M – 28

MODELLING THE AVERAGE SCORES OF NATIONAL EXAMINATION IN WEST JAVA

Karin Amelia Safitri, Khairil Anwar Notodiputro, Anang Kurnia
Department of Statistics, FMIPA, Bogor Agricultural University, Indonesia

Abstract

Formal education in Indonesia is commonly divided into stages such as preschool, primary school (*SD*), Secondary School (*SMP-SMA*), and universities/colleges. Indonesian government has been taking serious efforts on how to improve the quality of education in Indonesia. The roadmap for continous improvement of education quality can be designed based on the results of National Examination (*UN*) taken regularly by high school students.

This research was aimed at exploring informations on how the scores of *UN* can be linked with other explanatory variables. A panel data which consists of average scores of *UN* for all public senior high schools (SMA Negeri) in West Java Provinces during 2011-2013 and other related variables such as total scores of accreditation, regional domestic product, human development index, scores of school's facilities and its infrastructure, scores of school's educators, average scores of final school exams, were used in this research. The average scores of UN in this case were dependent on variations between high schools and time periods as well as other explanatory variables in which the effects were either fixed or random. The data of this research was modelled with linear mixed models and using the Generalized Estimating Equation (GEE) approach. Both linear mixed models and GEE have been commonly used to analyse the panel data.

This paper showed that the GEE provided a model of better performance than the linear mixed models in explaining the variability of the response variable which was the average scores of *UN*. The GEE also showed significant correlation between explanatory variables and the response.

Key words: fixed effects, GEE, linear mixed model, national examination, random effects.

INTRODUCTION

Background

Education plays important roles for the national development. Quality of human resources can be improved through better quality of education continuously. In Indonesia, there are several types of education namely formal, non-formal and informal education which are complemented each other. The level of formal education comprises of elementary education, high school or intermediate education, and higher education. A good quality school must provide quality teaching, curriculum, management, and facilities. The quality of a school can be reflected by the national examination (*UN*) scores and the school examination (*US*) scores.

UN is a standard evaluation system of primary and secondary education in Indonesia. In order to control the quality of education nationwide, *UN* has been conducted once in a year. The results was not only used as a consideration for students to continue to the next level of study but also for a basis of government policy in assisting education units as well as for the quality

of education mapping (*Permendikbud No.5*).

This research aimed at analyzing the data of *UN* scores for all public senior high schools, commonly abbreviated in Indonesia as *SMA*, in West Java. These *SMA*s have been be referred as observation units. This kind of data is panel data, since the data is a combination of cross sectional data and time series data. Data was taken in four years from 2011-2014. A Linear Mixed Model has been used since the response variable is assumed normally distributed. Linear mixed models have been widely used in analysis of panel data, where each time series constitues an individual curve as a cluster. The model accommodates both fixed effects and random effect, since the unit and time effects can be assumed to be random.

Another approach used in this research was the Generalized Estimating Equation (GEE). The GEE has been commonly used to model correlated data from repeated measures such as panel data. The quasi likelihood method is usually employed for estimating the model parameters. GEE covers the extentions of Generalized Linear Model (GLM) to panel data.

It is important to mention this research considered that *UN* scores can be affected by internal factors as well as external factors. The external factors are the factors within school, education system, and the effect of school policies. Economics condition of each regency/city in West Java where the schools are located may affect the quality of education. Hence, it will affects the UN scores of the schools. This economic condition can be measured by the gross regional income per capita (commonly mentioned as *PDRB* in Indonesia) just like what Ahmad (2011) has stated that the government funds allocated for education sector can affect economic growth or *PDRB* and vice versa. According to Statistics Indonesia (2008), the growth rate of *PDRB* also influences human development index (*IPM*) through household expenditures for daily primary needs including foods, medicines, and school stuffs. The growth rate of *PDRB* may also influence *IPM* through government domestic expenditure policies including priority funds for social aspects. Moreover *IPM* may influence *PDRB* through qualified human resources in terms of good health and quality education.

Based on the above discussion it is interesting to find the best model for the average scores of *UN* all public *SMA*s that could be explained by its explanatory variables such as *IPM*, *PDRB*, average scores of *US*, total accreditation scores, educators scores, school facilities scores, teaching content scores, teaching process scores, school management scores, school finance scores, and graduates competence scores. Moreover, we need to know which ones of the explanatory variables have significant effects on the average scores of *UN*.

Objectives

1. To develop the best model for modelling between average scores of *UN* all public *SMA*s for both natural science major and social science major and another explanatory variables mentioned above.
2. To evaluate which explanatory variables contribute significantly to the average scores of *UN*.

LITERATURE REVIEWS

Panel Data

A panel data set is one that follows a given sample of individuals over time, and thus provides multiple observations on each individual on the sample (Hsiao, 2003). Observing a panel data set is basically observing a broad section of subjects over time, and thus allows us to study dynamic, as well as cross sectional, aspects of a problem (Frees, 2004).

The general panel data model for observation $\boldsymbol{y}_{it}$ of cross section data *i* at time *t* can be defined as

$$\boldsymbol{y}_{it} = \alpha + \boldsymbol{X}_{it}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{it}$$

α is a scalar of overall intercept coeffisient, $\boldsymbol{\beta}$ is a vector of slope coeffecient with K x 1 where K is number of explanatory variables, $\boldsymbol{y}_{it}$ is a vector response variable observed for individual unit *i* at time *t*, $\boldsymbol{X}_{it}$ is a matrix of explanatory variables individual unit *i* at time *t*, $\boldsymbol{\varepsilon}_{it}$ is the error

term vector ~ $N(0, \sigma_\varepsilon{}^2)$.

Linear Mixed Model

Jiang (2007) used a linear mixed model which takes a general form as

$$\boldsymbol{y} = \mathbf{X\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \quad (2.1)$$

where $\boldsymbol{y}$ is observation vector dimension n_i, **X** is known covariate matrix**, β** is regression coefficient vector (fixed effect), **Z** is design matrix that contains only 1 and 0, α is normally distributed random effect vector, $\boldsymbol{\varepsilon}$ is error vector.

Basic assumption that should not be violated in linear mixed model is the first, random effect and error have mean zero and certain varians, e.g. var(**α**) = matrix **G** and var(**ε**)= matrix **R**. The second assumption, **α** and **ε** must be uncorrelated. If random effect and error are assumed normally distributed, then the assumptions can be written as

$$\alpha \sim N(0, \mathbf{G})$$
$$\varepsilon \sim N(0, \mathbf{R})$$

and the distribution of response variable is **y**~N(**Xβ,V)**, with **V** = Var (**y**) = **ZGZ'**+**R.** Moreover if the random effect and error is not assumed normally distributed, then model (2.1) becomes non Gaussian linear mixed model.

Jiang (2007) also stated that estimation in linear mixed model consists of estimating both the fixed and random effects as explained below.

Estimating fixed effect is Best Linear Unbiased Estimator (BLUE):

$$\widehat{\boldsymbol{\beta}}_{BLUE} = (\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{X'V^{-1}y}$$

and random effect prediction which is Best Linear Unbiased Prediction (BLUP) can be written as

$$\tilde{\boldsymbol{\alpha}} = BLUP(\boldsymbol{\alpha}) = \boldsymbol{GZ'V^{-1}}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$$

In fact, variance component **V** is unknown therefore, in estimating fixed effect and random effect, **V** is replaced by the estimator $\widehat{\boldsymbol{V}}$ and the estimation result is empirical best linear unbiased predictor (EBLUP).

$$\widehat{\boldsymbol{V}} = \boldsymbol{Z}\widehat{\boldsymbol{G}}\boldsymbol{Z'} + \widehat{\boldsymbol{R}}$$

where $\widehat{\boldsymbol{G}}$ and $\widehat{\boldsymbol{R}}$ are estimators using Maximum Likelihood Estimation (MLE) or Restricted Maximum Likelihood (REML).

Henderson (1959) introduced the solutions which can produce simultaneously the GLS estimator of $\boldsymbol{\beta}$ and BLUP of $\boldsymbol{\alpha}$**.**

$\boldsymbol{y}|\boldsymbol{\alpha} \sim N(\boldsymbol{X\beta} + \boldsymbol{Z\alpha}, \boldsymbol{R})$ and $\alpha \sim N(0, G)$ the joint density of $\boldsymbol{y}$ and $\boldsymbol{\alpha}$ is

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{\alpha}) &= f(\boldsymbol{y}|\boldsymbol{\alpha}).f(\boldsymbol{\alpha}) \\ &= (2\pi)^{-\frac{n}{2}}|\boldsymbol{R}|^{-\frac{1}{2}}exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Z\alpha})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Z\alpha})\right\} \\ &\quad \times (2\pi)^{-\frac{q}{2}}|\boldsymbol{G}|^{-\frac{1}{2}}exp\left\{-\frac{1}{2}\boldsymbol{\alpha'G^{-1}\alpha}\right\} \\ &= \frac{exp\left\{-\frac{1}{2}\left[(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Z\alpha})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Z\alpha}) + \boldsymbol{\alpha'G^{-1}\alpha}\right]\right\}}{(2\pi)^{\frac{n+q}{2}}|\boldsymbol{R}|^{\frac{1}{2}}|\boldsymbol{G}|^{\frac{1}{2}}} \end{aligned}$$

find the log of $f(y, \alpha)$:

$$\begin{aligned} \log f(\boldsymbol{y}, \boldsymbol{\alpha}) &= -\frac{n+q}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{R}| - \frac{1}{2}\log|\boldsymbol{G}| \\ &\quad -\frac{1}{2}\left[(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Z\alpha})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Z\alpha}) + \boldsymbol{\alpha'G^{-1}\alpha}\right] \end{aligned}$$

then calculate the partial derivatives of log of $f(y, \alpha)$ with respect to β and α:

$$\frac{\partial \log f(\boldsymbol{y}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} = \boldsymbol{X'R^{-1}}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Z\alpha})$$

$$\frac{\partial \log f(\boldsymbol{y}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \boldsymbol{Z'R^{-1}(y - X\beta - Z\alpha) - G^{-1}\alpha}$$

Those results are set to zero and the equations that are obtained:

$$\begin{cases} \boldsymbol{X'R^{-1}X\beta + X'R^{-1}Z\alpha = X'R^{-1}y} \\ \boldsymbol{Z'R^{-1}X\beta + Z'R^{-1}Z\alpha + G^{-1}\alpha = Z'R^{-1}y} \end{cases}$$

write the equations above in matrix form:

$$\begin{bmatrix} \boldsymbol{X'R^{-1}X} & \boldsymbol{X'R^{-1}Z} \\ \boldsymbol{Z'R^{-1}X} & \boldsymbol{Z'R^{-1}Z + G^{-1}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'R^{-1}y} \\ \boldsymbol{Z'R^{-1}y} \end{bmatrix}$$

Estimating The Model Parameters

Some approaches used to estimate the model parameters in linear mixed model there are the traditional approach, MLE, and REML.

Restricted Maximum Likelihood

REML can produce the unbiased or nearly estimator for variance and it doesn't require more complex computation just like the ML estimators. Searle (1992) stated that the REML approach is now more preferable to estimate the variance parameters in mixed model.

y is response vector that normally distribute which is transformed for **z** = **A'y** which has distribution **z ~ N(0, A'VA)**. Define a non-zero *n x (n-p)* matrix **A** for **A'X** = **0** so then **A A'** = **Q** where $\boldsymbol{Q = I - X(X'X)^{-1}X'}$ and **AA' = I t**hen define **z** = **A'y** = $\boldsymbol{A'(y - X\hat{\beta}_{GLS})}$. The expected value of z is **0** and covariance matrix is **A'VA.** The density of **z** is

$$f_{\boldsymbol{z}}(\boldsymbol{z}) = (2\pi)^{-\frac{n-p}{2}} |\boldsymbol{A'VA}|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}\boldsymbol{z'(A'VA)^{-1}z}\right\}$$

$$= (2\pi)^{-\frac{n-p}{2}} |\boldsymbol{A'VA}|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}\boldsymbol{(y - X\hat{\beta}_{GLS})'(A'VA)^{-1}(y - X\hat{\beta}_{GLS})}\right\}$$

According to Rao (1973) that

$$\boldsymbol{(y - X\hat{\beta}_{GLS})'(A'VA)^{-1}(y - X\hat{\beta}_{GLS}) = (y - X\hat{\beta}_{GLS})'V^{-1}(y - X\hat{\beta}_{GLS})}$$

then we come to the final expression of the desity of **z**.

$$f_{\boldsymbol{z}}(\boldsymbol{z}) = (2\pi)^{-\frac{n-p}{2}} |\boldsymbol{X'X}|^{-\frac{1}{2}} |\boldsymbol{X'V^{-1}X}|^{-\frac{1}{2}} |\boldsymbol{V}|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}\boldsymbol{(y - X\hat{\beta}_{GLS})'V^{-1}(y - X\hat{\beta}_{GLS})}\right\}$$

the final density of **z** leads to the restricted or residual log likelihood whcih written below.

$$\log L_{REML}(\boldsymbol{\sigma}) = const - \frac{1}{2}\log|\boldsymbol{X'V(\sigma)^{-1}X}| - \frac{1}{2}\log|\boldsymbol{V(\sigma)}| - \frac{1}{2}\boldsymbol{r(\sigma)'V(\sigma)^{-1}r(\sigma)}$$

Just like estimating using MLE, log likelihood REML is maximized also requires numerical technique for example Newton Raphson with minor adaptations (Longford 2003). By subtituting the estimator $\boldsymbol{\hat{V}_{MLE} = V(\hat{\sigma}_{MLE})}$ into GLS formula.

$$\boldsymbol{\hat{\beta}_{REML} = (X'\hat{V}_{REML}^{-1}X)^{-1}X'\hat{V}_{REML}^{-1}y}$$

$\boldsymbol{\hat{\beta}_{REML}}$ is not identical $\boldsymbol{\hat{\beta}_{MLE}}$ eventhough **y** is normal distribution.

Generalized Estimating Equation

Zeger & Liang (1986) introduced an alternative approach to estimate the parameter in maximum likelihood case. This approach is usually known as generalized estimating equation (GEE). Basically, GEE is an extention of generalized linear model using quasi likelihood estimation.

$\boldsymbol{y_{it}, x_{it}}$ are observations with $t = 1,2, \ldots n_i$ and $i = 1,2, \ldots k$. $\boldsymbol{y_{it}}$ is response variable and $\boldsymbol{x_{it}}$ is covariate vector $p \times 1$ so that $\boldsymbol{Y_i(n \times 1)}$ and $\boldsymbol{x_i(n \times p)}$ for object i. Assumes that response variable is a member of exponential family distribution, so that

$$E(\boldsymbol{y}_{it}) = b'(\theta_{it}) = \mu_{it}$$

$Var(\boldsymbol{y}_{it}) = b''(\theta_{it})a_{it}(\phi)$, where ϕ a possibly unknown scale parameter.

For $\boldsymbol{Y}_i$, the variance of $\boldsymbol{y}_{it}$j as a function of the mean, $\boldsymbol{V}_i = \left(\boldsymbol{A}_i^{1/2}\boldsymbol{R}_i(\alpha)\boldsymbol{A}_i^{1/2}\right)\phi$, where $\boldsymbol{A}_i$ is a diagonal matrix $n \times n$ with diagonal element t is $b''(\theta_{it})$ and $\boldsymbol{R}_i(\alpha)$ is $n \times n$ working correlation matrix.

Where α can be obtained by approaching by

$$\hat{\alpha} = \phi \sum_{i=1}^{k} \sum_{t=t'} \frac{\hat{r}_{it}\hat{r}'_{it}}{\left[\sum \frac{1}{2n_i(n_i-1)} - P\right]},$$

Some general forms of structure of working correlation matrix can be known in the table 1 as follow.

Table 1. Structures of working correlation matrix

Stuctures Name	Structures	Stuctures Name	Structures
Independence	$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$	Autoregressive	$\begin{pmatrix} 1 & \rho & \cdots & \rho^{t-1} \\ \rho & 1 & \cdots & \rho^{t-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{t-1} & \rho^{t-2} & \cdots & 1 \end{pmatrix}$
Exchangeable	$\begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$	M-dependent	$\begin{pmatrix} 1 & \rho_1 & \cdots & \rho^{t-1} \\ \rho_1 & 1 & \cdots & \rho^{t-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{t-1} & \rho^{t-2} & \cdots & 1 \end{pmatrix}$
Unstructured	$\begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,t} \\ \rho_{1,2} & 1 & \cdots & \rho_{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,t} & \rho_{2,t} & \cdots & 1 \end{pmatrix}$	Fixed	$\begin{pmatrix} 1 & r_{1,2} & \cdots & r_{1,t} \\ r_{1,2} & 1 & \cdots & \rho_{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1,t} & r_{2,t} & \cdots & 1 \end{pmatrix}$

GEE is arranged as $\boldsymbol{U}_G(\beta) = \sum_{i=1}^{k} \boldsymbol{D}'_i\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \mu_i)$ with $\boldsymbol{D}_i = \frac{\vartheta\mu_t\vartheta\theta_t\vartheta\eta_t}{\vartheta\theta_t\vartheta\eta_t\vartheta\beta}$.

RESEARCH METHOD

Average score of *UN* and *US* all *SMAN*s in West Java 2011-2014. Data was accessed from Educational Research Center, Ministry of Education and Culture. Data of total accreditation scores, educators scores, school facilities scores, teaching content scores, teaching process scores were accessed in National Accreditation Institution. Data of human development index and gross regional product in West Java were accessed through Statistics Indonesia Website. Variables used in this research are

Y : Average scores of *UN* all *SMAN*s for both natural and social science major (UN).
X1 : Educators scores all *SMAN*s in West Java (EDS)
X2 : School facilities scores all *SMAN*s in West Java (FSS)
X3 : Average scores of *US* all *SMAN*s in West Java (US)
X4 : *IPM* scores all regencies/cities in West Java (IPM)
X5 : *PDRB* all regencies/cities in West Java (PDRB)
X6 : Total accreditation scores all *SMAN*s in West Java (TAS)
X7 : Teaching content scores all *SMAN*s in West Java (CSS)
X8 : Teaching process scores all *SMAN*s in West Java (TPS)

Steps in data analysis process in this research are.

1. Data Exploration
2. Building the initial model.
   The model which has been built as the initial model based on the linear mixed model:
   $\boldsymbol{y}_{it} = \boldsymbol{X\beta} + \boldsymbol{u}_i + \boldsymbol{v}_{it}$

$$UN = \beta_0 + \beta_1 \text{EDS} + \beta_2 \text{FSS} + \beta_3 \text{US} + \beta_4 \text{IPM} + \beta_5 \text{PDRB} + \beta_6 \text{TAS} + \beta_7 \text{CSS} + \beta_8 \text{TPS} + \boldsymbol{u}_i + \boldsymbol{v}_{it}$$

3. Modelling the data with linear mixed model, the model parameters were estimated with Restricted Maximum Likelihood.
4. Modelling the data with GEE, the working correlation matrix is autoregressive.
5. Comparing the models

RESEARCH RESULT AND ANALYSIS

This chapter shows the result of this research which is according to the mixed model data analysis. There are four initial models established which are divided into two kinds of models. Those two models built are based on the data of natural science diciplines and the rest are the models of social science diciplines. Evaluating the model and selecting the best model are done by comparing the BIC values and deviance of each model.

Data was explored by plotting between averages scores of UN versus years. Data of average scores of *UN* for all SMANs, natural scince major in West Java shown in scatter plot on the below

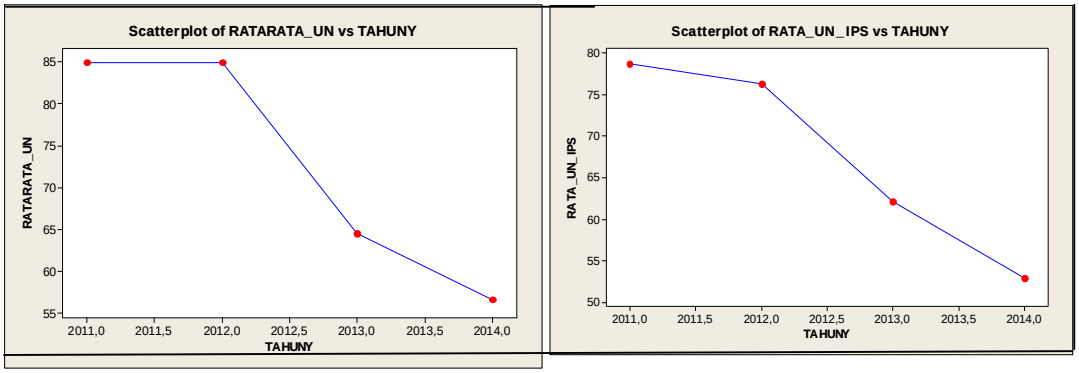

Figure 1. Scatterplot of Average score of UN versus years for a) natural science diciplines b) social science diciplines

It shows that there's such an incive decline on the slope between year 2012 and 2013. The most current years which are 2013 and 2014 the average scores of UN for natural science diciplines got decreased. Same as on the average of UN plot of natural science diciplines, sharp slope happens between year 2012 and 2013 on the plot of social science major.

Started from 2013 both in natural and social science diciplines, there was a newer system that has been implemented for accomplishing UN that caused the average score of UN got down. New system of UN which was like each task paper has been made exactly different for each pupil was implemented in 2013 in order to control every kind of cheat while doing the test of UN.

Modelling With Linear Mixed Model

Panel data which is both of natural and social science diciplines is modelled with linear mixed model. The intention of using linear mixed model instead of using another method for the analysis is the response variable (the average scores of *UN*) is normally distributed $\text{N}(0, \sigma_y{}^2)$ and the model has random effect which is school gives an effect randomly in this reasearch. The model which has been built as the initial model based on the linear mixed model:

$$\boldsymbol{y}_{it} = \boldsymbol{X\beta} + \boldsymbol{u}_i + \boldsymbol{v}_{it}$$

$$\widehat{UN}_{it} = \beta_0 + \beta_1 \text{EDS} + \beta_2 \text{FSS} + \beta_3 \text{US} + \beta_4 \text{IPM} + \beta_5 \text{PDRB} + \beta_6 \text{TAS} + \beta_7 \text{CSS} + \beta_8 \text{TPS} + \boldsymbol{u}_i + \boldsymbol{v}_{it}$$

The estimation result by modelling the data for the natural science diciplines which is analysed using R software, package 'lme4' represents as follows.

Table 1. Modelling in Random Effect Model for Natural Science Major

Parameter	Estimate	Std.Error	t.value	p.z
(Intercept)	136.7899325	13.4292312	10.185983	0.000000e+00
rata_US	-1.02293870	0.10902801	-9.382348	0.000000e+00
akre	0.193228429	0.17103648	1.1297498	2.585817e-01
ipm	-0.020623462	0.14636648	-0.140903	8.879466e-01
pdrb	-0.144725934	0.02886559	-5.013788	5.336877e-07
ISI	-0.004182185	0.08082400	-0.051744	9.587324e-01
PRS	-0.058661240	0.06268280	-0.935843	3.493542e-01
TEN	0.090488615	0.06256899	1.446221	1.481151e-01
SAR	0.070093554	0.05291827	1.324563	1.853163e-01

Table 1. shows that the estimated value of the parameters in linear mixed model. Judging by the software output, it seems that there are only the scores of *US* and *PDRB* which affect significantly the scores of *UN*.

Based on the table 1. The model which is used to predict the estimation of the scores of *UN* in a certain school *i* and at a certain year *t.*

$$\widehat{UN}_{it} = 136{,}8 + 0{,}09\text{EDS} + 0{,}07\text{FSS} - 1{,}02\text{US} - 0{,}02\text{IPM} - 0{,}14\text{PDRB} + 0{,}19\text{TAS} - 0{,}004\text{CSS} - 0{,}06\text{TPS} + \boldsymbol{u}_i + \boldsymbol{v}_{it}$$

The estimation result by modelling the data for the social science diciplines which is also analysed using R software, package 'lme4' represents as follows

Table 2. Modelling in Random Effect Model for Social Science Major

Parameter	Estimate	Std.Error	t.value	p.z
(Intercept)	95.5113357	11.5978112	8.2352898	2.220446e-16
rata_US	-0.6946077	0.09168786	-7.5757874	3.574918e-14
akre	0.22027012	0.15009680	1.4675204	1.422345e-01
ipm	0.05405338	0.12738235	0.4243396	6.713182e-01
pdrb	-0.13995148	0.02537485	-5.5153623	3.480622e-08
ISI	-0.02559474	0.07096874	-0.3606481	7.183625e-01
PRS	-0.01861390	0.05514847	-0.3375234	7.357224e-01
TEN	0.09839509	0.05501004	1.7886751	7.366715e-02
SAR	0.04780935	0.04703454	1.0164731	3.094041e-01

Table 2. shows that the estimated value of the parameters in linear mixed model. Judging by the software output, it seems that there are only the scores of Educator, *US* and *PDRB* which affect significantly the scores of *UN*.

Based on the table 2. The model which is used to predict the estimation of the scores of *UN* in a certain school *i* and at a certain year *t.*

$$\widehat{UN}_{it} = 95{,}5 + 0{,}1\text{EDS} + 0{,}05\text{FSS} - 0{,}7\text{US} + 0{,}05\text{IPM} - 0{,}14\text{PDRB} + 0{,}22\text{TAS} - 0{,}03\text{CSS} - 0{,}01\text{TPS} + \boldsymbol{u}_i + \boldsymbol{v}_{it}$$

Modelling With GEE

Modelling with GEE estimation approach does not have coefficient or variable which represents the random effect as school effect ($\boldsymbol{u}_i$) like either in the fixed effect model or random effect model.

Model which is obtained by using GEE in this research converts the school effect into the form of dummy variables. The number of dummy variables is k -1, where k is the number of schools which have been observed that can be seen in the table 3. Those variables are included into the model.

Table 3. Dummy Variables For Each School

School	D1	D2	D3	D4	D5	D6	D7	...	D420
Sch1	1	0	0	0	0	0	0	...	0
Sch2	0	1	0	0	0	0	0	...	0
...	...	...	...	...	...	...	...	...	...
Sch420	0	0	0	0	0	0	0	...	1
Sch421	0	0	0	0	0	0	0	...	0

The result of estimation with GEE for the data of natural science and social science can be seen in the table 4 and table 5, working correlation matrix which is used this model estimation is autoregressive because of the panel data which is analysed in this research is the data that the research objects were observed at several time points.

The estimation result by modelling the data for the natural science diciplines with GEE which is analysed using SAS software represents as follows.

Table 4. Dummy Variables For Each School

Parameter		Estimate	StdError	Limits		Z	Pr > |Z|
Intercept		1884.384	38.6640	1808.604	1960.164	48.74	<.0001
no	1	88.3460	2.6646	83.1236	93.5684	33.16	<.0001
no	2	89.6727	3.0777	83.6405	95.7048	29.14	<.0001
...	...	...	...	...	...	...	...
no	420	0.7095	5.5203	-10.1101	11.5291	0.13	0.8977
no	421	0.0000	0.0000	0.0000	0.0000	.	.
rata_US		-0.2257	0.1032	-0.4280	-0.0235	-2.19	0.0287
akre		-0.1197	0.2182	-0.5474	0.3079	-0.55	0.5832
ipm		-24.9299	0.5400	-25.9882	-23.8715	-46.17	<.0001
pdrb_per_kapita		0.4479	0.0655	0.3195	0.5764	6.84	<.0001
ISI		-0.2715	0.1686	-0.6019	0.0589	-1.61	0.1073
PRS		0.1193	0.1270	-0.1296	0.3682	0.94	0.3476
TEN		0.2425	0.1861	-0.1222	0.6073	1.30	0.1925
SAR		0.2155	0.1457	-0.0700	0.5010	1.48	0.1389

GEE yields the form of model as follows.

$$\widehat{UN}_{it} = 1884{,}4 + 88{,}3D1 + 89{,}7D2 + \cdots + 0{,}71D420 + 0{,}24\text{EDS} - 0{,}21\text{FSS} - 0{,}23\text{US} - 24{,}9\text{IPM} + 0{,}45\text{PDRB} - 0{,}11\text{TAS} - 0{,}27\text{CSS} - 0{,}12\text{TPS} + \boldsymbol{v}_{it}$$

There are only 419 public schools which have the specific diciplines like social science. The estimation result by modelling the data for the social science diciplines with GEE which is also analysed using SAS software represents as follows.

Table 5. Dummy Variables For Each School

Parameter		Estimate	StdError	95% confLimits		Z	Pr > |Z|
Intercept		1061.450	62.524	938.906	1183.99	16.98	<.0001
no	1	80.9964	3.2585	74.6099	87.3828	24.86	<.0001
no	2	82.2968	3.2588	75.9097	88.6839	25.25	<.0001
...	...	...	...	...	...	...	...
no	418	-0.1974	0.3697	-0.9220	0.5272	-0.53	0.5935
no	419	0.0000	0.0000	0.0000	0.0000	.	.
rata_US		-0.3822	0.0923	-0.5631	-0.2013	-4.14	<.0001
akre		0.2354	0.1679	-0.0936	0.5645	1.40	0.1607
ipm		-13.34	0.9180	-15.143	-11.544	-14.54	<.0001
pdrb		-0.797	0.1575	-1.1056	-0.4882	-5.06	<.0001
ISI		-0.102	0.0908	-0.2800	0.0758	-1.12	0.2608
PRS		0.0166	0.0721	-0.1247	0.1579	0.23	0.8180
TEN		0.0207	0.0739	-0.1242	0.1655	0.28	0.7797
SAR		0.0384	0.0583	-0.0759	0.1527	0.66	0.5103

GEE yields the form of model as follows.

$$\widehat{UN}_{it} = 1061{,}45 + 80{,}99D1 + 82{,}3D2 + \cdots - 0{,}19D418 + 0{,}02\text{EDS} + 0{,}04\text{FSS} - 0{,}38\text{US} - 13{,}34\text{IPM} - 0{,}79\text{PDRB} + 0{,}24\text{TAS} - 0{,}1\text{CSS} + 0{,}02\text{TPS} + \boldsymbol{v}_{it}$$

Model Comparison

The comparison between the first model which is built with linear mixed model and the second model with GEE has been done in seeing its value of R-square and MSE. For the natural science diciplines, R-square of the first model is 76,7 % and its MSE value is 1,46 however R-square of the second model is 84,5% and its MSE value is 1,06. For the social science diciplines, R-square of the first model is 72,3 % and its MSE value is 2,12 however R-square of the second model is 80,92% and its MSE value is 1,9. It means that GEE is the best model that can be implemented for this research case of UN Both for natural science diciplines and social science diciplines.

Conclusion

It was concluded from this research that the *UN* scores could be modeled using the GEE approaches. Factors which were significantly affecting the score include the average scores of US, accreditation scores, and *IPM*. This model could be used for predicting future results of *UN* as well as for a basis of improvement of school learning activities

REFERENCES

Ahmad A. (2011). *Analisis Hubungan Pengeluaran Pendidikan Dan Pertumbuhan Ekonomi Dengan Menggunakan Pendekatan Kausalitas Granger.* Jurnal Ekonomi & Pendidikan, Volume 8 Nomor 2, November 2011

[BPS] Badan Pusat Statistik. (2009). *Toward A New Concencus: Democracy and Human Development Report.*

[BPS] Badan Pusat Statistik. (2008). *Index Pembangunan Manusia 2007-2008*. Jakarta :BPS.

Frees EW. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge University Press, New York.

Hsiao C. (2003). *Analysis of Panel Data (2nd edition).* Cambridge University Press, New York.

Henderson CR, Kempthorne O, Searle SR, Von KCN. (1959). *Estimationof environmental and genetic trends from records subject to culling*. Biometrics.

Jiang J. (2007). *Linear, Generalized Linear Mixed Models and Their Applications.* New York : Springer.

Juanda B, Junaidi. (2012). *Ekonometrika Deret Waktu*. Bogor : IPB Press.

Kurnia A. (2000). *Pendekatan GEE dan Quasi-Likelihood Dalam Generalized Linear Mixed Models* [Tesis]. Bogor (ID): Institut Pertanian Bogor.

Longford NT. (1993). *Random coefficient models*. Oxford: Oxford University Press.

Mongi CE. (2014). *Penggerombolan Dan Pemetaan Kabupaten/Kota Di Provinsi Jawa Barat Berdasarkan Nilai Ujian Nasional SMA Dan Akreditasi Sekolah* [Tesis]. Bogor (ID): Institut Pertanian Bogor.

Permana AY. (2012). *Analisis Pengaruh PDRB, Pengangguran, Pendidikan, dan Kesehatan Terhadap Kemiskinan di Jawa Tengah 2004-2009* [Skripsi]. Semarang (ID): Universitas Dipenogoro.

Rao CR. (1973). *Linear Statistical inference and Its Applications.* Second Edition.New York: Wiley.

Searle SR, Casella G, McCulloch CE. (1992). *Variance Components.* New York: Wiley.

Sinaga SS. (2011). *The Uses of Canonical Analysis to Know The Pattern of Relationship among Scores Of National Exam , Scores of School Exam , and Progress Report (Case Study at SMA Budhi Warman II Jakarta)* [Tesis]. Bogor (ID): Institut Pertanian Bogor.

Yulianingsih KA. (2012). *Penerapan Regresi Poissom Untuk Mengetahui Faktor Yang Mempengaruhi Ketidaklulusan Siswa SMA/SSMK di Bali.* E-Jurnal Matematika Vol. 1

No. 1 Agustus 2012, 59-63.

Zeger SL, Liang KY. (1986). *Longitudinal Data Analysis For Discrete And Continous Outcomes*. Biometrics 42, 121-130.

Zeger SL, Liang KY,& Albert PS. (1988). *Models for Longitudinal Data: A Generalized Estimating Equation Approach.* Biometrics 44, 1049-1060.