Estimated Spline in Nonparametric Regression with a Generalized Cross Validation and Unbiased Risk Approach

Agustini Tripena and Agung Prabowo
Department of Mathematics, Faculty of Mathematics and Natural Sciences,
Universitas Jenderal Soedirman, Indonesia
agustini.brsurbakti@unsoed.ac.id , agung.prabowo@unsoed.ac.id

Yosita Lianawati
Department of Information Systems,
Sekolah Tinggi Ilmu Komputer Yos Sudarso Purwokerto, Indonesia
yosita.lianawati@stikomyos.ac.id

Abdul Talib Bon
Department of Production and Operations,
University Tun Hussein Onn Malaysia, Malaysia
talibon@gmail.com

Abstract

Regression analysis is widely used to determine the pattern of the relationship between predictor variables and response variables. Nonparametric regression is a regression analysis in which the regression function is unknown and the response variable is correlated. Splines are polynomials that are segmented and flexible so they can adapt effectively to the local properties of data. The shape of the spline estimation is influenced by the lambda parameter when determining the location of the knots point. This study examines how to develop a spline estimation form in nonparametric regression cases by selecting the optimal node using the methods of generalized cross validation (GCV) and unisex risk (UBR). From the data used, the optimal knot point is 0.000130245 and 0.000402165 for the number of splines 5 and 20, respectively. The results of this study indicate that UBR tends to be smaller than the GCV value.

Keywords:

non-parametric regression, spline regression, knots point, generalized cross validation, unbiased risk.

1. Introduction

Regression analysis is a statistical method that can be done to determine the pattern of the relationship between one or more variables (Budiantara *et al*., 2009; Januaviani *et al*., 2020). In addition to knowing the pattern of relationships, regression analysis can also be used to predict. The form of the relationship pattern between the predictor variables and the response variables can be identified based on past information or by using a scatter plot (Hardle, 1990). There are three approaches in estimating the regression curve, namely the parametric, nonparametric, and semi-parametric approaches. In the parametric regression approach, there are very strong and rigid assumptions, namely the shape of the regression curve is known, for example linear, quadratic, cubic polynomial degree *p*, exponent, and others.

In the nonparametric approach, the regression curve is only assumed to be smooth in the sense within a certain function (Winarti and Sony, 2010). The advantage when using a nonparametric approach is that it has high flexibility, meaning that the data is expected to be able to find its estimation form, without any influence from the researcher's subjectivity (Budiantara, 2005b). Several methods have been widely used to model regression using a nonparametric approach, namely: kernel, spline, K-nearest neighborhood, histogram, Fourier series estimator, MARS, orthogonal series, wavelets, and neural network (Eubank, 1999).

Of the various methods that are widely used, a spline is one method that has many advantages. Splines are polynomial pieces that have segmented and truncated properties (Eubank, 1999). Various advantages of using a spline include: having an interpretation. Good visuals are flexible and can handle smooth functional characters. Also, spline can overcome data patterns and can describe changes in the behavior of the data that fluctuate in certain sub-intervals (Doksum and Koo, 2000). Spline also has the advantage of overcoming data patterns that show sharp rises/falls with the help of knot points and the resulting curve is relatively smooth (Budiantara, 2006). Spline is obtained based on optimization which is an extension of the optimizations used in parametric regression.

One of the advantages of the spline as explained is that it is flexible, meaning that this model tends to seek its data estimation of the ability of the data pattern to move. This happens because in the spline there are knots points. A knot point is a point of association that indicates a change in data behavior patterns. With spline, knots point can provide better flexibility than polynomial so it is possible to adapt effectively to local characters (Budiantara, 2005a). There are several methods for selecting the optimal knot point in nonparametric regression

Splines include the cross-validation (CV) and generalized cross-validation (GCV) method (Craven and Wahba, 1997), unisex risk (UBR) (Wang, 1998), generalized maximum likelihood (GM) (Wahba, 1998). This study will compare the results of selecting the optimal knot point with the GCV method and the UBR method. GCV and UBR are methods of selecting optimal knot points that have many advantages. The advantages possessed by GCV include simple and efficient calculation, optimal asymptotically, invariant to transformation, and not requiring information on variance σ^2. Meanwhile, the selection of knot points using the UBR method will be better if it is used on data that is not normally distributed. Based on this statement, the study aims to determine the form of the spline estimator in the nonparametric regression model and to select the optimal λ smoothing parameter. With the UBR method, then the GCV and UBR methods will be compared to choose the optimal smoothing parameter λ in the spline estimator using simulation data.

2. Research Methodology

To complete this research, the following steps were taken:

1. Examine the spline estimator in nonparametric regression.
2. Assessing the selection of smoothing parameters in the spline estimator using the UBR method.
3. Comparing the GCV and UBR methods to select the optimal smoothing parameter in the spline estimator using simulation data based on the MSE value.

3. Discussion

3.1 Nonparametric Regression

Nonparametric regression is a statistical method used to determine the relationship between the response variable and the predictor if the form of the relationship between the response variable and the predictor is unknown or previous information is not obtained. If given a data pair (X_i, Y_i), i = 1, 2, 3, …n and the relationship between the response variable Y_i and the independent X_i follows the model

$$Y_i = f(X_i) + \varepsilon_i, X_i \in [a, b], i = 1, , 2, 3, \dots n \quad (1)$$

where $f(Xi)$ is a regression curve of unknown shape and independent random error ε_i is normally distributed with zero mean and variance σ^2. In nonparametric regression flexibility is highly maintained, the function $f(Xi)$ is assumed to be smooth in a continuous and differentiable sense (Draper and Smith, 1998; Sirait et al., 2020b).

3.2 Splines in Nonparametric Regression

Spline in nonparametric regression has high flexibility properties and has the ability to estimate the behavior of data that tend to be different at different intervals (Budiantara, 2006; Eubank, 1999). The ability to estimate the behavior of this data is shown by the truncated function attached to the estimator and these pieces, called knots, are joint points that show changes in the behavior pattern of the function at different intervals. A spline is a type of piecewise polynomial, which is a polynomial that has segmented properties. This segmented nature provides more flexibility than ordinary polynomials, making it possible to adapt effectively to the location characteristics of a function or data. In the spline function, there is a knot point which is a point of association that shows changes in curve behavior at different hoses (Hardle, 1990). The spline function of degree m is any function that can generally be presented in the following form:

$$f(X_i) = \sum_{j=0}^{m} \beta_j X_i^j + \sum_{j=1}^{J} \beta_{j+m} (X_i - k_j)_+^m \quad (2)$$

where β_j is the real constant and

$$(X_i - k_j)_+^m = \begin{cases} (X_i - k_j)^m ; & X \geq k_j \\ 0 \quad ; & X < k_j \end{cases}$$

If m = 1, 2, and 3 are obtained respectively linear spline, quadratic spline, and cubic spline and k_j j is the point of knots. If it is assumed that the error ε_i has an independent distribution with a mean of zero and variance σ^2, then Y_i in the regression model is also normally distributed with a mean $f(X_i)$ and variansi σ^2. As a result, the estimation for parameter β by using the least square method is obtained, namely by minimizing the number of squares the error is as follows:

$$\sum_{i=1}^{n} \varepsilon_i^2 = (Y_i - f(X_i))^2$$
$$= \left(Y_i - \left(\sum_{j=0}^{m} \beta_j X_i^j + \sum_{j=1}^{J} \beta_{j+m\,(X_i - k_j)_+^m}\right)\right)^2 \quad (3)$$

By presenting the matrix, we get:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \tilde{\varepsilon}'\tilde{\varepsilon}$$
$$= \left(\tilde{Y} - X\tilde{\beta}\right)'\left(\tilde{Y} - X\tilde{\beta}\right)$$
$$= \tilde{Y}'\tilde{Y} - 2\tilde{\beta}'X'\tilde{Y} + \tilde{\beta}'X'X\tilde{\beta} \quad (4)$$

If equation (4) is derived concerning the vector $\tilde{\beta}$ ^ 'and the result is equal to zero, it is obtained

$$\tilde{\beta} = (X'X)^{-1}X'\tilde{Y} \quad (5)$$

with

$$X = \begin{bmatrix} 1 & X_1 & \dots & X_1^m & (X_1 - k_1)_+^m & \dots & (X_1 - k_j)_+^m \\ 1 & X_2 & \dots & X_2^m & (X_n - k_1)_+^m & \dots & (X_n - k_j)_+^m \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & X_n & \dots & X_n^m & (X_n - k_1)_+^m & \dots & (X_n - k_j)_+^m \end{bmatrix} \quad (6)$$

spline estimator in nonparametric regression

The spline function is the sum of the polynomial functions with a truncated function. In this section, we know about the nonparametric regression model, where the f curve estimation is done using a spline. Given paired data (X_j, Y_j) and the relationship between X_j and Y_j is assumed to follow a nonparametric regression model:

$$Y_i = f(X_i) + \varepsilon_i, j = 1,2, \dots, n \quad (7)$$

In this study, a study was conducted with the regression curve f approached with the spline function f with knot K. In the form of a matrix, it is presented as follows:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} f(X_1) \\ f(X_2) \\ \vdots \\ f(X_n) \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (8)$$

If the spline regression model is presented in the form of a matrix, it is obtained:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{bmatrix} 1 & X_1 & X_1^2 & \dots & X_1^m & (X_1 - k_1)_+^m & \dots & (X_1 - k_r)_+^m \\ 1 & X_2 & X_1^2 & \dots & X_2^m & (X_2 - k_1)_+^m & \dots & (X_2 - k_r)_+^m \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_n & X_1^2 & \dots & X_n^m & (X_n - k_1)_+^m & \dots & (X_n - k_r)_+^m \end{bmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \\ \beta_1 \\ \vdots \\ \beta_r \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (9)$$

or it can be written with

$$\tilde{Y} = X[K_1, K_2 \dots K_r]\tilde{\beta} + \varepsilon \quad (10)$$

Furthermore, the parameter estimation $\tilde{\beta} = (\alpha_0 \quad \alpha_1 \quad \alpha_2 \quad \cdots \quad \alpha_m \quad \beta_1 \quad \cdots \quad \alpha_{r)})'$ obtained by the least square method, by completing the optimization:

$$\underset{\substack{m \\ \beta \epsilon R = 1+r}}{\text{Min}} \left\{\tilde{\varepsilon}'\tilde{\varepsilon}\right\} = \underset{\substack{m \\ \beta \epsilon R = 1+r}}{\text{Min}} \left\{ \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}' \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \right\}$$
$$= \underset{\substack{m \\ \beta \epsilon R = 1+r}}{\text{Min}} \left\{\left(\tilde{Y} - X[K_1, K_1, \dots K_1,]\tilde{\beta}\right)'\left(\tilde{Y} - X[K_1, K_1, \dots K_1,]\tilde{\beta}\right)\right\} \quad (11)$$

The sum of the squares of the error with the matrix translation is given as follows:

$$\begin{aligned}\sum_{i=1}^{n}\varepsilon_i^2 &= \tilde{\varepsilon}'\tilde{\varepsilon}\\ &= \left(\tilde{Y} - X[K_1,K_1,\dots K_1,]\tilde{\beta}\right)'\left(\tilde{Y} - X[K_1,K_1,\dots K_1,]\tilde{\beta}\right)\\ &= \left(\tilde{Y} - X[K]\tilde{\beta}\right)'\left(\tilde{Y} - X[K]\tilde{\beta}\right)\\ &= \tilde{Y}'\tilde{Y} - 2\tilde{\beta}'X[K]'\tilde{Y} + \tilde{\beta}'X[K]'\tilde{X}[\mathrm{K}]\tilde{\beta}' \end{aligned} \quad (12)$$

If equation (12) is derived with respect to the vector and the result is equalized to zero, we get:

$$\frac{\partial(\tilde{\varepsilon}'\tilde{\varepsilon})}{\partial\tilde{\beta}'} = \frac{\partial\left(\tilde{Y}'\tilde{Y} - 2\tilde{\beta}'X[K]'\tilde{Y} + \tilde{\beta}'X[K]'X[\mathrm{K}]\,\tilde{\beta}'\right)}{\partial\tilde{\beta}'} = 0$$

We obtained

$$\tilde{\beta} = (X[K]'X[K])^{-1}\mathrm{X[K]}'\,\tilde{Y} \quad (13)$$

where $X[K] = X[K_1, K_2, \dots K_r]$

$$= \begin{bmatrix} 1 & X_1 & X_1^2 & \dots & X_1^m & (X_1-K_1)_+^m & \dots & (X_1-K_r)_+^m \\ 1 & X_2 & X_2^2 & \dots & X_2^m & (X_2-K_1)_+^m & \dots & (X_2-K_r)_+^m \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_n & X_n^2 & \dots & X_n^m & (X_n-K_1)_+^m & \dots & (X_n-K_r)_+^m \end{bmatrix}$$

Consequently, the estimate for the spline regression curve with knots K is given by:

$$\hat{f}(X_I) = X[K]\,\underset{\sim}{\tilde{\beta}} \quad (14)$$

$$= X[K]'(X[K]')X[K])^{-1}X[K]'\tilde{Y}$$

So that:

$$\hat{f}(X_I) = A[K\,]\tilde{Y} \quad (15)$$

where A[K] = X[K $A[K] = X[K](X[K]'(X[K]')X[K])^{-1}$X[K]' represents the points of knots (Tripena, 2011).

3.3 Testing of Spline Nonparametric Regression Model Parameters

3.3.1 Concurrent Test

Simultaneous testing of model parameters is a simultaneous regression curve parameter test using the *F*-test. The hypothesis in the simultaneous test for n = 75 as follows:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{m+p} = 0$

$H_1 :$ *at least one* $\beta_k \neq 0$; k = 1, 2, …., $m + p$

The $m + p$ value represents many parameters in the nonparametric spline regression except β_0

Test Statistic: $F_{count} = \frac{MSR}{MSE}$ (16)

where

$MSR = \frac{SSR}{df_{reg}} = \frac{\sum_{j=1}^{n}(\hat{Y}_i - \bar{Y})^2}{m+p}$ and

$MSE = \frac{MSE}{df_{error}} = \frac{\sum_{j=1}^{n}(Y_i -)\hat{Y}_i^{\,2}}{n-(m+p)-1}$

Obtained the value $F_{obtained\ the\ value} = 0.91763327$,

For n = 125, we obtained the value $F_{obtained\ the\ value} = 0.2946812$.

H_0 is rejected if $F_{count} > F_{n-(m+p)-1}$ or $p - value < \alpha$.

3.3.2 Individual Test

Individual testing is conducted to determine whether individual parameters have a significant effect on the response variable. The hypothesis on individual tests for n = 75 is as follows:

$H_0 : \beta_k = 0$

$H_1 : \beta_k \neq 0$; k = 1, 2,…, $m + p$

Individual testing is done using the *t*-test (Draper and Smith, 1998):

Test statistic: $t_{count} = \frac{\beta_k}{\sqrt{var(\beta_k)}}$ (17)

where $var(\beta_k) = diag[(\mathbf{X}'\boldsymbol{X})^{-1}\hat{\sigma}^2]$, where $\hat{\sigma}^2$ is MSE.

Obtained the value $t_{count\ constants} = 0.957055$, $t_{count\ c2} = 0.821415$ and $t_{count\ c3} = 0.957414$

For n = 125 we obtained the value

$t_{count\ constants} = -0.10252531$, $t_{\text{count } c2} = 0.69122977$ and $t_{\text{count } c3} = 0.3447888$
H_0 rejected if $|t_{count}| > t_{(n(m+p)-1}\ or\ p - value < \alpha$

3.4 Optimal Knot Point Selection

The knot point is a joint point where there is a change in behavior in the data. The best spline regression model depends on the optimal knot point (Eubank, 1998 in Tripena, 2011). Methods for finding optimal knot points that are often used are general cross-validation (GCV), mean square error (MSE), and unbiassed risk (UBR).

3.4.1 Mean Square Error (MSE)

The simple criterion used as a measure of performance over a good estimator is MSE is (Eubank, 1999; Sirait et al., 2020a; Sirait et al., 2020c):

$$MSE\ (K) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \tilde{f}\ (X_i))^2 \tag{18}$$

where:

X_i : variable independen / predictor
Y_i : variable dependen / respon
n : number of observations

3.4.2 General Cross-Validation (GCV)

Other criteria that can be used as performance measures over a good estimator are GCV and. GCV method is a method that is often used to select the most optimal knot points. Assuming that trace $\left[\boldsymbol{A}(\tilde{K})\right] < n$, then the GCV criteria are defined as follows (Craven and Wahba, 1997):

$$GCV(K) = \frac{MSE(\tilde{K})}{(n^{-1}trace[I-\boldsymbol{A}(K)])^2} \tag{19}$$

where:

I : identity matrix
n : number of observations
$A(\tilde{K})$ is the matrix $X(X^TVX)^{-1}X^TV$
$MSE(\tilde{K}) = n^{-1}\sum_{i=1}^{n}(Y_i - \tilde{\tilde{f}}\ (X))^2$
$\tilde{f}(X) = A(\tilde{K})\tilde{Y} = [X(X^TVX)^{-1}X^T]\tilde{Y}$

The GCV equation can be written with the following equation:

$$GCV(\tilde{K}) = n^{-1}\sum_{i=1}^{n}(Y_i - \tilde{\mu}_{K(J)})^2 \quad \left\{\frac{1-\boldsymbol{A}(K_1,K_2,\dots K_J)}{n^{-1}trace[\boldsymbol{I}-\boldsymbol{A}(K_1,K_2,\dots K_J)]}\right\}^2 \tag{20}$$

Equation (20) can be seen that GCV is the result of a weighted generalization of the CV method the CV equation is as follows:

$$CV(\tilde{K}) = n^{-1}\sum_{i=1}^{n}\left(\frac{(Y_i-\tilde{\mu}_{K(J)}}{\boldsymbol{I}-\boldsymbol{A}(K_1,K_2,\dots K_J)}\right)^2 \tag{21}$$

To obtain optimal knot points using the GCV method, the following optimizations are carried out:

$$\begin{aligned}
&\underset{K_k\in R}{\text{Min}}\{GCV(K_1,K_2,\dots K_J)\} = \underset{K_k\in R}{\text{Min}}\left\{\frac{MSE(K_1,K_2,\dots K_J)}{(n^{-1}(trace[\boldsymbol{I}-\boldsymbol{A}(K_1,K_2,\dots K_J)]))^2}\right\}\\
&= \underset{K_k\in R}{\text{Min}}\left\{\frac{n^{-1}\sum_{i=1}^{n}(Y_i-\tilde{f}(X))^2}{(n^{-1}(trace[\boldsymbol{I}-\boldsymbol{A}(K_1,K_2,\dots K_J)]))^2}\right\} = \underset{K_k\in R}{\text{Min}}\left\{\frac{n^{-1}\sum_{i=1}^{n}(Y_i-\tilde{f}(X))^2}{\{n^{-1}(trace[\boldsymbol{I}-\boldsymbol{A}(K_1,K_2,\dots K_J)]\}^2}\right\}\\
&= \underset{K_k\in R}{\text{Min}}\left\{\frac{n^{-1}\sum_{i=1}^{n}(Y_i-\tilde{f}(X))^2}{\{1-n^{-1}trace\ \boldsymbol{A}(\underbrace{\boldsymbol{K}}_{\sim})\}^2}\right\} = \underset{K_k\in R}{\text{Min}}\left\{\frac{n^{-1}(\tilde{Y}-A[\tilde{K}]\tilde{Y})^T(\tilde{Y}-A[\tilde{K}]\tilde{Y})}{\{1-n^{-1}trace\ \boldsymbol{A}(\tilde{\boldsymbol{K}})\}^2}\right\}\\
&= \underset{K_k\in R}{\text{Min}}\left\{\frac{n^{-1}\tilde{Y}^T(\boldsymbol{I}-\boldsymbol{A}[\tilde{K}])^T(\boldsymbol{I}-\boldsymbol{A}[\tilde{K}]\tilde{Y})}{\{1-n^{-1}trace\ \boldsymbol{A}(\tilde{\boldsymbol{K}})\}^2}\right\}
\end{aligned} \tag{22}$$

The smallest GCV value generated will provide the optimal knot point.

3.4.3 Unbiassed Risk (UBR)

The following describes the definition of the loss function and the risk function referred to in (Eubangk, 1999) as follows:

Definition 1: Loss Function

If $\underset{\sim}{\tilde{g}}_{\lambda} = (\tilde{g}_{\lambda 1},\ \tilde{g}_{\lambda 2,} \ldots \tilde{g}_{\lambda n,})'$ *is the estimator for* $\underset{\sim}{g} = (g_{1,}\, g_{2,} \ldots g_{n,})'$ *'then the squared loss function is defined as*

$$L(\lambda) = n^{-1} \sum_{i=1}^{n} (g_i - \tilde{g}_{\lambda i})^2 \tag{23}$$

Definition 2: Risk Function

If $\underset{\sim}{\tilde{g}}_{\lambda} = (\tilde{g}_{\lambda 1},\ \tilde{g}_{\lambda 2,} \ldots \tilde{g}_{\lambda n,})'$ *is the estimator for* $\underset{\sim}{g} = (g_{1,}\, g_{2,} \ldots g_{n,})'$*, then the expectation of the quadratic risk function is defined as*

$$R(\lambda) = E[L(\lambda)] = \sum_{i=1}^{n} E(g_i - \tilde{g}_{\lambda i})^2 \tag{24}$$

L(λ) and R(λ) are measures of the performance of an estimator.

3.5 Selection of Refining Parameters in the Spline Estimator

The spline estimator is $\underset{\sim}{\tilde{g}}_{\lambda} = S(\lambda)\, \underset{\sim}{Y}$ is very dependent on λ (smoothing parameter) (Wahba, 1990; Eubank, 1999) states that the smoothing parameter is very dependent on λ has a very important role in nonparametric regression. For small λ (λ → 0) will give a very rough estimator. Conversely, a very large value λ (λ → ∞) will produce a very smooth estimator. In nonparametric regression, smoothing λ must be selected as the most optimal. There are various methods to select the optimal smoothing parameter, one of which is the UBR method. In this research, it will be derived to select the smoothing parameter in the spline estimator. Define the quadratic loss function:

$$L(\lambda) = n^{-1} \sum_{i=1}^{n} (\tilde{g}_{\lambda}(X_i) - g(X_i))^2 \tag{25}$$

In connection with the Loss function the Risk function is defined:

$$R(\lambda) = n^{-1} \sum_{i=1}^{n} E(\tilde{g}_{\lambda}(X_i) - g(X_i))^2 \tag{26}$$

The risk function $R(\lambda)$ equation (22) is presented in vector form:

$$\begin{aligned}
R(\lambda) &= n^{-1} \sum_{i=1}^{n} E([(\tilde{g}_{\lambda}(X) - g(X)'[(\tilde{g}_{\lambda}(X) - g(X)]) \\
&= n^{-1} E([S(\lambda)\, \underset{\sim}{Y} - g(X)]'[S(\lambda)\, \underset{\sim}{Y} - g(X)]) \\
&= n^{-1} E([S(\lambda)(\underset{\sim}{g}\left(\underset{\sim}{X}\right) + \varepsilon) - g(X)]'[S(\lambda)\, \underset{\sim}{g}\left(\underset{\sim}{X}\right) + \varepsilon) - \underset{\sim}{g}\,(\underset{\sim}{X})]) \\
&= n^{-1} E([S(\lambda)(\underset{\sim}{g}\left(\underset{\sim}{X}\right) + S(\lambda)\, \underset{\sim}{\varepsilon} - g(X)]'[S(\lambda)\, \underset{\sim}{g}\left(\underset{\sim}{X}\right) + S(\lambda)\, \underset{\sim}{\varepsilon} - \underset{\sim}{g}\,(\underset{\sim}{X})]) \\
&= n^{-1} E([S(\lambda)(\underset{\sim}{g}\left(\underset{\sim}{X}\right) - \underset{\sim}{g}\left(\underset{\sim}{X}\right) + S(\lambda)\, \underset{\sim}{\varepsilon}]' - [S(\lambda)\, \underset{\sim}{g}\left(\underset{\sim}{X}\right) - \underset{\sim}{g}\left(\underset{\sim}{X}\right) + S(\lambda)\, \underset{\sim}{\varepsilon}]) \\
&= n^{-1} E([S(\lambda) - I)(\underset{\sim}{g}\left(\underset{\sim}{X}\right) + S(\lambda)\, \underset{\sim}{\varepsilon}]' - [S(\lambda)\, \underbrace{-I)g}\left(\underset{\sim}{X}\right) + S(\lambda)\, \underset{\sim}{\varepsilon}]) \\
&= n^{-1} E([\left(\underset{\sim}{g}\left(\underset{\sim}{X}\right)'\, S(\lambda) - I\right)' + S(\lambda)\, \underset{\sim}{\varepsilon}]' - [S(\lambda)\, \underbrace{-I)g}\left(\underset{\sim}{X}\right) + S(\lambda)\, \underset{\sim}{\varepsilon}]) \\
&= n^{-1} E([\left(\underset{\sim}{g}'\left(\underset{\sim}{X}\right)\quad S(\lambda) - I\right)' + \underset{\sim}{\varepsilon}\,' S'(\lambda)] - [S(\lambda)\, \underbrace{-I)g}\left(\underset{\sim}{X}\right) + S(\lambda)\, \underset{\sim}{\varepsilon}]) \\
&= n^{-1} E(\left(\underset{\sim}{g}'(X)S(\lambda) - I\right)' + S(\lambda) - I) - \underset{\sim}{g}\,(X) + \underset{\sim}{g}'(X)S(\lambda)\, \underbrace{-I)}' S(\lambda)\, \underset{\sim}{\varepsilon} + \underset{\sim}{\varepsilon}\,' S'(\lambda) - S(\lambda)\, \underbrace{-I)g}\left(\underset{\sim}{X}\right) \\
&\qquad + \underset{\sim}{\varepsilon}\,' S'(\lambda) S(\lambda)\, \underset{\sim}{\varepsilon})) \\
&= n^{-1} \left(\underset{\sim}{g}'(X)S(\lambda) - I\right)' + S(\lambda) - I) - \underset{\sim}{g}\,(X) + n^{-1}\, \underset{\sim}{g}'(X)S(\lambda)\, \underbrace{-I)}' S(\lambda) E(\underset{\sim}{\varepsilon}) \\
&\qquad + n^{-1} E\, \underset{\sim}{\varepsilon}\,' S'(\lambda) S(\lambda)\, \underbrace{-I)g}\left(\underset{\sim}{X}\right) + n^{-1}\, E(\underset{\sim}{\varepsilon}\,' S'(\lambda) S(\lambda)\, \underset{\sim}{\varepsilon}) \\
&= n^{-1} \left(\underset{\sim}{g}'(X)S(\lambda) - I\right)' S(\lambda) - I) - \underset{\sim}{g}\,(X) + 0 + 0 + n^{-1} E\, \underset{\sim}{\varepsilon}\,' S'(\lambda) S(\lambda)\, \underset{\sim}{\varepsilon} \\
&= n^{-1} \left(\underset{\sim}{g}'(X)S(\lambda) - I\right)' S(\lambda) - I) - \underset{\sim}{g}\,(X) + 0 + n^{-1} trace(S'(\lambda)S(\lambda)\sigma^2) \\
&= n^{-1} \left(\underset{\sim}{g}'(X)S(\lambda) - I\right)' S(\lambda) - I) - \underset{\sim}{g}\,(X) + n^{-1}\sigma^2 trace(S'(\lambda)S(\lambda))
\end{aligned}$$

From the description above, the risk function R(λ) in vector form is obtained:

$$R(\lambda) = n^{-1} \left(\underset{\sim}{g}'(X)S(\lambda) - I\right)' S(\lambda) - I) - \underset{\sim}{g}\,(X) + n^{-1}\sigma^2 trace(S'(\lambda)S(\lambda)) \tag{27}$$

$$E\big(UBR(\lambda)\big) = R(\lambda)$$

Suppose it is given

$$\text{UBR}(\lambda) = n^{-1} Y'(S(\lambda) - I)'S(\lambda) - I)\, \underset{\sim}{Y}) + n^{-1}\sigma^2 trace(S'(\lambda)S(\lambda)) + n^{-1}\sigma^2 trace(S(\lambda) - I)'S(\lambda) - I))$$

Furthermore, it will be proven that UBR (λ) is unbiased for R(λ)

$$E[\mathrm{UBR}(\lambda)] = E([\, n^{-1}Y'(S(\lambda)-I)'S(\lambda)-I)\,\underset{\sim}{Y}] + n^{-1}\sigma^2 trace[S'(\lambda)S(\lambda)] + n^{-1}\sigma^2 trace[S(\lambda)-(I)'S(\lambda)-])$$

$$= E([\, n^{-1}Y'\,(S(\lambda)-I)'S(\lambda)-I)\,\underset{\sim}{Y}] + n^{-1}\sigma^2 trace[S'(\lambda)S(\lambda)] + n^{-1}\sigma^2 trace[S(\lambda)-I)'S(\lambda)-])] \quad (28)$$

It will be calculated first

$$= E([\, n^{-1}Y'(S(\lambda)-I)'S(\lambda)-I)\,\underset{\sim}{Y}]$$
$$= n^{-1}E[(\underset{\sim}{g}'\left(\underset{\sim}{X}\right)+\underset{\sim}{\varepsilon})'(S(\lambda)-I)'S(\lambda)-I)(\underset{\sim}{g}\,(\underset{\sim}{X})+\underset{\sim}{\varepsilon})]$$
$$= n^{-1}E[(\underset{\sim}{g}'\left(\underset{\sim}{X}\right)(S(\lambda)-I)'S(\lambda)-I)(\underset{\sim}{g}\,(\underset{\sim}{X})+\underset{\sim}{g}'\left(\underset{\sim}{X}\right)(S(\lambda)-I)'S(\lambda)-I)(\underset{\sim}{g}\,(\underset{\sim}{X})+\underset{\sim}{\varepsilon})]$$
$$+\underset{\sim}{\varepsilon}'(S(\lambda)-I)'S(\lambda)-I)(\underset{\sim}{g}\left(\underset{\sim}{X}\right)+\underset{\sim}{\varepsilon}'(S(\lambda)-I)'S(\lambda)-I)\,\underset{\sim}{\varepsilon}]$$
$$= n^{-1}\,\underset{\sim}{g}'\left(\underset{\sim}{X}\right)(S(\lambda)-I)'S(\lambda)-I)(\underset{\sim}{g}\,(\underset{\sim}{X})+0+0+n^{-1}trace[(S(\lambda)-I)'S(\lambda)-I)\,\underset{\sim}{\varepsilon}]$$
$$= n^{-1}\,\underset{\sim}{g}'\left(\underset{\sim}{X}\right)(S(\lambda)-I)'S(\lambda)-I)(\underset{\sim}{g}\,(\underset{\sim}{X})+n^{-1}\sigma^2 trace[(S(\lambda)-I)'S(\lambda)-I)]$$

As a result:

$$E[UBR((\lambda)] = n^{-1}\,\underset{\sim}{g}'\left(\underset{\sim}{X}\right)(S(\lambda)-I)'S(\lambda)-I)(\underset{\sim}{g}\,(\underset{\sim}{X})+n^{-1}\sigma^2 trace[(S(\lambda)-I)'S(\lambda)-I)]$$
$$= R(\lambda)$$

So we get an unbiased UBR (λ) for R(λ)

Thus the UBR method for selecting smoothing parameters in the spline estimator is given by:

$$UBR(\lambda) = n^{-1}\,\underset{\sim}{Y}'\left(\underset{\sim}{X}\right)(S(\lambda)-I)'S(\lambda)-I)\,\underset{\sim}{Y}+n^{-1}\sigma^2 trace[S'(\lambda)S(\lambda)+n^{-1}\sigma^2 trace[S'(\lambda)-I)'S(\lambda)-I) \quad (29)$$

The λ value optimal is obtained from the value of which minimizes UBR (λ)

If σ^2 is not known, it can be estimated by:

$$\tilde{\sigma}^2 = \frac{\underset{\sim}{Y}'(I-S(\lambda))'((I-S(\lambda))\underset{\sim}{Y}}{trace[(I-S(\lambda))} \quad (30)$$

3.6 UBR and GCV Methods in Spline Estimation using Simulation Data

In this research, a simulation was conducted to provide an overview of the spline model. The research simulation was conducted to evaluate the goodness of the UBR and GCV methods to compare the goodness of the two methods. The measurement reliability is based on the smallest MSE value obtained in the UBR and GCV methods. We use MATLAB software for simulation. The simulation method is described and builds a regression model $Y_i = f(X_i) + \varepsilon$ with a sample size of n = 75, n = 125 with a value of K = 5, K = 20. Given $X_i = \frac{i}{n}$ with i =1, 2, 3,…,n. Generating $\varepsilon_i \sim N(0, \sigma^2)$, where $\sigma^2 = 0.1$. Calculated matrix $\boldsymbol{X}$, matrix $\boldsymbol{D}$ and matrix hat

$$\boldsymbol{S}(\lambda) = \boldsymbol{X}(n^{-1}\boldsymbol{X}'\boldsymbol{X} + \lambda\mathbf{D})^{-1}n^{-1}\boldsymbol{X}' \quad (31)$$

Furthermore, the estimator is calculated

$$\underset{\sim}{\tilde{\boldsymbol{a}}}\,(\lambda) = (n^{-1}\boldsymbol{X}'\boldsymbol{X} + \lambda\mathbf{D})^{-1}n^{-1}\boldsymbol{X}'Y \quad (32)$$

and determine the spline estimator with $\underset{\sim}{\tilde{f}}_{(\lambda)}(t) = \boldsymbol{S}(\lambda)$, so we get

$$\underset{\sim}{\tilde{f}}_{(\lambda)}(X) = \tilde{\boldsymbol{b}}(\lambda)X + \frac{1}{2}\tilde{a}_0(\lambda) + \sum_{k=1}^{K}\widetilde{a_k}(\lambda)kX_i \quad (33)$$

3.7 Simulation Results for Optimal λ Value

In this study, the sample size was n = 75 and n = 125 with the variance value σ^2, namely 0, 1. In addition, the K value used was divided into two, namely $K = 5$ and $K = 20$. The results of optimal λ and MSE analysis using two methods, namely UBR and GCV in spline estimation, are shown in Table 1 as follows:

Table 1. Simulation results of optimal λ values and MSE with UBR method and GCV on Spline estimation with n =75 and n =125, σ^2 = 0, 1 and K = 5, K = 20

n	Var	K	UBR method		GCV method	
			MSE UBR	λ_{optimal}	MSE GCV	λ_{optimal}
75	0.1	5	0.000130245	0.00100320	0.000402165	0.00341503

125	0.1	20	0.001011035	0.00522312	0.005120243	0.00758340
		5	0.000300134	0.00203631	0.000610113	0.00510231
		20	0.002234200	0.01132043	0.0122305437	0.02286325

Based on the results in Table 1, the MSE value for the UBR method at size $n = 75$ with $K = 5$ and $K = 20$, respectively, is obtained at 0.000130245 and 0.001011035. Meanwhile, the MSE value for the GCV method at size $n = 75$ with $K = 5$ and $K = 20$, respectively, was obtained at 0.000402165 and 0.005120243. Based on the MSE value of the two methods, the UBR method has the smallest MSE value compared to the GCV method. In addition, the MSE value for $n = 125$ with $K = 5$ and $K = 20$ in Table 1, for the UBR method also has the smallest MSE value compared to the GCV method. In Table 1, it can also be seen that from the two methods, the MSE value continues to increase along with the increase in n and K values. The optimal value of λ is obtained when $n = 75$ and $K = 5$.

In the results of the analysis in Table 1, determining the MSE value and λ value using the Minitab 17 software tool. The results of the analysis of the two methods for the Histogram, Scatter Plot and Time Series Plot for parameter $n = 75$ are given in Figures 1, Figure 2 and Figure 3, respectively. Meanwhile, for the parameter $n = 125$ from both methods, the histogram, Scatter Plot and Time Series Plot results were obtained, respectively, as shown in Figure 4, Figure 5 and Figure 6.

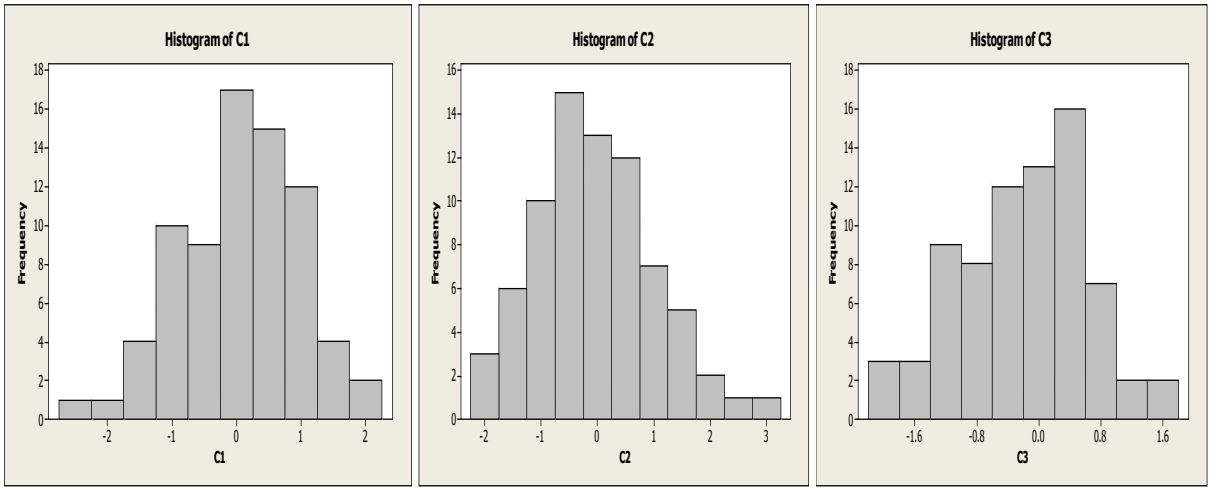

Figure 1. Histogram for $n = 75$

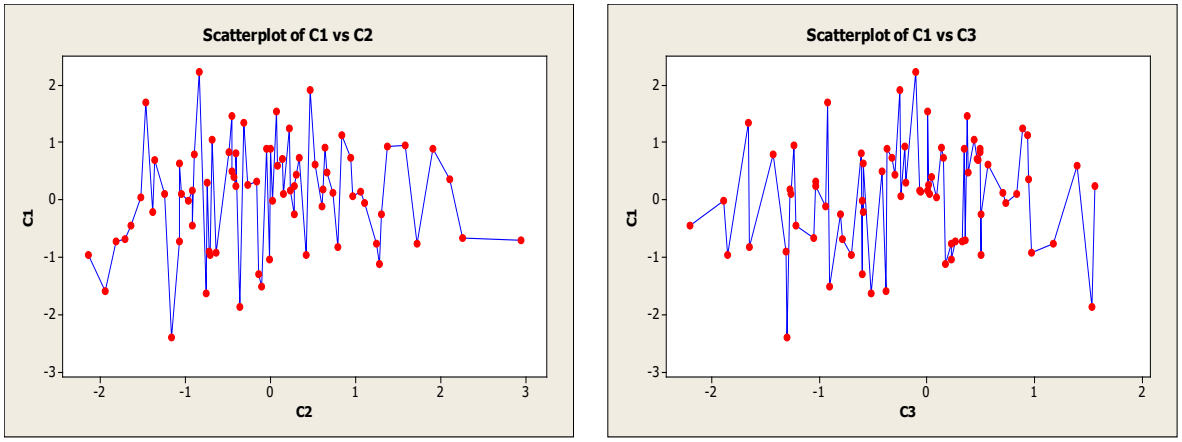

Figure 2. Scatter Plot for $n = 75$

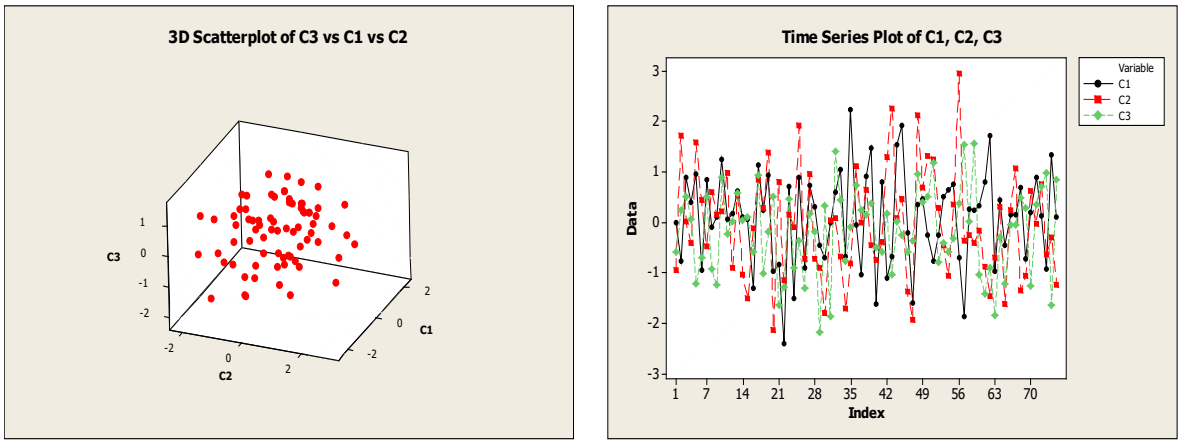

Figure 3. Time Series Plot for $n = 75$

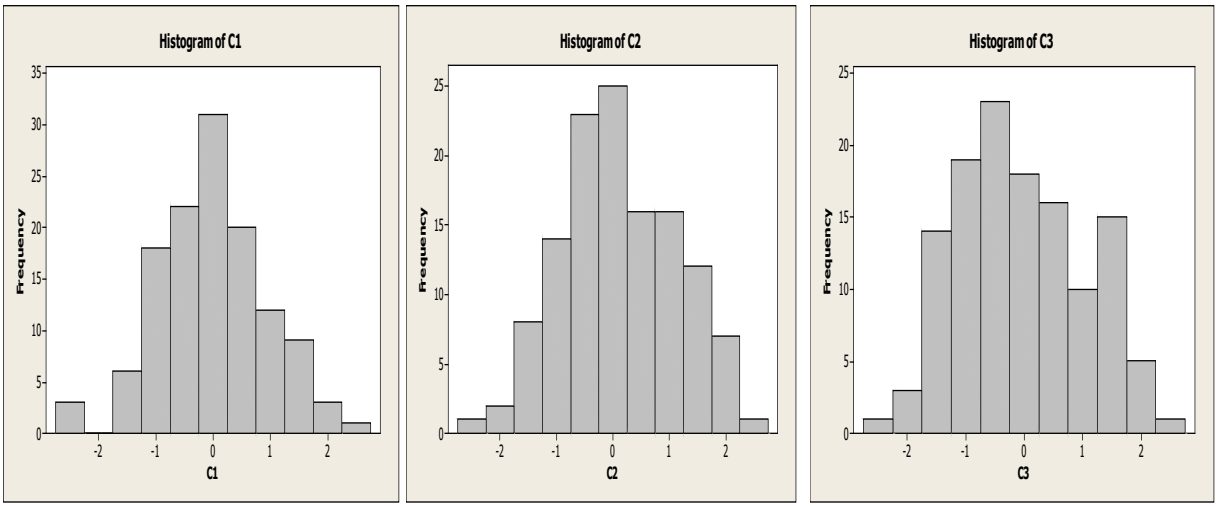

Figure 4. Histogram for $n = 125$

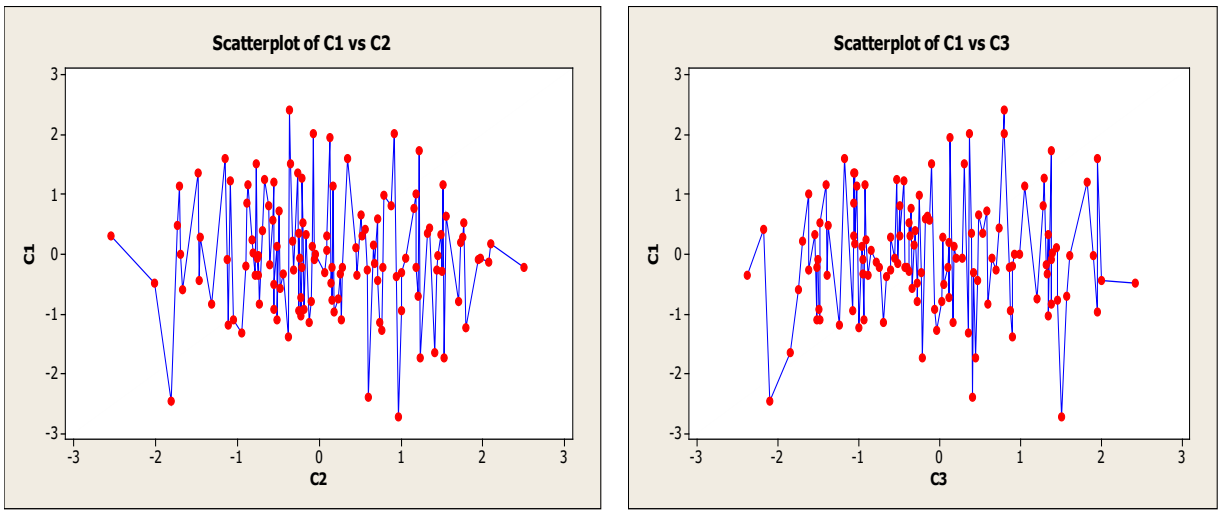

Figure 5. Scatter Plot for $n = 125$

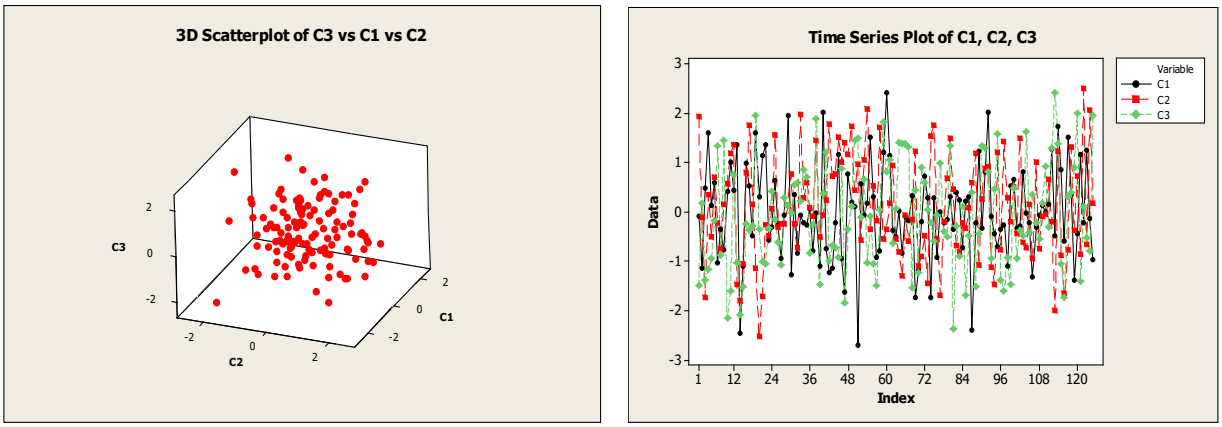

Figure 6. Time Series Plot for $n = 125$

4. Conclusions

Based on the analysis and discussion, the conclusions are as follows:

1. The optimization result of spline estimator in nonparametric regression is obtained

   $$\hat{f}_\lambda(X_i) = \hat{b}(\lambda)X_i + \frac{1}{2}\hat{a}_0(\lambda) + \sum_{k=1}^{K}\hat{a}_i(\lambda)k\,X_i$$

   where $\underset{\sim}{\hat{a}}(\lambda) = (\hat{b}(\lambda), \frac{1}{2}\hat{a}_0(\lambda), \frac{1}{2}\hat{a}_1(\lambda), \ldots, \frac{1}{2}\hat{a}_K(\lambda))'$
2. The spline estimator $\underset{\sim}{\hat{f}}_\lambda(X)$ is linear in the $\underset{\sim}{Y}$ observation and has a biased nature for the $\underset{\sim}{f}(X)$. Curve. Estimator spline $\underset{\sim}{\hat{f}}_\lambda(X)$ is normally distributed if the model error is also normally distributed
3. Selection of smoothing parameters in the optimal knot point spline estimator using the GCV method

   $$\underset{K_k \in R}{\text{Min}}\left\{\frac{n^{-1}\tilde{Y}^T(\boldsymbol{I} - \boldsymbol{A}[\tilde{K}])^T(\boldsymbol{I} - \boldsymbol{A}[\tilde{K}]\tilde{Y})}{\{1 - n^{-1}trace\,\boldsymbol{A}(\tilde{\boldsymbol{K}})\}^2}\right\}$$
4. Selection of smoothing parameters in spline estimator with UBR method is given by:

   $$UBR(\lambda) = n^{-1}\underset{\sim}{Y}'\left(\underset{\sim}{X}\right)(S(\lambda) - I)'S(\lambda) - I)\,\underset{\sim}{Y} + n^{-1}\sigma^2 trace[S'(\lambda)S(\lambda) + n^{-1}\sigma^2 trace[S'(\lambda) - I)'S(\lambda) - I)$$
5. The simulation results for n =75 and n =125, σ^2= 0, 1 and K = 5, K = 20 it is found that the MSE value of the UBR method tends to be smaller than the MSE value of the GCV method in each simulation model. The greater the K value, the greater the MSE value for both the UBR and GCV methods. Selection of the optimal smoothing parameter λ using the UBR method tends to be better than the GCV method.

Acknowledgments

The authors would like to thank Jenderal Soedirman University (UNSOED) and the Ministry of Research, Technology and High Education of Republic of Indonesia. This work was an individual research.

References

Budiantara, I.N., Penentuan titik-titik knots dalam regresi spline, *Jurnal Jurusan Statistika FMIPA-ITS*, 2005a.

Budiantara, I.N., Model keluarga spline polinomial truncated dalam regresi semiparametrik, *Berkala MIPA-ITS*, 2005b.

Budiantara, I.N., Lestari, B., Islamiyati, A., Pemilihan knot optimal dalam estimator sline terbobot pada regresi nonparametrik heteroskedastik data longitudinal, *Proceeding on Seminar Nasional Statistika IX Institut Teknologi Sepuluh Nopember Surabaya*, 2009.

Budiantara, I.N., Model spline dengan knots optimal, *Jurnal Ilmiah Dasar FMIPA Universitas Jember*, 7: 77-85, 2006.

Craven, P., and Wahba, G., Smooting noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Number Math*, 31: 377-403, 1997.

Doksum, K., and Koo, Y.J., On spline estimators and prediction intervals in nonparametric regression", *Computational Statistics and Data Analysis*, 35: 57 – 82, 2000.

Draper, N.R and Smith, H., *Applied regression analysis*, Third Edition, Canada, John Wiley & Sons, Inc., 1998

Eubank, R.L., *Spline smoothing and nonparametric regression*, New York, Marcel Deker, 1999.

Hardle, W., *Applied nonparametric regression*, New York, Cambridge University Press, 1990.

Januaviani, T.M.A., Sukono, Lesmana, E., Kalfin. Modeling of money supply using lasso regression and simulated annealing. *Journal of Advanced Research in Dynamical and Control Systems*, 12(6), pp. 840–848. 2020

Sirait, H., Sukono, Sundari, S., and Kalfin. (2020a). Ratio Estimator of Population Mean using Quartile and Skewness Coefficient. *International Journal of Advanced Science and Technology*, 29(6), pp. 3289 – 3295

Sirait, H., Sukono, Syahara, E., Prabowo, A., and Kalfin. (2020b). Modification of Regression Rating to Estimate Population Average Using Quartiles and Population Variations Coefficient. *International Journal of Advanced Science and Technology*, 29(7s), pp. 3494-3500 .

Sirait, H., Sukono, Karolin, S., Kalfin and Bon, A. T. (2020c). Ratio Estimator for Population Variations Using Additional Information on Simple Random Sampling. *Proceedings of the International Conference on Industrial Engineering and Operations Management* Detroit Michigan, USA, pp. 2512– 2521.

Tripena, A., Penentuan model regresi spline terbaik, *Jurnal Program Studi Matematika FMIPA dan Teknik Jenderal Soedirman*, 2011.

Wahba, G., A. Comparison of GCV and GML for choosing the smoothing parameter in generalized spline smoothing problem, *Annal of Statistic*, 13: 1378-1402, 1985.

Wahba, G. (1990). *Spline model for observation data*, Pensylvania, SIAM, 1990.

Wang, Y., Smoothing spline models with correlated errors, *Journal of the American Statistics Association*, 93: 343-348, 1998.

Winarti & Sony, S. Pendekatan semiparametrik spline pada data nilai Ujian Nasional Siswa SMKN I Nguling Pasuruan, *Jurnal Sains dan Seni*, 3(2): 194-199, 2010.

Biographies

Agustini Tripena is a lecturer in the Department of Mathematics, Universitas Jenderal Soedirman, with the field of research is spline regression.

Agung Prabowo is a lecturer in the Department of Mathematics, Universitas Jenderal Soedirman, with the field of research are: financial mathematics, survival model analysis and ethnomathematics.

Yosita Lianawati is a lecturer in Department of Information Systems, with the field of research is computer science.

Abdul Talib Bon is a professor of Production and Operations Management in the Faculty of Technology Management and Business at the Universiti Tun Hussein Onn Malaysia since 1999. He has a PhD in Computer Science, which he obtained from the Universite de La Rochelle, France in the year 2008. His doctoral thesis was on topic Process Quality Improvement on Beltline Moulding Manufacturing. He studied Business Administration in the Universiti Kebangsaan Malaysia for which he was awarded the MBA in the year 1998. He's bachelor degree and diploma in Mechanical Engineering which his obtained from the Universiti Teknologi Malaysia. He received his postgraduate certificate in Mechatronics and Robotics from Carlisle, United Kingdom in 1997. He had published more 150 International Proceedings and International Journals and 8 books. He is a member of MSORSM, IIF, IEOM, IIE, INFORMS, TAM and MIM.