

# Statistical Inference

*by* Dwi Rizky

---

**Submission date:** 15-May-2023 09:01PM (UTC+0700)

**Submission ID:** 2093768379

**File name:** 10.\_Statistical\_Inference\_on\_Nonparametric.pdf (1.15M)

**Word count:** 4100

**Character count:** 19637

PAPER • OPEN ACCESS

2

## Statistical Inference on Nonparametric Spline Models and its Applications

4

To cite this article: D R Arifanti and R Hidayat 2021 *J. Phys.: Conf. Ser.* **2123** 012023

View the [article online](#) for updates and enhancements.

### You may also like

11

- [Enhanced etching resolution of self-assembled PS-b-PMMA block copolymer](#)

11

- [by ionic liquid additives](#)  
Hongbo Hao, Shuangjun Chen, Jiaxing Ren et al.

5

- [Effects of acid diffusion and resist molecular size on line edge roughness for chemically amplified resists in EUV lithography: computational study](#)  
Masanori Koyama, Kyohei Imai, Masamitsu Shirai et al.

- [Efficient and improved qualification method for patterns with irregular edges in printed electronics](#)

Ting-Jeng Liu, Shao-Min Hsu, Meng-Jhu Wu et al.



The Electrochemical Society  
Advancing solid state & electrochemical science & technology

243rd Meeting with SOFC-XVIII

Boston, MA • May 28 – June 2, 2023

Accelerate scientific discovery!

Learn More & Register



## Statistical Inference on Nonparametric Spline Models and its Applications

D R Arifanti<sup>1</sup> and R Hidayat<sup>2\*</sup>

<sup>1</sup>Department of Tadris Matematika, IAIN Palopo, Indonesia

<sup>2</sup>Master of Mathematics Education Program, Cokroaminoto Palopo University, Indonesia

\*Email: dayatmath@gmail.com

**Abstract.** One of the components of the Human Development Index which is still a problem and concern in the world today is the Life Expectancy Rate (LER). United Nations Development Program (UNDP). United Nations Development Program (UNDP) uses the LER to measure community health status as well as a benchmark for development success. LER in Indonesia continues to increase almost throughout the year. That is, the hope of a newborn baby to be able to live longer is getting higher. LER data modelling with parametric regression is not necessarily suitable to be applied because the LER relationship pattern has a pattern that varies at certain age intervals. Spline regression is a regression method that can handle data whose pattern changes at certain intervals. Spline is one of the models in nonparametric regression that has a very special and very good visual statistical interpretation. In addition, splines are also able to handle data characters or functions that are smooth (smooth). This study aims to derive the form of the estimator and the shortest confidence interval for the quadratic spline model and model the LER data in Indonesia.

**Keywords:** Human development index, spline regression, nonparametric regression, life expectancy rat

### 1. Introduction

If discussed in probabilistic theory, Life Expectancy Rat (LER) is a quadratic function, because the graph pattern is U-shaped and almost like the shape of an inverted normal distribution curve but not symmetrical. In statistics, the quadratic function is U-shaped in the case of LER data, the curve tends to be quadratic but the shape is slightly skewed to the right, not smooth and, not truncated. Curves that are not necessarily smooth in shape, need to be proven and analysed using a method that can sensitively detect the existence of points that are pieces. One method that can be used to solve modelling problems is regression analysis. Regression analysis is one of the methods used to determine the pattern of the relationship between response variables and predictor variables. There are three approaches in the regression analysis method, namely the parametric approach, the nonparametric approach, and the semiparametric approach. If a data regression curve forms a known pattern such as linear, quadratic, cubic, or polynomial of degree k then the appropriate approach is the parametric approach. The nonparametric regression approach is used when the shape of the regression curve of a data is unknown. Meanwhile, the semiparametric approach is used in the shape of the regression curve is partly known and the pattern is partly unknown. Not all of the data contained in the field are known to be related patterns, so a nonparametric regression approach can be used to solve these problems.

According to [1] said that in the nonparametric regression approach the data is expected to find its own form of estimation without being influenced by the subjectivity of the researcher. In other words,



Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

nonparametric regression approach has high flexibility. Nonparametric spline regression approach is more suitable to be applied to LER data than parametric approach. Several nonparametric regression approaches that are often used are Fourier Series, Kernel, Spline, and Wavelets. The spline approach has advantages such as the model tends to find its own estimate wherever the data moves. This advantage is caused because in the Spline there are knot points which are joint points of integration that indicate changes in data behaviour patterns [2]. In general, the Spline is an estimator obtained by minimizing the Penalized Least Square (PLS), namely the optimization criteria for combining goodness of fit and curve smoothness. This optimization solution is obtained by using the RKHS or Gateaux method which is mathematically relatively difficult to do. To overcome this, another simpler estimator method is used, namely Least Square (LS) optimization, which is expected to obtain relatively easier and simpler mathematical calculations and mathematical interpretations [3]. The statistical inference that is very important in spline regression is the confidence interval with the concept of the shortest confidence interval with Pivotal Quantity.

The population is one of the wealth owned by a nation. Indonesia is one of the countries that occupy the top five positions with the largest population in the world. In 2015, Indonesia was in fourth place after China, India and the United States. As a large nation, Indonesia should have the ability to carry out development by making the population the subject and object of development. The success of nation-building cannot be separated from the achievement of human development as a real wealth owned by a nation. In principle, there are three most basic choices at all levels of development, namely the community wants to live a long and healthy life, wants to get an education, and wants to have access to the resources needed to live a decent life. The concept of HDI became the forerunner to the formation of the Human Development Index which was first launched by the United Nations Development Program (UNDP) in 1990. Therefore, the success of a nation's development can also be seen from the achievement of the Human Development Index. One of the components of the HDI that is still a problem and concern for the world today is the Life Expectancy Rate (LER). UNDP uses LER to measure community health status as well as a benchmark for development success. Low LER in an area must be followed by health development programs and other social programs such as environmental health, nutritional adequacy and calories including poverty eradication programs. LER in Indonesia continues to increase almost throughout the year. That is, the hope of a new-born baby to be able to live longer is getting higher. In 2010 Indonesia's LER was 69.81 years and continued to increase until 2015 to 70.78 years. According to Bappenas, the government targets Indonesia's LER to reach 73.7 years by 2025. However, if we look closely, the growth of LER in Indonesia is still very low. In the period 1990 – 2015, the average growth of LER was only about 0.3 percent per year. This is in line with WHO's observations on LER in developing countries which move slowly and even slower when compared to underdeveloped countries.

Along with the increase in LER nationally, provinces in Indonesia also experienced positive LER growth. However, there are still LER disparities between provinces. This fact is certainly not in line with the government's development target stated in the NAWACITA program, namely the equitable distribution of development between regions in Indonesia. In 2015, provincial LER in Indonesia ranged from 64 – 74 years. Of the 34 provinces, there is still 73.53 percent of provinces that have LER under the national LER. Most of these provinces are located in the eastern part of Indonesia, including West Papua Province 65.19 years, followed by Papua Province 65.09 years and the lowest is West Sulawesi Province 64.22 years. Meanwhile, 26.47 percent of provinces already have LER above the national LER. Most of these provinces are located in the western part of Indonesia. The province with the highest LER was DI Yogyakarta Province, which was 74.68 years old, followed by Central Java Province at 73.96 years and East Kalimantan Province at 73.65 years.

When compared with the conditions five years ago, in 2010 the lowest LER was in West Sulawesi Province, which was 62.50 years, while the LER in DI Yogyakarta Province had reached 74.17 years in the same year. This means that babies born in 2015 in DI Yogyakarta Province on average have a life expectancy of about 12 years longer than babies born in West Sulawesi Province in the same year. The increase in life expectancy nationally in Indonesia still places Indonesia's LER under the LER of neighbouring countries. Sourced from the United Nations in its publications, BPS recorded the LER of several countries in the world during the period 1990 - 2015. In that period, Indonesia's LER was still

below Singapore (82.2), Malaysia (74.9), Thailand (74.3), and Cambodia (71.6). Indonesia's LER achievement is slightly better than the Philippines (68.6) and India (66.3). According to data from the Central Intelligence Agency (CIA) World, LER in Indonesia is still in the low category when compared to countries around the world.

Based on the resulting study, LER Indonesia is ranked 137th out of 223 countries in the world. This is a reflection of the low quality of life of the population in Indonesia. The study is in line with the problems and challenges stated in the National Long-Term Development Plan (RPJPN) 2005-2025, namely improving the quality of life in various development fields. The Global Burden Disease (GBD) study under the supervision of the Institute for Health Metrics and Evaluation (IHME) said that developing countries are currently facing serious challenges, namely lifestyle and deadly diseases, namely heart disease, stroke and diabetes that can affect all ages. According to the records of the World Health Organization (WHO), one of the countries that has succeeded in overcoming this deadly disease is South Korea, so that in 2015 South Korea was able to achieve LER above 80 years and in the same year succeeded in creating a community of people aged almost reaching the age of 80. 100 years. As an indicator of the quality of health and well-being, LER is of course influenced by many factors. Several studies that examine the factors that affect LER include literacy rates, average years of schooling, and per capita expenditure. Other studies also mention that the factors that influence LER are the Human Development Index, the percentage of deliveries with non-medical assistance, the average age at first marriage for women, the average length of schooling, the average household expenditure, the percentage of areas with village status, the average length of exclusive breastfeeding, and the percentage of households using clean water.

## 2. Nonparametric regression

The nonparametric regression approach is used if the spread of the data pattern forms a certain pattern, whether linear, quadratic or cubic. To be able to use parametric regression models, past experience is needed or there are other sources available that can provide detailed information about the data. In nonparametric regression there is a firm assumption, namely the shape of the regression curve is known [1]. The method used for parameter estimation is Least Square and Maximum Likelihood [4]. The approach model generated by these methods is equivalent to the parameter estimates in the model. In general, linear parametric regression can be written in the form [5][6]

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_r x_{ri} + \varepsilon_i \text{ for } i=1,2,\dots,n, \quad (1)$$

where  $y$  response variable,  $x_1, x_2, \dots, x_r$  as predictor variable,  $\varepsilon_i$  is an independent random error with normal distribution with zero mean and variance  $\sigma^2$ . If the assumed model is correct then parametric regression estimation is very efficient, but otherwise it will lead to misleading data interpretation [7].

In general, the form of multivariable parametric regression with  $r$  predictor variables in equation (1) can be written in matrix form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (2)$$

The vector and matrix forms in equation (2) are [8]:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1r} \\ 1 & x_{21} & x_{22} & \dots & x_{2r} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nr} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_r \end{pmatrix}, \text{ dan } \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_r \end{pmatrix}$$

$\boldsymbol{\beta}$  parameters are estimated using the Ordinary Least Square (OLS) method which minimizes  $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$  where  $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ . By lowering the partial  $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$  to  $\boldsymbol{\beta}$  and equates to zero, then we get the estimator  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ . For example, given paired data  $(x_{1i}, x_{2i}, \dots, x_{qi}, y_i)$ , relationship between  $(x_{1i}, x_{2i}, \dots, x_{qi})$  with  $y_i$  modeled by additive multivariable nonparametric regression:

$$y_i = f_1(x_{1i}) + f_2(x_{2i}) + \dots + f_q(x_{qi}) + \varepsilon_i, i = 1, 2, \dots, n \quad (3)$$

Regression curve shape  $f_k$  with  $k = 1, 2, \dots, q$  assumed to be unknown. Nonparametric regression curve  $f_k(x_k)$ ,  $k = 1, 2, \dots, q$  estimated by least square method [9].

### 3. Spline model

Many studies of spline estimators on nonparametric regression models have been carried out such as [10,11,12,13]. The nonparametric regression model which states the relationship between the predictor variable and the response variable can be written as follows:

$$y_i = f(x_i) + \varepsilon_i, i = 1, 2, \dots, n, \quad (4)$$

Equation (4) can be changed in the form of matrix multiplication

$$\mathbf{f}(x) = \mathbf{X}(\lambda)\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (5)$$

with  $\mathbf{X}(\lambda)$  is a matrix that depends on  $\lambda$ , whereas  $\boldsymbol{\beta}$  is a vector that depends on  $\boldsymbol{\theta}$ , and  $\phi$ . To get the estimator of  $f$ , The Maximum Likelihood Estimation (MLE) method is also used. The first estimation is to estimate  $\boldsymbol{\beta}$ , with likelihood function:

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2 | \lambda) &= \prod_{i=1}^n \left( \frac{1}{\sqrt{2\pi\sigma^2}} \text{Exp} \left( -\frac{\varepsilon_i^2}{2\sigma^2} \right) \right) \\ &= (2\pi\sigma^2)^{-n/2} \text{Exp} \left( -\frac{1}{2\sigma^2} \|\boldsymbol{\varepsilon}\|^2 \right) \\ &= (2\pi\sigma^2)^{-n/2} \text{Exp} \left( -\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}(\lambda)\boldsymbol{\beta}\|^2 \right) \end{aligned} \quad (6)$$

so that obtained:

$$\hat{\boldsymbol{\beta}}(\lambda) = \left[ (\mathbf{X}(\lambda))' \mathbf{X}(\lambda) \right]^{-1} (\mathbf{X}(\lambda))' \mathbf{y} \quad (7)$$

Causes an estimate for  $\mathbf{f}(x)$  is

$$\hat{\mathbf{f}}(x) = \mathbf{X}(\lambda)\hat{\boldsymbol{\beta}} = \mathbf{X}(\lambda) \left[ (\mathbf{X}(\lambda))' \mathbf{X}(\lambda) \right]^{-1} (\mathbf{X}(\lambda))' \mathbf{y} \quad (8)$$

The knot points are the controller of the balance between the smoothness of the curve and the suitability of the curve to the data. The selection of knot points can be done by using the Generalized Cross Validation or GCV [14,15] method.

$$\text{GCV}(\lambda) = \frac{\text{MSE}(\lambda)}{\left( n^{-1} \text{tr}(\mathbf{I} - \mathbf{S}(\lambda)) \right)^2}, \quad (9)$$

where  $\text{MSE}(\lambda)$  in equation (9) are:

$$\text{MSE}(\lambda) = n^{-1} \sum_{i=1}^n (y_i - f(x_i))^2 \quad (10)$$

The selection of optimal knot points is done by looking at the smallest GCV value.

#### 4. Method

The steps of this research are described in detail as follows:

1. The nonparametric regression model which states the relationship between the predictor variable and the response variable can be written as follows:

$$y_i = f(x_i) + \varepsilon_i, i = 1, 2, \dots, n, \quad (11)$$

2. Looking for confidence interval  $1-\alpha$  in nonparametric spline quadratic regression with Pivotal Quantity approach. The steps are as follows:
  - a. Determine the basis of the quadratic spline function.
  - b. Estimating the parameters of the nonparametric spline quadratic regression model by using the MLE method.
3. Calculating the length of the confidence interval using the concept of the shortest confidence interval.
4. Deriving the shortest confidence interval for the nonparametric regression curve.

#### 5. Result

By using the concept of Pivotal Quantity, which is a statistic whose distribution does not contain a regression curve  $f(x_i)$ ,

$$Z_i(x_1, x_2, \dots, x_n) = \frac{\hat{f}(x_i) - f(x_i)}{\sqrt{\text{var}(\hat{f}(x_i))}} \sim N(0,1), i = 1, 2, \dots, n \quad (12)$$

with probability function

$$f(Z_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z_i^2}, -\infty < z < \infty \quad (13)$$

does not load regression curve  $f(x_i)$ .

So,  $Z_i(x_1, x_2, \dots, x_n)$  is the Pivotal Quantity for the  $f$ . Confidence interval  $1-\alpha$  obtained from solving the probability equation:

$$P\left(-z_{\alpha/2} < Z(x_1, \dots, x_n) < z_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(-z_{\alpha/2} < \frac{\hat{f}(x_i) - f(x_i)}{\sqrt{\text{var}(\hat{f}(x_i))}} < z_{\alpha/2}\right) = 1 - \alpha$$

The above equation is equivalent to

$$-z_{\alpha/2} < \frac{\hat{f}(x_i) - f(x_i)}{\sqrt{\text{var}(\hat{f}(x_i))}}, \text{ or } f(x_i) < \hat{f}(x_i) + z_{\alpha/2} \sqrt{\text{var}(\hat{f}(x_i))} \quad (14)$$

$$\frac{\hat{f}(x_i) - f(x_i)}{\sqrt{\text{var}(\hat{f}(x_i))}} < z_{\alpha/2}, \text{ or } f(x_i) > \hat{f}(x_i) - z_{\alpha/2} \sqrt{\text{var}(\hat{f}(x_i))} \quad (15)$$

From equations (14) and (15) it is obtained that the confidence interval for the regression curve  $f$

$$P\left(\left[\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \sum_{k=1}^r \hat{\beta}_{2+k} (x_i - k_+)^2\right] - z_{\alpha/2} \sqrt{\text{var} \hat{f}(x_i)} < f(x_i)\right)$$

$$\left[\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \sum_{k=1}^r \hat{\beta}_{2+k} (x_i - k_+)^2\right] + z_{\alpha/2} \sqrt{\text{var} \hat{f}(x_i)} = 1 - \alpha \quad (16)$$

where  $\text{var} \hat{f}(x_i)$  is a diagonal element from matrix:

$$\begin{aligned} \text{var} \hat{f}(x) &= \text{var}(X\hat{\beta}) \\ &= \text{var}(X(XX)^{-1}XY) \\ &= X(XX)^{-1}X \text{var}(Y)X(XX)^{-1}X \\ &= \sigma^2 X(XX)^{-1}X \end{aligned} \tag{17}$$

### 6. Quadratic spline model for life expectancy rate

To see the pattern of the relationship between LER and IMR, the plot is given as follows:

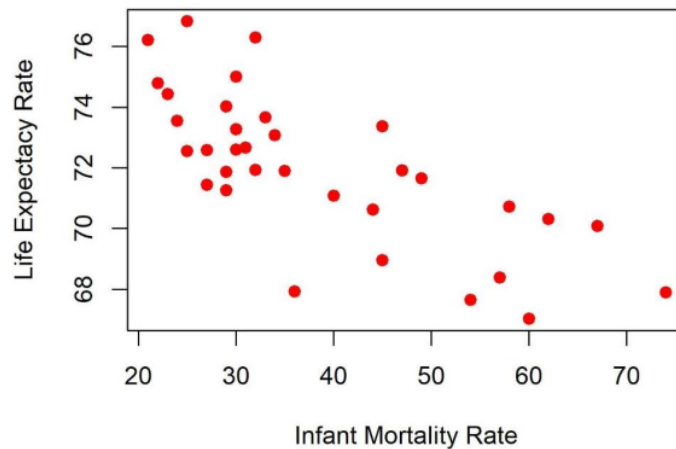


Figure 1. Scatter plot of the relationship between LER and IMR

From the picture above, there are indications of changes in data behaviour patterns. In addition, the data pattern tends to be quadratic. Therefore, in this study the author will try to approach using nonparametric spline quadratic regression.

In using Spline regression, the first step that must be done is to determine the location of the knot point. The method of selecting knots is by experimenting in places where there is a change in data behaviour at certain intervals. From various knots, the minimum GCV value is chosen. To get the best results in spline regression, a quadratic spline model was tried with one knot to three knots. Then calculate the estimated parameter value and create an estimation curve. The GCV calculation is carried out using the R Program facility. The GCV will be searched for 1 knot, 2 knots and 3 knots quadratic spline model, as follows:

a. Quadratic model 1 knot point:

$$y = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 (x_i - k_1)_+^2 + \varepsilon$$

b. Quadratic model 2 knot point:

$$y = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 (x_i - k_1)_+^2 + \beta_4 (x_i - k_2)_+^2 + \varepsilon$$

c. Quadratic model 3 knot point:

$$y = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 (x_i - k_1)_+^2 + \beta_4 (x_i - k_2)_+^2 + \beta_5 (x_i - k_3)_+^2 + \varepsilon$$



**Table 1.** GCV value for 1,2 and 3 knots in the quadratic model

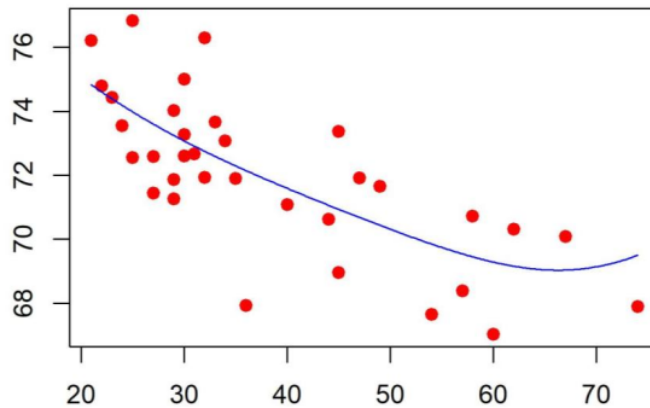
Number of Knots	Knot location		GCV
1 Knot	45		23,863
	46		21,536
	47		25,780
2 Knot	35	49	7,863
	35	50	<b>5,832</b>
	35	51	7,531
3 Knot	35	50	18,635
	35	50	19,836
	35	50	19,736

Based on Table 1, it can be seen that the smallest GCV value is 5.832 in the quadratic spline model with three knot points. After obtaining the minimum GCV score for the quadratic spline model, the next step is to calculate the estimate for the quadratic spline model.

The estimation of the quadratic spline model with two knot points is as follows:

$$y = 25,63 - 20,55x_i + 0,72x_i^2 - 4,63(x_i - 35)_+^2 + 6,33(x_i - 50)_+^2 \quad (18)$$

The estimation of the spline model above can be interpreted that in Indonesia with infant mortality rates above 35 infants, life expectancy decreases. A gradual decline will be more visible when the infant mortality rate is above 50. It is indicated by the shape of the curve that is slowly decreasing. The estimation of the quadratic spline model with two knot points (35;50) for LER in Indonesia has a coefficient of determination  $R^2 = 89.17\%$  and Mean Square Error = 28.56943. If shown in the figure, the visualization of the quadratic spline curve estimation model is shown in Figure 2



**Figure 2** Quadratic spline curve with knot points 35 and 50

With a quadratic spline model with two knot points, namely at the infant mortality, a 95% confidence interval will be built with the equation:

$$P\left(\hat{f}(x_i) - z_{\alpha/2} \sqrt{\text{var } \hat{f}(x_i)} < f(x_i) < \hat{f}(x_i) + z_{\alpha/2} \sqrt{\text{var } \hat{f}(x_i)}\right) = 0,95$$

$$P\left[\left(25,63 - 20,55x_i + 0,72x_i^2 - 4,63(x_i - 35)_+^2 + 6,33(x_i - 50)_+^2 - 1,96\sqrt{\text{var } \hat{f}(x_i)} < f(x_i) < \right.\right.$$

$$\left.\left. \left(25,63 - 20,55x_i + 0,72x_i^2 - 4,63(x_i - 35)_+^2 + 6,33(x_i - 50)_+^2 + 1,96\sqrt{\text{var } \hat{f}(x_i)}\right)\right) = 0,95$$

The quadratic spline model obtained is

$\hat{y} = 25,63 - 20,55x_i + 0,72x_i^2 - 4,63(x_i - 35)_+^2 + 6,33(x_i - 50)_+^2$ , It is known that at the 95% significance level, LER in Indonesia is still within the mean range.

## 7. Conclusion

Based on the analysis and discussion that has been carried out, the following conclusions can be drawn:

1. The confidence interval in nonparametric spline quadratic regression is built using the concept of Pivotal Quantity. The Pivotal Quantity for the regression curve  $f$  is:

$$Z_i(x_1, x_2, \dots, x_n) = \frac{\hat{f}(x_i) - f(x_i)}{\sqrt{\text{var}(\hat{f}(x_i))}} \sim N(0,1), i = 1, 2, \dots, n$$

The confidence interval  $1 - \alpha$  is obtained by solving the probability equation:

$$P\left(-z_{\alpha/2} < \frac{\hat{f}(x_i) - f(x_i)}{\sqrt{\text{var}(\hat{f}(x_i))}} < z_{\alpha/2}\right) = 1 - \alpha$$

Confidence Interval for the  $f$  regression curve

$$P\left(\left[\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \sum_{k=1}^r \hat{\beta}_{2+k}(x_i - k_k)_+^2\right] - z_{\alpha/2} \sqrt{\text{var} \hat{f}(x_i)} < f(x_i) < \left[\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \sum_{k=1}^r \hat{\beta}_{2+k}(x_i - k_k)_+^2\right] + z_{\alpha/2} \sqrt{\text{var} \hat{f}(x_i)}\right) = 1 - \alpha$$

2. The best quadratic spline model to estimate the relationship between LER and IMR is a quadratic spline model with two knot points

$$y = 25,63 - 20,55x_i + 0,72x_i^2 - 4,63(x_i - 35)_+^2 + 6,33(x_i - 50)_+^2$$

3. After getting the best spline model for LER, then the confidence interval will be built. With a quadratic spline model for LER with a quadratic spline model with two knot points, namely 35 and 50, a 95% confidence interval will be built with the equation:

$$P\left(\left[\left(25,63 - 20,55x_i + 0,72x_i^2 - 4,63(x_i - 35)_+^2 + 6,33(x_i - 50)_+^2\right) - 1,96\sqrt{\text{var} \hat{f}(x_i)} < f(x_i) < \left(25,63 - 20,55x_i + 0,72x_i^2 - 4,63(x_i - 35)_+^2 + 6,33(x_i - 50)_+^2 + 1,96\sqrt{\text{var} \hat{f}(x_i)}\right)\right] = 0,95$$

## 8. References

- [1] Eubank R L 1988 *Spline Smoothing and Nonparametric Regression* (New York: Marcel Deker)
- [2] Hidayat R, Budiantara I N, Otok B W and Ratnasari V 2020 *Comm. in Stat. - Theory and Methods* **50** 17
- [3] Hidayat R, Budiantara I N, Otok B W and Ratnasari V 2019 *App. Math. and Inf. Sci.* **13** 1
- [4] Wahba G 1990 *Spline Models for Observational Data* (Pennsylvania: SIAM)
- [5] Dagum E B and Bianconcini S 2009 *Bus. and Eco. Stat. Sec. Proceedings of the American Statistical Association* **2** 2
- [6] Galtchouk L and Pergamenschikov S 2009 *J. of the Korean Stat. Soc.* **38**
- [7] Lee T C M 2002 *J. Stat. Com. Sim.* **72** 8
- [8] Budiantara I N and Purnomo D T 2011 *Int. J. Basic and App. Sci.* **11** 1

- [9] Wang Y 1998 *J. of The American Stat. Ass.* **93**
- [10] Dagum E B and Bianconcini S 2009 *Busin. and Eco. Stat. Sec.* Proceedings of the American Stat. Ass. **2** 1
- [11] Yue Y R, Simpson D, Lindgren F and Rue H 2014 *Bayesian Anal.* **397** 424
- [12] Oehlert G W 1992 *The Annals of Stat.* **20** 1
- [13] Becher H, Kauermann G, Khomski P and Kouyate B 2009 *Bio. Journal.* **51** 1
- [14] Wang Y, Guo W and Brown M B 2000 *Stat. Sin.* **10** 2
- [15] Lee T C M 2004 *Stat. and Prob. Lett.* **67** 1

# Statistical Inference

---

## ORIGINALITY REPORT

---

17%

SIMILARITY INDEX

14%

INTERNET SOURCES

11%

PUBLICATIONS

5%

STUDENT PAPERS

---

## PRIMARY SOURCES

---

- |   |   |    |
|---|---|----|
| 1 | M. Sato, T. Hiroto, Y. Matsushita, K. Nozawa. "First-principles study of the initial stage of Pentacene adsorption on the twofold surface of the Ag-In-Yb quasicrystal", Journal of Physics: Conference Series, 2023<br>Publication | 3% |
| 2 | <a href="http://www.sciencegate.app">www.sciencegate.app</a><br>Internet Source   | 2% |
| 3 | <a href="http://repository.its.ac.id">repository.its.ac.id</a><br>Internet Source   | 2% |
| 4 | <a href="http://eprints.covenantuniversity.edu.ng">eprints.covenantuniversity.edu.ng</a><br>Internet Source   | 2% |
| 5 | Toshio Sakamizu, Hiroshi Shiraishi. "Acid-Breakable Resin-Based Chemical Amplification Positive Resist for Electron-Beam Mastering: Design and Lithographic Performance", Japanese Journal of Applied Physics, 2004<br>Publication  | 2% |
| 6 | <a href="http://scik.org">scik.org</a><br>Internet Source   | 1% |
-

7	Submitted to Columbia University Student Paper	1 %
8	Submitted to Universitas Brawijaya Student Paper	1 %
9	Faisol, Tony Yulianto, Moh. Yaqin, Achmad Basuki, Muhammad Agus Zainuddin. "Prediction of seawater salinity using truncated spline regression method", AIP Publishing, 2022 Publication	1 %
10	ijeais.org Internet Source	1 %
11	Ali Masud, Wenjie Wu, Maninderjeet Singh, Wafa Tonny et al. "Solvent Processing and Ionic Liquid-Enabled Long-Range Vertical Ordering in Block Copolymer Films with Enhanced Film Stability", Macromolecules, 2021 Publication	1 %
12	Muhammad Yahya Matdoan, Marno Wance, Abdul Malik Balami. "Ridge regression modeling in overcoming multicollinearity problems in multiple linear regression models (case study: Life expectancy in Maluku Province)", AIP Publishing, 2021 Publication	1 %

13

Muhammad Ghazali, Ita Fitriati, Ramdani Purnamasari. "Support Vector Regression for Modeling Effect of Education Rate on Life Expectancy Rate in Indonesia", IJECA (International Journal of Education and Curriculum Application), 2018

Publication

1 %

14

[www.mdpi.com](http://www.mdpi.com)

Internet Source

1 %

Exclude quotes On

Exclude matches < 1%

Exclude bibliography On