



Modeling Student Mathematics Achievement in Senior High School Based on Selection Results Using Gee 2 Method with Natural Spline

Erfan Syahuri¹, I Made Tirta², Budi Lestari², Dian Anggraeni²

¹ISMAN Lumajang

²FMIPA Universitas Jember, Jl. Kalimantan no.37 Jember econd Affiliation

Email : erfansyahuri@gmail.com

ARTICLE INFO

Article History:

Received Date: 15th April 2017

*Received in Revised Form Date:
30th April 2017*

Accepted Date: 15th May 2017

*Published online Date: 01st
August 2017*

Key Words:

**GEE2, Admission Selection,
High School Achievement,
Natural Spline.**

ABSTRACT

Every school has a vision and mission to become the superior institution so that it can compete and gain trust from the public. To achieve that, one of the efforts of the school is doing the selection of new students at the beginning of each academic year. In Lumajang region, admission of new students (PPDB) are selected using several components, such as national test scores (NUN) and Mapping/Placement test (MP). This research explores the best model of the relationship between selection components (and other conditions of students at the time of selection) and academic achievement during high school (in the form semester mathematics grade) starting from semester 1 till 5 at 3 schools in Lumajang regions. We apply Generalized Estimating Equation order 2 (GEE2) with Natural Spline. The results show that (i) the three schools, have different model and PGRI has the highest mean, followed by SMA1 and SMA3, as shown by significant negative estimates of the coefficients. (i) Although it is relatively small, distance from school has negative contribution to the mathematics grade as shown by negative (but significant) coefficient; (ii) The Junior High School NUN has nonlinear (and nonparametric) contribution as shown by the graphical representation and coefficient of natural spline. (iii) Score of Placement Test contribute positively and significantly to the the semester mathematics grade.

Copyright © Erfan et al, 2017, this is an open access article distributed under the terms of the Pancaran Pendidikan Journal license, which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited

INTRODUCTION

Every school has a vision and mission to become a superior institution so that it can compete and gain high trust from the people. To realize that, one of the efforts is to conduct a recruitment of new students (PPDB) at the beginning of each new school year. In Lumajang PPDB uses several requirements or criteria such as the National Test Score (NUN) of Junior High School and Mapping Test. There are other conditions that need to be considered for students who have been accepted in recruitment of new students selection that may affect students' achievement during in senior high school such as distance of residence and income of parents.

To model the relationship between selection scores (served as predictors) and students achievement on mathematics during 5 semesters in high school (served as responses). The common statistical methods to model relationship between responses and predictors are various regression analyses (or statistical models). Statistical models have been developed from linear model (having one or more predictors) but with one univariate/independent response (having Gaussian/Normal distribution). The linear model have developed to Generalized Linear Model (GLM) to accomodate data with response which are not normally distributed but still independent (Nelder and Wedderburn 1972). For data that may have non normally distributed and correlated response, Liang and Zeger in 1986 introduced a method called Generalized Estimating Equation (GEE) which is a multivariate generalization of GLM. In the GEE method, it is necessary to select the appropriate correlation structure and can describe the correlation among response. The selection of the best model on the GEE method uses Quasi-Likelihood Under the Independence Information Criterion (QIC). GEE2 is an extension of GEE to introduce the equation of scale parameter estimation which is completed simultaneously with the first estimation equation as in ordinary GEE (Tirta et al, 2016) . To accomodate nonlinearity (indicated by the pattern of data that tends to rise or fall sharply) and to produce a good curve shape, then GEE2 may be combined with natural spline or B-spline methods to include a nonparametric component.

This research aims to (i) find the best model for describing the relationship between semester mathematics achievement during high school, and various components of students' selection commonly done in Lumajang regions, (ii) describe the components which are significantly related to semester mathematics achievement, by applying GEE2 method with possible extension using natural spline. The finding will help school.

METHODS

The population of the research are senior high school students in Lumajang regions, where the sample (respondents) are 150 students each randomly sampled from 3 senior high schools in Lumajang (SMAN 1, SMAN 3 and SMA PGRI Lumajang). The response variables in this research are mathematics achievement in 5 semesters during senior high school $Y = (NR1, NR2, NR3, NR4, NR5)$, while the predictors are national exam scores ($NUN=X1$), Mapping/Placement Tests ($MP=X2$), Distance from Schools ($X4$), Parents Income ($X5$), and 1 factor (School) as predictors ($X6$).

Since the responses are vector (assumed to be correlated each others) the most appropriate method to apply is GEE, especially GEE2 with Natural spline extension. With GEE2 we can model the mean and the scale parameters (dependant upon the distribution of responses variable), correlation structure to model correlation among responses, and component nonparametric in the form of natural spline. The selection of the best (most

appropriate) model in term of number of predictors, correlation structure, distribution and link for responses and nonparametric components are done by computing QIC and choose the model with the smallest value of QIC.

The fitting of GEE2 is done by utilizing online (web-based) version of GEE at <http://statslab-rshiny.fmipa.unej.ac.id/RProg/MSI/> and <http://statslab-rshiny.fmipa.unej.ac.id/JORS/GEE/> (Tirta et al, 2016)..

RESULTS AND DISCUSSION

We begin with exploring the correlation among the semester grade (RS1,RS2, RS3, RS3, RS4, RS5). As shown in Figure 1 the correlation among (RS1, RS2, RS3, RS4) are relatively constant arround 0.7, but correlation with SR5 are mostly ‘dropp’ to arround 0.5. Therefore two type of correlation structure, AR-1 and exchangeable, are worth to be considered.

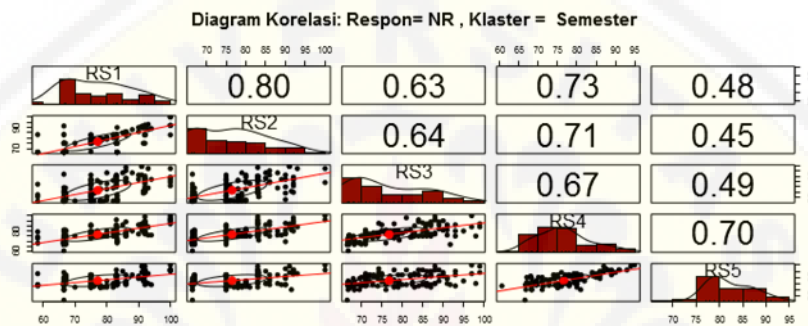


Figure 1 . Correlation among Mathematic Grade semester 1 to 5

In order to find the most suitable predictors, we start the initial model by choosing all the available Xs as predictor and the results (the estimate and their p values) are as follows.

Coefficients	estimate	p
(Intercept)	6.09e+01	0.00e+00
SekolahSMAN1	-6.54e+00	1.00e-04
SekolahSMAN3	-1.38e+01	0.00e+00
Jarak	-1.84e-01	9.30e-03
Penghasilan	1.63e-07	6.61e-01
NUN	4.17e-02	2.66e-01
TP	2.93e-01	1.73e-05

From the initial model, we find that among the predictors, coefficients of NUN (X1) and income (Parents’ Income) = X5 are not significant (pval>5%). Since NUN is a primary component in students selection, we will consider it to contribute as nonparametric component using natural spline smoother, and consider Parent Income as predictor of scale parameters. The smoother graphics for NUN is given in Figure 2.

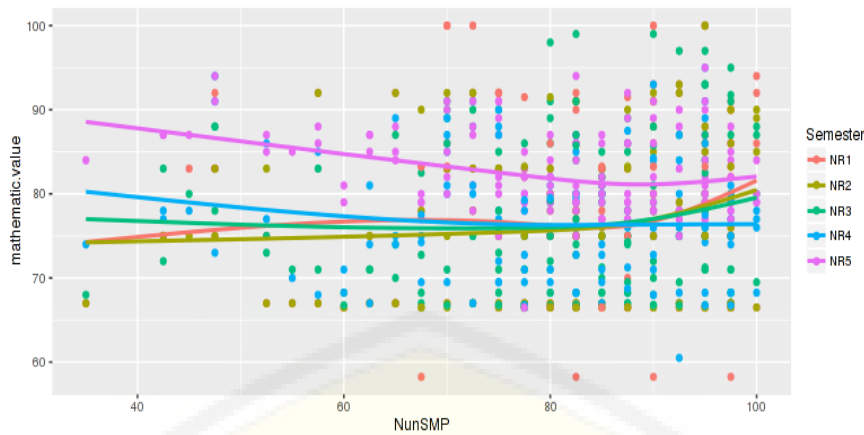


Figure 2. Correlation between the junior Figure 1

Figure 1. The plot of NUN and Mathematics Grade (for 5 semesters) using Natural Spline on 3 degree of freedom.

Figure 2. indicates nonlinearity and nonparametric relationship between NUN and Mathematics Grade. So for the next model exploration we will considered NUN as nonparametric component (smoother) using natural spline with degree of freedom 3.

After considering the smoother and the predictor for scale parameter, we then check some candidate for distributions and correlation structures. Since the response are continuous variable we consider Gamma (log) and Gaussian as candidates for distribution, while for correlation structure the candidates are Exchangeable and AR-1. So we have some candidate models to explore. We compute the QIC value for each model and choose the model with smallest QIC as the best (most appropriate model). Based on this criteria, we find that the QIC are similar for Gaussian and Gamma where for AR1 the QIC= 3017.701 and for exchangeable the QIC=3017.701. Therefore we choose Gaussian (identity) and Exchangeable as the most appropriate (final) model. The fitting give rises to the following estimates (we only report the estimates and their p-values, since the space is limited). Although the Parents' income does not contribute significantly neither in mean model nor in scale model, we retain it in the scale model since it enhanced the model in term of reducing (a little bit) in se (standard error) of the estimates.

Famili = **gaussian** Link= **identity**
 Struktur Korelasi= **exchangeable**

	estimate	p
(Intercept)	61.389	0.00e+00
ns(NUN, df = 3)1	1.0834	5.863e-01
ns(NUN, df = 3)2	9.808	5.280e-02
ns(NUN, df = 3)3	2.176	2.463e-01
SekolahSMAN1	-7.0438	1.484e-05
SekolahSMAN3	-13.704	3.330e-16
Jarak	-0.175	1.404e-02
TP	-5.8282e-07	4.997e-05

correlation		
	estimate	p
alpha	0.38	0

scale		
	estimate	p
(Intercept)	5.5401e+01	0.000

Penghasilan -5.828e-07 0.809

The results show that student achievement as shown by semester mathematical grade are positively correlated following exchangeable structure, which mean they are significant and relatively constant. The 3 schools have different models, SMA1 has mean of mathematical grade lower than SMA PGRI and SMA3 has the lowest among all, as shown by significant negative estimates of the coefficients. However this finding does not indicate the quality of learning in each school, since except the NUN, the predictors and the grade are local and may not comparable each other. Although it is relatively small, distance from school has negatif contribution to the mathematics grade as shown by negatif (but significant) coefficient. This results agrees with the results found by Surani (2012), that the closer the students stay to the school, the more opportunity for students to get access to school facilities and may also related to students fitness in attending the class. Perhaps the unexpected result is that the Junior High School NUN no significant linear contribution, but has a bit nonlinear (and nonparametric) contribution as shown by the graphical representation and coefficient of natural spline. This finding actually needs further investigation since the NUNs are national level and the semester mathematics grades are local, school based. In contrary to national wide NUN score, the score of Local Test (Placement Test) contribute positively and significantly to the the smester mathematics grade.

The final model can be formulated as follows

The mean model

Mathematics Grade = 61.3898+ 1.0834* ns(NUN,3)¹ + 9.8083* ns(NUN ,3)² + 2.1766* ns(NUN,3)-7.0438 * SMA1 -13.704* SMA3 -0.1753*Distance +0.294* Placement test.

Te scale model

Scale=5.540+-5.828e-07 *Parents' _Income.

CONCLUSION

The fitting of students learning achievement on mathematics using GEE order 2 gives results in:

1. Student achievement during high school in the form of mathematics grade report from semester 1 (one) to 5 (five) are positively correlated with exchangeable structure.
2. The 3 schools have different models, SMA1 has mean of mathematical grade lower than SMAPGRI and SMA3 has the lowest among all.
3. Distance of student residence from school has negative contribution to the mathematics grade as shown by negatif (but significant) coefficient.
4. The Junior High School National Test Scores have no significant linear contribution, but have a bit nonlinear (and nonparametric) contribution.
5. Local Test (Placement Test) scores contribute positively and significantly to the the smester mathematics grade.

Recommendations

Further study is needed to model high school local achievements on mathematics (semester grade, together with other non academic factors) and Senior high school national test score or acceptance/ rejection to university, to find out how various local/ internal score related to national or regional test.

ACKNOWLEDGEMENT

The author thanks to anonymous referees for constructive comments and suggestion for improvement of the manuscript. Thanks also due to SMA PGRI, SMAN1 and SMAN3 Lumajang for kindly giving access to data. Finally, thanks The University of Jember for facilitating the Virtual Statistics Laboratory so that we have more alternative ways to access the statistical methods.

REFERENCES

- A Surani.2012.Pengaruh dukungan akademik dan jarak tempat tinggal siswa terhadap prestasi belajar, UK Satya Wacana.
- Budiantara, I. N.2009. Spline dalam Regresi Nonparametrik: Sebuah Pemodelan Statistika Masa Kini dan Masa Mendatang. Surabaya: Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Sepuluh Nopember.
- Budiantara, I. N., (2001b), “Estimasi Parametrik dan Nonparametrik untuk Pendekatan Kurva Regresi”, Makalah Pembicara Utama pada Seminar Nasional Statistika V, Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Sepuluh Nopember (ITS), Surabaya.
- Dahar,R.W.1989.Teori-teoriBelajar.Jakarta:Erlangga
(<https://nurussyahid.blogspot.co.id/2013/12/ptk-prestasi-hasil-belajar-siswa.html>)
- Dwi P.I.S.2013. Model Aditif Terampat dengan pendekatan Spline.Jember: FMIPA Unej.
- Eubank, R.L.1988. Spline Smoothing and Nonparametric Regression, Marcel Dekker: New York.
- Ernawati.2008. Multigroup Structural Equation Model untuk membandingkan Prestasi Belajar siswa yang berasal dari Sekolah Negeri dan Sekolah Swasta, Tesis, FMIPA, ITS.
- Friedman, J.H. 1990. Estimating functions of mixed ordinal and categorical variables using Multivariate Adaptive Regression Splines. Technical Report LCS 107, Statistics Department, Stanford University.
- Friedman, J.H.1991. Multivariate Adaptive Regression Splines (With Discussion). Stanford California 94309.
- Hardin, J.W. and Hilbe, J.M. (2012). Generalized Estimating Equations, 2nd Edition, Chapman and Hall/CRC: New York.
- Henaulu, M.H.2009. Pemodelan Nilai UNAS Siswa SMA Negeri 11 Ambon dengan Pendekatan Regresi Nonparametrik Spline, Tesis, FMIPA, ITS.
- Kusdaryanti, W.1999. Analisis Faktor Prediksi yang Mempengaruhi Hasil Belajar Siswa Kelas IV – V SD Muhammadiyah 4 Pucang, TA, FMIPA, ITS.
- Lestari, B., (2008b). Penalized weighted least- squares estimator for bivariate nonparametric regression model with correlated errors. Proceeding of the National Seminar on Mathematics and Statistics, (MS’08), Airlangga University, Surabaya, pp: 83-95.
- Oktafiani, L.C.2015. Pengembangan Web Interaktif Rshiny untuk GEE2. Jember: FMIPA Universitas Jember
- Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57, 120-125.
- Purwayuningsih, W.2011. Penerapan Regresi Semiparametrik Spline Pada Pemodelan

- Nilai UN siswa SMKN 1 Nguling Pasuruan, Tesis, ITS.
- Scheumaker, L.L.1981. *Spline Functions: Basic Theory*, John Wiley & Sons, Inc: Canada.
- Sunaryo, S., dan Purwahyuningsih, W. (2010), "Pendekatan Regresi Semiparametrik Spline (Pada data nilai Ujian Nasional siswa SMKN 1 Nguling Pasuruan)", Surabaya, Seminar Nasional Pascasarjana X.
- Surani.2012.Pengaruh dukungan akademik dan jarak tempat tinggal siswa terhadap prestasi belajar, UK Satya Wacana.
- Sutarsih, S.2008. Pendekatan Regresi Spline untuk Memodelkan Nilai Unas Siswa SMK Negeri 3 Buduran Sidoarjo, Tesis, IPB.
- Srinadi, I.A.M. 2002. Estimator Spline pada Model Semiparametrik, Tesis. Surabaya: Institut Teknologi Sepuluh Nopember.
- Tirta, I. M. 2009. Analisis Regresi dengan R. Jember: Jember University Press.
- Tirta, I.M. Anggraeni, D. Oktafiani, L.C., 2016. On the Development of Web-based Graphical User Interface, for Learning and Fitting Generalized Estimating Equation. The First International Basic Science Conference (IBSC), 2016. (Submitted for *Pertanika*)
- Trimarjoko, B., 2011. Pengaruh Tingkat Pendidikan Dan Tingkat Pendapatan Orang Tua Terhadap Prestasi Belajar Siswa Kelas XI Jurusan Akuntansi di SMK Yayasan Pembinaan Pembangunan Masyarakat (YPPM) Boja Tahun Ajaran 2010/2011. Jurusan Akuntansi. Fakultas Ekonomi. Universitas Negeri Semarang.
- Wibowo, W., Haryatmi, S., Budiantara, I.N. 2009. Least Square Methods for Estimating Curve of Spline Semiparametric Regression, Proceeding of National Seminar on Mathematic and Mathematic Education, Yogyakarta State University, December 5 th 2009, p. 633-645, ISBN : 978-979-16353-3-2
- Wibowo, W., Haryatmi, S., Budiantara, I.N. 2010. Inference And Confidence Interval For Regression Curve In Spline Semiparametric Model, Proceeding of National Seminar on Mathematic and Mathematic Education, University of Muhammadiyah Malang, January 30 th, 2010
- Swan, T., (2006), Generalized Estimating Equations when The Response Variable Has a Tweedle Distribution: In Application for Multisite Rainfall Modelling, <http://pdf-search-engine.com/GEE.pdf>, diakses tanggal 11 Oktober 2014.