Application of Confidence Intervals for Parameters of Nonparametric Spline Truncated Regression on Index Development Gender in East Java

Rifani Nur Sindy Setiawan[1], I Nyoman Budiantara[1], Vita Ratnasari[1]

***Abstract*—The Gender Development Index (GDI) is an index that measures the achievement of human basic capability development for the health, education and economic sectors within a region by considering equality between men and women. In this research GDI with the factors that are suspected to affect it will be modeled using nonparametric spline truncated regression, because the results of scatterplot pattern of GDI with some predictor variables not to follow a certain pattern. Determination of predictor variables that significantly influence GDI by using confidence interval obtained high school enrollment rate of female population, morbidity of female population, percentage of last aid of birth by medical, and female labor-force participation rate have significantly influenced to GDI in East Java.**

***Keywords*—Gender Development Index (GDI), Nonparametric Regression, Spline Truncated, Confidence Interval.**

I. INTRODUCTION

Regression analysis is method used to determined the relationship between response variable with one or more predictor variable. The regression approach is divided into three approaches: parametric regression, semiparametric regression, and nonparametric regression. Parametric regression approach requires assumptions such as shape of the curve must be known, error normal distribution and homogeneous variance. Nonparametric regression approach is a statistical method used to determine the relationship between the response variable with predictor variables of unknown function form, simply assumed to be smooth in the sense contained in a particular function space. Nonparametric regression approach is very flexible regression to model the data pattern [1]. While the semiparametric regression approach is a regression that contains parametric components and nonparametric components.

Nonparametric regression has been developed, such as using kernel, spline, polynomial local, and deret fourier. Spline has several advantages including having a simple and good statistical interpretation and good visual representation. Spline univariable regression is nonparametric regression analysis if there is one response variable and one predictor variable. If there is one response variable and more than one predictor variable then it is called Spline multivariable regression [2].

One of the most important parts of statistical inference is the confidence interval. The confidence intervals for parameters of nonparametric spline truncated regression. Confidence intervals for parameters of nonparametric regression can be used to determine predictor variables that significantly influence response variables. The conclusion is based on the confidence interval for parameter that contains a zero value. If the confidence interval contains a zero value, then the predictor variable has no significant effect on the response variable.

National development is human development and development of Indonesian society fairly and equally. However, in order to create these conditions, there are several problems, including the existence of the gap in development achievement between women and men as well as the low quality of life and the role of women in development. The gender gap in various areas of development is also marked by the low opportunities women have for work, as well as low access to economic resources such as technology, information, markets, credit and working capital. The different gender roles that exist in Indonesia is a matter of social injustice that places women as the main victims. Forms of gender inequality and gender justice are known as gender disparities that will cause gender issues [3].

To improve gender equality, the women's basic needs such as health, education, employment and participation must be considered. These basic needs reflect the quality of human resources. The Government has strived to realize gender equality and justice in the life of society and the state through several policies and programs. But in practice there are still many obstacles and challenges [4].

To evaluate the development already accommodating the gender aspects can be used related indicators, such as the

[1]Rifani Nur Sindy Setiawan, I Nyoman Budiantara, Vita Ratnasari are with Department of Statistics, Faculty of Mathematics, Computing, and Data Science, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia. E-mail: setiawan15@mhs.statistika.its.ac.id; i_nyoman_b@statistika.its.ac.id; vita_ratna@statistika.its.ac.id.

Gender Development Index (GDI). GDI was introduced by the United Nations Development Programs (UNDP) in the 1995 Human Development Report. This GDI number is expected to provide information on the development outcomes that already accommodate gender aspects [6]. Based on data released by UNDP in BPS publications, GDI Indonesia still occupies a low position compared to ASEAN countries (excluding Vietnam and Myanmar) which is the third lowest position after Timor Leste and Cambodia.

The position of East Java GDI achievement in 2014 separated 20 districts/cities under the achievement of the provincial IPG and 18 districts/cities above the achievement of provincial GDI. This condition illustrates that there are still many districts/cities that need improvement in programs that lead to gender mainstreaming. In 2014 there is a disparity of GDI numbers between provinces in Indonesia. On a nationwide scale of IPG achievements from 34 provinces, East Java ranks 16th. In Java Island, East Java occupies the second lowest position after West Java. If seen the development of East Java IPG value from 2010 until 2014 has increased.

[5] has discussed gender differences in East Java, which concluded that factors affecting the gender gap are the junior secondary enrollment rate for the female population, the percentage of the female population with junior secondary education, and the percentage of the female population employed in the formal sector. [6] also studied the components of the Gender Development Index in East Kalimantan and South Kalimantan Province in 2011. From these studies, it was found that factors affecting GDI components in the provinces of East Kalimantan and South Kalimantan for sex were population density, ratio's facility health, percentage of educated population above junior high, and unemployment rate. While the factor that affects the female's GDI is the percentage of educated population above junior high.

Based on the description, GDI with the factors that are suspected to affect it will be modeled using nonparametric spline truncated regression, because the results of scatterplot pattern of GDI with some predictor variables ie high school enrollment rate of female population, morbidity of female population, percentage of last aid of birth by medical, and female's labor-force participation rate not to follow a certain pattern. The determination of the variabel that have significantly influence using confidence interval.

II. Literature Review

A. Nonparametrik Spline Truncated Regression

Spline is a piece of polynomials, ie polynomials having continuous segmented nature. Spline has a high flexibility and has the ability to estimate data behavior that tends to differ at different intervals [1]. If given data $x_{1i}, x_{2i}, \dots, x_{pi}, y_i$ and the relationship between $x_{1i}, x_{2i}, \dots, x_{pi}$ and y_i follow multivariable:

$$y_i = f(x_{1i}, x_{2i}, \dots, x_{pi}) + \varepsilon_i, i = 1,2, \dots n \quad (1)$$

Where f is the unknown curve of the regression form. If the regression curve f is assumed to be additive and approached with linear spline function, then obtained regression model.

$$y_i = f(x_{1i}) + f(x_{2i}) + \cdots + f(x_{pi}) + \varepsilon_i = \sum_{j=1}^{p} f(x_{1i}) + \varepsilon_i, i = 1,2, \dots, n \quad (2)$$

where

$$f(x_{ji}) = \beta_0 + \beta_{j1}x_{ji} + \sum_{u=1}^{r} \beta_{j(1+u)} (x_{ji} - K_{ju})_+^1 \quad (3)$$

with

$$(x_{ji} - K_{ju})_+^1 = \begin{cases} (x_{ji} - K_{ju})_+^1, x_{ji} \geq K_{ju} \\ 0 \quad , x_{ji} \geq K_{ju} \end{cases} \quad (4)$$

With $K_{j1}, K_{j2}, \dots K_{jr}$ are knot points showing the pattern of behavioral changes of functions at different sub-intervals.

B. Selection of Optimal Knots Point

The knot point is a common fusion point where there is a change in function behavior at different intervals [7]. One method used to select the optimal knot point is to use the GCV (Generalized Cross Validation) method [8]. The best Spline model is obtained from the smallest GCV value.

$$GCV(K_1, K_2, \dots K_r) = \frac{MSE\ (K_1, K_2, \dots K_r)}{(n^{-1} tr[I - A(K_1, K_2, \dots K_r)])^2} \quad (5)$$

where $GCV(K_1, K_2, \dots K_r) = n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, $(K_1, K_2, \dots K_r)$, are knot point, and matrix $A(K_1, K_2, \dots K_r) = \boldsymbol{X(X'X)^{-1}X'}$ [9].

C. Estimation for parameters of Nonparametric Spline Truncated Regression

The estimation for parameters of nonparametric spline truncated regression. If given model of nonparametric spline truncated regression with r knot

$$y_i = \beta_0 + \sum_{j=1}^{p} \left(\beta_{j1}x_{ji} + \sum_{u=1}^{r} \beta_{j(1+u)} (x_{ji} - K_{ju})_+^1 \right) + \varepsilon_i, \quad i = 1,2, \dots, n \quad (6)$$

If equation (5) is expressed in matrix form $\underset{\sim}{y} = X(K)\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$, with

$$\underset{\sim}{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \underset{\sim}{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\mathbf{X(K)} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} & (x_{11} - K_{11})_+^1 & \cdots & (x_{11} - K_{1r})_+^1 & \cdots & (x_{p1} - K_{11})_+^1 & \cdots & (x_{p1} - K_{pr})_+^1 \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} & (x_{12} - K_{11})_+^1 & \cdots & (x_{12} - K_{1r})_+^1 & \cdots & (x_{p2} - K_{11})_+^1 & \cdots & (x_{p2} - K_{pr})_+^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} & (x_{1n} - K_{11})_+^1 & \cdots & (x_{1n} - K_{1r})_+^1 & \cdots & (x_{pn} - K_{11})_+^1 & \cdots & (x_{pn} - K_{pr})_+^1 \end{pmatrix}$$

$$\underset{\sim}{\beta}' = (\beta_0\ \beta_{11}\ \beta_{21}\ \dots\ \beta_{p1}\ \beta_{12}\ \beta_{13}\ \dots\ \beta_{1(1+r)} \dots\ \beta_{p2}\ \beta_{p3}\ \dots\ \beta_{p(1+r)})$$

If assumed $\underset{\sim}{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$, sin ce $\underset{\sim}{y}$ is a linier combination of $\underset{\sim}{\varepsilon}$ then $\underset{\sim}{y}$ is also normally distributed

$$\underset{\sim}{y} \sim N\left(E\left(\underset{\sim}{y}\right), Var\left(\underset{\sim}{y}\right)\right) \quad (7)$$

$$E\left(\underset{\sim}{y}\right) = E\left(\mathbf{X(K)}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}\right) = \mathbf{X(K)}\underset{\sim}{\beta} + E\left(\underset{\sim}{\varepsilon}\right) = \mathbf{X(K)}\underset{\sim}{\beta}$$

$$\begin{aligned} Var\left(\underset{\sim}{y}\right) &= Var\left(\mathbf{X(K)}\underset{\sim}{\beta}+\underset{\sim}{\varepsilon}\right) \\ &= 0+Var\left(\underset{\sim}{\varepsilon}\right) \\ &= \sigma^2\mathbf{I} \end{aligned} \quad (8)$$

Based on equation (7) and (8) obtained $\underset{\sim}{y}$ is normally distributed with mean $\mathbf{X(K)}\underset{\sim}{\beta}$and variance $\sigma^2\mathbf{I}$. One method that can be used to get point estimation of $\underset{\sim}{\beta}$ is by using the Maximum Likelihood Estimation (MLE). The probability distribution of $\underset{\sim}{\varepsilon}$ is

$$g\left(\underset{\sim}{\varepsilon}\right)=\frac{1}{\sqrt{2\Pi\sigma^2}}\exp\left(\frac{1}{2\sigma^2}\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}\right) \quad (9)$$

$$g\left(\underset{\sim}{y},\underset{\sim}{\beta}\right)=\frac{1}{\sqrt{2\Pi\sigma^2}}\exp\left(\frac{1}{2\sigma^2}\left(\underset{\sim}{y}-\mathbf{X(K)}\underset{\sim}{\beta}\right)'\left(\underset{\sim}{y}-\mathbf{X(K)}\underset{\sim}{\beta}\right)\right)$$

Based on equation (9) obtained likelihood function

$$\begin{aligned} L(\underset{\sim}{\beta}) &= \prod_{i=1}^{n} g(y_i,\underset{\sim}{\beta}) \\ &= \left(\sqrt{2\Pi\sigma^2}\right)^{-\frac{n}{2}}\exp\left(\frac{1}{2\sigma^2}\left(\underset{\sim}{y}-\mathbf{X(K)}\underset{\sim}{\beta}\right)'\left(\underset{\sim}{y}-\mathbf{X(K)}\underset{\sim}{\beta}\right)\right) \end{aligned} \quad (10)$$

If equation (10) is transformed to a logarithmic form, it will be obtained

$$\begin{aligned} \ell\left(\underset{\sim}{\beta}\right) &= \log L\left(\underset{\sim}{\beta}\right) \\ &= \log\left(\left(\sqrt{2\Pi\sigma^2}\right)^{-\frac{n}{2}}\exp\left(\frac{1}{2\sigma^2}\left(\underset{\sim}{y}-\mathbf{X(K)}\underset{\sim}{\beta}\right)'\left(\underset{\sim}{y}-\mathbf{X(K)}\underset{\sim}{\beta}\right)\right)\right) \end{aligned} \quad (11)$$

Using a partial derivative of $\underset{\sim}{\beta}$ is obtained:

$$\begin{aligned} \frac{\partial\ell\left(\underset{\sim}{\beta}\right)}{\partial\underset{\sim}{\beta}} &= \frac{\partial\left(-\frac{n}{2}\log\left(\sqrt{2\Pi\sigma^2}\right)-\frac{1}{2\sigma^2}\left(\underset{\sim}{y}-\mathbf{X(K)}\underset{\sim}{\beta}\right)'\left(\underset{\sim}{y}-\mathbf{X(K)}\underset{\sim}{\beta}\right)\right)}{\partial\underset{\sim}{\beta}} \\ &= \frac{\partial\left(-\frac{n}{2}\log\left(\sqrt{2\Pi\sigma^2}\right)-\frac{1}{2\sigma^2}\underset{\sim}{y}'\underset{\sim}{y}-2\underset{\sim}{\beta}'\mathbf{X(K)}'\underset{\sim}{y}+\underset{\sim}{\beta}'\mathbf{X(K)}'\mathbf{X(K)}\underset{\sim}{\beta}\right)}{\partial\underset{\sim}{\beta}} \\ 0 &= -\frac{1}{2\sigma^2}\left(-2\mathbf{X(K)}'\underset{\sim}{y}+2\mathbf{X(K)}'\mathbf{X(K)}\underset{\sim}{\hat{\beta}}\right) \\ \underset{\sim}{\hat{\beta}} &= \left(\mathbf{X(K)}'\mathbf{X(K)}\right)^{-1}\mathbf{X(K)}'\underset{\sim}{y} \end{aligned} \quad (12)$$

D. Confidence Intervals for parameters of Nonparametric Spline Truncated Regression

Confidence intervals for parameters of nonparametric spline truncated regression divided into two forms, when σ^2 known and σ^2 unknown. The confidence intervals for parameters of nonparametric spline truncated regression when σ^2known is

$$P\left(\underset{\sim}{\hat{\beta}}_{ju}-z_{\alpha/2}\sqrt{\sigma^2 a_{jj}}\le\beta_{ju}\le\underset{\sim}{\hat{\beta}}_{ju}+z_{\alpha/2}\sqrt{\sigma^2 a_{jj}}\right)=1-\alpha \quad (13)$$

and the confidence intervals for parameters of nonparametric spline truncated regression when σ^2 unknown is

$$P\left(\underset{\sim}{\hat{\beta}}_{ju}-t_{\alpha/2,(n-p(r+1))}\sqrt{\frac{\underset{\sim}{y}'\left[I-\mathbf{X(K)}\left(\mathbf{X(K)}'\mathbf{X(K)}\right)^{-1}\mathbf{X(K)}'\right]\underset{\sim}{y}}{n-p(1+r)}}\right. \quad (14)$$

$$\left(\mathbf{X(K)}'\mathbf{X(K)}\right)^{-1}_{jj}\le\beta_{ju}\le\underset{\sim}{\hat{\beta}}_{ju}-t_{\alpha/2,(n-p(r+1))}$$

$$\left.\sqrt{\frac{\underset{\sim}{y}'\left[I-\mathbf{X(K)}\left(\mathbf{X(K)}'\mathbf{X(K)}\right)^{-1}\mathbf{X(K)}'\right]\underset{\sim}{y}}{n-p(1+r)}}\right)=1-\alpha$$

E. Checking the Assumption of Residuals

Checking the assumption of residuals in this research include checking the assumption of independence residuals, identical residuals, and normality residuals.

F. The Assumption of Independence Residuals

Checking the assumption of independence residuals is used to detect correlation between residuals. The independence assumption on residual is indicated by the covariat value between ε_i and ε_j equal to zero. For checking the assumption can be seen from the plot is on the boundary of significant area that is $\pm z_{\alpha/2}/\sqrt{n}$. So it indicated there is no case of autocorrelation [9].

G. The Assumption of Identical Residuals

Statistic Glejser is one of of the methods that can be used to detect heterogenity variance of residuals [10]. The hypothesis used is

$H_0: \sigma_1^2=\sigma_2^2=\cdots=\sigma_n^2=\sigma^2$ (residual identical)

H_1 : at least there is one $\sigma_i^2\neq\sigma^2$, *i=1,2,…,n* (residual not identical)

With statistic

$$F_{value}=\frac{\sum_{i=1}^{n}(|\hat{e}_i|-|\bar{e}_i|)^2/k-1}{\sum_{i=1}^{n}(|e_i|-|\hat{e}_i|)^2/n-k} \quad (14)$$

H_0 rejected if $F_{value} > F_{table}(F_{\alpha;(k-1,n-k)})$.

H. The Assumption of Normality Residuals

Checking the assumption of normality residuals used to check the residual is normally distributed or not. The hypothesis used is

H_0: residual is normally distributed

H_1: residual is not normally distributed

With the statistics Kolmogorov smirnov

$$z_{value}=Sup_x|F_n(x)-F_0(x)| \quad (15)$$

H_0 rejected if $z_{value} > z_\alpha$.

I. Gender Development Index

The Gender Development Index (GDI) is one of the indicators to measure the success rate of development achievements that already accommodate gender issues. GDI is a direct measurement of the inequality of the genders in the achievement of Human Development Index (HDI). GDI is the ratio of female HDI to male HDI.

$$\text{GDI}=\frac{\text{HDI}_\text{P}}{\text{HDI}_\text{L}} \quad (16)$$

When the number of IPG approaches the number 100, then the development of gender is more balanced or evenly distributed. However, if the more away from the number 100, then the development of gender increasingly unbalanced between the sexes. The dimensions used to measure the quality of life in GDI is longevity and healthy living, knowledge, and decent living standard / welfare [3]. In this study, the dimensions of longevity and healthy living were measured by the morbidity of female population [4] and the last medical birth attendant, the dimensions of knowledge were measured by the School Enrollment Rate of Women's Senior High Schools [11], and the standard of living worth measuring with Female Labor-Force Participation Rate.

III. METHOD

A. Data Source

The data is used in this research is secondary data obtained from the publication of the Badan Pusat Statistik (BPS) of East Java province. The unit of observation in this study includes 29 counties and nine cities in the province of East Java in 2014.

B. Reseacrh Variables

The research variables used in this research are high school enrollment rate of female population (x_1), morbidity of female population (x_2), percentage of last aid of birth by medical (x_3), and labor-force participation rate of female population (x_4).

C. Step of Analysis

The steps to solve the problems and achieve the goals in this research are as follows: (1) Creating scatter plot data between response variables with each predictor variable; (2) Modeling Using Nonparametric Spline Truncated Regression with one, two, three, and combination knot point; (3) Determining the best model using GCV method; (4) Calculating the coefficient of determination R^2; (5) Checking the assumtion of Residuals. Such as ; (6) Determining the confidence interval for parameters of nonparametric spline truncated regression; and (7) Geting the conclusion which was determination of the variabel that have significantly influence using confidence interval.

IV. RESULTS AND DISCUSSION

A. Modeling Using Nonparametric Spline Truncated Regression

For the first we will discuss about the charateriristics of each variable,

TABLE 1.
STATISTIC DESCRIPTIVE OF GDI WITH THE VARIABLES PREDICTORS

Variabel	Minimum	Maximum	Average
y	76,63	98,23	90,06
x_1	40,60	89,45	70,88
x_2	9,78	26,40	14,99
x_3	69,63	100	94,84
x_4	43,58	72,11	54,28

From table 1 we know that the average IPG (y) in East Java Province in 2014 amounted to 90,06. The highest IPG value in East Java province was 98,23 is Blitar, while Sumenep had the lowest IPG value of 76.63.

The position of East Java GDI achievement in 2014 separated 20 districts/cities under the achievement of the provincial IPG and 18 districts/cities above the achievement of provincial GDI. The spread of GDI in the province of East Java shown in figure 1.

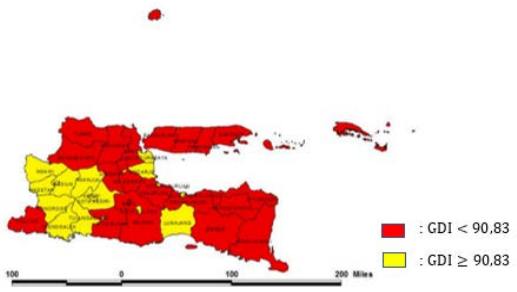

Figure 1. GDI in East Java

Based on figure 1 it can be seen that there are 20 regions that have IPG under the achievement of provincial GDI, indicated by red color. While the yellow area shows its GDI value is already above the achievement of GDI province. The scatterplot between GDI with each of the variables expected to influence shown in figure 2

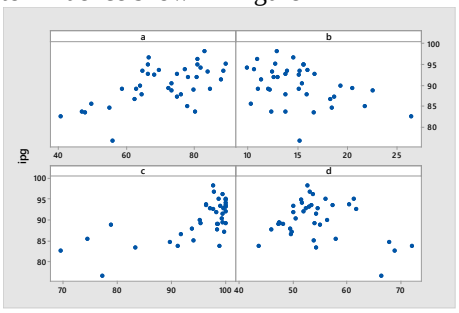

Figure 2. Scatterplot between GDI with the variables predictors

Based on figure 2 can be seen that scatterplot pattern of GDI with some predictor variables ie high school enrollment rate of female population (a), morbidity of female population (b), percentage of last aid of birth by medical (c), and female's labor-force participation rate (d) not to follow a certain pattern. The first step in modeling using nonparametric spline truncated regression is to select the optimal knot point with one knot point, two knots, three knots, and a combination of knot points.

1) Selection of optimal knot point with one knot point

In this section will be discussed about the selection of optimal knot point on GDI and four predictor variables that are suspected to affect it. The model of nonparametric spline truncated regression with one-point knot using four variables predictor as follows

$$y_i = \hat{\beta}_{01} + \hat{\beta}_{11}x_{1i} + \hat{\beta}_{12}(x_{1i} - K_{11})_+^1 + \hat{\beta}_{21}x_{2i} + \hat{\beta}_{22}(x_{2i} - K_{21})_+^1 + \hat{\beta}_{31}x_{3i} + \hat{\beta}_{32}(x_{3i} - K_{31})_+^1 + \hat{\beta}_{41}x_{4i} + \hat{\beta}_{42}(x_{4i} - K_{41})_+^1 \quad (18)$$

The value of GCV generated with a one knot point is shown in table 2.

TABLE 2.
GCV VALUE WITH ONE POINT KNOT

Knot				GCV
x_1	x_2	x_3	x_4	
67,52	18,94	86,36	57,53	8,30
68,51	19,28	86,98	58,05	8,19
69,51	19,62	87,60	58,57	8,16
70,51	19,96	88,22	59,08	8,22
71,51	20,29	88,84	59,60	8,37

Based on table 2, the minimum GCV value with one knot point is 8,16.

2) Selection of optimal knot point with two knot point

The next step after getting the minimum GCV with one point knots, then select the optimal knots with two point knots. The model of nonparametric spline truncated regression with two-point knot using four variables predictor as follows

$$y_i = \hat{\beta}_{01} + \hat{\beta}_{11}x_{1i} + \hat{\beta}_{12}(x_{1i} - K_{11})_+^1 + \hat{\beta}_{21}x_{2i} + \hat{\beta}_{22}(x_{2i} - K_{21})_+^1 + \hat{\beta}_{31}x_{3i} + \hat{\beta}_{32}(x_{3i} - K_{31})_+^1 + \hat{\beta}_{41}x_{4i} + \hat{\beta}_{42}(x_{4i} - K_{41})_+^1 \quad (19)$$

The value of GCV generated with a one knot point is shown in table 3

TABLE 3.
GCV VALUE WITH TWO POINT KNOT

Knot				GCV
x_1	x_2	x_3	x_4	
72,50	20,63	89,46	62,21	8,84
73,50	20,97	90,08	62,79	
72,50	20,63	89,46	62,21	8,53
74,50	21,31	90,70	63,38	
72,50	**20,63**	**89,46**	**62,21**	**8,49**
75,49	**21,65**	**91,32**	**63,96**	
72,50	20,63	89,46	62,21	8,67
76,49	21,99	91,94	64,54	
72,50	20,63	89,46	62,21	9,23
77,49	22,33	92,56	65,12	

Based on table 3, the minimum GCV value with one knot point is 8,49.

3) Selection of optimal knot point with three knot point

The next step after getting the minimum GCV with two point knots, then select the optimal knots with three point knots. The model of nonparametric spline truncated regression with three-point knot using four variables predictor as follows

$$y_i = \hat{\beta}_{01} + \hat{\beta}_{11}x_{1i} + \hat{\beta}_{12}(x_{1i} - K_{11})_+^1 + \hat{\beta}_{13}(x_{1i} - K_{12})_+^1 + \hat{\beta}_{14}(x_{1i} - K_{13})_+^1 + \hat{\beta}_{21}x_{2i} + \hat{\beta}_{22}(x_{2i} - K_{21})_+^1 + \hat{\beta}_{23}(x_{2i} - K_{22})_+^1 + \hat{\beta}_{24}(x_{2i} - K_{23})_+^1 + \hat{\beta}_{31}x_{3i} + \hat{\beta}_{32}(x_{3i} - K_{31})_+^1 + \hat{\beta}_{33}(x_{3i} - K_{32})_+^1 + \hat{\beta}_{34}(x_{3i} - K_{33})_+^1 + \hat{\beta}_{41}x_{4i} + \hat{\beta}_{42}(x_{4i} - K_{41})_+^1 + \hat{\beta}_{42}(x_{4i} - K_{42})_+^1 + \hat{\beta}_{43}(x_{4i} - K_{43})_+^1 \quad (20)$$

The value of GCV generated with a one knot point is shown in table 4

TABLE 4.
GCV VALUE WITH THREE POINT KNOT

Knot				GCV
x_1	x_2	x_3	x_4	
42,59	10,46	70,87	44,74	
47,58	12,15	73,97	47,66	10,22
73,50	20,97	90,08	62,79	
42,59	10,46	70,87	44,74	
47,58	12,15	73,97	47,66	10,32
74,50	21,31	90,70	63,38	
42,59	**10,46**	**70,87**	**44,74**	
47,58	**12,15**	**73,97**	**47,66**	**7,19**
75,49	**21,65**	**91,32**	**63,96**	
42,59	10,46	70,87	44,74	
47,58	12,15	73,97	47,66	7,50
76,49	21,99	91,94	64,54	
42,59	10,46	70,87	44,74	
47,58	12,15	73,97	47,66	8,14
77,49	22,33	92,56	64,54	

Based on table 4, the minimum GCV value with one knot point is 7,19.

4) Selection of optimal knot point with combination knot point

The next step after getting the minimum GCV with one, two, and three point knots, then select the optimal knots with combination point knots. The value of GCV generated with a one knot point is shown in table 5.

TABLE 5.
GCV VALUE WITH SPLINE LINIER COMBINATION KNOT

Knot				GCV
x_1	x_2	x_3	x_4	
69,51	**20,63**	**89,46**	**58,56**	
79,48	**21,65**	**91,32**		**6,13**
80,47				
69,51	20,64	89,46	62,21	
79,48	21,65	91,32	63,95	7,57
80,48				
69,51	20,63	89,46	60,46	
79,48	21,65	91,32	66,29	6,76
80,47			66,87	

Based on table 5, the minimum GCV value with combination knot point is 6,13.

5) Modeling with Knot Optimum Point

After obtaining the minimum GCV value by using one, two, three, and a combination of knot points, the next is to select the best model by comparing the smallest GCV value of each knot. In table 6 we will show the minimum GCV value for each knot point.

TABLE 6.
COMPARISON OF GCV VALUES

Knot Number	GCV Minimum
1	8,15
2	8,48
3	7,18
Combination of Knot Point	6,13

Based on table 6 obtained minimum GCV value is 6.13, that is at the combination of knot point. This result will be used in GDI modeling in East Java. The model of nonparametric spline truncated regression as follows

$$\hat{y}_i = -18{,}05 + 0{,}3x_{1i} - 0{,}81(x_{1i} - 69{,}51)_+^1 + 4{,}75(x_{1i} - 79{,}48)_+^1 - 4{,}68(x_{1i} - 80{,}47)_+^1 - 0{,}54x_{2i} - 0{,}17(x_{2i} - 20{,}63)_+^1 + 5{,}42(x_{2i} - 21{,}65)_+^1 + 0{,}81x_{3i} - 3{,}18(x_{3i} - 89{,}46)_+^1 + 2{,}62(x_{3i} - 21{,}93)_+^1 +$$

$$0{,}54x_{4i} - 1{,}58(x_{4i} - 58{,}56)_+^1 \quad (21)$$

The model of nonparametric spline truncated regression with the combination of 3, 2, 2, 1 knot points has R^2 equal to 87,48. This value can be interpreted that this model can explain the GDI 87,48%.

B. *Confidence Interval for Parameters of Nonparametric Spline Truncated Regression*

After we get the model of nonparametric spline truncated regression, then a 95% confidence interval will be constructed with the formula given in equation (13). Here are the result or a confidence interval nonparametric regression.

TABLE 6.
CONFIDENCE INTERVAL FOR PARAMETERS 95%

Variable	Parameters	Upper Limit	Lower Limit
-	$\hat{\beta}_{01}$	-57,68	21,58
x_1	$\hat{\beta}_{11}$	0,07	0,54
	$\hat{\beta}_{12}$	-1,29	-0,33
	$\hat{\beta}_{13}$	1,29	8,20
	$\hat{\beta}_{14}$	-8,12	-1,23
x_2	$\hat{\beta}_{21}$	-0,88	-0,20
	$\hat{\beta}_{22}$	-5,41	5,05
	$\hat{\beta}_{23}$	-1,14	11,98
x_3	$\hat{\beta}_{31}$	0,37	1,26
	$\hat{\beta}_{32}$	-6,48	0,10
	$\hat{\beta}_{33}$	-0,45	5,70
x_4	$\hat{\beta}_{41}$	0,28	0,79
	$\hat{\beta}_{42}$	-2,09	-1,07

If the confidence interval contains a zero value, then the parameter does not significantly affect the model. Based on Table 6 are obtained from the 13 parameters 8 parameters that significantly influence the model. But overall, the four predictor variables ie high school enrollment rate of female population, morbidity of female population, percentage of last aid of birth by medical, and female's labor-force participation rate have significantly influence to the response variable (GDI).

C. *Checking the Assumption of Residuals*

1) *The Assumption of Independence Residuals*

The result ACF plot of the residuals is shown in figure 3.

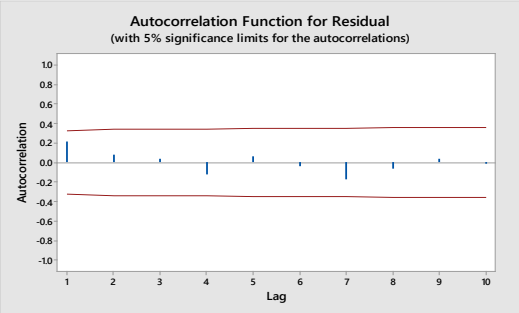

Figure 3. ACF plot of Residual

Based on figure 3 it can be seen that the residual autocorrelation value is at a significant limit or in other words no lag is out of bounds. So it can be concluded that there is no correlation between residuals.

2) *The Assumption of Identical Residual*

The hypothesis used is

$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2$ (residual identical)

H_1 : at least there is one $\sigma_1^2 \neq \sigma^2, i = 1,2,\ldots,n$, (residual not identical)

TABEL 7.
ANOVA

Source	df	SS	MS	Fvalue
Regresi	12	11,32839	0,9440321	0,8804611
Error	25	26,80505	1,072202	
Total	37	38,13344		

Based on Table 7 we know that F_{value} = 0,88 < F_{table} = 2,16. Rejection H_0 if $F_{value} > F_{table}$. So, it fails to reject H_0, Then it can be concluded that the variance of residual is homogeneous.

3) *The Assumption of Normality Residual*

The hypothesis used is.

H_0: residual is normally distributed

H_1: residual is not normally distributed

Based on the normality test with Kolmogorov obtained p-value equal to 0.27 > α = 0.05 then failed to reject H_0. So it can be concluded that the residual is normally distributed.

V. CONCLUSION

The best nonparametric regression model is as follows

$$\hat{y}_i = -18{,}05 + 0{,}3x_{1i} - 0{,}81(x_{1i} - 69{,}51)_+^1 + 4{,}75(x_{1i} - 79{,}48)_+^1 - 4{,}68(x_{1i} - 80{,}47)_+^1 - 0{,}54x_{2i} - 0{,}17(x_{2i} - 20{,}63)_+^1 + 5{,}42(x_{2i} - 21{,}65)_+^1 + 0{,}81x_{3i} - 3{,}18(x_{3i} - 89{,}46)_+^1 + 2{,}62(x_{3i} - 21{,}93)_+^1 + 0{,}54x_{4i} - 1{,}58(x_{4i} - 58{,}56)_+^1$$

which has R^2 value of 87,48%. Determination of predictor variables that significantly influence GDI by using confidence interval obtained high school enrollment rate of female population, morbidity of female population, percentage of last aid of birth by medical, and female labor-force participation rate have significantly influenced to GDI in East Java.

ACKNOWLEDGEMENT

The authors say thanks to the major of Statistics Intitut Teknologi Sepuluh Nopember, Surabaya.

REFERENCES

[1] R. L. Eubank, *Spline smoothing and nonparametric regression*. New York: Marcel Dekker Inc, 1988.

[2] I. N. Budiantara, "Model Spline Multivariabel dalam Regresi Nonparametrik," in *Prosiding Seminar Nasional Matematika*, 2004.

[3] Badan Pusat Statistik, *Pembangunan Manusia Berbasis Gender Provinsi Jawa Timur 2015*. Surabaya, 2015.

[4] Badan Pusat Statistik, *Profil Gender Provinsi Jawa Timur 2015*. Surabaya, 2015.

[5] U. Q. Hafizh and V. Ratnasari, "Pemodel Dispartas Gender di Jawa Timur dengan Pendekatan Model Regresi Probit Ordinal," Institut Teknologi Sepuluh Nopember, 2013.

[6] L. J. Hakim, "Analisis Komponen Indeks Pembangunan Gender

Dengan Geographically Weighted Multivariate Regression Model di Provinsi Kalimantan Timur dan Kalimantan Selatan Tahun 2011," Institut Teknologi Sepuluh Nopember, 2014.

[7] I. N. Budiantara, "Metode UBR, GML, CV dan GCV dalam Regresi Nonparamtrik Spline," *Maj. Ilm. Himpun. Mat. Indones.*, vol. 6, pp. 285–290, 2000.

[8] I. N. Budiantara, "Model Spline dengan Knots Optimal," *J. Ilmu Dasar, FMIPA Univ. Jember*, vol. 7, pp. 77–85, 2006.

[9] I. A. S. Intansari, "Inferensi Statistik Untuk Kurva Regresi Nonparametrik Spline Kuadratik dan Aplikasinya Pada Data ASFR (Age Specific Fertility Rate) di Bali," Institut Teknologi Sepuluh Nopember, 2016.

[10] Setiawan and Dwi Endah Kusrini, *Ekonometrika*. Andi, 2010.

[11] D. Aryanto, "Pendugaan Area Kecil Terhadap Defisit Kesempatan Kerja Produktif Pada Level Kecamatan di Provinsi Maluku Denga Pendekatan Empirical Bayes," Institut Teknologi Sepuluh Nopember, 2014.