PROCEEDINGS OF THE 1st STEEEM 2019
Volume 1, Number 1, 2019, pp. 244-252.
Universitas Ahmad Dahlan, Yogyakarta-Indonesia, December 30, 2019

Spline Nonparametric Regression Approach For Modeling Factors Affecting Vocational National Exam Results in Surabaya

Harun Al Azies[1] **and Alfisyahrina Hapsery** [2]
[1] Statistics Department, PGRI Adi Buana University, Indonesia
[2] Statistics Department, PGRI Adi Buana University, Indonesia

Abstract. Government policy in the field of evaluation is to hold a National Examination (UN). There are various factors that affect student national exam results, one of which is the teacher-student ratio. The relationship of the results of students' national examinations with their causal factors can be known one of them by modeling using the method of regression analysis. The regression method approach used is nonparametric regression that is Spline regression, with the advantages of the model tends to look for estimates wherever the data moves. Based on the results of the analysis shows that the best spline nonparametric regression model is a linear spline model with one knot point with the resulting GCV value is 0.044. The teacher-student ratio factor at the level of vocational education affects the achievement of national student exam results in vocational schools in the city of Surabaya. This study shows that a low teacher ratio (one teacher teaches a small number of students) does not guarantee students get good test scores. In general, schools that have a high UN score have a low teacher ratio (1: ≤15). However, there are still many schools with a small teacher ratio that also has a low UN score.

1. Introduction

Education plays a very important role in improving the quality of human resources (HR). Evaluation of school education is an integral part of controlling school education because it is necessary to understand the implementation and the results of coordination that need to be evaluated. Educational evaluation, evaluation of results, implementation process, and managerial factors supporting educational processes. Government policy in the field of evaluation is the holding of the National Examination (UN). The National Examination is carried out under the legal umbrella of Law Number 20 Year 2003 concerning the National Education System.

Therefore, various improvements in the quality or quality of education through national examinations need to be done so that the quality of human resources really materializes as expected. There are various factors that improve the quality of education, one of which is the teacher-student ratio. To find out fully about these factors on education, this study examines the level of secondary education specifically vocational education. Related stated by SUSENAS that the level of individual investment from vocational education (SMK) is greater than general education (high school). The research variables consist of teacher-student ratio as the independent variable and the average value of the National Examination as a substitute variable. The average value of the National Examination here is one of the indicators used in measuring the quality of education. The relationship of the results of students' national

exams with their causal factors can be considered wrong by modeling using the regression analysis method.

Modeling can be done using regression analysis method, where there are three questions raised about parameters, asking nonparametric, and using semiparametric [1]. In expecting nonparametric data, it is expected to find the form of estimation by itself without being agreed by the subjectivity researcher [2]. One that uses nonparametric regression that is often used is Spline. Research on spline that has been done by Al Azies [6] about the effect of DPT immunization on infant mortality in East Java using obtaining nonparametric spline regression. Based on the results of analysis and discussion using Spline analysis, it is known that the factors that influence the incidence of IMR in East Java are toddlers receiving type 3 DPT immunization. Another study was conducted by Yolanda [3] on the analysis of factors affecting the quality of vocational high school education in the district of sijunjung in this study the data were analyzed descriptively and multiple linear regression analysis was used as a method. The results of the study prove that the teaching experience of teachers and infrastructure shows a significant effect on the quality of vocational education in Sijunjung Regency.

Based on the background, there are problems in this study, namely how the characteristics of Vocational Schools in the City of Surabaya along with factors that are thought to influence and how the modeling uses univariable spline nonparametric regression approach. In this study the conclusions are limited to the relationship between the variables of the national examination results with the expected factors.

2. Literature review

2.1. Nonparametric Regression

Nonparametric regression is a suggested method of regression in which the curve shape of the regression function is unknown. In nonparametric regression, the regression curve is only assumed to be smooth (smooth) in the sense that it is contained in a particular function space so that it has high features [2]. The nonparametric regression model is the result of Equation 1 as follows.

$$y_i = f(x_i) + \varepsilon_i, i = 1,2,3,...,n \quad (1)$$

y_i : response variable

$f_{(xi)}$: unknown smooth function

x_i : predictor variable

ε_i : random errors are assumed to be identical, independent and normally distributed with zero mean and variance σ^2

2.2. Spline Regression

Spline is a segmented polynomial (piecewise polynomial) piece that has flexibility. The joint point of the pieces or points that indicate changes in curve behavior at different intervals [2]. The model of spline regression is as Equation 2 as follows.

$$f(x_i) = \sum_{j=0}^{m} \beta_j {x_i}^j + \sum_{j=0}^{p} \beta_{(m+l)} (x_i - k_l)^m \, , \, (x_i - k_l)^m = \begin{cases} (x_i - k_l)^m ; x_i \geq k_l \\ 0; x_i < k_l \end{cases} \quad (2)$$

$f(x_i)$ = spline regression function

$k_1, k_2, ..., k_k$ = Knot point

x = Predictor variable
β = Constant

2.3. Estimation of Spline Regression Parameters
To estimate parameters, we can use the Maximum Likelihood Estimator (MLE) Method, thus the Maximum Likelihood Estimator (MLE) is a technique that is often used in parametric models both to find parameter estimators and test statistical constructs [4]. The likelihood function can be written as in Equation 3 as follows.

$$L(x_1, x_2, ..., x_n; \mu) = \prod_{i=1}^{n} \left(\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \right) \exp(-\frac{1}{2\sigma^2}(x_i - \mu)^2) \quad (3)$$

The log-likelihood function can be written as in Equation 4 as follows.

$$\ln[L(\mu)] = \ln\left\{ \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right] \right\} \quad (4)$$

2.4. Selection of Optimal Spline Regression Knots Points
Knots are joint fusion points where there is a change in behavior in the data. The best spline regression model depends on the optimal knot point [5]. The method to find the optimal knot point that is often used is Generalized Cross Validation (GCV) and Mean Squared Error (MSE). The optimal knot point is obtained from the minimum GCV value.

2.4.1 Selection of Optimal Spline Regression Knots Points
The simple criterion used as the main measure of a good estimator is the Mean Squared Error (MSE) which can be seen in Equation 5 as follows.

$$MSE(k) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2 \quad (5)$$

x_i : independent variable / predictor
y_i : dependent variable / response
n : number of observation

2.4.2 Generalized Cross Validation (GCV)
Another criterion that can be used as a performance measure for a good estimator is Generalized Cross Validation which can be seen in Equation 6 as follows.

$$GCV(k) = \frac{MSE(k)}{(n^{-1} trace[I - A(k)])^2} \quad (6)$$

I : matiks identity
n : number of observation
A(k) is a matrix $X(X'VX)^{-1}X'V$

2.5. Testing the Significance of Spline Regression Model Parameters
Testing the model parameters simultaneously is a test of regression curve parameters simultaneously using the F test. Here is the hypothesis of the simultaneous test

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

H_1 : there is at least one $\beta_k \neq 0; k = 1,2,...,m+p$ the m + p value represents many parameters in spline nonparametric regression in addition. The m + p value represents many parameters in the nonparametric spline regression except β_0.

$$F_{hitung} = \frac{MSE}{MSR};\ MSR = \frac{SSR}{df_{regresi}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{m+p};\ MSE = \frac{SSE}{df_{error}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(m+p)-1} \quad (7)$$

After testing simultaneously, then further testing is carried out individually. Individual testing is performed to determine whether individual parameters have a significant effect on the response variable. The following are hypotheses from individual tests

$H_0 : \beta_k = 0$

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

$$t_{hitung} = \frac{\hat{\beta}_k}{\sqrt{\text{var}(\hat{\beta}_k)}};\ \text{var}(\hat{\beta}_k) = diag[(X^{'}X)^{-1}\hat{\sigma}^2] \quad (8)$$

3. Research Methodology

3.1. Data Sources

This study uses secondary data obtained from the Publication Center for Education research and the Secondary Education Basic Data of the Ministry of Education and Culture. The response variable in this study is the National Exam Results of Surabaya City Vocational School Students 2018/2019 Academic Year. While the predictor variable used in this research is the teacher-student ratio in the Vocational School of Surabaya

3.2 Analysis Steps

The steps of the analysis used to answer the objectives of the study include the following.

1. Describe descriptive statistics of each variable to find out the characteristics of each vocational school in the city of Surabaya.
2. Make a scatter plot between the predictor variable and the response variable to determine the behavior of the data pattern.
3. Modeling the National Examination Results of Vocational Students in Surabaya with 1 knots, 2 knots and 3 knots linear spline
4. Select the optimal knot point using the Generalized Cross Validation (GCV) method where the optimal knot point is related to the smallest GCV.
5. Calculate the minimum MSE value of the model with the optimal knots produced.
6. Model the National Examination Results of Vocational Students in Surabaya using a spline with optimal knots.
7. Interpreting the model
8. Test the model assumptions
9. Draw conclusions.

4. Results and Discussion

4.1. Overview of Vocational High School Conditions in Surabaya

the following discussion explains in descriptive statistics the Conditions of Vocational High Schools in Surabaya

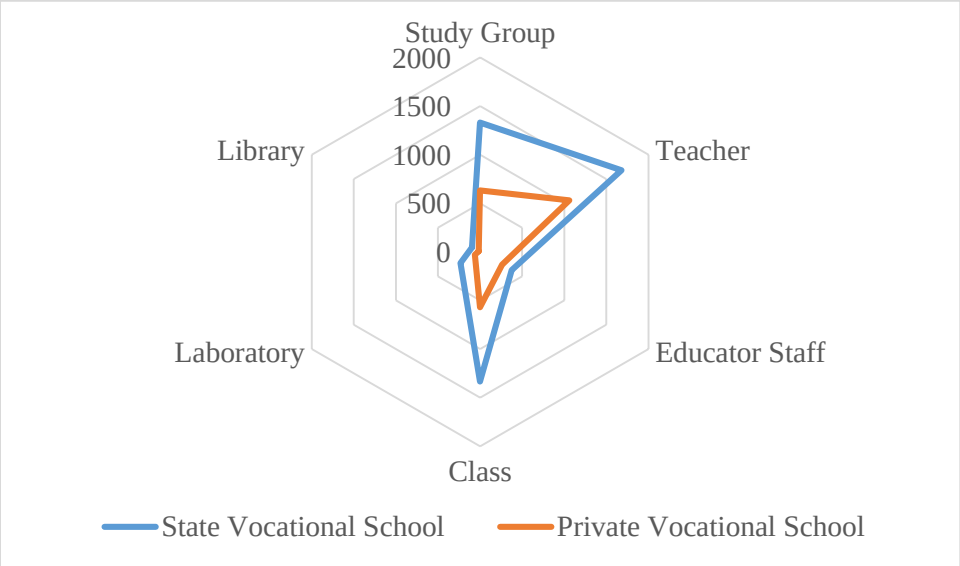

Figure 1. Graph of Difference in Vocational High School Facilities in Surabaya

Based on Figure 1, public and private vocational schools consider the number of teachers to be sufficiently considered, besides that public and private vocational schools do not pay attention to the number of laboratories, libraries and teaching staff

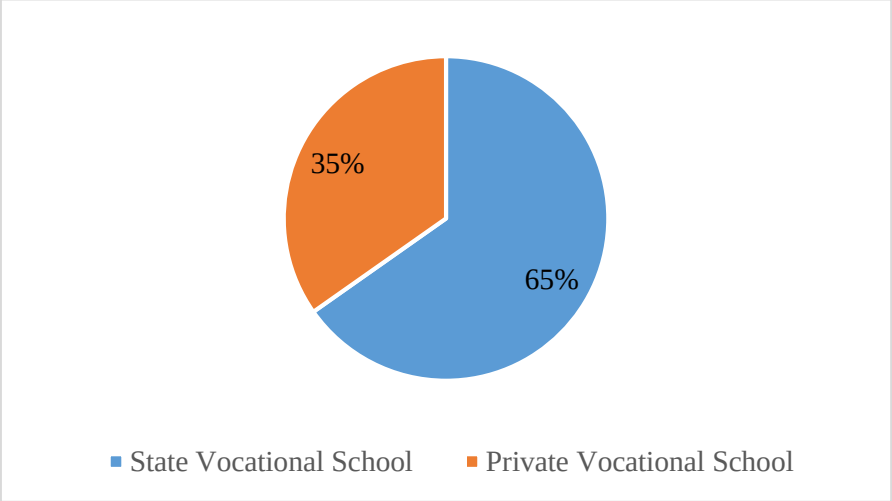

Figure 2. The Number of Vocational High School Student In Surabaya

Based on Figure 2 it can be seen that 65 percent of vocational students in Surabaya are public high school students with a total of 39,343 from 10 schools, while 20,962 students or 35 percent of the total vocational students in Surabaya are vocational students from 93 schools with private status.

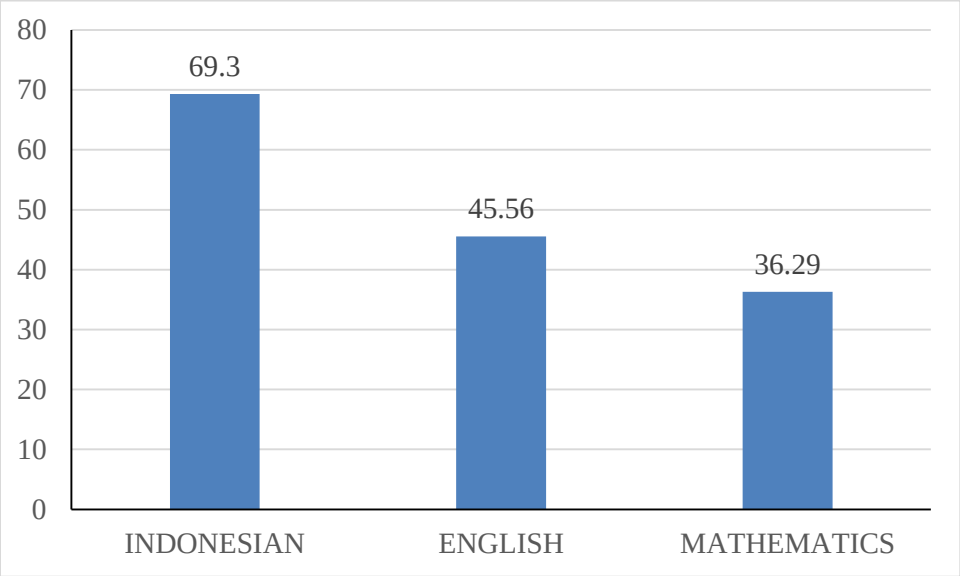

Figure 3. National Exam Results Based on Subjects

Figure 3 explains that the highest average student test results are Indonesian subjects, 69.3 and 45.56 English and the lowest are 36.29 mathematics subjects.

4.2. Description of the Relationship between Vocational National Examination Results with Teacher and Student Ratios

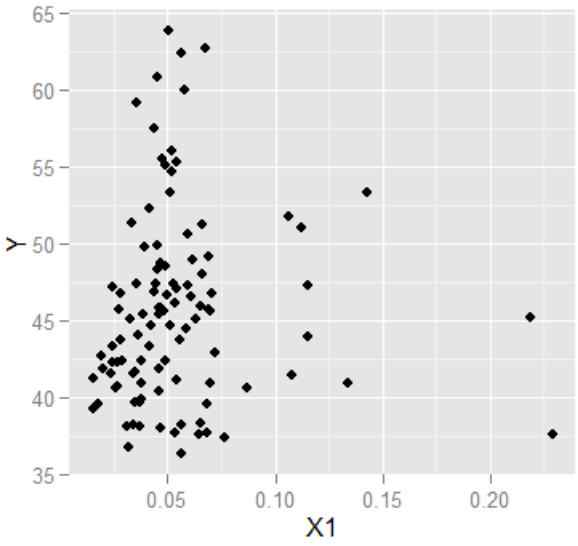

Figure 4. Scatterplot between National Examination Results and Teacher-Student Ratios

Figure 4 shows the pattern of relationships formed between the National Examination Results and the Teacher-Student Ratio that is thought to be influential. Based on the scatterplot formed shows that the pattern of relationships that occur does not form a particular relationship pattern. This indicates that to

solve this problem a nonparametric regression approach is used, where the function of the formed regression curve is unknown.

4.3. Univariable Spline Regression Model

Based on the results of scatter plots, the modeling of the results of the national exams for vocational high school students in Surabaya was carried out using the spline nonparametric regression method. Following are the stages of univariable spline regression analysis

4.3.1Selection of the Best Model

In this modeling used a linear model using 1 point knot, 2 point knot, and 3 point knots. Selection of the Best Model In regression analysis, one of the goals to be achieved is to get the best model that is able to explain the relationship between predictor variables and response variables based on certain criteria such as GCV. The resulting optimal knot point can be seen from the GCV value. The model with the minimum GCV value is said to be the best model. Following are the GCV and MSE values generated using 1 point knot, 2 point and 3 point knots are shown in Table 1.

Table 1. GCV Values Using 1 Knot Point, 2 Knot Points and 3 Knot Points

1 Knot		2 Knot		3 Knot	
Knot Point	GCV	Knot Point	GCV	Knot Point	GCV
0.044	36.79 [b]	0.044 ; 0.051	37.44	0.044 ; 0.036 ; 0.051	37.52
0.051	37.52	0.044 ; 0.058	37.52	0.044 ; 0.036 ; 0.058	37.19
0.058	37.86	0.044 ; 0.036	36.89	0.044 ; 0.036 ; 0.065	36.89
0.036	38.34	0.044 ; 0.065	37.37	0.044 ; 0.036 ; 0.029	37.5
0.065	38.91	0.044 ; 0.029	37.25	0.044 ; 0.036 ; 0.072	36.81

[b] Optimum GCV

Table 1 shows that the minimum GCV and MSE values generated for each knot point. Based on the minimum GCV value, the model selection for the achievement of national exam scores of vocational students in Surabaya uses one optimal knot point at the 0.044 knot point, with a minimum GCV value of 36.79.

4.3.2 Testing Parameters of the Selected Spline Regression Model

Parameter testing is done in two ways, namely simultaneous testing and individual testing. The hypothesis used to determine the effect of parameters simultaneously (simultaneously) on the model that has been obtained is as follows.

Table 2. Test the Significance of Parameters

Db	F_{hitung}	*P-value*
4;33	6,19	0.002*

Based on the ANOVA test results in Table 2, the F-count value of 6.19 was obtained with a p-value of 0.002. So from these results can be taken from the starting H_0 in accordance with the simultaneous test and partially produce significant model parameters to the linear spline regression model.

4.3.3 Interpretation of Results of Linear Spline Regression Model

Using a significance level of 5%, it was concluded that the teacher-student ratio factor influenced the national exam results of vocational students in Surabaya. Spline Linear Regression Model and Interpretation for Spline model are as follows.

$$\begin{aligned}\hat{y} &= 35{,}006 + 285{,}321X - 314{,}949\,(X - 0{,}044) \\ &= \begin{cases} 35{,}006 + 285{,}321X & , X < 0{,}044 \\ -29{,}628 - 21{,}148X & , X \geq 0{,}044 \end{cases}\end{aligned} \tag{9}$$

1. If the teacher-student ratio in vocational students is less than 0.044

Based on the model built above, it can be explained that when the teacher-student ratio in vocational schools is less than 0.044, it means that when the teacher-student ratio in vocational school increases, the results of the national exam scores of vocational school students will increase by 285,321.

2. If the teacher-student ratio in vocational students is more than 0.044

Based on the model built above can explain when the teacher-student ratio in vocational school is more than 0.044, it means that when the teacher-student ratio in vocational schools increases, the results of the national exam scores of vocational school students in the city of Surabaya will decrease by 21,148.

5. Conclusion

The best Spline nonparametric regression model is the linear Spline model with one knot point. The resulting GCV value is 0.044. The teacher-student ratio factor at the level of vocational education affects the achievement of national student exam results in the city of Surabaya in the 2018/2019 academic year, this shows that a low teacher ratio (one teacher teaches a few students), does not guarantee students get a test score that is well. In general, schools that have a high UN score have a low teacher ratio (1: ≤15). However, there are still many schools with a small teacher ratio that also has a low UN score, under 50. This study still uses a linear spline regression program with a combination of one, two, and three knots. It is necessary to develop the program into a quadratic and cubic order using a combination of knots.

Acknowledgments

Authors wishing to acknowledge assistance or encouragement from colleagues. special work by technical staff or financial support from organizations should do so in an unnumbered Acknowledgments section immediately following the last numbered section of the paper.

References

[1] Budiantara, I.N., 2005, Model Keluarga Spline Polinomial Truncated dalam Regresi Semiparametrik, Berkala MIPA, Institut Teknologi Sepuluh Nopember, Surabaya.

[2] Eubank, R.L., 1991, Nonparametric Regression and Spline Smoothing, Mercel Dekker, New York.

[3] Yolanda, Fani Yaly.,2016, Analisis Faktor-Faktor Yang Mempengaruhi Mutu Pendidikan Sekolah Menengah Kejuruan Di Kabupaten Sijunjung, Thesis, Perencanaan Pembangunan, Fakultas Ekonomi Universitas Andalas.

[4] Ardiyanti, S.T., 2010, Pemodelan Angka Kematian Bayi dengan Pendekatan Geographically Weighted Poisson Regression di Provinsi Jawa Timur, Tugas Akhir, Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Sepuluh Nopember, Surabaya

[5] Riskiyanti, R., 2010, Analisis Regresi Multivariat Berdasarkan Faktor-faktor yang Mempengaruhi Derajat Kesehatan di Provinsi Jawa Timur, Tugas Akhir, Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Sepuluh Nopember, Surabaya.

[6] Al Azies H and Trishnanti D 2019 *Pemodelan Pengaruh Imunisasi DPT Terhadap Angka Kematian Bayi di Jawa Timur Tahun 2016 Menggunakan Pendekatan Regresi Nonparametrik Splin*e vol 12, J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika (Surabaya: Universitas PGRI Adi Buana Surabaya) p 26-31