit.

3. The deduction function for the second nonparametric predictor, JP, is expressed as

$$f^{(2)}(z_2) = 25.4053z_2 - 0.1871z_2^2 + 0.0123(z_2 - 44.46)_+^2 + 0.0730(z_2 - 63.16)_+^2 + 0.1439(z_2 - 79.87)_+^2 + 0.2048(z_2 - 126.93)_+^2 - 0.3480(z_2 - 192.47)_+^2 + 0.1081(z_2 - 423.35)_+^2$$

$$f^{(2)}(z_2) = \begin{cases} 25.4053z_2 - 0.1871z_2^2, & 0 \leq z_2 < 44.46 \\ 24.3116z_2 - 0.1748z_2^2 + 24.3133, & 44.46 \leq z_2 < 63.16 \\ 15.0902z_2 - 0.10181z_2^2 + 315.5238, & 63.16 \leq z_2 < 79.87 \\ -7.8964z_2 + 0.0421z_2^2 + 1233.4931, & 79.87 \leq z_2 < 126.93 \\ -59.8869z_2 + 0.2469z_2^2 + 4533.0720, & 126.93 \leq z_2 < 192.47 \\ 74.0722z_2 - 0.1011z_2^2 - 8358.4840, & 192.47 \leq z_2 < 423.35 \\ -17.4561z_2 + 0.0070z_2^2 + 11015.7625, & z_2 \geq 423.35. \end{cases}$$

Based on the model equation, it is known that if JP of a province in Indonesia is between 44.46 thousand to less than 79.87 thousand increases by one unit, it will increase KH case. Meanwhile, if JP of a province in Indonesia is between 79.87 and less than 192.47 increasing, then there will be fewer cases of KH. If JP of a province in Indonesia between 192.47 thousand and less than 423.35 thousand increases by one unit, it will increase KH case. If JP of a province in Indonesia between 192.47 thousand and less than 423.35 thousand increases by one unit, it will increase KH case. Meanwhile, if JP of a province in Indonesia is more than 423.35 thousand, an increase of one unit will reduce the KH case.

4. The deduction function for the third nonparametric predictor, PKRB, is expressed as

$$f^{(2)}(z_3) = 128.46948z_3 + 0.2334 \times 10^{-4}(z_3 - 1.74)_+$$

$$f^{(2)}(z_3) = \begin{cases} 128.4694z_3, & z_3 < 1.74 \\ -0.0004 + 128.4694z_3, & z_3 \geq 1.74. \end{cases}$$

Based on the model equation, if provinces in Indonesia with PKRB less than 1.74 percent increase by one unit, it will increase the KH case by 128.4694 unit. Meanwhile, if provinces in Indonesia with PKRB more than 1.74 percent increase by one unit, KH cases will increase by 128.4694 unit.

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

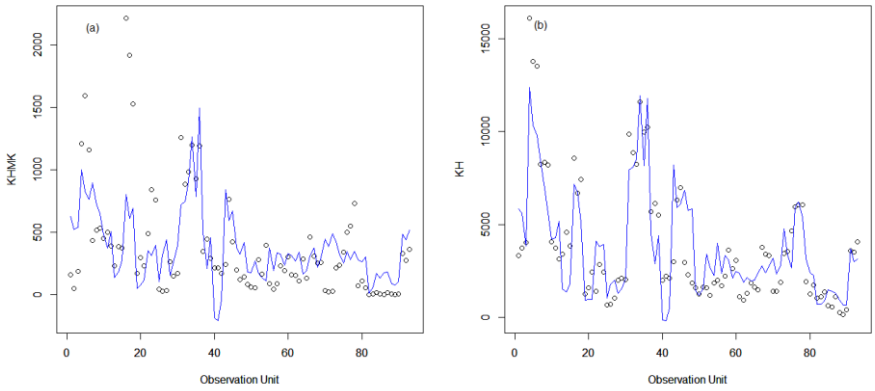

Figure 5. Plot of Actual Data and Prediction Data (a) first response (b) second response

Based on Figure 5, it can be seen that the prediction data tends to follow the movement of the actual data, even though the curve shows that the predicted value is not exactly the same as the actual data. This means that the penalized spline semiparametric regression model obtained has the ability to adjust itself more effectively in overcoming data patterns that have high increases and decreases.

4. CONCLUSION

We have implemented a semiparametric regression approach with a penalized spline on the biresponse longitudinal data. We calculated the semiparametric regression approach using crime data in each province in Indonesia. Based on the results, the predicted data tends to follow the actual data movement, even though the predicted values are not exactly the same as the actual data. In addition, the resulting R^2 value is 0.8318, a values close to 1. Therefore, we conclude that the model result of the semiparametric regression with the penalized spline is generally good.

REFERENCES

[1] Astiti, D., Sumarjaya, I., Susilawati, M. (2016). Analisis Regresi Nonparametrik Spline Multivariat untuk Pemodelan Indikator Kemiskinan Di Indonesia. E-Jurnal Matematika. 5(3): 111-116.

[2] Budiantara, I. (2014). Pemodelan Regresi Nonparametrik dan Semiparametrik Spline (Konsep, Metode Dan Aplikasinya). Prosiding Seminar Nasional Matematika, 1-15.

[3] Agustina, N., Suparti, Mukid, M. (2015). Pemodelan Data Indeks Harga Saham Gabungan Menggunakan Regresi Penalized Spline. Jurnal Gaussian. 4(3): 603 – 612.

[4] BPS. (2017). Statistik kriminal 2017. Jakarta: Badan Pusat Statistik.

[5] Kartono. (1999). Patologi Sosial. Jakarta: Raja Grafindo Persada.

[6] Nurdiani, N., Herrhyanto, N., Dasari, D. (2017). Regresi Nonparametrik Birespon Spline. Jurnal EurekaMatika 5(1): 106 - 121.

[7] Draper, N., Smith, H. (1992). Applied Regression Analysis, Second Edition. New York: John Wiley and Sons.

[8] Ruppert, D., Wand, M., Carrol R. (2003). Semiparametric Regression. New York: John Willey and Sons.

[9] Diggle, P., Liang, K., Zeger, S. (1994). Analysis of Longitudinal Data. New York: Oxford University Press.

[10] Wu, H., Zhang, J. (2006). Nonparametric Regression Methods for Longitudinal Data Analysis. New Jersey: John Willey and Sons.