# Semiparametric Spline Truncated Regression on Modelling AHH in Indonesia

Dewi Fitriana[1], I Nyoman Budiantara[1], Vita Ratnasari[1]

***Abstract*—Life expectancy (AHH) is an indicator that can reflect the health status of a region, whether from infrastructure, access, and health quality. As one of the dimensions of the Human Development Index (HDI), AHH is deemed to need global attention. Today, AHH growth in developing countries is slow even slower when compared to underdeveloped countries. In Indonesia, the trend of life expectancy at birth continues to increase, but the achievement of national AHH is still lagging behind the AHH of neighboring countries. In addition, as an archipelagic country, the achievement of AHH among provinces still shows a disparity.**

***This study modeled the factors that determine AHH with semiparametric spline truncated regression approach. The method of selecting optimum knots using Generalized Cross Validation (GCV) method. The best model that is formed is a model using three knots with the coefficient of determination of 84.70 percent. Significant variables were percentage of the poor population, percentage of households using clean drinking water, percentage of population who have health complaints, the percentage of under five years who have been complete immunized, and Mean Years of Schooling (MYS). The results of this study are expected to be an input for the Government to take policy in order to improve the national AHH as a whole.***

***Keywords*—Life Expectancy, Semiparametric, Spline Truncated, Likelihood Ratio Test (LRT).**

## I. INTRODUCTION

Regression analysis is one of the statistical methods are often used to determine the relationship between the response variable and the predictor variable. If the pattern of the relationship between the response variable and the predictor variables partly known and partly unknown pattern, it is advisable to use a semiparametric regression approach [1]. One model of semiparametric regression approach is spline. Spline models have very special and very nice statistical interpretation and visual interpretation, and it has a very good ability in handling data that behavior change in subgroups specified interval [2]. Therefore, the spline method developed in the last decade. Budiantara [3] developed a spline estimator in nonparametric regression by using a base spline function family. Truncated approach using spline basis family function truncated gives mathematical calculations easier and simpler, and optimization is used without involving *a penalty* that optimization Least Square (LS). Engle [4] introduced a semiparametric regression to estimate the relationship between weather and electricity sales to approach a linear spline. Ruliana [5] conducted a study on simultaneous hypothesis testing spline models on Structural Equation Modeling (SEM) Nonlinear.

This study will be modeling the factors that influence AHH in Indonesia by spline models truncated on regression semiparametric. The population is one of the assets owned by the nation. Indonesia is one of the countries that occupy the top five with the largest population in the world. In 2015, Indonesia was ranked fourth after China, India, and the United States. As a great nation, Indonesia should have the ability to perform development by making the component of HDI to measure the level of public health as well as a benchmark for the success of development.

According to observations by the WHO, AHH growth in developing countries slow moving even slower when compared with under developed country. Study Global Burden Disease (GBD) under the supervision *of the Institute for Health Metrics and Evaluation* (IHME) said that the developing countries are currently facing serious challenges that lifestyle and deadly diseases that are heart disease, stroke and diabetes that can affect all walks of life. Increased life expectancy nationally in Indonesia still puts Indonesia under AHH of neighboring countries. Sourced from the United Nations in its publications, BPS recorded AHH of few countries in the world during the period of 1990 - 2015. In the period of the year, AHH Indonesia was still under Singapore (82.2), Malaysia (74.9), Thailand (74.3), and Cambodia (71.6). Along with the increase AHH nationally, the provinces in Indonesia also experienced positive growth AHH. However, there appears disparity of AHH between provinces.

## II. LITERATURE REVIEW

### A. Spline Truncated in Semiparametric Regression

If given pairwise data $(x_i, t_i, y_i), i = 1,2, \dots, n$, where is the response variable, is the predictor variable following the parametric pattern and is the predictor variable following the nonparametric pattern, so the relationship patterns, and can be accepted in the regression model such as equations.

$$y_i = \tilde{x}_i{}'\tilde{\beta} + f(t_i) + \varepsilon_i, i = 1,2, \dots, n \quad (1)$$

Furthermore, if the regression curve is approximated by a spline regression curve with knots $K_1$, $K_2$, …, $K_r$ then:

$$f(t_i) = \sum_{j=1}^{m} \gamma_j t_i^j + \sum_{k=1}^{r} \gamma_{k+m}(t_i - K_k)_+^m \quad (2)$$

Where is parameter nonparametric component and *truncated* function $(t_i - K_k)_+^m$ is:

[1]Dewi Fitriana, I Nyoman Budiantara, Vita Ratnasari are with Department of Statistics, FMIPA, Institut Teknologi Sepuluh Nopember (ITS), Kampus ITS Sukolilo, Surabaya 60111, Indonesia. E-mail: dewi.fitriana@bps.go.id; i_nyoman_b@statistika.its.ac.id; vita_ratna@statistika.its.ac.id.

$$(t_i - K_k)_+^m = \begin{cases} 0 & , t_i < K_k \\ (t_i - K_k)^m & , t_i \geq K_k \end{cases} \quad (3)$$

The regression curve $f(t_i)$ is a truncated spline nonparametric regression curve with $m$ degree and with many of $r$ knots point, $m$ degree is a degree in a polynomial equation. The $K_1, K_2, \dots, K_r$ knots point are the knots that show the changes on pattern behavior of curves in the different interval subgroups, which $K_1 < K_2 < \cdots < K_r$. So, the truncated spline semiparametric regression equation in equation (1) becomes:

$$y_i = \tilde{x}_i{}' \tilde{\beta} \sum_{j=1}^{m} \gamma_j t_i^j + \sum_{k=1}^{r} \gamma_{k+m} (t_i - K_k)_+^m + \varepsilon_i \quad (4)$$

The truncated spline semiparametric regression above, consists of a response variable with one or more parametric predictor variables and only one nonparametric predictor variable. If the semiparametric regression consists of a response variable with more than one predictor variables, which is parametric and nonparametric components, with the composition of the data such as $(x_{i1}, \dots, x_{ip}, t_{i1}, \dots t_{iq}, y_i)$, then the relationship between $(x_{i1}, \dots, x_{ip}, t_{i1}, \dots t_{iq},)$ and $y_i$ can be written such us:

$$v_i = \mathbf{x_i'\beta_i} \sum_{l=1}^{q} \left( \sum_{j=1}^{m} \gamma_j t_i^j + \sum_{k=1}^{r} \gamma_{(k+m)} (t_{il} - K_{kl})_+^m \right) \quad (5)$$

Equation (5) also can be written as follows:

$$\tilde{y} = \mathbf{Z}(\mathbf{X}, \mathbf{T}, \tilde{k})\tilde{\delta} + \tilde{\varepsilon} \quad (6)$$

Response $\tilde{y}$ is vector $n \times 1$, matrix $\mathbf{Z}(\mathbf{X}, \mathbf{T}, \tilde{k}) = (\mathbf{X}: \mathbf{T})$, $\mathbf{X}$ is a matrix that contains predictors of parametric components of size $n \times (p+1)$ and $\mathbf{T}$ is a matrix that contains predictors of nonparametric components of size $n \times ((m+r)q)$ that depend on knots point $\tilde{k}$, where $\tilde{k}$ is vector of knots point of size $r \times 1$. $\tilde{\delta}' = (\beta_0, \dots, \beta_p, \gamma_{11}, \dots, \gamma_{m1}, \gamma_{1q}, \dots, \gamma_{mq}, \gamma_{(m+1)q}, \dots, \gamma_{(m+r)q})$ is parameter vector of size $((p+1)+(m+r)q) \times 1$ and $\tilde{\varepsilon}$ is error vector.

### *B. Parameter Estimation*

Getting the parameter estimation under $H(\Omega)$ using the Maximum Likelihood Estimation (MLE) method. Likelihood function under $H(\Omega)$ is

$$L(\tilde{\delta}_\Omega, \sigma_\Omega^2) = \prod_{i=1}^{n} \left( \frac{1}{\sqrt{2\pi\sigma_\Omega^2}} exp\left( -\frac{1}{2\sigma_\Omega^2} \left( \tilde{y} - \sum_{h=0}^{p} \beta_h x_{hi} + \sum_{l=1}^{q} \left( \sum_{j=1}^{m} \gamma_{jl} t_{li}^j + \sum_{k=1}^{r} \gamma_{(m+k)l} (t_{li} - K_{kl})_+^m \right) \right)^2 \right) \right)$$

$$\log L\left(\tilde{\delta}_n, \sigma_\Omega^2\right) = -\frac{n}{2}\log(2\pi_\Omega^2) - \frac{1}{2\pi_\Omega^2} [\tilde{y}'\tilde{y} - 2\tilde{\delta}'_\Omega \mathbf{Z}'(\mathbf{X}, \mathbf{T}, \tilde{k})\tilde{y} + \tilde{\delta}'_\Omega \mathbf{Z}'(\mathbf{X}, \mathbf{T}, \tilde{k})\mathbf{Z}(\mathbf{X}, \mathbf{T}, \tilde{k})\tilde{\delta}_\Omega] \quad (7)$$

The results of partial derivatives of the equation (7) are

$$\tilde{\delta}_\Omega = \left( \mathbf{Z}'(\mathbf{X}, \mathbf{T}, \tilde{k})\mathbf{Z}(\mathbf{X}, \mathbf{T}, \tilde{k}) \right)^{-1} \mathbf{Z}'(\mathbf{X}, \mathbf{T}, \tilde{k})\tilde{y} \quad (8)$$

and

$$\hat{\sigma}_\Omega^2 = \frac{(\tilde{y} - \mathbf{Z}(\mathbf{X}, \mathbf{T}, \tilde{k})\delta_\Omega)'(\tilde{y} - \mathbf{Z}(\mathbf{X}, \mathbf{T}, \tilde{k})\delta_\Omega)}{n} \quad (9)$$

Getting the parameter estimation under using the method of Lagrange Multiplier Function (LM). The LM function given that

$$F(\tilde{\delta}_\omega, \theta) = V(\tilde{\delta}_\omega) + 2\theta(\tilde{c}_j'\tilde{\delta}_\omega)$$
$$V(\tilde{\delta}_\omega) = (\tilde{y} - \mathbf{Z}(\mathbf{X}, \mathbf{T}, \tilde{k})\tilde{\delta}_\omega)'\mathbf{Z}(\mathbf{X}, \mathbf{T}, \tilde{k})\tilde{\delta}_\omega \quad (10)$$

with constraint $\tilde{c}_j'\tilde{\delta}_\omega = 0$

So that would be obtained $\hat{\tilde{\delta}}_\omega$ as below:

$$\hat{\tilde{\delta}}_\omega = \hat{\tilde{\delta}}_\Omega - \left( \mathbf{Z}'(\mathbf{X}, \mathbf{Z}, \tilde{k})\mathbf{T}(\mathbf{X}, \mathbf{Z}, \tilde{k}) \right)^{-1} \tilde{c}_j \left( \tilde{c}_j' \left( \mathbf{Z}'(\mathbf{X}, \mathbf{T}, \tilde{k})\mathbf{Z}(\mathbf{X}, \mathbf{T}, \tilde{k}) \right)^{-1} \tilde{c}_j \right)^{-1} \tilde{c}_j' \tilde{\delta}_\Omega \quad (11)$$

### *C. Formulation of Partial Hypothesis Testing*

Formulation of partial hypothesis used to test the significance of the parameters on semiparametric spline truncated regression model is as below:

$$H_0: \tilde{c}_j'\tilde{\delta} = \tau, H_1: \tilde{c}_j'\tilde{\delta} \neq \tau \quad (12)$$

Where

$\tilde{c}_j' = (0\ 0 \dots 1 \dots 0\ 0)_{(1 \times p+1+(m+r)q)}$

$\tilde{\delta} = (\beta_0, \dots, \beta_{h-1}, \dots \beta_{h+1}, \dots, \beta_p, \gamma_{11}, \dots, \gamma_{(m+1)q}, \dots, \gamma_{(m+r)q})$

Define the parameters space under $H(\Omega)$ is as below:

$$\Omega = \{\delta = (\beta_0, \dots, \beta_p, \gamma_{11}, \dots, \gamma_{(r+m)1}, \dots, \gamma_{1q}, \dots, \gamma_{(m+r)q}), \sigma_\Omega^2\} \quad (13)$$

Define the parameters space under $H_0(\omega)$ is as below:

$$\omega = \{\delta = (\beta_0, \dots, \beta_{h-1}, \dots \beta_{h+1}, \dots, \beta_p, \gamma_{11}, \dots, \gamma_{(r+m)1}, \gamma_{1q}, \dots, \gamma_{(r+m)q}), \sigma_\Omega^2 | \tilde{c}_j'\tilde{\delta} = 0\} \quad (14)$$

### *D. Statistics Test and Rejection Area for Partial Hypothesis*

Furthermore, to obtain a statistical hypothesis test of equation (12) were completed using the LRT.

The maximum likelihood function under space $H(\Omega)$ is

$$L(\hat{\Omega}) = (2\pi\sigma_\Omega^2) - \frac{n}{2e} - \frac{n}{2} \quad (15)$$

The maximum likelihood function under space $H(\omega)$ is

$$L(\hat{\omega}) = \left(2\pi\hat{\sigma}_\omega^2\right) - \frac{n}{2e} - \frac{n}{2}$$

$$L_{ratio} = \lambda(y) = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \frac{\left(2\pi\hat{\sigma}_\omega^2\right) - \frac{n}{2e} - \frac{n}{2}}{\left(2\pi\sigma_\Omega^2\right) - \frac{n}{2e} - \frac{n}{2}}$$

$$= \left( \frac{1}{1 + \frac{(\tilde{c}_j'\hat{\tilde{\delta}}_\Omega)' \left( c'(\mathbf{Z}'(\mathbf{X}, \mathbf{Z}, \tilde{k})\mathbf{T}(\mathbf{X}, \mathbf{Z}, \tilde{k}))^{-1} c_j \right)^{-1} c_j' \hat{\tilde{\delta}}_\Omega}{(\tilde{y} - \mathbf{Z}(\mathbf{X}, \mathbf{T}, \tilde{k})\hat{\tilde{\delta}}_\Omega)'(\tilde{y} - \mathbf{Z}(\mathbf{X}, \mathbf{T}, \tilde{k})\hat{\tilde{\delta}}_\Omega)} } \right)^{\frac{n}{2}} = \left( \frac{1}{1 + \frac{Q_1}{Q_2}} \right)^{\frac{n}{2}} \quad (16)$$

Then the statistic test is

$$W^* = \frac{Q_1/(1)}{Q_2/(n-(p+1+(r+m)q))} \quad (17)$$

Distribution of statistical test in equation (17) is

$$W^* \sim F_{(1,(n-(p+1+(r+m)q)))} \quad (18)$$

The critical area is $\lambda < k$, where $0 < \lambda < 1$,

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} < k$$

$$\frac{Q_1/(1)}{Q_2/(n-(\mathrm{p}+1+(r+m)q))} > \left(\left(\frac{1}{k}\right)^{\frac{n}{2}} - 1\right)\frac{(n-(\mathrm{p}+1+(r+m)q))}{1} \quad (19)$$

So, $W^* > K^*$

Critical area for hypothesis test $H_0: \tilde{c}_j'\tilde{\delta} = \tau, H_1: \tilde{c}_j'\tilde{\delta} \neq \tau$ is

$$C = \{(x_1, x_2, \dots, x_p, t_1, t_2, \dots t_q, y); W^* > k^*\}$$

Where $k^* = \left(\left(\frac{1}{k}\right)^{\frac{2}{n}} - 1\right)\frac{(n-(\mathrm{p}+1+(r+m)q))}{1}$

### *E. Selection of Optimal Knot Point*

The important thing in semiparametric spline truncated regression is the selection of optimal knot point. One commonly used method of choosing an optimal knot point is the Generalized Cross Validation (GCV).

$$CGV(\tilde{k}) = \frac{n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\left[n^{-1}trace\left(\mathbf{I}-\mathbf{A}(\tilde{k})\right)\right]^2} \quad (20)$$

Where $y_i$ is variable response, $\hat{y}_i$ is the estimated value of variable response, $i = 1,2,\dots,n$ number of observation $\tilde{k} = (K_1, K_2, \dots, K_1)^1$, knots point, Matrix $A(\tilde{k}) = Z(\tilde{k})\left(Z'(\tilde{k})Z(\tilde{k})\right)^{-1}Z'(\tilde{k})$ and **I** identity matrix.

### *F. Life Expectancy (AHH)*

Life Expectancy at birth by the World Bank is the average number of years of life expectancy of a group people born in the same year, assuming deaths at each age remain constant in the future [6].

## III. Research Methodology

### *A. Source Data and Research Variables*

The data used in this research is secondary data, i.e derived data from the National Socio Economic Survey (Susenas) 2015 published by the Statistics Indonesia (BPS) in Welfare Statistics 2015 and education data has been published in the Human Development Index 2015. The observation unit used in this study was all provinces in Indonesia.

### *B. Variables Used*

In this study, the variables used are the response variable that life expectancy in Indonesia, the predictor variables, i.e percentage of poor population, the percentage of households using clean drinking water the percentage of population who had health complaints, the percentage of under-fives who have been complete immunized, and Mean Years of Schooling.

### *C. Step of Analysis*

Stages of research as follows:

1. Create plot the response variable with each predictor variable.
2. Determine the variable component of parametric and nonparametric components.
3. Modeling the relationship between the response variable and the predictor variables using semiparametric spline truncated estimator for 1 knots point, two knots point, 3 knots point and knots point combinations
4. Choosing the optimal knots point based on GCV method.
5. Testing the significance of parameters simultaneously.
6. Perform a partial significance testing parameters.
7. Examine the assumptions are independent, identical and normal distribution for residuals.
8. Make interpretation semiparametric spline truncated regression of the model AHH in Indonesia.

## IV. Results and Discussion

### *A. Analysis Descriptive*

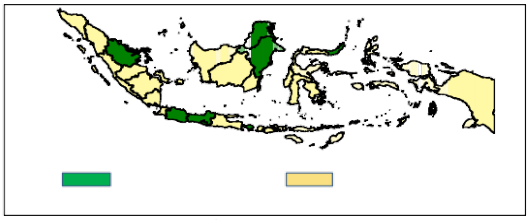

**Figure 1.** Map of AHH by Province in indonesia

Figure 1 shows that most provinces in Indonesia have under the national AHH, 70.78 years. The provinces that have reached over the national AHH are Riau, North Sulawesi, Bali, North Kalimantan, West Java, DKI Jakarta, East Kalimantan, Central Java and DIY Province. Almost all areas of eastern Indonesia have under the national AHH.

Characteristics of each variable, both response and predictor variables can be seen in table 1.

TABLE 1.
STATISTIC DESCRIPTIVE OF RESPONSE VARIABLE AND PREDICTOR VARIABLES

| Variable | Minimum | Maximum | Range | Mean | Standar Deviasi |
|---|---|---|---|---|---|
| $y$ | 64,22 | 74,68 | 10,64 | 69,32 | 2,66 |
| $x_1$ | 3,92 | 27,33 | 23,40 | 11,83 | 6,16 |
| $x_2$ | 41,08 | 93,4 | 52,32 | 68,62 | 11,04 |
| $x_3$ | 16,71 | 39,58 | 22,87 | 28,42 | 6,11 |
| $x_4$ | 40,36 | 76,01 | 35,65 | 59,17 | 8,73 |
| $x_5$ | 5,99 | 10,70 | 4,71 | 8,02 | 0,96 |

Based on table 1, the response variable (y) of AHH Indonesia has an average value of 69.32 years with a standard deviation of 2.66. The range of AHH value in Indonesia involve 34 provinces is 10.46. The highest AHH is 74.68 years is in the Province of DIY. While thelowest AHH is 64.22 years in the Province of West Sulawesi.

The highest percentage of poor population Indonesia in 2015 is located in Papua province and the lowest is in DKI Jakarta Province, that is 3.93 percent. The average percentage of poor population Indonesia in 2015 is 11.83 percent with a standard deviation 6.16.

The lowest percentage of households using clean drinking water in 2015 was Bengkulu Province 41.08 percent and the highest was DKI Jakarta at 93.40 percent. So the percentage range of households using clean drinking water is 52.32 percent. The average percentage of households

using clean drinking water in Indonesia is 68.62 percent with standard deviation of 11.04.

The average of percentage population who had health complaints in 2015 is 28.42 percent with a standard deviation of 6.11. The percentage of population who had health complaints in Indonesia is 22.87 percent with the highest is DIY Province is 39.58 percent and the lowest percentage is North Maluku Province is 16.71 percent.

Percentage of population aged 15 years and over in 2015 are on average educated for 8 years. So that the population aged 15 years and over in general have education up to the Junior High School level. The MYS range in Indonesia is 4.71 years, which is between 5.99 years and 10.70 years. MYS lowest in Papua Province, while the highest MYS in DKI Jakarta Province.

### *B. Determine Parametric and Nonparametric Variables Component*

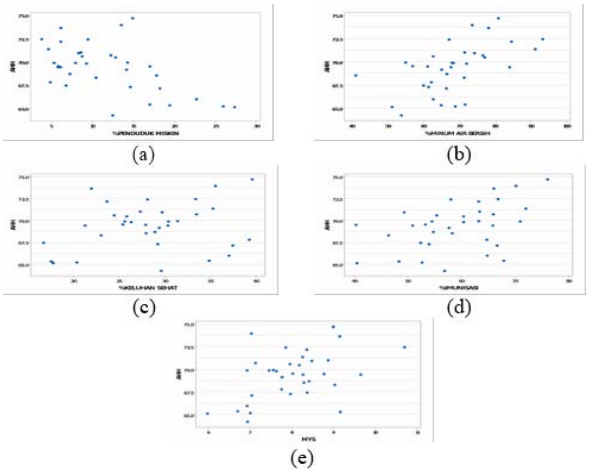


**Figure 2.** Plot the response variable and predictor variables

In figure 2 there are five *scatter plot* between the response variable and each predictor variable. Plot (a) is between AHH with the percentage of poor population, plot (b) between AHH with the percentage of households using clean drinking water, plot (c) between AHH with the percentage of population who had health complaints, plot (d) between AHH with percentagesof under-fives who have been complete immunized and plot (e) between AHH with Mean Years of Schooling (MYS). From the results of the plot shows that the only plot (a) which tends to form a certain pattern, a pattern that looks relationships tend to follow a straight line or linear. While the plot (b), (c), (d) and (e) looks less likely to form a specific pattern and the pattern of relationships which looks likely to change the behavior of sub intervals. So the conclusion of variables including parametric and nonparametric components as shown in table 2.

TABLE 2.
SUMMARY OF PARAMETRIC AND NONPARAMETRIC VARIABLE COMPONENT

| Notation | Variable | Note |
|---|---|---|
| $x_1$ | Percentage of Poor Population | Parametric |
| $t_1$ | Percentage of households using clean drinking wat | Nonparametri |
| $t_2$ | Percentage of Population Who Had Health Complaint | Nonparametri |
| $t_3$ | Percentage of Under-Fives Who Have Been Complete Immunized | Nonparametri |
| $t_4$ | MYS | Nonparametri |

### *C. Modeling AHH with Semiparametric Spline Truncated*

Modeling AHH with semiparametric spline truncated depend on knot that used, there are tried one knot, two knots, three knots, and combination knots.

TABLE 3. GCV VALUE WITHT ONE KNOT POINT

| Knot | | | | GVC |
|---|---|---|---|---|
| $t_1$ | $t_2$ | $t_3$ | $t_4$ | |
| 61.37 | 25.58 | 54.18 | 7.82 | 5.04 |

Minimum GCV value generated with one knot is equal to 5.04. The point of knots on percentage of households using clean drinking water variable $(t_1)$ is 61.37 $(K_{11})$ percentage of population who had health complaints variable is $(t_2)$ 25.58 $(K_{12})$ percentage of underfives who have been complete immunized variable is $(t_3)$ 54.18 $(K_{13})$ and MYS variable is $(t_4)$ 7.82 $(K_{12})$. GCV value generated using semiparametric regression spline truncated with two knots are presented in Table 4.

TABLE 4.
GCV VALUE WITH TWO KNOTS POINT

| Knot | | | | GVC |
|---|---|---|---|---|
| $t_1$ | $t_2$ | $t_3$ | $t_4$ | |
| 47.49 | 19.51 | 44.73 | 6.57 | 4.76 |
| 52.83 | 21.84 | 48.36 | 7.05 | |

Shown in Table 4, the value of minimum GCV generated is equal to 4.76. The point of knots on a variable percentage of households using clean drinking water $(t_1)$ is 47.49 $(K_{11})$ and $(K_{21})$ 52.83 percentage of the population who had health complaints variable $(t_2)$ is 19.51 $(K_{12})$ and $(K_{22})$ 21.84 percentage of under-fives who have been complete immunized variable $(t_3)$ is 44.73 $(K_{13})$ and $(K_{23})$ 48.36 and MYS variable $(t_4)$ is 6.57 $(K_{14})$ and 7.05 $(K_{24})$. GCV value is generated by using spline semiparametric regression truncated to three knots are presented in Table 5. Based on Table 5 minimum GCV value generated is equal to 4.19. The point of knots on percentage of households using clean drinking water variable is 60.30 82.72 and 83.79 percentage of population who had health complaints variable is 25.11 34.91 and 35.38 percentage of under-fives have been complete immunized children variable is 53.46 68,73 and 69.46 and MYS variable is 7.72 9.74 and 9.83).

TABLE 5.
GCV VALUE WITH THREE KNOTS POINT

| Knot | | | | GVC |
|---|---|---|---|---|
| $t_1$ | $t_2$ | $t_3$ | $t_4$ | |
| 60.30 | 25.11 | 53.46 | 7.72 | 4.19 |
| 82.72 | 34.91 | 68.73 | 9.74 | |
| 83.79 | 35.38 | 69.46 | 9.83 | |

GCV value generated using semiparametric regression spline truncated by a combination of knots point are presented in table 6.

TABLE 6.
CGV VALUE WITH COMBINATION KNOTS POINT

| Combination | Knot | | | | CGV |
|---|---|---|---|---|---|
| | $t_1$ | $t_2$ | $t_3$ | $t_4$ | |
| (3,1,2,2) | 60,3 | 25,58 | 44,75 | 6,57 | 4,47 |
| | 82,72 | | 48,36 | 7,05 | |
| | 83,79 | | | | |

Based on table 6 minimum GCV value generated is equal to 4,47. The point of knots on percentage of households using clean drinking water variable ($t_1$) is 60,30 ($K_{11}$), 82,72 ($K_{21}$) , and 83,70 ($K_{31}$), percentage of population who had health complaints variable ($t_2$) is 25,58 ($K_{12}$), percentage of under-fives have been complete immunized variable ($t_3$) is 44,73 ($K_{13}$) and 48,36 ($K_{23}$) and MYS variable ($t_4$) yaitu 6,57 ($K_{14}$) dan 7,05 ($K_{24}$) .

TABEL 7.
CGV VALUE EACH MODEL

| Knot Model | GCV |
|---|---|
| One Knot Poin | 5,04 |
| Double Knots Point | 4,76 |
| Three Knots Point | 4,19 |
| Combination Knots Point | 4,47 |

Table 7 shows that there is a minimum GCV at three knots points in the amount 4.19. So that the best semiparametric regression model spline truncated is model with three-point knots on percentage of households using clean drinking water variable, three points knots on percentage of the population who had health complaints variable, three point knots on percentage of under-five have been complete immunized variable and three point knots on MYS variable. So semiparametric regression model spline truncated formed as follows:

$$y = \beta_0 + \beta_1 x_1 + \gamma_{11} t_1 + \gamma_{21}(t_1 - K_{11})_+^1 + \gamma_{31}(t_1 - K_{21})_+^1 + \gamma_{41}(t_1 - K_{31})_+^1 + \gamma_{12} t_2 + \gamma_{22}(t_2 - K_{12})_+^1 + (t_2 - K_{22})_+^1 + \gamma_{42}(t_2 - K_{32})_+^1 + \gamma_{13} t_3 + (t_3 - K_{13})_+^1 + \gamma_{33}(t_3 - K_{23})_+^1 + \gamma_{43}(t_3 - K_{33})_+^1 + \gamma_{14} t_4 + \gamma_{24}(t_4 - K_{14})_+^1 + \gamma_{34}(t_4 - K_{24})_+^1 + \gamma_{44}(t_4 - K_{34})_+^1 + \varepsilon$$

## D. *Testing The Significant Parameter Simultanously*

Testing hypothesis to test the significance of parameters simultaneously using the following hypothesis:

$H_0: \beta_h = 0 ; H_1: \beta_h \neq 0 ; h = 0,1$

$H_0: \gamma_{jl} = 0 ; H_1: \gamma_{jl} \neq 0 ; j = 1 ; l = 1,2,3,4$

$H_0: \gamma_{(m+)l} = 0 ; H_1: \gamma_{(m+)l} \neq 0 ; m = 1 ; k = 1,2,3 ; l = 1,2,3,4$

The result of partial testing hypothesis is shown in table 9. Rejection of $H_0$ if $W^* > F_{tabel}$ or $p - value < \alpha$

TABLE 9.
THE RESULT OF PARTIAL TESTING

| Variabel | Parameter | Estimator | W* | p-value |
|---|---|---|---|---|
| Konstan | $\beta_0$ | 68,65 | 44,21 | 0,00 |
| $x_1$ | $\beta_1$ | -0,23 | 9,02 | 0,01 |
| $t_1$ | $\gamma_{11}$ | -0,12 | 1,82 | 0,19 |
| | $\gamma_{21}$ | 0,30 | 4,57 | 0,04 |
| | $\gamma_{31}$ | -1,81 | 1,39 | 0,25 |
| | $\gamma_{41}$ | 0,60 | 0,11 | 0,74 |
| $t_2$ | $\gamma_{12}$ | 0,06 | 0,13 | 0,73 |
| | $\gamma_{22}$ | -0,33 | 1,40 | 0,25 |
| | $\gamma_{32}$ | 17,09 | 11,22 | 0,00 |
| | $\gamma_{42}$ | -19,66 | 11,12 | 0,00 |
| $t_3$ | $\gamma_{13}$ | -0,36 | 7,9 | 0,01 |
| | $\gamma_{23}$ | 0,57 | 8,27 | 0,01 |
| | $\gamma_{33}$ | -6,94 | 3,45 | 0,08 |
| | $\gamma_{43}$ | 8,93 | 4,12 | 0,06 |
| $t_4$ | $\gamma_{14}$ | 3,36 | 9,13 | 0,00 |
| | $\gamma_{24}$ | -6,19 | 9,12 | 0,00 |
| | $\gamma_{34}$ | 9,42 | 12,23 | 0,00 |
| | $\gamma_{44}$ | 8,48 | 12,23 | 0,00 |

Table 9 shows that there are 11 parameters significant of 18 parameters. All predictor variables affect the response variable.

## E. *Interpretation*

The spline truncated semiparametric regression model with three knots has fulfilled the residual assumption of IIDN, so the resulting model can be interpreted further. The best spline truncated semiparametric regression model generated is as follows:

$$\hat{y} = 68,65 - 0,23x_1 - 0,12t_1 + 0,30(t_1 - 60,30)_+ - 1,81(t_1 - 82,72)_+ + 0,60(t_1 - 83,79)_+ + 0,06t_2 - 0,33(t_2 - 25,11)_+ + 17,09(t_2 - 34,91)_+ - 19,66(t_2 - 35,38)_+ - 0,36t_3 + 0,57(t_3 - 53,46)_+ - 6,94(t_3 - 68,73)_+ + 8,93(t_3 - 69,46)_+ + 3,63t_4 - 6,19(t_4 - 7,72)_+ + 9,42(t_4 - 9,74)_+ + 8,48(t_4 - 9,83)_+$$

Model interpretation of the variables that significantly influence is as follows:

a. The percentage of poor population with other assumptions of constant variables is as follows:

$$\hat{y} = 68,65 - 0,23x_1$$

b. If there is a percentage of poor population increase as much as one percent so AHH will decrease by 0.23 percent. The percentage of households using clean drinking water with assumption other variables are constant as follows:

$$\hat{y} = \begin{cases} -0,12t_1 & ,t_1 < 60,30 \\ 0,18t_1 - 18,27 & ,60,30 \leq t_1 < 82,72 \\ -1,63t_1 + 131,6 & ,82,72 \leq t_1 < 83,79 \\ -1,03t_1 + 81,54 & ,t_1 \geq 83,79 \end{cases}$$

In the percentage of households using clean drinking water, there are four sub-intervals that have behavioral changes. For the percentage of households using clean drinking water between 60.30 and 82.7 percent, each

increase of one percent then AHH will increase by 0.18 years.

c. The percentage of population who had health complaints with assumption other variables are constant as follows:

$$\hat{y} = \begin{cases} 0,06t_2 & ,t_2 < 25,11 \\ -0,27t_2 + 8,41 & ,25,11 \le t_2 < 34,91 \\ 16,82t_2 - 588,26 & ,34,91 \le t_2 < 35,38 \\ -2,84t_2 + 107,31 & ,t_2 \ge 35,38 \end{cases}$$

For provinces with a percentage of population who had health complaints between 25.11 percent and 34.91 percent, any increase of one percent of the population with health complaints will cause AHH to decrease by 0.27 years. As for the percentage of the population who had health complaints more than 35.38 percent, if the percentage of people who had health complaints increased 1 percent then AHH will decrease by 2.84 years.

d. The percentage of under-fives who have been immunized with assumption other variables are constant as follows:

$$\hat{y} = \begin{cases} -0,36t_3 & ,t_3 < 53,46 \\ 0,21t_3 - 14,22 & ,53,46 \le t_3 < 68,73 \\ -6,73t_3 + 227,92 & ,68,73 \le t_3 < 69,46 \\ 2,2t_3 - 88,02 & ,t_3 \ge 69,46 \end{cases}$$

In the percentage of under-five who have been complete immunized, there were four sub-intervals of behavior change. For the percentage of under fives whohave been complete immunized from 53.46 percent to 68.73 percent, if an increase in percentage of underfives who have been complete immunizied by one percent, AHH will rise 0.21 years. There are 21 provinces in Indonesia that have this kind of behavior. As for the percentage of under- fives who have been complete immunizied above 69.46 percent, if there is an increase in percentage of under- fives who have been complete immunizied of one percent then the value of AHH will rise 2.2 years. Areas that have this behavior are Central Java, Bangka Belitung, Bali and DIY.

e. The MYS with assumption other variables are constant as follows:

$$\hat{y} = \begin{cases} 3,63t_4 & ,t_4 < 7,72 \\ -2,56t_4 + 47,79 & ,7,72 \le t_4 < 9,74 \\ 6,86t_4 - 43,96 & ,9,74 \le t_4 < 9,83 \\ 15,34t_4 - 127,32 & ,t_4 \ge 9,83 \end{cases}$$

The MYS variable has four behavioral sub-intervals. For MYS value less than 7.72 years, if the value MYS rise one year, AHH will rise 3.63 years. Areas with such patterns of behavior are the provinces of Papua, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, West Sulawesi, West Papua, Central Java, Gorontalo, East Java, Bangka Belitung, Lampung and South Sulawesi. Policies to improve the quality of life such as AHH through improved education in these provinces are appropriate. Meanwhile, there is one province that has a pattern of every 1-year increase of MYS, it will increase AHH by 15.34 years, that is DKI Jakarta Province.

## V. Conclusion

The best model is model with three knots. Of the five variables used have a significant effect on the model. The coefficient of determination ($R^2$) obtained is 84.70 percent, so the model is feasible to use.

## Acknowledgement

This study was financially supported by the Statistics of Indonesia.

## References

[1] G. Wahba, *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

[2] R. L. Eubank, *Nonparametric Regression and Spline Smoothing, Second Edition*. New York: Taylor & Francis, 1999.

[3] I. N. Budiantara, "Model Keluarga Spline Polinomial Truncated dalam Regresi Semiparametrik," *Berk. Ilm. MIPA*, vol. 15, no. 3, 2005.

[4] R. F. Engle, C. W. J. Granger, J. Rice, and A. Weiss, "Semiparametric Estimates of the Relation Between Weather and Electricity Sales," *J. Am. Stat. Assoc.*, vol. 81, no. 394, p. 310, Jun. 1986.

[5] Ruliana, I. N. Budiantara, Otok B. W., and Wibowo W., "Simoltaneus Hypothesis Testing of Spline Truncated Model in Nonlinear Structural Equation Modeling (SEM)," *J. Theor. Appl. Inf. Technol.*, vol. 3189, no. 2, 2016.