Nonparametric truncated spline regression model on the data of human development index (HDI) in Indonesia

Dewi Retno Sari Saputro¹, Kornelius Ronald Demu², Purnami Widyaningsih³,

¹Department of Mathematics, Universitas Sebelas Maret, Surakarta, 57126, Indonesia
²Department of Mathematics, Universitas Sebelas Maret, Surakarta, 57126, Indonesia
³Department of Mathematics, Universitas Sebelas Maret, Surakarta, 57126, Indonesia
dewiretnoss@staff.uns.ac.id

Abstract: The standard measurement for country’s human development is Human Development Index (HDI). Several factors allegedly affect it, such as life expectancy, gross domestic regional product (GDP), the number of poor people, and the percentage of an illiterate people. If the HDI’s data and those factors is plotted, then the plot do not form a specific pattern, so the HDI’s data can be applied with nonparametric truncated spline regression model. The best of nonparametric truncated spline regression model is influenced by the order and knot points selection. In this article, nonparametric truncated spline regression model on the data of HDI in Indonesia is applied at first order with 3, 4 and 5 knot points. Based on the result, the best nonparametric truncated spline regression model on the data of HDI in Indonesia is obtained by the combination of 5-5-5-4 knot points. There were 2 of 4 allegedly factors significant to HDI in Indonesia based on parameter significance test followed in this research.

1. Introduction
Human Development Index (HDI) is a standard measurement of country’s human development by considering the aspects of health, education, and life expediency. In recent years, the HDI in Indonesia has increased. There are 4 factors allegedly affected the HDI in Indonesia such as life expectancy, gross domestic product (GDP), the number of poor people, and the percentage of an illiterate people. (UNDP [9]). If the HDI’s data and those 4 factors are plotted, then the plot do not form the specific pattern. The regression model approach used on the unspecified pattern is nonparametric regression model (Eubank [5]). According to Hardle [7], there are 4 methods used in nonparametric regression model such as kernel, k-NN, orthogonal series, and spline. Spline is a piecewise of the segmented polynomial function. Truncated spline is a modification from spline function. Nonparametric truncated spline regression model is affected by the knot points (Budiantara [3]). The knot points is the point of data pattern’s change on a particular subinterval. The best nonparametric truncated spline regression model is obtained based on the selection of order and optimal knot points (Budiantara [3]). The optimal knot points is obtained based on the minimum value of generalized cross validation (GCV) (Wahba [10]). The Estimation of parameters on nonparametric truncated spline regression model are using least squares method (LS) (Lee [8]). The purposes of this research are to apply nonparametric truncated spline regression model to the data of HDI in Indonesia and to determine factors that affect the HDI in Indonesia.
2. Experimental Details
This research was an application while had applied the HDI’s data for 34 provinces in Indonesia using nonparametric truncated spline regression model. Research data classified as secondary data from BPS ([2]) with the HDI’s data in Indonesia in 2015 (Y), life expectancy by province (X₁), GRDP by provincial annual expenditure (X₂), the number of poor people by province (X₃), and the percentage of an illiterate people by province (X₄). The steps used in the research started from determining the relationship between response variable and each predictor variable through the scatter plot. After that, the knot points are determined in each predictor variable. Furthermore, the best nonparametric truncated spline regression model is obtained based on the optimal knot points. Next step, the model parameters are estimated with least square method (LS). The final step of this research were testing the significance of model parameters, assumption for residuals, and also determining the coefficient determination (R^2) on the nonparametric truncated spline regression model.

3. Results and Discussion

![Figure 1a. Y vs X1 relationship plot, 1b. Y vs X2 relationship plot](image)

Figure 1a. Y vs X1 relationship plot, 1b. Y vs X2 relationship plot

The reason for using nonparametric regression model in the attempt of modeling the HDI’s data in Indonesia is caused by the relationship plots between Y with each X₁, X₂, X₃, and X₄ were not form a specific pattern. Two relationship plots between Y with each X₁ and X₂ were shown in Figure 1(a) and1(b) respectively. Figures 1(a) and 1(b) showed that 2 relationship plots between Y with each X₁ and X₂ were not form the specific pattern. In the same way, the relationship plots between Y with each X₃ and X₄ were not also form a specific pattern, so the approach of regression model used for this case is nonparametric regression model. The minimum value of GCV is affected by the selection of order and the number of knot points. Based on the research, order one is selected. The number of knot points started from 3, 4, and 5 knot points. According to Table 1, the value of minimum GCV is 7.6307. That value is resulted from the combination of 5-5-5-4 optimal knot points. After obtaining the optimal knot point in each predictor variable, 24 model parameters are estimated with least square methods (LS). The result of parameter estimation on nonparametric truncated spline regression model were shown in Table 2.

| No. | the number of knot points | minimum’s GCV |
|-----|--------------------------|---------------|
| 1   | 3 - 3 - 3 - 3            | 8.3966        |
| 2   | 4 - 4 - 4 - 4            | 9.3048        |
| 3   | 5 - 5 - 5 - 5            | 9.9343        |
| 4   | 3-5-4-5, 5-3-4-5, etc.   | 7.6307        |

Table 1. The minimum value of GCV based on a number knot points used

Based on the result of parameter estimation in Table 2, the nonparametric spline regression mode on the data of human development index (HDI) in Indonesia is written by $y\hat{} = -150.029 +$
The model (1) can be written in the form of nonparametric truncated spline regression model based on the interval of optimal knot points at each predictor variable. The following nonparametric truncated spline regression model is written based on the interval of optimal knot point at $x_1$. The optimal knot point $x_1$, such as 66.7, 67.4, 68.4, 69.5, and 72.5.

\[ \hat{y} = \begin{cases} 
-150.029 + 3.45006x_1, & x_1 < 66.7; \\
1059.168 - 14.67884x_1, & 66.7 \leq x_1 < 67.4; \\
-970.84 + 15.44x_1, & 67.4 \leq x_1 < 68.4; \\
400.49 - 4.6087x_1, & 68.4 \leq x_1 < 69.5; \\
2.1 + 1.12348x_1, & 69.5 \leq x_1 < 72.5; \\
1359.39 - 17.5223x_1, & x_1 \geq 72.5. 
\end{cases} \]

Nonparametric truncated spline regression model based on the interval of optimal knot points at $X_2, X_3$, and $X_4$ is written in the same way.

Significance parameter test is determining which predictor variable that significantly influence the response variable (Gujarati [6]). The significance parameter test consist of the simultaneous and partial test. The following hypothesis for this test is written by $H_0: \beta_1, \beta_2, \beta_3, \ldots, \beta_{23} = 0$ and $H_1: \text{at least one } k \neq 0; k = 1, 2, 3, \ldots, 23$. The level of significance = 0.05. The conclusion $H_0$ is rejected if $DK = \{F_{hitung} > F_{\alpha; k^{\text{model}}-k-1} = F(0.05; 23; 10) = 2.75\}$. The statistic test $F_{hitung} = 8.6640$. Because 8.6640 $\in$ DK, $H_0$ is rejected, which means there is at least one model parameter that significant to the regression model. In the attempt to determine that significant model parameter, partial test is performed. The following hypothesis for this test is written by $H_0: \beta_k = 0, k = 1, 2, \ldots, 23$ and $H_1: \beta_k \neq 0$. The level of significance = 0.05. The conclusion $H_0$ is rejected if $DK = \{F_{hitung} > F_{\frac{k^{\text{model}}-k}{2}} = 2.20\}$ Based on the calculation, there are 11 model parameters that significant to regression model. Then $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_{19}, \beta_{20}, \beta_{21}, \beta_{22}$ and $\beta_{23}$ declared as

| Parameter | Estimation | Parameter | Estimation |
|-----------|------------|-----------|------------|
| $\beta_0$ | -150.029   | $\beta_{12}$ | 0.000430924 |
| $\beta_1$ | 3.45006    | $\beta_{13}$ | 0.00808703  |
| $\beta_2$ | -18.1289   | $\beta_{14}$ | -0.0357964  |
| $\beta_3$ | 30.1189    | $\beta_{15}$ | 0.0365047   |
| $\beta_4$ | -20.0487   | $\beta_{16}$ | -0.0205421  |
| $\beta_5$ | 5.73218    | $\beta_{17}$ | 0.003598    |
| $\beta_6$ | -18.646    | $\beta_{18}$ | 0.0229005   |
| $\beta_7$ | -0.0000868926 | $\beta_{19}$ | -19.6633    |
| $\beta_8$ | 0.000217312 | $\beta_{20}$ | 20.2997     |
| $\beta_9$ | -0.000173979 | $\beta_{21}$ | -0.58512    |
| $\beta_{10}$ | 0.0000890236 | $\beta_{22}$ | 4.7376      |
| $\beta_{11}$ | -0.000086224 | $\beta_{23}$ | -5.54316    |
the significant model parameter to regression model. This indicates life expectancy and the percentage of an illeterate people are the significant predictor variables to HDI in Indonesia.

3.1. Residual Assumption Test

The following hypothesis for normality test is written by $H_0$ : residuals are normally distributed and $H_1$ : residuals are not normally distributed. The level of significance = 0.05. The conclusion $H_0$ is rejected if $DK = \{KS|KS > q_{1-\alpha;n} = q_{0.95;34} = 0.152\}$. The Kolmogorov Smirnov statistic test KS = 0.076. Because 0.076 $\notin$ DK, then $H_0$ is not rejected. This means residuals are normally distributed. The assumption of normality is fulfilled.

The assumption of independence is fulfilled.

The following hypothesis for heteroscedasticity test is written by $H_0$ : there is no heteroscedasticity against residuals dan $H_1$ : there is heteroscedasticity against residuals. The level of significance = 0.05. The conclusion $H_0$ is rejected if $DK = \{F_{hitung} | F_{hitung} > F_{(a,k-1,n-k)} = 2.60\}$. $F_{hitung} = 0.1979$. Because 0.1979 $\notin$ DK, then $H_0$ is not rejected. There is no heteroscedasticity against residuals. The assumption of heteroscedasticity is fulfilled.

4. Conclusion

The following nonparametric truncated spline regression model on the data of HDI in Indonesia based on the equation (1). Life expectancy and percentage of an illeterate people influenced the HDI in Indonesia.

References

[1]. Bintariningrum, M. F., dan I. N. Budiantara., 2014 Jurnal Sains dan Seni Vol. 3,7–12
[2]. BPS, Badan Pusat Statistik, 2015 Indeks Pembangunan Manusia di Negara Indonesia, Jakarta
[3]. Budiantara, I. N., 2014 Seminar Nasional FMIPA, Universitas Pendidikan Ganesha, Bali
[4]. Drapper, N.R., and H. Smith, 1992 Applied Regression Analysis, 2 ed., John Wiley and Sons. Inc., New York
[5]. Eubank, R. L., 1999 Spline Smoothing and Nonparametric Regression, 2 ed., Marcel Dekker.Inc, New York
[6]. Gujarati, N. D.,2009 Basic Econometrics, 4 ed., McGraw-Hill.Inc, New York
[7]. Hardle, W., 1994 Applied Nonparametric Regression, Cambridge University Press, New York
[8]. Lee, T. C. M., 2002 Statistica Vol.72, no. 8, 647–663
[9]. UNDP, United Nations Development Programee, 1990 Human Development Report, New York
[10]. Wahba, G., 1990 Spline Models for Observational Data, SIAM, Pennsylvania