Estimation of Biresponse Multipredictor Model using Local Linear Estimator in Case of Scholastic Aptitude and Islamic Tests Modeling

Nidhomuddin\textsuperscript{1}, N Chamidah\textsuperscript{2*} and A Kurniawan\textsuperscript{2}

\textsuperscript{1}PhD Student, Faculty of Science and Technology, Airlangga University, Indonesia
\textsuperscript{2}Department of Mathematics, Faculty of Science and Technology, Airlangga University, Indonesia

*\textit{nur-c}@fst.unair.ac.id

\textbf{Abstract.} Scholastic aptitude test (SAT) and Islamic test are tests that used by state Islamic colleges (SIC) in Indonesia to select new students on the admission test. Both of them are taken from the same subjects, so they are correlated each other. The behavior patterns of both test scores are change because they were observed from 2016 to 2018. In that period, SAT and Islamic test scores of SIC in western Indonesia are higher than those in eastern Indonesia. So the locally modeling is appropriate for this case. There is a significant correlation between both of the tests then they can be modeled simultaneously by using biresponse nonparametric model. In this research, estimating parameter of biresponse multipredictor nonparametric regression model by using weighted least square method. The results show that in average, the increasing public senior high schools percentage and first choice percentage would increase the SAT scores and Islamic test scores of SIC in western Indonesia. Otherwise, in eastern Indonesia the increasing public senior high school percentage and first choice percentage would decrease the SAT scores and Islamic test scores. It means that there are different patterns of SAT and Islamic test scores between western Indonesia and eastern Indonesia.

1. Introduction
Scholastic aptitude test (SAT) and Islamic test are tests that used by state Islamic colleges (SIC) in Indonesia to select new students on the admission test. Both of them are taken from the same subjects, so they are correlated each other. The SAT has an important effect, especially to predict students achievement in the college [1]. The national education system is expected to create a religious character for each students in Indonesia. Therefore, SAT and Islamic tests become an important part to create excellent students who have good morality.

Several factors can determine students achievement in the college. Senior high school status has an influence for students who take courses of introduction to basic mathematics at the Faculty of Education, Mataram State Islamic University [2]. Achievements of Faculty of Science and Technology students of Sunan Kalijaga State Islamic University are influenced by the status of senior high school [3]. Physics education students at Alauddin Makassar State Islamic University have a good GPA and can be categorized as cum laude when admission test choose physics education as first choice [4].
The behavior patterns of SAT and Islamic test scores changed because they were observed from 2016 to 2018. In that period, SAT and Islamic test scores in western Indonesia are higher than those in eastern Indonesia. So the locally modeling is appropriate for this case.

In average, there is a correlation between SAT and Islamic tests. Modelling of SAT scores and Islamic test scores based on public senior high school percentage obtained mean square error (MSE) 2901.498 [5]. SAT and Islamic test are influenced not only by public senior high school percentage but also first choice percentage and fresh graduated percentage. Because there is a significant correlation between SAT and Islamic test scores also there are more than 1 predictor variable then they can be modeled simultaneously by using bireponse multipredictor nonparametric model.

The researches about smoothing techniques for estimating nonparametric regression models have been studied by several researchers. Local polynomial estimators have been discussed by [6-7], spline estimators have been discussed by [8-12], local linear estimators for estimating bireponse nonparametric regression model have been studied by [13-14]. In addition, [15] have designed the standard growth charts for the weight of children in East Java by using local linear estimator, [16] have modelled maternal mortality and infant mortality in East Kalimantan by using local linear estimator. The regression function can be estimated by using local linear estimator at each point of observation. So, we can approach the real pattern of data and estimate the model that does not require many data [17].

2. Longitudinal Data
Data observations were obtained from \( n \) independent subjects observed several times in the different periods and correlated each other is called as longitudinal data. Researcher [18] showed behavior changes of the subjects in the certain period in longitudinal data analysis. Next, suppose given paired longitudinal data \( (x_{ijk}, y_{ijk}) \) which follows the multipredictor bireponse nonparametric regression model for longitudinal data:

\[
y_{ijk} = f_{(1)}(x_{ijk}) + f_{(2)}(x_{ijk}) + \varepsilon_{ijk}, \quad i = 1, 2, ..., n; \quad j = 1, 2, ..., p; \quad k = 1, 2, ..., m; \quad r = 1, 2
\]

where \( y_{ijk} = (y^{(1)}_{ijk}, y^{(2)}_{ijk})^T \), \( f_{(1)}(x_{ijk}) = (f_{(1,1)}(x_{ijk}), f_{(1,2)}(x_{ijk}))^T \), and \( \varepsilon_{ijk} = (\varepsilon^{(1)}_{ijk}, \varepsilon^{(2)}_{ijk})^T \) is random error assumed to be normal distribution with mean \( \theta \) and variance \( \Sigma_1 \).

3. Local Linear Estimator
Local linear estimator is used to estimate regression function \( f \) in equation (1). The regression function in the bireponse multipredictor model at a certain point \( (x_{01}, x_{02}, ..., x_{0p}) \) can be written in vector notation:

\[
f^{*}(x_{01}, x_{02}, ..., x_{0p}) = \begin{bmatrix} f_{(1)}(x_{01}) \\ f_{(2)}(x_{02}) \end{bmatrix} = \begin{bmatrix} x^{(1)}_{01} \\ 0 \end{bmatrix} \beta^{(1)} + \begin{bmatrix} x^{(2)}_{02} \end{bmatrix} \beta^{(2)}(x_{02})
\]

We can be expressed equation (2) and (3) as follows:

\[
f = X^*(x_{01}, x_{02}) \beta
\]

where \( f = \begin{bmatrix} f_{(1)}(x) \\ f_{(2)}(x) \end{bmatrix} ; \quad X^* = \begin{bmatrix} x^{(1)}_{01} \\ 0 \end{bmatrix} ; \quad X^{(2)}_{\tau} = \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix} (x_1 - x_{01}) (x_2 - x_{02}) \cdots (x_p - x_{0p})
\]

\[
\beta = \begin{bmatrix} \beta^{(1)} \\ \beta^{(2)} \end{bmatrix} \quad \beta^{(1)} = \begin{bmatrix} \beta^{(1)}_{01}(x_{01}) & \beta^{(1)}_{02}(x_{02}) & \cdots & \beta^{(1)}_{0p}(x_{0p}) \end{bmatrix}^T
\]

\[
\beta^{(2)} = \begin{bmatrix} \beta^{(2)}_{01}(x_{01}) & \beta^{(2)}_{02}(x_{02}) & \cdots & \beta^{(2)}_{0p}(x_{0p}) \end{bmatrix}^T
\]

2
To estimate $\hat{\beta}(x_0)$ we use weighted least square (WLS) method by minimizing the function:

$$Z(x_0) = (y - X(x_0)\hat{\beta}(x_0))^T \sum^{-1} K_h(x_0)(y - X(x_0)\hat{\beta}(x_0))$$

(5)

So, we get:

$$\hat{\beta}(x_0) = [X^T(x_0) \sum^{-1} K_h(x_0) X(x_0)]^{-1} X^T(x_0) \sum^{-1} K_h(x_0) y$$

(6)

Based on equation (4) and (6), we get the local linear estimator for $f(x)$:

$$\hat{f}(x) = x^T(x_0) [X^T(x_0) \sum^{-1} K_h(x_0) X(x_0)]^{-1} X^T(x_0) \sum^{-1} K_h(x_0) y$$

(7)

4. Optimal Bandwidth

Bandwidth ($h$) is smoothing parameter in local linear regression that controls the smoothness of fit. Optimal bandwidth is needed because if $h$ is too small then variance value will increase and give very noisy estimation of regression function. While if $h$ is too large, the estimation of regression function will gives large bias and over smoother [19].

Selection method used to obtain optimal bandwidth is generalized cross validation (GCV). According to [20] the optimal bandwidth is obtained by minimizing the GCV:

$$GCV(h) = \frac{n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\left(n^{-1} \text{tr}(I - A(h))\right)^2}$$

(8)

where $n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ is called mean square error (MSE).

5. Research Variables and Steps of Analysis

In this research, we use scores of SAT as first response variable and Islamic test scores as second response variable, and we use public senior high school percentage, first choice percentage and fresh graduated percentage as predictor variables. The data consists of 32 SIC in western Indonesia and 12 SIC in eastern Indonesia.

Steps to model the SAT scores and Islamic test scores are:

a. Determining descriptive statistic of SAT scores and Islamic test scores in western and eastern Indonesia.
b. Calculating the correlation between two response variables (SAT scores and Islamic test scores).
c. Determining the minimum GCV value to get the optimal bandwidth in equation (8).
d. Estimating SAT scores and Islamic test scores by using local linear estimator based on the optimal bandwidth (step c).
e. Plotting both observation and estimation of SAT scores and Islamic test scores.
f. Comparing SAT scores and Islamic test scores model of SIC in western and eastern Indonesia.

6. Results and Discussion

Firstly, we give the descriptive statistics of SAT scores and Islamic test scores in western and eastern Indonesia. Descriptive statistics are given by bar charts in Figures 1 and 2.
From Figures 1 and 2, in average, the SAT scores and Islamic test scores in western Indonesia are higher than those in eastern Indonesia. In average, we can conclude that SAT scores and Islamic test score in western and eastern Indonesia are different. The SAT scores and Islamic test scores of SIC in western and eastern Indonesia decrease in 2018, so we need to increase the test score in order to the test score can be better.

Secondly, calculate the correlation between two response variables (SAT scores and Islamic test scores). The correlation between SAT scores and Islamic test scores is 0.865 and p-value is 0.000. Based on the Pearson correlation test with $\alpha = 0.05$, the SAT scores and Islamic test scores have significant correlation.

Next, we determine minimum value of GCV to get optimal bandwidth ($h$). Plot bandwidth ($h$) versus GCV is given in Figure 3.

Based on Figure 3, we obtain the optimal bandwidth of 6.2, the minimum GCV value of 1528.107 and MSE value of 944.0105.

Plot of observation and estimation of SAT scores for 44 SIC in Indonesia is given in Figure 4.
Figure 4. Plot of observation and estimation of SAT scores

Plot of observation and estimation of Islamic test scores for 44 SIC in Indonesia is given in Figure 5.

Figure 5. Plot of observation and estimation of Islamic test scores

As illustration example for SIC in western Indonesia, we give the estimated model of SAT score ($\hat{y}^{(1)}$) and Islamic test score ($\hat{y}^{(2)}$) for “IAIN Jember” as follows:

$$\hat{y} = \begin{cases} 
\hat{y}^{(1)}_{41} = 541.80 + 0.66(x_1 - 45.77) + 5.90(x_2 - 92.37) + 2.72(x_3 - 77.71) \\
\hat{y}^{(2)}_{41} = 544.98 + 0.51(x_1 - 45.77) + 4.09(x_2 - 92.37) + 2.82(x_3 - 77.71) \\
(39.57 < x_1 < 51.97, 86.17 < x_2 < 98.57, 71.51 < x_3 < 83.91)
\end{cases}$$  

(9)

Based on equation (9), the estimated model of “IAIN Jember” shows that if the public senior high school percentage ($x_1$) increases one percent then the SAT score will increase of 0.66 and Islamic test score will increase of 0.51. Further, if first choice percentage ($x_2$) increases one percent then the SAT score will increase of 5.90 and Islamic test score will increase of 4.09. Next, if the fresh graduated percentage
increases one percent then the SAT score will increase of 2.72 and Islamic test score will increase of 2.82. As illustration example for SIC in eastern Indonesia, we give the estimated model of SAT score \( \hat{y}^{(1)} \) and Islamic test score \( \hat{y}^{(2)} \) for “IAIN Palopo” as follows:

\[
\hat{y} = \begin{cases} 
\hat{y}^{(1)}_{351} = 481.06 - 1.46(x_1 - 84.04) - 1.82(x_2 - 90.57) - 1.01(x_3 - 83.01) \\
\hat{y}^{(2)}_{351} = 473.30 - 1.71(x_1 - 84.04) - 1.69(x_2 - 90.57) + 0.45(x_3 - 83.01) 
\end{cases}
\]

(10)

Based on equation (10), if the public senior high school percentage \( x_1 \) increases one percent then the SAT score will decrease of 1.46 and Islamic test score will decrease of 1.71. Therefore, if first choice percentage \( x_2 \) increases one percent then the SAT score will decrease of 1.82 and Islamic test score will decrease of 1.69. Next, if the fresh graduated percentage \( x_3 \) increases one percent then the SAT score will decrease of 1.01 and Islamic test score will increase of 0.45.

7. Conclusion
In the same case of [5], by adding two predictor variables, we get the smaller MSE than [5]. In average, the SAT scores and Islamic test scores will increase if the public senior high schools percentage and first choice percentage increase for the SIC in western Indonesia. Otherwise, SIC in eastern Indonesia, the SAT scores and Islamic test scores will decrease if the public senior high schools percentage and first choice percentage increase. It means that there are different patterns of SAT scores and Islamic test scores between SIC in western and eastern Indonesia so that local linear estimator approach is appropriate in this case. According to the estimated models in this research, the ministry of religious affairs can make policies to increase SAT scores and Islamic test scores for improving the quality of SIC in Indonesia.

8. References
[1] Lohman D F 2004 Aptitude for College: The Importance of Reasoning Tests for Minority Admissions. In R. Zwick (Ed.) Rethinking the SAT: The Future of Standardized Testing in University Admissions (New York: Routledge Falmer ) pp. 41-55.
[2] Andriani P 2010 Pengaruh Asal Sekolah dan Jurusan Terhadap Hasil Belajar Pengantar Dasar Matematika Mahasiswa Fakultas Tarbiyah IAIN Mataram Beta 3(2) 118-133.
[3] Qudratullah M F 2014 Pengaruh Jalur Penerimaan Mahasiswa dan Asal Sekolah Terhadap Prestasi Mahasiswa di Fakultas Sains dan Teknologi UIN Sunan Kalijaga Fourier 3(1) 12-23.
[4] Kasmawati 2017 Gambaran Prestasi Belajar Mahasiswa Fisika dalam Kaitannya dengan Peminatan Pilihan dalam Proses Penerimaan Mahasiswa Baru Idaarah 1 152-161.
[5] Nidhomuddin, Chamidah N and Kurniawan A 2019 Admission Test Modelling of State Islamic College in Indonesia Using Local Linear for Bivariate Longitudinal Data IOP Conference Series: Materials Science and Engineering 546 (5) 052047.
[6] Chamidah N, Budiantara I N, Sunaryo S and Zain I 2012 Designing of child growth chart based on multi-response local polynomial modeling Journal of Mathematics and Statistics 8(3) 342-347.
[7] Chamidah N and Lestari B 2019 Estimation of covariance matrix using multi-response local polynomial estimator for designing children growth charts: A theoretically discussion Journal of Physics: Conference Series 1397 012072.
[8] Chamidah, N and Lestari B 2016 Spline estimator in homoscedastic multiresponse nonparametric regression model in case of unbalanced number of observations, Far East Journal of Mathematical Sciences, Vol.100 (9), pp. 1433-1453
[9] Lestari, B. Fatmawati-, Budiantara, I.N.Chamidah, N. 2018 Estimation of Regression Function in
Multi-Response Nonparametric Regression Model Using Smoothing Spline and Kernel Estimators Journal of Physics: Conference Series, Vol. 1097 012091.

[10] Islamiyati, A. Fatmawati, Chamidah, N. Estimation of Covariance Matrix on Bi-Response Longitudinal Data Analysis with Penalized Spline Regression. 2018 Journal of Physics: Conference Series, Vol. 979 012093.

[11] Adiwati, T and Chamidah, N. 2019 Modelling of Hypertension Risk Factors Using Penalized Spline to Prevent Hypertension in Indonesia IOP Conf. Series: Materials Science and Engineering 546 052003.

[12] Massaid A Hanif M, Febrianti D and Chamidah N 2019 Modelling of Poverty Percentage of Non-Food Per Capita Expenditures in Indonesia Using Least Square Spline Estimator IOP Conf. Series: Materials Science and Engineering 546 052044.

[13] Chamidah N and Rifada M 2016 Estimation of median growth curves for children up two years old based on bireponse local linear estimator AIP Conference Proceedings 1718 110001.

[14] Darnah, Utoyto M I and Chamidah N 2019 Estimation of the Bi-response Poisson Regression Model Based on Local Linear Approach International Journal of Academic and Applied Research 3(5) 14-18. ISSN: 2000-005X.

[15] Chamidah N, Tjahjono E, Fadilah A R and Lestari B 2018 Standard Growth Charts for Weight of Children in East Java Using Local Linear Estimator Journal of Physics: Conf. Series 1097 012092.

[16] Darnah, Utoyto M I and Chamidah N 2019 Modeling of Maternal Mortality and Infant Mortality Cases in East Kalimantan using Poisson Regression Approach Based on Local Linear Estimator IOP Conference Series: Earth and Environmental Science 243(1), 012023.

[17] Nottingham Q J and Cook D F 2011 Local Linear Regression for Estimating Time Series Data Journal of Computational Statistics and Data Analysis 37 209-217.

[18] Rajulton F 2001 The Fundamentals of Longitudinal Research: An Overview Special Issue on Longitudinal Methodology, Canadian Studies in Population 28(2) pp 169-185.

[19] Liang H and Cheng D Z 2005 Assessment of Esophageal Pressure in Gastroesophageal Reflux Disease by Local Regression Annals of Biomedical Engineering 33 847-853.

[20] Wu H and Zhang J T 2006 Nonparametric Regression Methods for Longitudinal Data Analysis (New Jersey: Willey-Interscience) p 67.