International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 6 Issue 2, February - 2022, Pages:77-81

Multilevel Nonparametric Regression Model with Truncated Linear Spline Estimator on Students' National Examination Scores

Hedi Kuswanto[1], Anna Islamiyati[2], Nirwan Ilyas[3]

[1]Graduate Department of Statistics, Hasanuddin University, Makassar 90245, Indonesia
[1]hedikuswanto454@gmail.com
[2,3] Department of Statistics, Hasanuddin University, Makassar 90245, Indonesia
[2]correspondence author: annaislamiyati@unhas.ac.id
[3]nirwanilyas@unhas.ac.id

Abstract: *Multilevel models can solve problems that arise from data with a hierarchical structure. One application of the use of the multilevel model that has been previously studied in the field of education is the value of the National Examination. The sample size of each district or city is different, causing the Maximum Likelihood estimation method to be appropriate. In addition, there is a tendency that the National Examination scores do not follow a parametric pattern so the truncated spline estimator approach is used. This research examines the application of the multilevel linear spline truncated model. The results obtained are when the number of students is below 76 people, the tendency of the average national exam score to increase, and when the number of students has reached 76 people, the average national exam score can decrease.*

Keywords— Linear; Multilevel Model; National Examination; Number of Students; Spline Truncated

1. INTRODUCTION

Spline is one of the estimators in nonparametric regression that can effectively adjust the data pattern [1]. The spline estimator gets a data estimate based on the movement of the data pattern, so the spline is called an estimator which has flexible estimation properties [2]. Spline estimators that have been developed by many researchers include truncated splines which are piecewise polynomial, which are polynomial pieces that have segmented properties in the intervals formed at knot knot points [3].

The development of the regression analysis method has experienced rapid progress, both in the aspect of the estimation method and the variation of the data used. One of the developments of simple linear regression analysis is the multilevel model. The multilevel model was introduced by Goldstein (1995) which states that the multilevel model can overcome problems that arise from data with a hierarchical structure. In a hierarchical structure, individuals in the same group have characteristics that tend to be similar, in other words between observations at lower levels are not independent of each other, if the violation of this assumption is ignored it will result in a violation of the assumption of freedom in conventional statistical approaches [4].

The most basic multilevel model is a two-level model, with individual data on the first level and group data on the second [5]. This study measures up to 2 levels. Nonparametric regression has also been developed in multilevel cases, including linear multilevel spline models [6], multivariate multilevel spline models [7], and kernel multilevel modeling [8]. The Maximum Likelihood (MLE) method is one of several parameters estimates that can be used in a multilevel model [4]. MLE produces an efficient and consistent regression coefficient estimator if the sample used is large so that the assumption violation can be ignored. [5].

Individuals and cities/districts are both tiered or hierarchical structures in general. A tiered population indicates that there are levels or levels in the data. One form of tiered data structure in education data is the average score of the National Examination [9]. The government always tries to make various innovations every year to realize the hope of improving the quality of education. One way to do this is to analyze information for the tiered data structure of the implementation of the National Examination at each level from different districts or cities so that the multilevel model is appropriate. The sample size of each district or city is different, causing the Maximum Likelihood estimation method to be appropriate.

Based on this description, the authors examine the estimation of the multilevel spline truncated regression model using the Maximum Likelihood method. The method was applied to the average data for the SMP UN scores for each school by taking into account the variance between districts/cities in South Sulawesi Province. The choice of data on the average UN score in this study was due to changes in values that varied at the district/city level in South Sulawesi Province. The diversity of data is getting more and more in the real world. The trend of increasingly large data at this time is a challenge for researchers to analyze big data into a decision reference. The problem, which is almost always unavoidable, is that the more data there is, the more likely there is to be an irregular pattern. Fluctuations up and down, data is spread unevenly, outliers are getting bigger, and there

International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 6 Issue 2, February - 2022, Pages:77-81

are still many data problems that will be encountered in real data. The parametric approach can only be used when the form of the function follows a parametric form causing limitations of the method. Therefore, a nonparametric regression approach has been developed that can be used for any data conditions.

2. LITERATUR REVIEW

2.1 Estimator Spline Truncated

Nonparametric regression models can be presented in the following way in general [10].

$$y_i = f(x_i) + \varepsilon_i\,, \qquad (1)$$

where y_i is the response variable on the i^{th} observation, x_i is the predictor variable on the i^{th} observation, $f(x_i)$ is a nonparametric regression function containing predictor variables, ε_i is a disturbance factor that cannot be explained by a model that can be called an error, which is assumed to be a random variable with a mean of zero and a variance of σ^2, and $i = 1,2, \dots, n$.

Spline in nonparametric regression has the ability to estimate the behavior of the data that tends to differ at different intervals [11]. The ability to estimate the behavior of this data is shown by the truncated (pieces) attached to the estimator, these pieces are called knot points. Suppose there is one predictor variable x_i, then the knot point is taken at intervals $a < k_r < b$, where a is the minimum value and b is the maximum value for the predictor x_i [12].

A spline truncated function of order q with knot points at$k_1, k_2, \dots, k_r$, in general, can be expressed as follows [13]:

$$f(x_i) = \sum_{l=0}^{q} \beta_l x_i^l + \sum_{h=1}^{r} \beta_{q+h}(x_i - k_h)_+^q, \qquad (2)$$

where $\beta_0, \beta_1, \dots, \beta_q, \beta_{q+1}, \dots, \beta_{q+r}$ is the regression parameter, k_h is the h^{th} knot point, $(h = 1,2, \dots, r)$. $(x_i - k_h)_+^q$ is a truncated polynomial function which is described as follows:

$$(x_i - k_h)_+^q = \begin{cases} (x_i - k_h)^q; & x_i \geq k_h \\ 0, & ; \ x_i < k_h \end{cases}.$$

If the spline function $f(x_i)$ in equation (2) is a function that expresses the relationship between p predictors and a single response, then it can be written as follows:

$$y_i = f(x_{1i}) + f(x_{2i}) + \cdots + f(x_{pi}) + \varepsilon_i \qquad (3)$$
$$= \sum_{j=1}^{p} f(x_{ji}) + \varepsilon_i; i = 1,2, \dots, n,$$

where

$$f(x_{ji}) = \beta_{0j} + \sum_{l=1}^{q} \beta_{jl} x_{ji}^l + \sum_{h=1}^{r} \beta_{j(q+h)}(x_{ji} - k_{jh})_+^q$$

where y_i is the response variable on the i^{th} observation, x_{ji} is the j^{th} predictor variable on the i^{th} observation, β_{0j} is the j^{th} predictor intercept, β_{jl} is the polynomial parameter on the j^{th} and l^{th} order predictor, k_{jh} is the value knot points on the j^{th} predictor and h^{th} knot points, r is the number of knot points, q is the order of the spline truncated polynomial, p is the number of predictor variables, $\beta_{j(q+h)}$ is the parameter truncated on the j^{th} predictor and knot points $(q+h)^{th}$, and ε_i is the error in the i-th observation which is assumed to be independent and normally distributed with mean 0 and variance σ^2 .

Equation (3) for n observational data can be expressed in the form of a matrix as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

where y is a vector that has size $n \times 1$, matrix X has size $n \times (1 + q + r)$, vector size $(1 + q + r) \times 1$, and vector size $n \times 1$ [14].

2.2 Model Multilevel Spline Linier Truncated

The multilevel regression model is part of the mixed linear model because there are two parameters, namely the fixed effect parameters and random effects which are combined into one equation [15].

In simple form, the equations for the mixed linear model are [16]:

$$y = \boldsymbol{X}\beta + \boldsymbol{Z}u + \varepsilon \qquad (4)$$

where

y : response variable vector
$\boldsymbol{X}$: predictor variable matrix for fixed parameters
β : fixed effect parameter vector
$\boldsymbol{Z}$: predictor variable matrix for random parameters
u : random effect vector
ε : random error vector

Assume that the errors ε and u are not independent of one another and are normally distributed with a variance of:

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{bmatrix}\right) \qquad (5)$$

with $\boldsymbol{G} = \boldsymbol{G}(\gamma)$ and $\boldsymbol{R} = \boldsymbol{\Sigma}$ with $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\phi)$. and are vectors of the variance parameters bound to u and ε while σ^2 is the variance scale parameter. The variance parameters are σ^2, ϕ γ. $\boldsymbol{G}, \boldsymbol{R}$ and $\boldsymbol{\Sigma}$ are variance matrices which are assumed to be positive definite.

$$\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \boldsymbol{H}) \qquad (6)$$

where

$$\boldsymbol{H} = \boldsymbol{ZGZ}' + \boldsymbol{R}$$

The definition of a level 1 model is a model that is compiled without taking into account the influence of the group level. For each group, multilevel modeling can be written as follows [9]:

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \ \text{with } \boldsymbol{\varepsilon}_j \sim N(0, \sigma^2 I) \qquad (7)$$

where

$$\boldsymbol{y}_j = [y_{1j} \quad y_{2j} \quad \cdots \quad y_{n_j j}]',$$
$$\boldsymbol{X}_j = \begin{bmatrix} 1 & x_{11j} & x_{21j} & \dots & x_{k1j} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n_j} & x_{2n_j} & \dots & x_{kn_j} \\ 1 & (x_{11}-k_{11})_+^q & (x_{21}-k_{21})_+^q & \dots & (x_{k1}-k_{k1})_+^q \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & (x_{1n}-k_{1r})_+^q & (x_{2n}-k_{2r})_+^q & \dots & (x_{kn}-k_{kr})_+^q \end{bmatrix},$$
$$\boldsymbol{\beta}_j = [\beta_{0j} \quad \cdots \quad \beta_{kj} \quad \beta_{(k+1)j} \quad \cdots \quad \beta_{(k+r)j}]',$$
$$\boldsymbol{\varepsilon}_j = [\varepsilon_{1j} \quad \varepsilon_{2j} \quad \cdots \quad \varepsilon_{n_j j}]'$$

The regression coefficient at level-1, β_{pj} with $p = 0,1,2,\dots,k$ in the level-1 model has different values between groups. Variations in the value of β_{pj} will be explained by forming a level 2 model. The formation of a level 2 model is carried out for each regression coefficient as the p-th response using predictor variables at level-2. The form of modeling at level-2 can be written as follows [9]:

$$\boldsymbol{\beta}_p = \boldsymbol{Z}\boldsymbol{\gamma}_p + \boldsymbol{u}_p \tag{8}$$

where

$$\boldsymbol{\beta}_p = [\beta_{p1} \quad \beta_{p2} \quad \cdots \quad \beta_{pm}]',$$
$$\boldsymbol{X}_j = \begin{bmatrix} 1 & Z_{11} & Z_{21} & \dots & Z_{l1} \\ 1 & Z_{12} & Z_{22} & \dots & Z_{l2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Z_{1m} & Z_{2m} & \dots & Z_{lm} \end{bmatrix},$$
$$\boldsymbol{\gamma}_p = [\gamma_{0p} \quad \gamma_{1p} \quad \cdots \quad \gamma_{lp}]',$$
$$\boldsymbol{u}_p = [u_{p1} \quad u_{p2} \quad \cdots \quad u_{pm}]'$$

As a matrix, it looks like this:

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{Z}_j\boldsymbol{\gamma} + \boldsymbol{X}_j\boldsymbol{u}_j + \boldsymbol{\varepsilon}_j \tag{9}$$

where $\boldsymbol{X}_j\boldsymbol{Z}_j\boldsymbol{\gamma}$ is a fixed effect and $[\boldsymbol{X}_j\boldsymbol{u}_j + \boldsymbol{\varepsilon}_j]$ is a random effect.

3. MATERIAL AND METHOD

3.1 Data Sources and Research Variables

Data on the level-1 predictor variable or individual level, namely schools, was obtained from the Primary Education Data with a total sample of 1584 schools and the level-2 predictor variable data or the group level, namely 24 districts. The variables used in this study consisted of the average value of the Junior High School National Examination as a response variable. Number of Junior High School Students as a variable in all South Sulawesi provinces in 2019

3.2 Steps Of Analysis

In this study, data analysis was carried out by making a multilevel regression model on the scores of the Junior High School National Examination between schools and districts/cities in South Sulawesi Province by estimating the parameters and u using the Maximum Likelihood method with a truncated spline estimator. Then choose the optimal knot point based on the minimum GCV value

4. RESULT AND DISCUSSION

4.1 Description Data

The data distribution pattern between the predictor variable and the response variable will be identified first before being analyzed using the multilevel linear spline truncated model. A boxplot is used to identify the subjects.

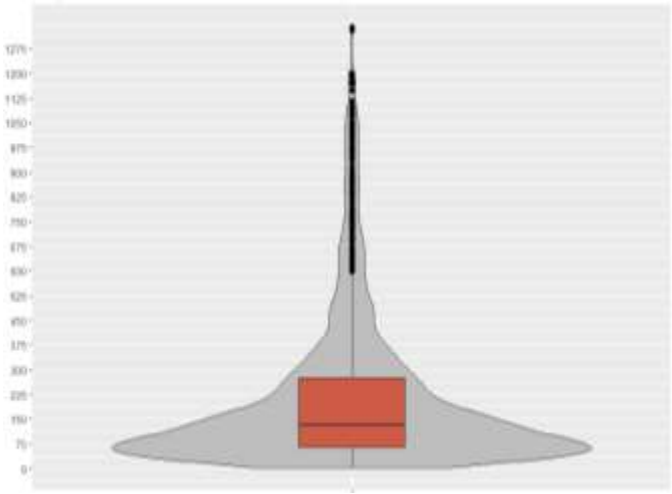

Fig 1: Boxplot number of students

Figure 1 shows that the data on the number of dominant students are spread between Q1 (63) and Q2 (132) so that changes in data patterns are likely to be in the range of Q1 to Q2, so the knot points chosen are in the vicinity of Q1 to Q2.

4.2 Application Of The Truncated Multilevel Spline Model

The estimation of the multilevel spline linear truncated regression model was carried out using a one-knot point approach. The optimal knot point selection is done by looking at the GCV value of each knot point on the predictor of the number of students. The following is a truncated linear spline multilevel regression model which is expressed in logit form with the one-knot point.

$$Y_{ij} = \beta_{11}x_{1ij} + \beta_{12}(x_{1i} - k_{11})_+ + \varepsilon_{ij}$$
$$\beta_{01} = \gamma_{00} + u_{0j}$$

The method that can be used to select the optimal knot point of the multilevel spline linear truncated regression model is the GCV (Generalized Cross Validation) method. The optimal knot point is obtained from the minimum GCV value. The GCV value from modeling using one-knot point is shown in Table 1 below:

Table 1: GCV with one knot point for predictor of Number of Students

Knot point (k_{11})	GCV Value	Knot point (k_{11})	GCV Value
61	75.61805009	77	75.51721187
62	75.61214413	78	75.51994253
63	75.60299917	94	75.67268404
64	75.59023501	95	75.68535235
65	75.57420349	96	75.69762123
66	75.55705725	97	75.70801046
67	75.54841858	98	75.71791092
74	75.5217016	99	75.72739127
75	75.51719392	100	75.73717023
76	**75.51508827**	101	75.74610758

In Table 1, the minimum GCV value is 75.51508827 which is at the knot point k_{11} = 76. The estimation results of the multilevel spline linear truncated regression model parameters are shown in Table 2 below:

Table 2: The results of the estimation of model parameters

Fixed Effect				
Parameter	Estimate	Std. Error	t value	Pr(>|t|)
β_{01}	46.5878	1.2258	38.0075	0.0000
β_{11}	0.0278	0.0132	2.1057	0.0354
β_{12}	-0.0266	0.0136	-1.9623	0.0499
Random Effect				
Group	Level	Variance	Std. Dev	
σ^2_{u0}	Level 2	18.0016	4.24282	
$\sigma^2_{\varepsilon_{ij}}$	Level 1	59.4353	7.70943	

According to Table 2, the multilevel spline linear truncated regression model with a one-knot point for the number of junior high school students is:

Level-1:
$Y_{ij} = 0.0278x_{1ij} - 0.0266(x_{1i} - 76)_+$
Level-2:
$\beta_{20} = 46.5878$
Mixed model:
$Y_{ij} = 46.5878 + 0.0278x_{1ij} - 0.0266(x_{1i} - 76)_+$

Figure 2 shows the graph of the multilevel spline truncated model on the predictor:

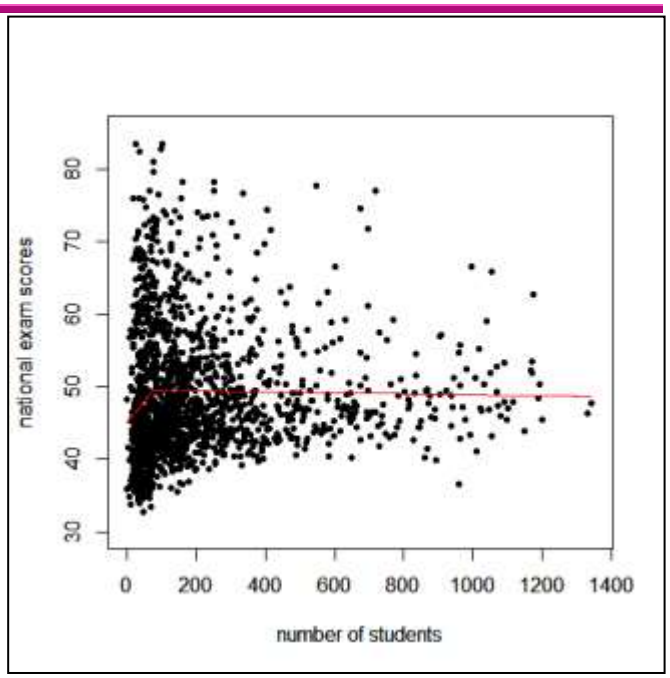

Fig 2: Graph of the multilevel spline truncated model

5. CONCLUSION

The multilevel linear spline truncated model on the data on the number of junior high school students has optimal knots located at 76 people, according to the results obtained, so the model formed is as follows:

$$Y_{ij} = 46.5878 + 0.0278x_{1ij} - 0.0266(x_{1i} - 76)_+$$

Based on the estimation results of the model, it can be concluded that through the multilevel spline truncated model the results are that when the number of students is below 76 people, the tendency of the average national exam score to increase, and when the number of students has reached 76 people, the average national exam score can be decrease.

6. REFERENCES

[1] A. Islamiyati, Raupong and Anisa "Use of Penalized Spline Linear to Identify Change in Pattern of Blood Sugar based on the Weight of Diabetes Patients," *International Journal of Academic and Applied Research (IJAAR), V*ol. 3, Issue 12, December, pp. 75-78, 2019.

[2] D.S. Salam, Anna Islamiyati and Nirwan Ilyas, "Binary Logistic Model in Nonparametric Regression Through Spline Estimator," *International Journal of Academic and Applied Research (IJAAR), V*ol. 5, Issue 10, October, pp. 50-53, 2021.

[3] J. Wang and Y. Lijian, "Polynomial spline confidence bands for regression curves," *Statistica Sinica*, pp. 325-342, 2009.

[4] J. J. Hox, Applied multilevel analysis, TT-publikaties, 1995.

[5] B. T. West and K. B. Welch, Linear Mixed Models, A Practical Guide Using Statistical Software., Chapel Hill: Wiley InterScience, 2007.

[6] L. D. Howe, K. Tilling, A. Matijasevich, E. S. Petherick, A. C. Santos, L. Fairley, J. W. J., I. Santos, A. Barros, R. Martin and M. Kramer, "Linear spline multilevel models for summarising childhood growth trajectories: a guide to their application using examples from five birth cohorts," *Statistical methods in medical research*, vol. 25, no. 5, pp. 1854-1874, 2016.

[7] C. Macdonald-Wallis, D. A. Lawlor, T. Palmer and K. Tilling, "Multivariate multilevel spline models for parallel growth processes: application to weight and mean arterial pressure in pregnancy," *Statistics in medicine*, vol. 31, no. 26, pp. 3147-3164, 2012.

[8] L. He, C.-T. Lu, H. Ding, S. Wang, L. Shen, P. S. Yu and A. B. Ragin, "Multi-way multi-level kernel modeling for neuroimaging classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[9] P. Zulvia, "Pemodelan Multilevel dan Analisis Data Panel pada Penelitian Pendidikan," *IPB e-Journal*, pp. 1-29, 2017.

[10] R. L. Eubank, Nonparametric regression and spline smoothing, Second Edition ed., CRC press, 1999.

[11] R. L. Eubank, Spline Smoothing and Nonparametric Regression, New York: Marcel Dekker Inc, 1988.

[12] W. Härdle, Applied nonparametric regression, Cambridge university press, 1990.

[13] G. Rodrıguez, Smoothing and non-parametric regression, Working paper, 2001.

[14] I. N. Budiantara, "Model Spline dengan Knot Optimal," *Jurnal Ilmu Dasar FMIPA Universitas Jember*, vol. 6, pp. 77-85, 2006.

[15] H. Goldstein, Multilevel Statistical Models, London: Arnold Publisher, 1995.

[16] A. Rencher and G. Schaalje, Linear Models in Statistics, 2nd Edition ed., New Jersey: John Willey and Sons Inc, 2007.