Estimation of Matrix Variance-Covariance on Nonparametric Regression Spline Truncated for Longitudinal Data

Mustain Ramli\textsuperscript{a}, Vita Ratnasari\textsuperscript{b} and I Nyoman Budiantara\textsuperscript{c}\textsuperscript{*}

Department of Statistics, Faculty of Science and Data Analytics, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia

E-mail: \textsuperscript{a}mustainramli1994@gmail.com
\textsuperscript{b}vitaratna70@gmail.com
\textsuperscript{c}nyomanbudiantara65@gmail.com

\textsuperscript{*}Corresponding author

Abstract. Regression analysis is a statistical analysis used to determine the pattern of relationships between predictor variables and response variables. There are two models estimation approaches in regression analysis, namely parametric regression and nonparametric regression. The parametric regression approach is used if the shape of the regression curve is known. If the data relationship pattern forms a linear pattern then the linear parametric regression approach is used. If the data relationship pattern forms a quadratic pattern, then the quadratic regression approach is used, and others \[1\]. Forms of relationship patterns can be identified based on past information or by using scatter plot data \[2\]. In practice, not all data follows certain patterns so that parametric regression which when forced to be used will give misleading results.

1. Introduction

Regression analysis is a statistical analysis used to find out the similarity in the pattern of relationships between predictor variables and response variables. There are two models estimation approaches in regression analysis, namely parametric regression and nonparametric regression. The parametric regression approach is used if the shape of the regression curve is known. If the data relationship pattern forms a linear pattern then the linear parametric regression approach is used. If the data relationship pattern forms a quadratic pattern, then the quadratic regression approach is used, and others \[1\]. Forms of relationship patterns can be identified based on past information or by using scatter plot data \[2\]. In practice, not all data follows certain patterns so that parametric regression which when forced to be used will give misleading results.
conclusions. Nonparametric regression is a model approach that assumes curve patterns are unknown and are contained within a particular function, such as Sobolev space, continuous function space, Hilbert space, Entropi space and others [3]. Hence, by using this approach the model obtained will be better because the data is expected to find its own estimation form without the subjectivity influenced by researchers and also this approach is very flexible [1]. Some of the nonparametric regression approach models that are of concern and most often used by many researchers include nonparametric kernel regression [2], spline nonparametric regression [4]-[3], Fourier series of nonparametric regression [5] and Wavelet [6]. Some of these approaches, especially spline nonparametric regression have been widely used in various fields of science, such as medicine, pharmacology, economics and others. Many researchers try to compare approaches in nonparametric regression, One of which is [7] compares the spline and kernel approach and concludes that spline is numerically better than the kernel. Spline is one of the nonparametric regression approaches that has very specific and very good statistical and visual interpretations [1], besides that spline is able to handle data characters that are smooth. Spline also has a very good ability to handle data whose behavior changes at certain sub-intervals [8].

Nonparametric regression modeling with spline truncated approach has been done by many researchers. The data used in the modeling also varies, some use cross section data and longitudinal data in applying the method. Researchers who use spline truncated nonparametric regression in their research include [9] modeling the percentage of poor people in Indonesia using spline nonparametric regression approach. [10] estimated the average number of children born alive per woman using spline truncated nonparametric regression. [11] conducted research on the application of confidence intervals for parameters of nonparametric spline truncated regression on index development gender in East Java. In these studies, the data used are cross section data, but in reality in our daily lives we are not only confronted with cross section data but we often encounter longitudinal data. In addition to research on spline truncated nonparametric regression using cross section data, there are also studies of spline truncated nonparametric regression using longitudinal data. Longitudinal data is data obtained from repeated observations of each subject at different time intervals. This data correlates to the same subject and is independent between different subjects [12]. Research with a nonparametric approach to longitudinal data is mostly done in the health sector, but can also be applied to other fields including the social and economic fields [13]. According to [14] research for longitudinal data is usually more complex and requires greater costs than research for cross section data, but is more reliable in finding answers about the dynamics of change. In addition, it has the potential to provide more complete information, depending on the operationalization of the theory and research methodology. According to [13] the advantage of using longitudinal data is to be able to know the changes that occur in individuals, do not require a lot of subjects because of repeated observations and also more efficient estimation because every observation is done. Some researchers who use longitudinal data, among others, [15] do spline truncated regression modeling for longitudinal data of monthly stock prices in the banking stock group. [16] conducted research on parameter interval estimation of semiparametric spline truncated regression model for longitudinal data.

In statistical analysis, besides using descriptive statistics it is also necessary to conduct an analysis in the form of statistical inference, one of which is the estimation of the regression curve. This regression curve estimation is very necessary in a good regression modeling to create a regression model or in testing hypotheses and confidence intervals. Research on estimation of multivariable spline truncated multivariable spline regression curves for longitudinal data has been done, but with the assumption that the variance-covariance matrix is given, so in this study will develop a study of the estimation of multivariable spline truncated multivariable regression curves for longitudinal data and estimation of matrix variance-covariance. This study is very important because it can be used to model cases with longitudinal data structures and more than one predictor variable.
2. Theoretical Review

In this section we will review some of the theories used.

2.1. Nonparametric Regression Spline Truncated

Nonparametric regression is used if the shape of the regression curve is unknown or information about the shape of the past data patterns is incomplete. One method of estimating nonparametric regression is spline. Spline is polynomial pieces that have segmented properties (piecewise polynomial) at the knots point [17]. This spline method is very good in modeling data whose patterns change at certain sub intervals [1]. Spline function $f(x)$ sequence $q$ with knots $K_1, K_2, ..., K_r$ can be written as follows.

$$f(x) = \sum_{h=0}^{q} \alpha_h x^h + \sum_{k=1}^{r} \beta_k (x - K_k)_+^q,$$  (1)

with truncated functions

$$(x - K_k)_+^q = \begin{cases} (x - K_k)^q & , x \geq K_k \ni \vspace{0.1cm} \\
0 & , x < K_k \ni \end{cases}.$$  (2)

In general, the truncated spline model can be presented in the form of equation (2) below.

$$y_i = \sum_{h=0}^{q} \alpha_h x_i^h + \sum_{k=1}^{r} \beta_k (x_i - K_k)_+^q + \epsilon_i.$$  (3)

Next, the model in equation (2) above can be written in the following form.

$$y_i = \alpha_0 + \alpha_1 x_i + ... + \alpha_{q-1} x_i^{q-1} + \alpha_q x_i^q + \beta_1 (x_i - K_1)_+^q + ... + \beta_r (x_i - K_r)_+^q + \epsilon_i.$$  (4)

Equation (3) above can be written in matrix notation as follows.

$$\mathbf{y} = \mathbf{X}(\mathbf{K}) \mathbf{B} + \mathbf{\epsilon}.$$  (4)

where,

$$\mathbf{y} = (y_1 \ y_2 \ \ldots \ y_n)^T,$$

$$\mathbf{X}(\mathbf{K}) = \begin{bmatrix} 1 & x_1 & \ldots & x_1^{q-1} & x_1^q & (x_1 - K_1)_+^q & \ldots & (x_1 - K_r)_+^q \\ 1 & x_2 & \ldots & x_2^{q-1} & x_2^q & (x_2 - K_1)_+^q & \ldots & (x_2 - K_r)_+^q \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^{q-1} & x_n^q & (x_n - K_1)_+^q & \ldots & (x_n - K_r)_+^q \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} \alpha_0 & \alpha_1 & \ldots & \alpha_{q-1} & \alpha_q & \beta_1 & \ldots & \beta_r \end{bmatrix}^T,$$

$$\mathbf{\epsilon} = (\epsilon_1 \ \epsilon_2 \ \ldots \ \epsilon_n)^T.$$  (5)

If $q = 1$ and the number of knots is one, then equation (3) above can be written as a linear spline function as follows.

$$y_i = \alpha_0 + \alpha_1 x_i + \beta_1 (x_i - K_1)_+^1 + \epsilon_i, \quad i = 1,2,\ldots,n.$$  (5)

2.2. Nonparametric Spline Truncated Regression for Longitudinal Data

Equation (5) above is a form of spline truncated nonparametric regression model for cross section data. There are as many as $n$ object of observations which is measured at a certain time. If $n$ the object is observed repeatedly over a certain period of time such as every week, every month, every three months, one year and so on, then the model can be developed into a spline truncated nonparametric regression model on
longitudinal data. In general, spline truncated nonparametric regression models on longitudinal data for one predictor variable can be written on the following form.

\[ y_{ij} = f(x_{ij}) + \epsilon_{ij}, \quad i = 1,2,...,n; \quad j = 1,2,...,t. \]  

In general, the spline truncated model on longitudinal data for one predictor variable can be presented in the form of equation (7) below.

\[ y_{ij} = \sum_{h=0}^{q} \alpha_{ih}x_{ij}^h + \sum_{k=1}^{r} \beta_{ik}(x_{ij} - K_{ki})^q + \epsilon_{ij}, \]  

with truncated functions

\[ (x_{ij} - K_{ki})^q = \begin{cases} (x_{ij} - K_{ki})^q, & x_{ij} \geq K_{(k+q)i}; \\ 0, & x_{ij} < K_{(k+q)i}. \end{cases} \]

Equation (7) is a form of spline truncated longitudinal data regression with one predictor variable. If the predictor variable is used as many as \( p \), then the resulting spline truncated multivariable regression model for longitudinal data is as follows.

\[ y_{ij} = \sum_{l=1}^{p} \left( \sum_{h=0}^{q} \alpha_{hil}x_{ij}^h + \sum_{k=1}^{r} \beta_{kil}(x_{ij} - K_{kli})^q \right) + \epsilon_{ij}. \]  

One of the methods to get the estimated parameters in the regression analysis is the Least Square method. There is a difference in using the Least Square method in cross section data and longitudinal data. On the cross section data the appropriate method used for estimating parameters of longitudinal data is the Weighted Least Square (WLS) method. Equation (8) if written into matrix notation is as follows.

\[ \mathbf{y} = \mathbf{X}(\mathbf{K})\mathbf{B} + \mathbf{\epsilon}. \]  

Based on equation model (9) estimator for parameters \( \mathbf{B} \) obtained by completing WLS optimization as follows.

\[ \min_{\mathbf{B} \in \mathbb{R}^{p(n+1)}} \left\{ \mathbf{y} - \mathbf{X}(\mathbf{K})\mathbf{B}^\top \mathbf{W}^{-1}(\mathbf{y} - \mathbf{X}(\mathbf{K})\mathbf{B}) \right\}, \]  

where, \( \mathbf{y} \) is a response vector with the size \( nt \times 1 \), \( \mathbf{X}(\mathbf{K}) \) is a predictor matrix of polynomial components and truncated components with the size \( nt \times n(1 + p(r + 1)) \) with the parameters vector \( \mathbf{B} \) with the size \( n(1 + p(r + 1)) \times 1 \) and \( \mathbf{\epsilon} \) is an error vector with the size \( nt \times 1 \), while \( \mathbf{W} \) a weighting matrix (variance-covariance matrix) with the size \( nt \times nt \) and contains a diagonal \( (\mathbf{W}_1, \mathbf{W}_2, ..., \mathbf{W}_n) \).

3. Result and Discussion

In this section we will discuss the estimation of parameters, curves and matrix variance-covariance on nonparametric regression spline truncated for longitudinal data.

3.1. Curve estimation of nonparametric regression spline truncated multivariable for longitudinal data

Models of nonparametric regression spline truncated multivariable for longitudinal data are given on equation (11) below.

\[ y_{ij} = \sum_{l=1}^{p} \left( \sum_{h=0}^{q} \alpha_{hil}x_{ij}^h + \sum_{k=1}^{r} \beta_{kil}(x_{ij} - K_{kli})^q \right) + \epsilon_{ij}, \quad i = 1,2,...,n; \quad j = 1,2,...,t, \]
this model loading \( n \) subjects with each subject as much as \( t \) observations with sequence as much as \( q \) and \( r \) is the number of knots. For sequence \( q = 1 \) and the number of knots is one, then the equation (11) above can be written into a linear spline function as follows.

\[
y_{ij} = \alpha_{0j} + \sum_{l=1}^{n} \left( \alpha_{il} x_{ijl} + \beta_{1lj} (x_{ijl} - K_{lj})_+ \right) + \varepsilon_{ij},
\]

(12)

with truncated functions

\[
(x_{ijl} - K_{lj})_+ = \begin{cases} 
(x_{ijl} - K_{lj})^+, & x_{ijl} \geq K_{lj} \\
0, & x_{ijl} < K_{lj}
\end{cases}
\]

where \( i = 1, 2, ..., n \) are subjects that is observed as much as \( n \), \( j = 1, 2, ..., t \) are observations which is done as much as \( t \) observations period, so the equation (12) can be written in the form of a special matrix as follows.

\[
y = X(K)B + \varepsilon, \quad \varepsilon \sim N(0, W),
\]

(13)

where,

\[
y = \begin{bmatrix} y_1 \\
y_2 \\
\vdots \\
y_n \end{bmatrix}, \quad X(K) = \begin{bmatrix} X_1(K) & 0 & \cdots & 0 \\
0 & X_2(K) & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & X_n(K) \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\
B_2 \\
\vdots \\
B_n \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\
\varepsilon_2 \\
\vdots \\
\varepsilon_n \end{bmatrix}.
\]

Next, using the weighting matrix \( W \), estimation of \( B \) on equation (13) can be obtained by completing the following Weighted Least Square (WLS) optimization.

\[
y = X(K)B + \varepsilon \quad \text{where} \quad \varepsilon = y - X(K)B,
\]

\[
\min_{B \in \mathbb{R}^{(r+1)\times l}} \left\{ \varepsilon^TW^{-1}\varepsilon \right\} = \min_{B \in \mathbb{R}^{(r+1)\times l}} \left\{ (y - X(K)B)^TW^{-1}(y - X(K)B) \right\},
\]

(14)

Completion of the optimization of equation (14) above is carried out by the elaboration as follows.

\[
(y - X(K)B)^TW^{-1}(y - X(K)B) = y^TW^{-1}y - 2B^TX(K)^TW^{-1}y + B^TX(K)^TW^{-1}X(K)B
\]

For example,

\[
Q = y^TW^{-1}y - 2B^TX(K)^TW^{-1}y + B^TX(K)^TW^{-1}X(K)B.
\]

(15)

Next, to get the estimation of \( B \) on equation (16) then a partial derivative is performed \( Q \) to \( B \) as follows.

\[
\frac{\partial Q}{\partial B} = -2X(K)^TW^{-1}y + 2X(K)^TW^{-1}X(K)B,
\]

minimum value of \( Q \) obtained by way of \( \frac{\partial Q}{\partial B} = 0 \) as follows.

\[
-2X(K)^TW^{-1}y + 2X(K)^TW^{-1}X(K)\hat{B} = 0,
\]

\[
\hat{B} = \left( X(K)^TW^{-1}X(K) \right)^{-1}X(K)^TW^{-1}y.
\]

(16)

Based on estimates \( \hat{B} \) on equation (16) above, then the estimation of curves on nonparametric regression spline truncated multivariable for longitudinal data with sequence \( q = 1 \) and the number of knots is one is obtained as follows.
\[ \hat{y} = X(K)\hat{B} \]
\[ = X(K)(X(K)^T W^{-1} X(K))^{-1} X(K)^T W^{-1} y \]
\[ = A y. \tag{17} \]
where \( A = X(K)(X(K)^T W^{-1} X(K))^{-1} X(K)^T W^{-1} \).

The estimator of spline truncated on nonparametric regression for longitudinal data is very dependent on the knot point. To get the best spline truncated estimator obtained from the optimal knot point, the optimal knot point can be obtained using the Generalized Cross Validation (GCV) method. The Generalized Cross Validation (GCV) function is given by:

\[ GCV = \frac{MSE}{(n^{-1}(tr[I - A(K)]))^2} \]
\[ = \frac{n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{(n^{-1}(tr[I - A(K)]))^2}. \tag{18} \]
The optimal knot point corresponds to the smallest Generalized Cross Validation (GCV) value.

### 3.2. Estimation of matrix variance-covariance on nonparametric regression spline truncated multivariable for longitudinal data

Research involving nonparametric spline truncated multivariable regression models for longitudinal data, has been extensively developed by researchers. Generally, these researchers assume the variance-covariance matrix of the random error is known, in fact the variance and covariance is unknown. As a result, it is necessary to estimate the variance-covariance matrix of random errors in the spline truncated multivariable nonparametric regression model for longitudinal data. For this purpose, the Maximum Likelihood Estimator (MLE) method will be used. \( \varepsilon_{ij} \) are random errors from the results of the prediction on the subject to \(-i\), observation to \(-j\) assumed to be multivariate normal distribution, with a mean \( E(\varepsilon) = 0 \) and variance-covariance matrix \( Var(\varepsilon) = W \), then the likelihood function is as follows.

\[ L(W) = \prod_{i,j=1}^{n} \frac{1}{(2\pi)^{\frac{n}{2}}|W|^\frac{n}{2}} \exp\left\{ -\frac{1}{2} (y - X(K)\hat{B})^T W^{-1} (y - X(K)\hat{B}) \right\} \]
\[ = \frac{1}{(2\pi)^{\frac{n}{2}}|W|^\frac{n}{2}} \exp\left\{ -\frac{1}{2} tr \left[ W^{-1} \sum_{i=1}^{n} (y_i - X_i(K)\hat{B}) (y_i - X_i(K)\hat{B})^T \right] \right\}. \tag{19} \]

Next, equation (19) made in the form of natural logarithms and obtained:

\[ \ln L(W) = -\frac{n}{2} (2\pi) - \frac{n}{2} |W| - \frac{n}{2} tr \left[ W^{-1} \sum_{i=1}^{n} (y_i - X_i(K)\hat{B}) (y_i - X_i(K)\hat{B})^T \right]. \tag{20} \]
To obtain the variance-covariance matrix \( W \), then the function \( \ln L(W) \) maximized by the way \( \frac{\partial \ln L(W)}{\partial W} = 0 \) as follows.
\[
\frac{\partial \ln L(W)}{\partial W} = -\frac{n}{2} W^{-1} W W^{-1} + \frac{n}{2} W^{-1} \frac{1}{n} \sum_{i=1}^{n} (y_i - X_i (K) \hat{B} ) (y_i - X_i (K) \hat{B})^T W^{-1},
\]
\[
-\frac{n}{2} W^{-1} W W^{-1} + \frac{n}{2} W^{-1} \frac{1}{n} \sum_{i=1}^{n} (y_i - X_i (K) \hat{B} ) (y_i - X_i (K) \hat{B})^T W^{-1} = 0.
\]

Hence, the estimator of matrix variance-covariance \( \hat{W} \) for the nonparametric spline truncated multivariable regression model for longitudinal data is following.

\[
\hat{W} = \frac{1}{n} \sum_{i=1}^{n} (y_i - X_i (K) \hat{B} ) (y_i - X_i (K) \hat{B})^T.
\]

4. Conclusions

Based on the analysis described above, then it can be concluded that:

1) The parameter estimation of the multivariable nonparametric spline truncated regression model for longitudinal data are obtained as follows.

\[
\hat{B} = (X(K)^T W^{-1} X(K))^{-1} X(K)^T W^{-1} y.
\]

2) The curve estimation of nonparametric regression spline truncated multivariable for longitudinal data is obtained as follows.

\[
\hat{y} = X(K) \hat{B} = X(K) \left[ X(K)^T W^{-1} X(K) \right]^{-1} X(K)^T W^{-1} y = A y,
\]

where \( A = X(K)^T W^{-1} X(K) \) and \( X(K)^T W^{-1} y \).

3) The estimation of matrix variance-covariance on nonparametric regression spline truncated multivariable for longitudinal data is obtained as follows.

\[
\hat{W} = \frac{1}{n} \sum_{i=1}^{n} (y_i - X_i (K) \hat{B} ) (y_i - X_i (K) \hat{B})^T.
\]

4) The optimal knot point is obtained from the minimum Generalized Cross Validation (GCV) value with the following formula:

\[
GCV = \frac{n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{(n^{-1} (tr[I - A(K)])^2).}
\]

References

[1] Eubank, R. L. 1999 Nonparametric Regression and Spline Smoothing Second Edition Marcel Dekker New York

[2] Hardle, W. 1990 Applied Nonparametric Regression Cambridge University Press New York

[3] Budiantara, I. N. 2000 Metode U, GML, CV dan GCV Dalam Regresi Nonparametrik Spline Majalah Ilmiah Himpunan Matematika Indonesia (MIHMI) No. 6 pp. 285-290

[4] Wahba, G. 1990 Spline Model for Observational Data SIAM XII Philadelphia
[5] Wisisono, I. N. R., Nurwahidah, A. I., and Andriyana, Y. 2018 Regresi Nonparametrik dengan Pendekatan Deret Fourier pada Data Debit Air Sungai Citarum *Jurnal Matematika*. Vol. 04 No. 02 pp. 75-82

[6] Antomiadis, A., Bigot., and Spatinas, T. 2001 Wavelet Estimator in Nonparametric Regression: A Comparative Simulation Study *Journal of Statistical Software* Vol. 6 pp. 1-83

[7] Laome, L. 2009 Model Regresi Semiparametrik Spline untuk Data Longitudinal pada Kasus Kadar CD4 Penderita HIV *Jurnal Paradigma* Vol. 13 No. 2 pp. 189-194

[8] Budiantara, I. N., Lestari, B., dan Isliamyati, A. 2010 Estimator Spline Terbobot dalam Regresi Nonparametrik dan Semiparametrik Heterokedastisitas untuk Data Longitudinal *Hibah Penelitian Kompetensi* DP2M-Dikti: Jakarta

[9] Budiantara, I. N., Ratna, M., Zain, I., and Wibowo, W. 2012 Modeling The Percentage of Poor People in Indonesia Using Spline Nonparametric Regression Approach *International Journal of Basic & Applied Sciences* IJBAS-IJENS Vol. 12 No. 06 pp. 119-124

[10] Fitriyani, N., Budiantara, I. N., Zain, I., and Ratnasari, V. 2016 Nonparametric Regression Spline in The Estimation of The Average Number of Children Born Alive Per Woman *The 1st International Conference on Science and Technology* Mataram University Press

[11] Setiawan, R. N. S., Budiantara, I. N, and Ratnasari, V. 2017 Application of Confidence Intervals for Parameters of Nonparametrk Spline Truncated Regression on Index Development Gender in East Java *Journal of Science* Vol. 2 No. 3 pp. 49-55

[12] Prahutama, A. 2013 *Model Regresi Nonparametrik Polynomial Lokal Birespon pada Data Longitudinal* Tesis Institut Teknologi Sepuluh Nopember Surabaya

[13] Wu, H. and Zhang, J. T. 2006 *Nonparametric Regression Methods for Longitudinal Data Analysis* A John Wiley and Son Inc. New Jersey

[14] Nurdini, A. 2006 Cross-Sectional Vs Longitudinal : Pilihan Rancangan Waktu dalam Penelitian Perumahan Permukiman *Dimensi Teknik Arsitektur* Vol. 34 pp. 52-58

[15] Fadhilah, K. N. 2016 Pemodelan Regresi Spline Truncated untuk Data Longitudinal (Studi Kasus: Harga Saham Bulanan pada Kelompok Saham Perbakan Periode Januari 2009 – Desember 2015 *Jurnal Gaussian* Vol. 5 pp. 447-454

[16] Prawanti, D. D., Budiantara, I. N., and Purnomo, J. D. T. 2019 Parameter Interval Estimation of Semiparametric Spline Truncated Regression Model for Longitudinal Data *9th Annual Basic Science International Conference* IOP Publishing

[17] Budiantara, I. N. 1999 Estimator Spline Terbobot dalam Regresi Semiparametrik *Majalah Ilmu Pengetahuan dan Teknologi* No. 10 pp. 103-109