IOP Conference Series: Materials Science and Engineering

Spline Truncated Estimator in Multiresponse Semiparametric Regression Model for Computer based National Exam in West Nusa Tenggara

To cite this article: Lilik Hidayati *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **546** 052029

Spline Truncated Estimator in Multiresponse Semiparametric Regression Model for Computer based National Exam in West Nusa Tenggara

Lilik Hidayati[1], Nur Chamidah[2*] and I Nyoman Budiantara[3]

[1]PhD Student, Faculty of Sciences and Technology, Airlangga University, Indonesia
[2]Department of Mathematics, Faculty of Sciences and Technology, Airlangga University, Indonesia
[3]Department of Statistics, Faculty of Math and Science, Sepuluh Nopember Institute of Technology, Indonesia

*Corresponding Author: nur-c@fst.unair.ac.id

Abstract. Multiresponse semiparametric regression model is a combination of parametric regression model and nonparametric regression model with response variables more than one and correlate.The estimate used in estimate the parameters is spline truncated. Excess spline truncated is a model that has excellent statistical and visual interpretation and can model data with changing patterns on certain sub-intervals, because spline is a kind of polynomial pieces. The data used in this study is the value of Computer Based National Examination (CBNE) Vocational High School (VHS) in the province of West Nusa Tenggara (NTB) in 2017, each subject tested on CBNE serve as response variables. Based on the significant correlation test results obtained p-value <0.05 so it can be concluded that there is correlation between the responses.The result of the multiresponse semiparametric regression model estimation is obtained by the best model with the value of MSE of 49,608; R^2 of 0.84 and minimum GCV value of 0.00000323 so it can be concluded that the value of CBNE VHS in NTB province satisfies goodness of fit criterions.

Keywords: Multiresponse semiparametric model, spline truncated, computer-based national exam, West Nusa Tenggara

1. Introduction

Regression analysis is an analysis to investigate the pattern of functional relationships between response variables and predictor variables. There are three types of regression developed by the researchers namely parametric regression, nonparametric regression, and semiparametric regression. Parametric regression is a regression analysis in which the regression curve pattern is assumed to be known as linear, quadratic, cubic and others. Nonparametric regression is a regression analysis in which the regression curve pattern is assumed to be unknown. Whereas if some of the regression curve patterns are assumed to be known and some are assumed to be unknown, the regression analysis used is semiparametric regression. One estimator that is often used in the last decade is the truncated spline. The spline is a model that can handle data or smooth functions. Truncated functions have advantages in overcoming data patterns whose behavior changes at certain sub-intervals [1].

Research on unirespon semiparametric regression models includes [2-4] using penalized spline estimators [5-6] use smoothing spline estimators [7-8]; Research semiparametric bi-response regression model [9]; use linear local estimator [10] use penalized spline estimators; [11] the paper convergence submitted using the Truncated Spline estimator. Whereas for the research of multiresponse semiparametric regression model using penalized spline estimators [12].

In modeling, many cases that cannot be solved only with one response regression analysis because there is a correlation between the response variables, so that this model can only be solved using a multiresponse regression model. The multiresponse regression model is a regression model with more than one response variable that correlates with one or more predictor variables. The multiresponse semiparametric regression model in this study is applied to the field of education because each student must experience it every year, especially those related to the value of the mid-level national exam. This national exam data is very important in Indonesia because it is a mandate of Law Number 20 of 2003 concerning the national education system which aims to measure the achievement of graduate competencies on subjects nationally by referring to Graduates' Competency Standards (GCS). Therefore, all students are required to take a national examination to measure the achievement of competency of student graduates nationally. National exam scores are very important for education units and local governments in making policy-making and mapping the achievement of student standards, and are used to develop coaching programs for education units in the region. Data on the National Examination of Vocational Schools in NTB from year to year is always the most prevalent compared to other provinces in Indonesia [13].

Based on the description, it is important that a parameter estimate is then used to model the value of a computer-based national exam so that it will be very beneficial for the government to determine improvement efforts and guidance programs in the education unit.

2. Methods

The data used in this study are secondary data about the value of computer-based National Examination (CBNE) in 2017 in West Nusa Tenggara Province, the steps taken in this study are:

a. Create a multiresponse semiparametric regression model using a truncated spline estimator.

$$\begin{aligned}
y_i^{(1)} &= \beta_{0i}^{(1)} + \beta_{1i}^{(1)} x_{1i}^{(1)} + \cdots + \beta_{pi}^{(1)} x_{pi}^{(1)} + \sum_{h=1}^{m} f_1(t_{hi}) + \varepsilon_i^{(1)} \\
y_i^{(2)} &= \beta_{0i}^{(2)} + \beta_{1i}^{(2)} x_{1i}^{(2)} + \cdots + \beta_{pi}^{(2)} x_{pi}^{(2)} + \sum_{h=1}^{m} f_1(t_{hi}) + \varepsilon_i^{(2)} \\
&\vdots \\
y_i^{(r)} &= \beta_{0i}^{(r)} + \beta_{1i}^{(r)} x_{1i}^{(r)} + \cdots + \beta_{pi}^{(r)} x_{pi}^{(r)} + \sum_{h=1}^{m} f_r(t_{hi}) + \varepsilon_i^{(r)}
\end{aligned} \tag{1}$$

With i=1,2,3, ...,n the number of observations each response variable r =1,2,3, ...,R and j= 1,2,3,..., p the number of predictor variables that is a parametric component and h = 1,2,3, ..., m the number of predictor variables which are nonparametric components.

b. Presents a multiresponse semiparametric regression model based on the spline truncated with γ knots as follows:

$$y_i^{(r)} = \sum_{j=0}^{p} \underset{\sim}{x}_{ji}^{(r)\prime} \beta_j^{(r)} + \sum_{h=1}^{m} \left[\sum_{c=0}^{d} \alpha_{ci}^{(r)} + \alpha_{ci}^{(r)} t_i^{(r)} + \sum_{k=1}^{K} \alpha_{c(1+k)}^{(r)} (t_{ci}^{(r)} - \gamma_K^{(r)})_+^1 \right] + \varepsilon_i^{(r)} \tag{2}$$

c. Presents a multiresponse semiparametric regression model based on the spline truncated in the form of a matrix:

$$\underset{\sim}{y} = \underset{\sim}{x}'\underset{\sim}{\beta} + T\underset{\sim}{\alpha} + \underset{\sim}{\varepsilon} \qquad (3)$$

d. Forms a new matrix notation from the multiresponse semiparametric regression model. For example, $C = \begin{bmatrix} X & Z \end{bmatrix}$ and $\underset{\sim}{\eta} = \begin{bmatrix} \underset{\sim}{\beta}' \, \underset{\sim}{\alpha}' \end{bmatrix}'$ the new regression model can be written in another form, namely $\underset{\sim}{y} = C\underset{\sim}{\eta} + \underset{\sim}{\varepsilon}$ (4)

e. Get estimates for parameters using the Weighted Least Square (WLS) method therefor we obtained: $\hat{\eta} = (C^T W^{-1} C)^{-1} y^T W^{-1} C$ (5)

f. Create program algorithms and scripts for multiresponse semiparametric regression models based on the truncated spline estimator.

g. Applying the program that has been made in step g on the data on the CBNE Value of Vocational Schools in West Nusa Tenggara Province.

h. Choose the optimal knot point using the GCV method defined as :

$$GCV(\gamma_1,\gamma_2,\gamma_3,\ldots,\gamma_K) = \frac{MSE(\gamma_1,\gamma_2,\gamma_3,\ldots,\gamma_K)}{\left(n^{-1}tr\left[I - A(\gamma_1,\gamma_2,\gamma_3,\ldots,\gamma_K)\right]\right)^2} \qquad (6)$$

Where $MSE(\gamma_1,\gamma_2,\gamma_3,\ldots,\gamma_K) = n^{-1}\sum_{i=1}^{n}\left(y_i - \hat{f}(t_i)\right)^2$

i. Calculate MSE and R2 as the criterions for the goodness of the model.

3. Result and discussion

Given data $(y^{(1)}, y^{(2)},\ldots, y^{(R)}; x_1, x_2,\ldots x_p; t_1, t_2,\ldots, t_m)$ as the semiparametric spline truncated multiresponse regression model that contains these variables is stated in equation (1) which can then be written in the matrix notation as follows:

$$\begin{bmatrix} \underset{\sim}{y}^{(1)} \\ \underset{\sim}{y}^{(2)} \\ \vdots \\ \underset{\sim}{y}^{(R)} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{x}_{1i} & 0 & 0 & 0 \\ 0 & \underset{\sim}{x}_{2i} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \underset{\sim}{x}_{pi} \end{bmatrix} \begin{bmatrix} \underset{\sim}{\beta}_1^{(1)} \\ \underset{\sim}{\beta}_2^{(2)} \\ \vdots \\ \underset{\sim}{\beta}_p^{(R)} \end{bmatrix} + \begin{bmatrix} f_{11} \\ f_{22} \\ \vdots \\ f_{mR} \end{bmatrix} + \begin{bmatrix} \underset{\sim}{\varepsilon}_1^{(1)} \\ \underset{\sim}{\varepsilon}_2^{(2)} \\ \vdots \\ \underset{\sim}{\varepsilon}_m^{(R)} \end{bmatrix} \qquad (7)$$

where

$$\underset{\sim}{y}^{(r)} = \begin{bmatrix} y_1^{(r)} \\ y_2^{(r)} \\ \vdots \\ y_n^{(r)} \end{bmatrix} \quad r = 1, 2, \ldots, R$$

where the parametric component parameters in the semiparametric regression model are as follows:

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

$$\underset{\sim}{\beta}_j^{(r)}=\begin{bmatrix}\beta_{01}^{(r)}\\ \beta_{12}^{(r)}\\ \vdots\\ \beta_{pj}^{(r)}\end{bmatrix}\quad \begin{matrix} r=1,2,\ldots,R\\ j=1,2,\ldots,p\end{matrix}$$

For the parameters of the parametric component in equation (1), while for the parameters of the nonparametric component in the multiresponse semiparametric regression model are as follows:

$$f(t_i)=\sum_{r=1}^{R}\sum_{h=1}^{m} f_R(t_{hi}) \tag{8}$$

For each f contains functions that are not known the shape of the regression curve pattern which is approached by the spline truncated function.

The semiparametric regression model of multiresponse spline truncated was applied to 2017 education data, especially at the CBNE of VHS in West Nusa Tenggara Province. Correlation testing is done aiming to determine the relationship between response variables, so that the data is feasible to be analyzed using multiresponse regression analysis. The formulation of the hypothesis is
H_0: There is no correlation between response variables $(\rho = 0)$
H_1: There is correlation between response variables $(\rho \neq 0)$

Based on the results of the correlation test obtained if the p-value <0.05 can be concluded that there is a correlation between the responses. In the multiresponse semiparametric regression model based on the spline truncated estimator there is a known point of knots. Optimal selection of knots is done to determine the best model formed. One method used to select the optimal knot point is to use the GCV (Generalized Cross Validation) method [14]. The spline model with the optimum knot point is obtained from the smallest GCV value.

Table 1. Results of analysis knots optimum.

Variables	Konts	Orde
t_1	4;5	2
t_2	5;7	2
$t_{3(1)}$	68 ; 69	1
$t_{3(2)}$	78.5 ; 79.5	1
$t_{3(3)}$	95 ; 96	2
$t_{3(4)}$	84 ; 85	2

Based on the results of the analysis of the multiresponse spline truncated semiparametric regression model obtained from the optimum knot points obtained:

Table 2. Results of analysis goodness of fit.

MSE	R^2	GCV
49.60792	0.8389319	0.000003229985

The estimated models based on multiresponse spline truncated semiparametric regression model are:

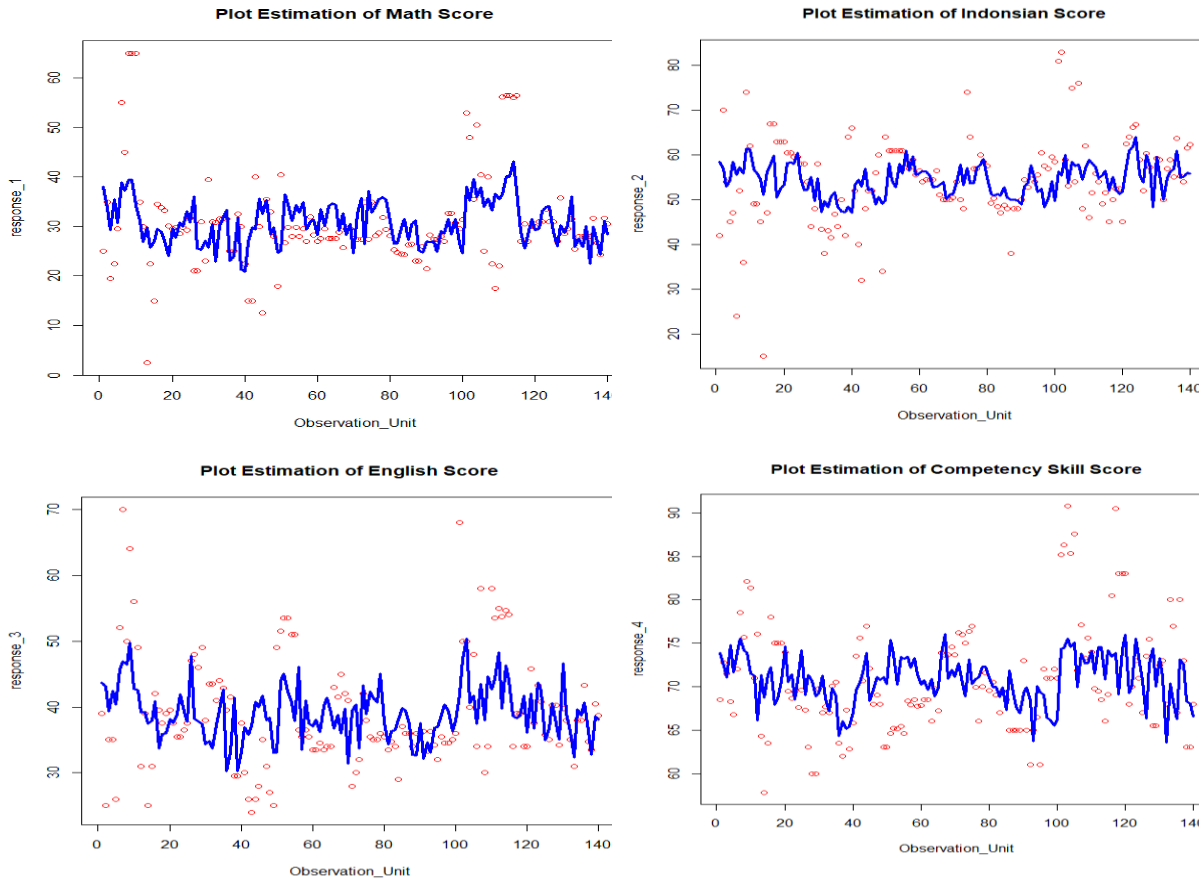
$$\begin{aligned}\hat{y}_1 = {} & 0.062630755 + 3.349289803\,D_1 + 4.218159410\,D_2 + 0.170590704\,D_3 + 0.234182869x_1 \\ & + 0.376260711\,x_2 + 0.140114511\,x_3 + 0.125395070\,x_4 + 0.062607995 + 5.635866199\,t_1 \\ & - 1.167786061\,t_1^{\,2} + 3.883784592\,(t_1 - 4)_+^1 - 3.129869242\,(t_1 - 5)_+^1 + 0.062607995 \\ & - 40.951052444\,t_2 + 4.610157592\,t_2^{\,2} - 6.472459207(t_2 - 5)_+^1 + 2.177573991(t_2 - 7)_+^1 \\ & + 0.062607995 + 3.441547322\,t_3 + 3.431878108(t_3 - 68)_+^1 - 3.156709768(t_3 - 69)_+^1 \\ & + 0.062607995 - 0.839843682t_4 + 11.589869383(t_4 - 78.5)_+^1 - 9.815915514383(t_4 - 79.5)_+^1 \\ & + 0.062607995 + 3.125226729\,t_5 - 0.032409825\,t_5^{\,2} + 1.398352974(t_5 - 95)_+^1 - 0.422108924 \\ & (t_5 - 96)_+^1 + 0.062631244 - 5.121235421\,t_6 + 0.035171527\,t_6^{\,2} - 1.869219940(t_6 - 84)_+^1 \\ & + 1.968389553(t_5 - 85)_+^1\end{aligned} \tag{9}$$

$$\begin{aligned}\hat{y}_2 = {} & -0.664361744 + 0.166644371\,D_1 + 5.151235810\,D_2 + 2.734096043\,D_3 + 0.125584690x_1 \\ & + 0.307114729\,x_2 + 0.136204032\,x_3 - 0.105639086\,x_4 - 0.664361744 - 4.644127984\,t_1 \\ & + 0.976453185\,t_1^{\,2} - 2.249178859\,(t_1 - 4)_+^1 + 0.766579085\,(t_1 - 5)_+^1 - 0.664361744 \\ & - 3.065207609\,t_2 + 0.191451136\,t_2^{\,2} + 0.253747568(t_2 - 5)_+^1 - 0.546542850(t_2 - 7)_+^1 \\ & - 0.664361744 - 10.826008282\,t_3 - 7.950737953(t_3 - 68)_+^1 + 12.634015669(t_3 - 69)_+^1 \\ & - 0.664361744 - 4.549862850t_4 - 0.989460162(t_4 - 78.5)_+^1 - 0.429185630(t_4 - 79.5)_+^1 \\ & - 0.664361744 + 18.209687268\,t_5 - 0.077167931\,t_5^{\,2} - 1.132595933(t_5 - 95)_+^1 + 3.304622759 \\ & (t_5 - 96)_+^1 - 0.664361744 + 1.726283031\,t_6 + 0.004592804\,t_6^{\,2} - 0.342465894(t_6 - 84)_+^1 \\ & + 0.425186279(t_5 - 85)_+^1\end{aligned} \tag{10}$$

$$\begin{aligned}\hat{y}_3 = {} & -0.046907435 + 1.570783284\,D_1 + 3.587634424\,D_2 + 2.190342297\,D_3 + 0.151952088x_1 \\ & + 0.185354796\,x_2 + 0.242902687\,x_3 + 0.371755432\,x_4 - 0.046907435 + 2.737390381\,t_1 \\ & - 0.643327429\,t_1^{\,2} + 2.536929784\,(t_1 - 4)_+^1 - 2.270809106\,(t_1 - 5)_+^1 - 0.046907435 \\ & - 40.401938401\,t_2 + 4.569586569\,t_2^{\,2} - 6.245262715(t_2 - 5)_+^1 + 1.875096233(t_2 - 7)_+^1 \\ & - 0.046907435 + 1.230645646\,t_3 + 2.080778634(t_3 - 68)_+^1 + 0.228786563(t_3 - 69)_+^1 \\ & - 0.046907435 - 1.266187961t_4 + 5.510804136(t_4 - 78.5)_+^1 - 5.679482208(t_4 - 79.5)_+^1 \\ & - 0.046907435 + 5.275740469\,t_5 - 0.040579187\,t_5^{\,2} - 0.994474570(t_5 - 95)_+^1 + 3.563631867 \\ & (t_5 - 96)_+^1 - 0.046907435 - 3.763019450\,t_6 + 0.028886845\,t_6^{\,2} - 0.867398661(t_6 - 84)_+^1 \\ & + 0.855546420(t_5 - 85)_+^1\end{aligned} \tag{11}$$

$$\begin{aligned}\hat{y}_4 = &-0.101097517 + 2.365358538\, D_1 + 3.968248143\, D_2 + 3.163131537\, D_3 + 0.109940412 x_1 \\ &+ 0.252416221 x_2 + 0.043099289\, x_3 + 0.214086264\, x_4 - 0.101097517 + 0.293222753\, t_1 \\ &- 0.226964814\, t_1^{\,2} + 1.751107357\, (t_1 - 4)_+^1 - 1.846420577\, (t_1 - 5)_+^1 - 0.101097517 \\ &- 18.524237082\, t_2 + 2.058015317\, t_2^{\,2} - 2.638892176 (t_2 - 5)_+^1 + 0.602913666 (t_2 - 7)_+^1 \\ &- 0.101097517 - 0.947748744\, t_3 - 4.700733115 (t_3 - 68)_+^1 + 2.771594609 (t_3 - 69)_+^1 \qquad (12) \\ &- 0.101097517 - 1.032663325 t_4 + 0.844369132 (t_4 - 78.5)_+^1 - 0.235129867 (t_4 - 79.5)_+^1 \\ &- 0.101097517 + 4.170353450\, t_5 - 0.015298034\, t_5^{\,2} + 4.566822878 (t_5 - 95)_+^1 - 10.593158437 \\ &(t_5 - 96)_+^1 - 0.101097517 - 1.117578075\, t_6 + 0.009658215\, t_6^{\,2} - 0.232865815 (t_6 - 84)_+^1 \\ &+ 0.207475034 (t_5 - 85)_+^1\end{aligned}$$

Function Based on the best model obtained it turns out that each increase in one unit score report in each subject resulted in an increase in the value of CBNE in each subject. The average score of CBNE for male students is better than female students. Predicate the value of school accreditation influences based on its level, namely the predicate A is better B; and accreditation B is better than schools with C accreditation. On the distance variable, parental education, the value of School Examination (SE) fluctuates on certain knots for the value of CBNE. The following estimation plot for each response.


Plot Estimation of Math Score

Plot Estimation of Indonsian Score

Plot Estimation of English Score

Plot Estimation of Competency Skill Score

Figure 1. Plots of Observation and Estimation for each response

4. Conclusion

The result of the semiparametric regression model, we get the R^2 of 0.84 and the MSE value of 49.61 that satisfies the goodness of fit criterions. The increase in the value of report cards, gender and the predicate value of school accreditation had an effect on the increase in the value of CBNE in each subject. Whereas distance, parental education and SE values experienced fluctuations in certain knots for the CBNE values in each lesson.

References

[1] Eubank, R. L. (1988), *Spline Smoothing and Nonparametrik Regression,* Marcel Dekker, New York

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

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

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

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

[6] Chen, M., and Song, Q., 2016 Semiparametric estimation and forecasting for exogenous log-GARCH models. *Journal of TEST* **25** pp. 93–112.

[7] 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 Reggression Model. *Proceeding 3rd International Seminar on Science and Technology (ISST).*

[8] 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* **6 (**1), pp.129-136

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

[10] Chamidah N, Kurniawan A, Zaman B and Muniroh L 2018 Least square spline estimator in multi-response semiparametric regression model for estimating median growth charts of children in East Java, Indonesia *Far East Journal of Mathematical Sciences (FJMS)* **107** (2) pp.295-307.

[11] Hidayati, L 2018. Bi-Respon Semiparametric Regression Model Based On Spline Truncated For Estimating Computer Based National Exam In West Nusa Tenggara. *Proceeding the 1st International Conference on Mathematics and Islam (ICMIs).*

[12] Wibowo, W., Haryatmi, S., and Budiantara, I.N. 2013. Modeling of Regional Banking Activitas using Spline Multiresponse Semiparametric Regression. *Journal of Applied Mathematics and Statistics.* **23** pp. 102-110.

[13] Puspendik Balitbang Kemendikbud, (2017). *Panduan Pemanfaatan Hasil Ujian Nasional Tahun Pelajaran 2016/201*. BSNP Jakarta.

[14] Budintara, I.N., 2005, *Model Pendekatan Spline dengan Titik Knot Optimal.* Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, ITS.