Application of semiparametric spline regression model in modelling human development index in 2016

S Amilia*, N Herrhyanto and E Puspita

Department of Mathematics Education, Universitas Pendidikan Indonesia, Bandung, Indonesia

*Corresponding author’s e-mail: sumi.amilia96@gmail.com

Abstract. Human development index is one of indicators to determine developed country. According to the data from Indonesian Central Bureau of Statistics in 2016, Human Development Index of Papua, West Papua, and East Nusa Tenggara are lower than others. Therefore, the aim of this study was to analyze the relationship between Human Development Index and four variables such as Life Expectancy, School Life Expectancy, Average School Duration, and Average per Capita Expenditure affecting an improvement of Human Development Index. Data of human development index and 4 variables used in this study was obtained from the article “Human Development Index 2016” published by Indonesian Central Bureau of Statistics. Method used in modelling Human Development Index in 2016 was semiparametric spline regression method. Linear spline, quadratic spline, and cubic spline were models applied in this study, whereas one point of knot was implemented as well. As a result, all variables significantly affected the Human Development Index in Papua, while in West Papua, variables of both Average School Duration and Average per Capita Expenditure related to Human Development Index significantly. In contrast to Papua and West Papua, 3 variables such as School Life Expectancy, Average School Duration, and Average per Capita Expenditure contributed on Human Development Index in East Nusa Tenggara.

1. Introduction
Indonesia ranks fourth in the order of the largest population in the world [1]. However, the rank of Indonesian population is not aligned with the rank of Indonesia’s Human Development Index (HDI). The United Nations Development Program (UNDP) in the Human Development Report in 2016 noted that the rank of Indonesia's HDI in 2015 was 113th from 188 countries and territories [2]. This indicates that the Indonesia’s HDI is still low compared with other countries. HDI explains how people can access development outcomes in obtaining income, health, education, etc. [3]. In 2016, 3 of 34 s in Indonesia with the lowest HDI were Papua, West Papua, and East Nusa Tenggara. Respectively, the HDI values of the three s were 58.05, 62.21, and 63.13[4]. In the fact, no study addresses the relationship between the HDI and the HDI determinants.

One of the statistical methods that can be used to determine the relationship between HDI and the HDI determinants is regression analysis. Regression analysis used is semiparametric spline regression method. The nonparametric component of the semiparametric regression is approximated by spline function. Spline is a seamless segmented polynomial model that can produce suitable regression functions with the data [5]. Spline is used because it has an advantage in overcoming the sharpness of an increase or a decrease of a data pattern with the knot.
In general, the aim of this study was to observe the best semiparametric spline regression model of Human Development Index in 2016 and to analyze variables affecting the Human Development Index in 2016 significantly.

2. Methods
The research method used is semiparametric spline regression. Semiparametric regression is a statistical method to analyze the relationship between response variables and predictor variables, where some of the patterns or curves of the relationships between the variables are known and unknown. Generally, semiparametric regression model is written as follows

\[ y_i = \beta_0 + \sum_{j=1}^{t} \beta_j x_{ij} + f(z_i) + \varepsilon_i \quad i = 1, \ldots, n \] (1)

where
- \( y_i \): the response variable on the i-th data
- \( \beta_0 \): the constant of the model
- \( \beta_j \): the unknown parameters
- \( f(z_i) \): the unknown i-th nonparametric function
- \( x_{ij} \): the j-th predictor variable of the i-th data for the parametric component
- \( \varepsilon_i \): the i-th error that is assumed to be identical, independent, and distributed \( N(0, \sigma^2) \)

If the unknown curve in the semiparametric regression is approximated by the spline function, so it is called semiparametric spline regression.

The \( p \)-spline model for any function \( f \) is written as follows

\[ f(z_i) = \sum_{m=0}^{p} \alpha_m z_i^m + \sum_{h=1}^{K} y_h (z_i - k_h)^{m^*} + \varepsilon_i \quad i = 1,2, \ldots, n \] (2)

where \( \sum_{m=0}^{p} \alpha_m z_i^m \) is a polynomial component and \( \sum_{h=1}^{K} y_h (z_i - k_h)^{m^*} \) is truncated component,

\[ (z_i - k_h)^{m^*} = \begin{cases} (z_i - k_h)^m & \text{if } z_i \geq k_h \\ 0 & \text{if } z_i < k_h \end{cases} \]

where
- \( \alpha_m \): coefficient of \( z_i^m \)
- \( z_i^m \): the predictor variable of the i-th data
- \( y_h \): coefficient of \( z_i^m \) on the h-th truncated knot of the p-spline
- \( k_h \): the h-th knot of \( z_i^m \).

Substitute the \( f(z_i) \) in equation (2) to equation (1), then

\[ Y_i = \beta_0 + \sum_{j=1}^{t} \beta_j X_{ij} + \sum_{m=0}^{p} \alpha_m z_i^m + \sum_{h=1}^{K} y_h (Z_i - k_h)^{m^*} + \varepsilon_i \quad i = 1,2, \ldots, n \] (3)

The knot is a common fusion point where there is a change in the behavioral pattern of the function or the curve [6]. The optimal knot is produced from the minimum Generalized Cross Validation (GCV) value. The GCV function is defined as follows

\[ GCV(p; k_1, k_2, ..., k_K) = \frac{n^{-1}RSS}{[n^{-1}\text{trace}(I-H(p;k_1,k_2,...,k_K))]^2} \] (4)

where \( H(p;k_1, k_2, ..., k_K) \) is HAT matrix, \( H(p;k_1, k_2, ..., k_K) = C(C^T C)^{-1} C^T \). \( C \) is combined matrix of parametric and nonparametric component with knots [7]. The parameter estimation used is least squares estimator, \( \hat{\alpha} = (C^T C)^{-1} C^T Y \).

3. Results and Discussion
To know the data belonged to the parametric component or nonparametric component, firstly made scatter plot between response variable and predictor variable. A predictor variable is a parametric component if the scatter plot between the response variable and the predictor variable forms a particular data pattern. Additionally, a predictor variable is a nonparametric component if the scatter plot between the response variable and the predictor variable does not form a particular data pattern.
3.1. Semiparametric spline regression model

After obtained the minimum GCV value and estimated parameters by software R, a semiparametric spline regression model can be formed for each data.

| Model Equations |
|------------------|
| Linear Semiparametric Spline Regression Model | $\hat{Y}_1 = -250.773 + 1.258X_{11} + 1.554X_{21} + 0.001X_{31} + 4.9374Z_{11}$ |
| | $-4.583(Z_{11} - 55.565)_+$ |
| Quadratic Semiparametric Spline Regression Model | $\hat{Y}_1 = -29794.880 + 1.261X_{11} + 1.550X_{21} + 0.001X_{31}$ |
| | $+1079.836Z_{11} - 9.776Z_{11}^2 + 9.777(Z_{11} - 55.210)_+^2$ |
| Cubic Semiparametric Spline Regression Model | $\hat{Y}_1 = -12331.180 + 1.253X_{11} + 1.555X_{21} + 0.001X_{31} + 625.864Z_{11}$ |
| | $-10.568Z_{11}^2 + 0.059Z_{11}^3 - 0.062(Z_{11} - 59.471)_+^3$ |
3.2. Selection of the best model

Selection of the best model in this study used mean squared error (MSE). MSE measures the dispersion between the actual values of the parameters. The model with the smallest MSE value is called to be the best model.

Table 4. Comparison of MSE values. The best model for the HDI in Papua was the linear semiparametric spline regression model with the MSE value 0.726, while the best model for the HDI in West Papua was the cubic semiparametric spline regression model with the MSE value 0.065, and the best model for the HDI in East Nusa Tenggara was cubic semiparametric spline regression model with the MSE value 1.823 x 10^{-7}.

| HDI of Papua | Linear Spline | 0.726 |
| HDI of Papua | Quadratic Spline | 0.759 |
| HDI of Papua | Cubic Spline | 0.884 |
| HDI of West Papua | Linear Spline | 0.118 |
| HDI of West Papua | Quadratic Spline | 0.088 |
| HDI of West Papua | Spline Cubic | 0.065 |
| HDI of East Nusa Tenggara | Quadratic Spline | 2.604x10^{-6} |
| HDI of East Nusa Tenggara | Spline Cubic | 1.823x10^{-7} |

3.3. Interpretation of the best semiparametric spline regression model

The following describes the interpretation of each the best semiparametric spline regression models for HDI of Papua, HDI of West Papua, and HDI of East Nusa Tenggara.
3.3.1. The best semiparametric spline regression model for the HDI in Papua. The best semiparametric spline regression model with one knot of the HDI in Papua was

\[
\hat{Y}_i = -250.773 + 1.258X_{1i} + 1.554X_{2i} + 0.001X_{3i} + 4.937Z_{1i} - 4.583(Z_{1i} - 55.565) \\
\text{If } Z_{1i} < 55.565 \text{ then } \\
\hat{Y}_i = -250.773 + 1.258X_{1i} + 1.554X_{2i} + 0.001X_{3i} + 4.937Z_{1i} \\
\text{While if } Z_{1i} \geq 55.565 \text{ then } \\
\hat{Y}_i = 3.868 + 1.258X_{1i} + 1.554X_{2i} + 0.001X_{3i} + 0.354Z_{1i} \\
\]  

If the other predictors are constant, if Life Expectancy is less than 55.565 years and the value increase one unit, then the HDI will elevate 4.937. But, if Life Expectancy is more than or same with 55.565 years and the value increase one unit, then the HDI will elevate 0.354.

3.3.2. The best semiparametric spline regression model for HDI in West Papua. The best semiparametric spline regression model with one knot for HDI in West Papua was

\[
\hat{Y}_i = -4.958 + 2.033X_{1i} + 0.001X_{2i} - 30.825Z_{1i} + 183.427Z_{2i} + 0.470Z_{1i}^2 - 15.945Z_{2i}^2 \\
-0.002Z_{1i}^3 - 0.462Z_{2i}^3 - 0.070(Z_{4i} - 64.691)^3 + 0.522(Z_{2i} - 12.420)^3 \\
\text{If } Z_{1i} < 64.691 \text{ and } Z_{2i} < 12.420 \text{ then } \\
\hat{Y}_i = -4.958 + 2.033X_{1i} + 0.001X_{2i} - 30.825Z_{1i} + 183.427Z_{2i} + 0.470Z_{1i}^2 - 15.945Z_{2i}^2 \\
-0.002Z_{1i}^3 - 0.462Z_{2i}^3 \\
\text{While if } Z_{1i} \geq 64.691 \text{ and } Z_{2i} \geq 12.420 \text{ then } \\
\hat{Y}_i = 19946.014 + 2.033X_{1i} + 0.001X_{2i} - 909.659Z_{1i} - 58.139Z_{2i} + 14.055Z_{1i}^2 + 3.505Z_{2i}^2 \\
-0.072Z_{1i}^3 - 0.060Z_{2i}^3 \\
\]  

If the other predictors were constant, the increase in Average School Duration by one unit conduce the HDI will elevate 2.033.

3.3.3. The best semiparametric spline regression model for HDI in East Nusa Tenggara. The best semiparametric spline regression model with one knot for HDI in East Nusa Tenggara was

\[
\ln(\ln Y_i) = -1.060 + 0.117\ln(\ln X_{1i}) + 0.061\ln(\ln X_{2i}) + 0.372\ln(\ln X_{3i}) \\
+0.463\ln(\ln Z_{1i}) + 1.239(\ln(\ln Z_{1i}))^2 - 0.574(\ln(\ln Z_{1i}))^3 \\
-(4.90 \times 10^{-4})(\ln(\ln Z_{1i}) - 1.435)^3 \\
\text{If } \\
\ln(\ln Z_{1i}) < 1.435 \\
Z_i < 66.663 \\
\text{then } \\
Y_i = 1.413 + 3.077X_{1i} + 2.895X_{2i} + 4.267X_{3i} + 4.899Z_{1i} + 31.563Z_{2i}^2 + 1.756Z_{1i}^3 \\
\text{While if } Z_{1i} \geq 66.663 \text{ then } \\
Y_i = 1.415 + 3.077X_{1i} + 2.895X_{2i} + 4.267X_{3i} + 4.874Z_{2i} + 31.785Z_{2i}^2 + 1.756Z_{1i}^3 \\
\]  

If the other predictors were constant, the increase in School Life Expectancy by one unit conduce the HDI will elevate 3.077.
4. Conclusion
Based on the study, it can be concluded that the best model of the HDI in Papua is linear semiparametric spline regression model,

$$\hat{Y}_i = -250.773 + 1.258X_{1i} + 1.554X_{2i} + 0.001X_{3i} + 4.937Z_{1i} - 4.583(Z_{1i} - 55.565)_+$$

with School Life Expectancy, Average School Duration, Average per Capita Expenditure, and Life Expectancy as the significant variables. In contrast to Papua, the best model of the HDI in West Papua is cubic semiparametric spline regression model,

$$\hat{Y}_i = -4.958 + 2.033X_{1i} + 0.001X_{2i} - 30.825Z_{1i} + 183.427Z_{2i} + 0.470Z_{3i}^2 - 15.945Z_{2i}^2 - 0.002Z_{1i}^3 + 0.462Z_{2i}^3 - 0.070(Z_{1i} - 64.691)_+^3 - 0.522(Z_{2i} - 12.240)_+^3$$

with both Average School Duration and Average per Capita Expenditure as the significant variable. School Life Expectancy, Average School Duration, and Average per Capita Expenditure are the significant variables using cubic semiparametric spline regression model as the best model for determining the HDI in East Nusa Tenggara,

$$\ln(Y_i) = -1.060 + 0.117\ln(X_{1i}) + 0.061\ln(X_{2i}) + 0.372\ln(X_{3i}) + 0.463\ln(Z_{1i}) + 1.239(\ln(Z_{1i}))^2 - 0.574(\ln(Z_{1i}))^3 - (4.90 \times 10^{-4})(\ln(Z_{1i}) - 1.435)_+^3$$

5. References
[1] Ministry of Culture and Education 2017 Hari Kependudukan Dunia 2017: Masa Depan Demografi Indonesia dan Keseimbangan Pertumbuhan Penduduk https://www.kemdikbud.go.id
[2] Indonesian United Nations Development Program 2017 Indeks Pembangunan Manusia Indonesia: Meningkat Tapi Kesenjangan Masih Tetap Ada http://www.id.unpd.org/content/indonesia/id/home/presscenter/pressreleases/2017/03/22/indonesia-s-human-development-index-rises-but-inequality-remains-.html
[3] Central Bureau of Statistics 2018 Laporan Bulanan Data Sosial Ekonomi Edisi 92 Subdirectorate of Publication and Statistical Compilation Jakarta Central Bureau of Statistics p 166
[4] Nugroho A 2017 Indeks Pembangunan Manusia 2016 Iswadi and Karyono Y (Jakarta: Central Bureau of Statistics) p 95
[5] Putra I M B, Srinadi I G A M and Sumarjaya I W 2015 Pemodelan Regresi Spline (Studi Kasus: Herpindo Jaya Cabang Ngaliyan) E-Jurnal Matematika https://ojs.und.ac.id/index.php/mtk/article/view/15103/10017 4 p 110-114
[6] Sugiantari A P and Budiantara I N 2013 Analisis Faktor-faktor yang Mempengaruhi Angka Harapan Hidup di Jawa Timur Menggunakan Regresi Semiparametrik Spline Jurnal Sains dan Seni POMITS http://ejurnal2.its.ac.id/index.php/sains_seni/article/view/3132/773 2 1 p D37-D41
[7] Eubank R L 1999 Nonparametric Regression and Spline Smoothing (New York: Marcel Dekker, Inc) p 43