Bi-Response Semiparametric Regression Model based on Spline Truncated for Estimating Computer based National Exam in West Nusa Tenggara

Lilik Hidayati[1], I Nyoman Budiantara[2] and Nur Chamidah[3]
[1]*Doctoral Student Majoring in Mathematics and Natural Sciences, Airlangga University, Surabaya, Indonesia*
[2]*Department of Statistics, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia*
[3]*Department of Mathematics, Faculty of Sciences and Technology, Airlangga University, Surabaya, Indonesia*
Kampus C Universitas Airlangga, Mulyorejo, Surabaya 60115, Indonesia

Keywords: Bi-Response Semiparametric Model, Spline Truncated, Computer-Based National Exam.

Abstract: Bi-response semiparametric regression model is a regression model consisting of two components, parametric and nonparametric with two response variables. The propose of this research, we estimate the parameters of bi-response semiparametric regression model based on spline truncated by using the weighted least square method. Then model it using bi-respon semiparametric regression model based on a spline truncated estimator. The joint point of combination of the truncated or the point that indicates the occurrence of changes in curve behavior at these intervals are called knots. The best model is determined by the optimal knot point, the method used to select the optimal knot point is to use the generalized cross validation method. The model is applied to the computer based national examination values of West Nusa Tenggara province. Based on the result of the estimation model, we get knot optimal brapathe determination coefficient (R^2) tends to one (i.e., 90%) and MSE tend to zero that it satisfies goodness of fit criterions.

1 INTRODUCTION

Regression analysis is an analysis to know the pattern of functional relationship between response variable (y) with predictor variable (x). If the regression curve is assumed to follow a certain pattern called parametric regression, while the regression curve is assumed not to follow a certain pattern called nonparametric regression. Semiparametric regression is a combination of parametric regression and nonparametric regression (Wahba, 1990). The development of research conducted by researchers who focused their research in semiparametric regression include using a penalized spline estimator (Bandyopadhyay and Maity, 2011; Tong et al., 2012; Yang and Yang, 2016); use smoothing spline estimators (Kim, 2013; Chen and Song, 2013); use the Truncated Spline estimator (Loklominet al., 2017; Pratiwiet al., 2017). Furthermore, the bi-response semiparametric regression model that has been conducted is using linear local estimators (Chamidah and Rifada, 2016); use a penalized spline estimator (Chamidah and Eridani, 2015). The development of research that has been carried out by previous researchers who focused their research in a semiparametric regression model that uses a truncated spline estimator is only limited to uni respon. So, the novelty of this research is the development of theory in estimating parameters and design a program algorithm for semiparametric Bi-response regression models based on a truncated spline estimator that is implemented on the data of the Computer Based National Examination (CBNE) Vocational High School (VHS) data in West Nusa Tenggara Province.

The development of industrial resources which gives more attention to vocational education, namely competency-based education (PP RI No. 41, 2015). Competency-based education is defined as VHS, so schools in VHS can be a solution for students to get an expertise after graduation. Vocational High School (VHS) graduation standards have been determined by the government in collaboration with the Department of Education and Culture (DEC) in each region (Permen RI No.3, 2017). The successful implementation of CBNE properly and smoothly must be supported by all parties, between DEC and

Hidayati, L., Budiantara, I. and Chamidah, N.
Bi-Response Semiparametric Regression Model based on Spline Truncated for Estimating Computer based National Exam in West Nusa Tenggara.
DOI: 10.5220/0008521903570361
In *Proceedings of the International Conference on Mathematics and Islam (ICMIs 2018)*, pages 357-361
ISBN: 978-989-758-407-7

schools as policy makers, teachers and students as implementers. Mahmud said that variables that affect student achievement in this case CBNE, consists of two factors, namely internal and external, among others, gender, accreditation value, distance traveled, parental education, report card grades, school examination scores (Mahmud, 1989). Internal and external factors are called predictor variables (x) while the response variable (y) is the value of CBNE in Mathematics subjects and competency skill. Based on the scatter plot data from these variables, some form certain patterns as parametric components and some do not form a particular pattern as a nonparametric component, so it is suitable to use a truncated spline estimator that can handle data patterns that experience behavior changes in certain sub-sub intervals.

The average value of CBNE VHS in West Nusa Tenggara Province nationally from year to year occupies the lowest position compared to other provinces in Indonesia (PuspendikBalitbangKemendikbud, 2017). Based on these facts, the research on the value of CBNE in 2017 is suitable using Bi-Response Semiparametric regression model based on spline truncated estimator

2 METHODS

Sources of data used in this study is data CBNE VHS on the Department of Computer Network Engineering (CNE) in West Nusa Tenggara Province in 2017. The response variable in this study is the data of CBNE values in the competency skill (y_1) and Mathematics subjects (y_2). Predictor variables in this study were the grades of math (x_1) and competency skill, gender (x_2), accreditation status of the department (x_3), the distance of students from home to school (t_1), the duration of parent education (t_2), and the joint school examination on the subject of competency skill (t_3) and Mathematics.

2.1 Estimating the Parameters of Semiparametric Birespon Regression Model based on Spline Truncated Estimator

1. Assume paired data is $(y^{(r)}, x_p, t_m)$ that meets semi parametric bi-response regression model based on spline truncated estimator as follows:

$$y_i^{(r)} = \beta_{0i}^{(r)} + \beta_{1i}^{(r)} x_{1i}^{(r)} + \cdots + \beta_{pi}^{(r)} x_{pi}^{(r)} + \sum_{h=1}^{m} f^{(r)}(t_{hi}) + \varepsilon_i^{(r)} \quad (1)$$

With assume $\varepsilon_i \sim IIDN(0, \sigma^2)$

Where $i=1,2,...,n$; $r=1,2$; $j=1,2,...,p$ and $h=1,2,...,m$

2. Approached the regression curve $y_i^{(r)}$ by using semiparametric bi-response regression model based on linear spline-truncated estimator with K knot

$$y_i^{(r)} = \sum_{j=0}^{p} \underset{\sim}{x}_{ji}^{(r)} \beta_{ji}^{(r)} + \sum_{h=1}^{m} \left[\sum_{c=0}^{d} \gamma_{ci}^{(r)} + \gamma_{ci}^{(r)} t_{hi}^{(r)} + \sum_{k=1}^{K} \gamma_{c(1+k)}^{(r)} (t_{hi}^{(r)} - K_k^{(r)})_+^1 \right] + \varepsilon_i^{(r)} \quad (2)$$

where $(t_i - K_k)_+^d = \begin{cases} (t_i - K_k)^1 & , t_i \geq K_k \\ 0 & , t_i < K_k \end{cases}$

The point of knots $(K_1, K_2, ..., K_K)$ are a point that shows the pattern of changes in behavior of functions at a certain sub-interval.

3. Equation (2) can be written in matrix

$$\underset{\sim}{y} = \boldsymbol{X}\underset{\sim}{\beta} + \boldsymbol{T}\underset{\sim}{\gamma} + \underset{\sim}{\varepsilon} \quad (3)$$

Where $\underset{\sim}{\varepsilon} \sim N(\underset{\sim}{0}, \Sigma)$ parametric component parameters $(\boldsymbol{X}\underset{\sim}{\beta})$ in the bi-respon semiparametric regression model for the parametric component are*n x 1= (n+1) x (1 + p)) + (1+p) x n+1*.The parameter $(\boldsymbol{T}\underset{\sim}{\gamma})$ of the nonparametric component in the bi-response semiparametric regression model is *n x 1= (n+1) x (1 + m)) + (1+p) x n+1*.

4. Form a new matrix notation of multi respon semiparametric regression model for statistical inference purposes. For example, $C = [\mathbf{X} \quad \mathbf{T}]$ $\underset{\sim}{\theta} = \left[\underset{\sim}{\beta} \quad \underset{\sim}{\gamma}\right]'$ so the new regression model can be written in another form that is

$$\underset{\sim}{y} = C\underset{\sim}{\theta} + \underset{\sim}{\varepsilon} \quad (4)$$

5. Obtain estimates for parameters $\underset{\sim}{\theta}$ using the Weighted Least Square (WLS) method with the following steps
   a) Forming function Q
   $$Q(\underset{\sim}{\theta}) = (\underset{\sim}{y} - C\underset{\sim}{\theta})'W^{-1}(\underset{\sim}{y} - C\underset{\sim}{\theta})$$
   b) Minimize the Q equation by solving the following equation

$$\frac{\partial L(\underset{\sim}{\theta})}{\underset{\sim}{\theta}} = 0$$

c) Get estimates from $\underset{\sim}{\theta}$ i.e $\hat{\theta}$

2.2 Create Algorithms and Programs of Semiparametric Bi-Response Regression Model Parameters based on Spline Truncated Estimators

1. Test the correlation between response variables
2. Determine optimal knot order and point based on minimum Generalized Cross Validation (GCV) criterion

$$GCV(K_1,...,K_K) = \frac{MSE(K_1,...,K_K)}{\left(n^{-1}tr\left[I - A(K_1,...,K_K)\right]\right)^2} \quad (5)$$

where, $MSE(K_1,...,K_K) = n^{-1}\sum_{i=1}^{n}\left(y_i - \hat{f}(t_i)\right)^2$

matriks $A(K_1,...,K_K)$ obtained from the equation:

$$\underset{\sim}{\hat{f}} = A(K_1,...,K_K)\,\underset{\sim}{y} \quad (6)$$

3. Determining the weighted matrix W
4. Estimate the function in equation (1)

3 RESULT AND DISCUSSION

Given data $(y^{(r)}, x_p, t_m)$ with the response variable $y^{(1)}, y^{(2)}$ are called Bi-response. Predictor variables for parametric components $x_1, x_2, ..., x_p$ while predictor variables for nonparametric components $t_1, t_2, ..., t_m$ .So that the spline semiparametric regression model is cut Bi-response which contains both components as stated in equation (1) can be written in matrix notation as follows:

$$\begin{bmatrix} \underset{\sim}{y}^{(1)} \\ \underset{\sim}{y}^{(2)} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{x}^{(1)} & 0 \\ 0 & \underset{\sim}{x}^{(2)} \end{bmatrix} \begin{bmatrix} \beta^{(1)} \\ \beta^{(2)} \end{bmatrix} + \begin{bmatrix} f(t)^{(1)} \\ f(t)^{(2)} \end{bmatrix} + \begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \end{bmatrix}$$

where

$$\underset{\sim}{y}^{(1)} = \begin{bmatrix} y_1^{(1)} & y_2^{(1)} & \cdots & y_n^{(1)} \end{bmatrix}^T,$$

$$\underset{\sim}{y}^{(2)} = \begin{bmatrix} y_1^{(2)} & y_2^{(2)} & \cdots & y_n^{(2)} \end{bmatrix}^T$$

$$\underset{\sim}{x}^{(1)} = \begin{bmatrix} 1 & x_{11}^{(1)} & x_{21}^{(1)} & \cdots & x_{p1}^{(1)} \\ 1 & x_{12}^{(1)} & x_{22}^{(1)} & \cdots & x_{p2}^{(1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n}^{(1)} & x_{2n}^{(1)} & \cdots & x_{pn}^{(1)} \end{bmatrix},$$

$$\underset{\sim}{x}^{(2)} = \begin{bmatrix} 1 & x_{11}^{(2)} & x_{21}^{(2)} & \cdots & x_{p1}^{(2)} \\ 1 & x_{11}^{(2)} & x_{22}^{(2)} & \cdots & x_{p2}^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n}^{(2)} & x_{2n}^{(2)} & \cdots & x_{pn}^{(2)} \end{bmatrix}$$

$$\beta^{(1)} = \begin{bmatrix} \beta_{01}^{(1)} & \beta_{12}^{(1)} & \cdots & \beta_{pn}^{(1)} \end{bmatrix}^T,$$

$$\beta^{(2)} = \begin{bmatrix} \beta_{01}^{(2)} & \beta_{12}^{(2)} & \cdots & \beta_{pn}^{(2)} \end{bmatrix}^T$$

$$f^{(1)}(t_{11}) = \gamma_{01}^{(1)} + \gamma_{11}^{(1)}t_{11}^{(1)} + \gamma_{21}^{(1)}\left(t_{11}^{(1)}\right)^2 + \cdots + \gamma_{d1}^{(1)}\left(t_{11}^{(1)}\right)^d + \gamma_{(d+1)1}^{(1)}(t_{11}^{(1)} - K_1^{(1)})_+^d + \gamma_{(d+2)1}^{(1)}(t_{11}^{(1)} - K_2^{(1)})_+^d + \cdots + \gamma_{(d+m)1}^{(1)}(t_{11}^{(1)} - K_k^{(1)})_+^d$$

$$f^{(2)}(t_{11}) = \gamma_{01}^{(2)} + \gamma_{11}^{(2)}t_{11}^{(2)} + \gamma_{21}^{(2)}\left(t_{11}^{(2)}\right)^2 + \cdots + \gamma_{d1}^{(2)}\left(t_{11}^{(2)}\right)^d + \gamma_{(d+1)1}^{(2)}(t_{11}^{(2)} - K_1^{(2)})_+^d + \gamma_{(d+2)1}^{(2)}(t_{11}^{(2)} - K_2^{(2)})_+^d + \cdots + \gamma_{(d+m)1}^{(2)}(t_{11}^{(2)} - K_k^{(2)})_+^d$$

The random vector error for each equation is:

$$\varepsilon^{(1)} = \begin{bmatrix} \varepsilon_1^{(1)} & \varepsilon_2^{(1)} & \cdots & \varepsilon_n^{(1)} \end{bmatrix}^T$$

$$\varepsilon^{(2)} = \begin{bmatrix} \varepsilon_1^{(2)} & \varepsilon_2^{(2)} & \cdots & \varepsilon_n^{(2)} \end{bmatrix}^T$$

Semiparametric bi-response truncated bi-response regression models can be formed like equations (4). The parameter estimation uses the Weighted Least Square (WLS) optimization method, so that by minimizing the goodness of fit of the semiparametric bi-response regression model in equation (4) is obtained:

$$\underset{\theta}{Min}\left[Q(\theta)\right] = \underset{\theta}{Min}\left[(y - C\theta)^T W^{-1}(y - C\theta)\right]$$

$$Q(\theta) = (y - C\theta)^T W^{-1}(y - C\theta)$$

The estimator of the parameter θ is obtained by decreasing each of the parameters θ so that it is obtained:

$$\frac{\partial Q(\theta)}{\partial \theta} = 0 \quad (7)$$

$$\hat{\theta} = (C^T W^{-1} C)^{-1} y^T W^{-1} C$$

Furthermore, the equation of the semiparametric biresponse regression model in equation (4) is used for the purposes of statistical inference based on the

truncated spline estimator. The data used for the implementation of the estimation of the semiparametric spline truncated bi-respon regression model is the CBNE Value in the Department of Computer Engineering Network of VHS in West Nusa Tenggara Province in 2017.The correlation test between the two response variables was carried out, namely the score of Mathematics CBNE with the score of competency skill CBNE, using the following hypothesis: Zero hypothesis (H0) that is if both variables do not have a linear relationship ($\rho = 0$) ; the alternative hypothesis (H_1) is if both have a linear relationship ($\rho \neq 0$). Based on the results of the correlation test obtained p-value <0.05 then reject H_0 so that it can be concluded that there is a correlation between responses.

The next step is modeling the value CBNE of VHS in the Province of West Nusa Tenggara in 2017 using a semiparametric Bi-response spline truncated regression model at each knot point. The knot point is a joint fusion point where data behavior changes. Optimal knot points are obtained from the minimum GCV value. Based on the analysis carried out, the best model is the bi-response semiparametric regression model based on the truncated spline estimator resulting in the minimum GCV of 0.00000691 with three knots. After obtaining the minimum GCV score, the next step calculates the estimate for the semiparametric bi-response spline truncated regression model with three knot points as follows:

$$y_1 = -0,24 + 3,66D_1 + 5,22D_2 + 1,74D_3 + 0,47x_1$$
$$+26,5t_1 - 7,89t_1^2 + 11,19(t_1 - 2)_+^1 - 3,88(t_1 - 3)_+^1$$
$$+0,64(t_1 - 5)_+^1 - 53,01t_2 + 6,26t_2^2 - 16,04(t_2 - 5)_+^1$$
$$+15,19(t_2 - 6)_+^1 - 5,43(t_2 - 7)_+^1 - 6,01t_3 + 0,11t_3^2$$
$$-15,57(t_3 - 66)_+^1 + 27,37(t_3 - 67)_+^1 - 12,07$$
$$(t_3 - 68)_+^1$$

$$y_2 = 136,37 + 2,14D_1 + 3,87D_2 + 3,08D_3 + 0,26x_1$$
$$-11,09t_1 + 3,51t_1^2 - 6,10(t_1 - 2)_+^1 + 3,5(t_1 - 3)_+^1$$
$$-1,42(t_1 - 5)_+^1 - 36,55t_2 + 4,49t_2^2 - 13,31(t_2 - 5)_+^1$$
$$+14,28(t_2 - 6)_+^1 - 5,59(t_2 - 7)_+^1 - 0,14t_3 + 1,08$$
$$(t_3 - 83)_+^1 - 0,84(t_3 - 84)_+^1 - 0,46(t_3 - 87)_+^1$$

Furthermore, for the criteria of goodness, the semiparametric bi-response regression model of truncated spline obtained MSE value of 48.92 with R^2 of 0.90.Based on the two best models, it can be interpreted as follows: 1) every increase in one unit of math report card and report card competency skill, it will result in an increase in CBNE scores in each of these subjects. 2) The value of CBNE in each subject is based on the gender of the students, so for male students it is more than female students. 3) Based on the value of school accreditation, the school exam scores on the competency skill subjects are increased based on the value of school accreditation, meaning that students of VHS with an accreditation at the school examination scores on the subjects of competence are higher than B accreditation; then B accreditation is higher than C accreditation. 4) then based on the school distance variable fluctuated at knots 66, 67, and 68 for the Computer-Based National Exams scores namely Mathematics CBNE and Competency Skill CBNE. 5) Likewise, the education variables of parents experienced fluctuations in knots 5, 6, and 7 for both UNBK scores namely Mathematics CBNE and Competency Skill CBNE. 6) Furthermore, for the mathematics school examination variable values fluctuated at 66, 67, and68 knots for both CBNE values namely Mathematics and Competency skill. 7) Likewise, the variable scores on school competency skills scores also fluctuated at knots 83, 84, and 87 for both CBNE scores namely Mathematics and Competency skill.

4 CONCLUSIONS

Based on the two best models, each increase in one unit of mathematical report value and competency skills, resulting in an increase in each CBNE value. Based on gender, male students are higher in value than female students. Whereas based on the school accreditation value, the accreditation value of A is higher than B and C. In school distance variables, parent education, and school exam scores on mathematics subjects and competency skill fluctuate on certain knots in each CBNE value. The result of the estimation Bi-response semiparametric regression model, we get the R^2 tends to one (i.e., 90%) and MSE tend to zero that it satisfies goodness of fit criterions.

ACKNOWLEDGEMENTS

Acknowledgment to the Excellence Scholarship of the Bureau of Planning and Foreign Cooperation of the Secretariat General of the Ministry of Education and Culture which has supported the funding of tuition fees for the doctoral study.

REFERENCES

Bandyopadhyay, S and Maity, A, Analysis of Sabine river flow data using semiparametric spline modeling, in Journal of Hydrology, vol.399, 2011, pp.274–280.

Chamidah, N and Rifada, M., 2016, Local Linier Estimator in Bi-Reaponse Semiparametric Regression Model for Estimating Median Growth Charts of Children, *Far East Journal of Mathematical Sciences (FJMS)* Volume 99, Number 8, pp.1233-1244.

Chamidah, N and Eridani, 2015. Designing of Growth Reference Chart by Using Birespon Semiparametric Regression Approach Based on P-Spline Estimator. *International Journal of Applied Mathematics and Statistics*, Int. J. Appl. Math. Stat, Vol.53, Issue No. 3.

Chen, M and Song, Q., 2016. Semiparametric estimation and forecasting for exogenous log-GARCHmodels. *Journal of TEST*, vol, 25, pp.93–112.

Hidayati, L and Budiantara, I, N., 2012. Multivariable Cubic Spline Regression in Score Modeling National Exam. *Proceeding2nd Basic Science International Conference*.s27-s30.

Kim, Y, J., 2013. A partial spline approach for semiparametric estimation of varying-coefficient partially linear models. *Journal of Computational Statistics and Data Analysis*, vol.62, pp.181-187.

Loklomin, S, B., Budiantara, I, N and Zain, I., 2017. Factor that influence the Human Development Index in Moluccas island using Interval Convidence approach for Parameters of Spline Truncated Semiparametric Regression Model. *Proceeding 3rd International Seminar on Science and Technology (ISST*.

Mahmud, D., 1989. *PsikologiPendidikan. Jakarta. Depdikbud*Dirjen.

PeraturanPemerintah RI No. 41, 2015. *Pembangunan SumberDayaIndustri*.

Permen RI No.3, 2017. *Penilaian Hasil Belajar oleh Pemerintah dan Penilaian Hasil Belajar oleh Satuan Pendidikan*.

Pratiwi, D. A., Budiantara, I N. and Wibowo, W., 2017, Pendekatan Regresi Semiparametrik Spline untuk Memodelkan Rata-rata Umur Kawin Pertama (UKP) di Provinsi Jawa Timur, *Jurnal Sains dan Seni ITS* : Vol. 6, No.1, hal.129-136.

Puspendik Balitbang Kemendikbud, 2017. *Panduan Pemanfaatan Hasil Ujian Nasional Tahun Pelajaran 2016/2017*. BSNP Jakarta.

Tong, T., Wu, and He, X., 2012. Coordinate ascent for penalized semiparametric regression on high-dimensional panel count data. *Journal of Computational Statistics and Data Analysis*, vol.56, pp.23-33.

Wahba, G., 1990. *Spline Model for Observational Data*, Society for Industrial and Applied Mathematics. Philadelphia.

Yang, J. and Yang, H., 2016. A robust penalized estimation for identification in semiparametric additive models. *Statistics and Probability Letters*, vol.110, pp. 268-277.