BAREKENG: Jurnal Ilmu Matematika dan Terapan
June 2022 Volume 16 Issue 2 Page 607–614
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss2pp607-614

RESTRICTED MAXIMUM LIKELIHOOD ESTIMATION FOR MULTIVARIATE LINEAR MIXED MODEL IN ANALYZING PISA DATA FOR INDONESIAN STUDENTS

Vera Maya Santi [1], Khairil Anwar Notodiputro[2*], Indahwati[3], Bagus Sartono[4]

[1]*Statistics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Negeri Jakarta*
Rawamangun Muka St., Pulo Gadung, Jakarta, 13220, Indonesia

[1,2,3,4]*Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University*
Dramaga St., Bogor, 16680, Indonesia

Corresponding author's e-mail: [2] khairil@apps.ipb.ac.id*

Abstract. *The Program for International Student Assessment (PISA), becomes one of the references or indicators used to assess the development of students' knowledge and skills in each member country of the Organization for Economic Cooperation and Development (OECD). The results of the PISA survey in 2018 placed Indonesia in the bottom 10, indicating that the implementation of the national education system has not been successful. This underlies the need for a more in-depth study of the factors that influence PISA data scores not only statistically qualitatively but also quantitatively which is still very rarely done. The data structure of the PISA survey results is complex, which involves multicollinearity, multivariate response variables, and random effects. Thus, it requires an appropriate statistical analysis method such as the multivariate mixed linear regression (MLMM) model. In this study, secondary data from the results of the 2018 PISA survey with Indonesian students as the smallest unit of observation were used as sample. School is used as an intercept random effect which is assumed to be normally distributed. Multicollinearity is overcome by selecting independent variables based on AIC and BIC values. Estimation of variance and random effect parameters was performed using the restricted maximum likelihood (REML) method. Based on the estimator of the variance of random effects for the response variables of mathematics, science, and reading literacy, it was obtained 1548.12, 1359.39, and 1082.48, respectively, which explains the significant effect of each school as a random effect on the three response variables.*

Keywords: *multicollinearity, MLMM, PISA, random effect, REML.*

Article info:

Submitted: 26th February 2022 *Accepted: 27th April 2022*

How to cite this article:
V. M. Santi, K. A. Notodiputro, Indahwati and B. Sartono, "RESTRICTED MAXIMUM LIKELIHOOD ESTIMATION FOR MULTIVARIATE LINEAR MIXED MODEL IN ANALYZING PISA DATA FOR INDONESIAN STUDENTS", *BAREKENG: J. Il. Mat. & Ter.*, vol. 16, iss. 2, pp. 607-614, June, 2022.

1. INTRODUCTION

Linear regression models with a single response variable frequently assume that all explanatory variables are fixed effects. However, in some cases, there are also those who assume the explanatory variable involved is a sample of the population of variables, so that this explanatory variable is said to be a random effect [1]. Linear regression models that include explanatory variables with fixed and random effects are referred to as linear mixed models (LMM).

Linear mixed regression models that involve more than one response variable are called Multivariate Linear Mixed Models. The random effect contained in the linear mixed model has a role to accommodate the possibility of clustering on the observed objects or the possibility of a correlation between the observed objects with one another in the same cluster [2]. Multivariate Linear mixed model can be applied to analyze fairly complex data in various fields such as biology, ecology, medicine, pharmaceutical science, and education.

Many studies on the application of Multivariate Linear Mixed Models have been conducted, including Jaffa, who used MLMM in the medical field to examine kidney function by involving three response variables, i.e., the average blood urea nitrogen content, the average serum creatinine content, and the glomerular filtration rate [3]. Gebregziabher et al. estimate a person's total health costs by involving the response variables for treatment costs, hospitalization costs, and outpatient costs [4]. Jaffa et al. conducted a study to identify cardiovascular risk in type-I diabetes diabetic patients [3]. Oskrochi et al. also applied the same model to examine shoulder complexity in breast cancer patients, involving four response variables in the form of muscle activity in the shoulder measured using electromyography (EMG) [5].

Multivariate Linear Mixed Models (MLMM) can be used to analyze data collected by the Program for International Student Assessment (PISA). This program has been established by the Organization for Economic Cooperation and Development (OECD) to evaluate the development of knowledge and skills of students aged around 15 years in a number of countries in the world that are OECD members [6]. PISA has been held regularly every three years since 2000 [7]. Currently, the PISA survey has been followed by 79 countries, including Indonesia, which has been part of PISA since 2000. This survey produces quite complex data sets involving many explanatory variables, several response variables, and even random effects. The need for statistical methods that can be used to analyze such complex data is real.

In the research on PISA conducted by Pakpahan [8] and Santi et al. [9], only one response variable was used and all explanatory variables were assumed to be fix, no random effects were involved in the analysis. Furthermore, Pakpahan's findings showed that 22 factors had a significant effect on the mathematical literacy score [8]. Meanwhile, Santi et al. [9] produced 11 factors that significantly influenced the scientific literacy score. Santi et al. [10] modeled PISA data using the Generalized Linear Mixed Model (GLMM) involving random effects on univariate response variables. Until now, studies on quantitative PISA data scores have been extremely rare. A more in-depth statistical analysis of PISA data scores involving multivariate response variables and random effects has also never been done. Therefore, this study uses three response variables simultaneously, which are the three scores on PISA, i.e., reading, math, and science literacy scores, and involves the school effect of each student, which is assumed to be a random effect using Multivariate Linear Mixed Models estimated through REML technique.

2. RESEARCH METHODS

2.1 Data

The target in the PISA program is students aged between 15 to 16 years or students who are nearing the end of compulsory education. According to the OECD, these students have acquired the knowledge and skills necessary to participate in modern society. Students' knowledge and skills are measured through the subject of the PISA instrument, which consists of science, mathematics, and reading literacies regardless of the curriculum system. The data used in this research are students aged around 15 to 16 years who are randomly selected through random sampling from the PISA [11]. According to Bluman, if it is found that the correlation value between variables is greater than 0.8, then this indicates the existence of multicollinearity [12]. One of the efforts made to overcome this multicollinearity is by excluding one of the correlated explanatory variables.

2.2 Multivariate Linear Mixed Model

The multivariate linear mixed model (MLMM) is a development of the linear mixed model (LMM) for modeling cases involving more than one response variable with multiple normal distributions [7], [13]. The following is the data structure for the Multivariate linear mixed model.

Table 1. MLMM Data Structure

Observation	Response (Y)			Explanatory (X)		
(i)	Y_1	…	Y_m	X_1	…	X_p
1	y_{11}	⋯	y_{1m}	x_{11}	⋯	x_{1p}
…	⋯	⋯	⋯	⋯	⋯	⋯
…	⋯	⋯	⋯	⋯	⋯	⋯
n	y_{n1}	⋯	y_{nm}	x_{n1}	⋯	x_{np}

Where i is the index for the observations, $i = 1, 2, \dots, n$; j is the index for the explanatory variable of constant effect, $j = 1, 2, \dots, p$; and k is the index for the response variable, $k = 1, 2, \dots, m$; where l index for random effect group, $l = 1, 2, \dots, q$. p is the number of explanatory variables, m is the number of multiple response variables, n is the number of observations, and q is the number of random effect groups. The multivariate linear mixed model can be written as follows [14][15].

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{Y}$ is a matrix of multiple response variables measuring $n \times m$, $\boldsymbol{X}$ is a design matrix of explanatory variables measuring $n \times (p + 1)$, $\boldsymbol{\beta}$ a fixed effect parameter matrix measuring $(p + 1) \times m$, $\boldsymbol{Z}$ is a random effect design matrix measuring $n \times q$, $\boldsymbol{u}$ is random effect group matrix $q \times m$ dan $\boldsymbol{e}$ is error or error matrix measuring $n \times m$. When translated into the form of a matrix, the following is obtained:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{e} = \begin{bmatrix} e_{11} & \cdots & e_{1m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{bmatrix}, \text{ and } \boldsymbol{Z} = \begin{bmatrix} z_{11} & \cdots & z_{1q} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nq} \end{bmatrix} \tag{2}$$

$\boldsymbol{X}$ is a matrix of explanatory variables measuring $n \times (p + 1)$, $\boldsymbol{e}$ is an error component matrix measuring $n \times m$ which is assumed to spread normal multivariate $\boldsymbol{e}$ ~ N($\mathbf{0}$, $\boldsymbol{\Sigma}$), and $\boldsymbol{Z}$ is a random effect design matrix measuring $n \times q$ which contains values 0 and 1, with a value of 1 for groups (clusters) of random effect which are the original group of observations, while 0 for other random effect groups which are not the original group of observations. Then, $\boldsymbol{\beta}$ is a fixed effect parameter matrix measuring $(p + 1) \times m$, and $\boldsymbol{u}$ is a random effect matrix measuring $q \times m$ as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{01} & \cdots & \beta_{0m} \\ \vdots & \ddots & \vdots \\ \beta_{p1} & \cdots & \beta_{pm} \end{bmatrix} \text{ and } \boldsymbol{u} = \begin{bmatrix} u_{11} & \cdots & u_{1m} \\ \vdots & \ddots & \vdots \\ u_{q1} & \cdots & u_{qm} \end{bmatrix} \tag{3}$$

If $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are expressed as vectors, they become as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{01} \\ \vdots \\ \beta_{p1} \\ \vdots \\ \beta_{0m} \\ \vdots \\ \beta_{pm} \end{bmatrix} \text{ and } \boldsymbol{u} = \begin{bmatrix} u_{11} \\ \vdots \\ u_{q1} \\ \vdots \\ u_{1m} \\ \vdots \\ u_{qm} \end{bmatrix} \tag{4}$$

where p is the number of fixed influence parameters, and q is the number of random effect clusters. The random effect on MLMM in this study is assumed to have a double normal distribution, $\boldsymbol{u}$ ~ *iid* MVN ($\mathbf{0}$, $\mathbf{D}$), where $\boldsymbol{u}_{lk}$ is the l-th random effect on the $k - th$ multiple response variable. If $i = 1, 2, \dots, n$ is the number

of observations and $k = 1, 2, \ldots, m$ is the number of response variables, the multiple response variable becomes the following:

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{bmatrix} \sim MVN(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{V}) \quad (5)$$

When written in vector form it becomes

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{n1} \\ \vdots \\ y_{1m} \\ \vdots \\ y_{nm} \end{bmatrix} \quad (6)$$

where $\boldsymbol{y_{ik}}$ is the i-th individual observation value for the k-th response variable. In MLMM, the random effect $\mathbf{u}$ has a normal multivariate distribution and the distribution of error $\boldsymbol{e}$ is also a normal multivariate spread, both of which are assumed to be independent. Thus, the multiple response variable $\boldsymbol{Y}$ has a normal multivariate distribution with a value of $\mathrm{E}(\boldsymbol{y}) = \mathbf{X}\boldsymbol{\beta}$, $\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{V} = \mathbf{Z}\boldsymbol{D}\boldsymbol{Z}' + \boldsymbol{\Sigma}$ is a matrix of variance, $\boldsymbol{y}$ and is required to be a positive definite. However, $\boldsymbol{D}$ and $\boldsymbol{\Sigma}$ matrices are not required to be positive definite matrix and $\mathbf{y}|\mathbf{u} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \boldsymbol{e})$.

The thing that distinguishes random effects from fixed effects is that information on random effects is only limited to the value of variance, while the permanent effect informs the value of the model parameter coefficients [15]. The purpose of adding random effects to multivariate linear mixed models is to accommodate the possibility of correlation between response variables. If the estimated variance of the random effect has a very small value or is close to zero, this indicates that the random effect in the model is not significant. Then, these multiple variable mixed linear models will approach the ordinary multiple variable linear model.

2.3 Restricted Maximum Likelihood (REML)

According to McCulloch and Searle, this REML method is a modification of ML by transforming the response variable vector $\mathbf{y}$ into $\mathbf{a}'\mathbf{y}$ where $\mathbf{a}'\mathbf{y}$ does not contain $\boldsymbol{\beta}$ fixed effect estimation result which means $\mathbf{a}'$ will result $\mathbf{a}'\mathbf{X} = 0$ [15]. If $\boldsymbol{A} = [\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_{(N-rank(X))}}]$ are independent with $\mathbf{A}'\mathbf{X} = \mathbf{0}$. Thus, if written as $\mathbf{A}'\mathbf{Y} \sim MVN(\mathbf{0}, \mathbf{A}'\mathbf{V}\mathbf{A})$, it has a density function opportunity [16].

$$f_{REML}(\boldsymbol{Y}) = (2\pi)^{-\frac{m}{2}} |\mathbf{A}'\mathbf{V}\mathbf{A}|^{-\frac{1}{2}} exp\left[-\frac{1}{2}([\boldsymbol{A'Y}]'\mathbf{A}'\mathbf{V}\mathbf{A}^{-1}[\boldsymbol{A'Y}])\right] \quad (7)$$

The form of the maximum likelihood and log-likelihood functions is as follows

$$L_{REML}(\theta) = \prod_{i=1}^{n-rank(x)} (2\pi)^{-\frac{m}{2}} |\boldsymbol{A'V_iA}|^{-\frac{1}{2}} exp\left[-\frac{1}{2}([\boldsymbol{A'Y_i}]'(\mathbf{A}'\mathbf{V_i}\boldsymbol{A})^{-1}[\boldsymbol{A'Y_i}])\right]$$

$$\ln L_{REML}(\theta) = -\frac{1}{2}\sum_{i=1}^{n-rank(x)}([\boldsymbol{A'Y_i}]'(\mathbf{A}'\mathbf{V_i}\boldsymbol{A})^{-1}[\boldsymbol{A'Y_i}]) - \frac{1}{2}\sum_{i=1}^{n-rank(x)} \ln|\boldsymbol{A'V_iA}| - c \quad (8)$$

with $c = \frac{m(n-rank(x))}{2}\ln(2\pi)$, estimation of variance is obtained by deriving equation 8 with respect to $\boldsymbol{\varphi_k}$ which is the *k*th element in the covariance matrix $\mathbf{V}$, with $k = 1, 2, \ldots, q$ [16], obtaining below:

$$\frac{\partial l}{\partial \boldsymbol{\varphi_k}} = \frac{1}{2}\left[(\boldsymbol{Y})'\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \boldsymbol{\varphi_k}}\boldsymbol{P}(\boldsymbol{Y}) - \mathrm{tr}\left(\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \boldsymbol{\varphi_k}}\right)\right] \quad (9)$$

with

$$\boldsymbol{P} = \mathbf{A}(\mathbf{A}'\mathbf{V}\mathbf{A})^{-1}\mathbf{A}' \quad (10)$$

In estimating parameters, the form of the log-likelihood function is not simple, so it cannot be easily evaluated. Therefore, a numerical iteration algorithm is used, i.e., Newton Raphson iteration.

2.4 Data Analysis Procedure

The following are the steps of data analysis used in this study:

1. Data input and cleaning.
2. Data exploration.
3. Building MLMM to model the three response variables simultaneously (literacy scores of reading, mathematics, and science) with the explanatory variables.
4. Estimating MLMM fixed effect parameters and variance using REML and Newton Raphson numerical iteration approach.
5. Testing the significance of the effect of explanatory variables on the three responses, literacy scores of math, science, and reading simultaneously using the Wald-test and partially using the t-test.
6. Interpreting the explanatory variables that have a significant effect on the three response variables simultaneously.

3. RESULTS AND DISCUSSION

The data used in this study is secondary data derived from survey results from the Organization for Economic Cooperation and Development (OECD) with its program, PISA. In the PISA survey, the reachable population of students aged around 15 years from 79 OECD member countries is around 31 million students. The sample in this study was 1,500 Indonesian students who were taken by the target population of all Indonesian students who took part in the PISA survey in 2018. In the sampling, multi-stage sampling method was used with stratified random sampling for school samples [17]. Samples of student observation units were taken by random sampling from each school, i.e., students aged around 15 years or nearing the end of compulsory education [11]. Students with the criteria for the age of 15-16 years include students who are currently studying in Junior High School (SMP) and Senior High School (SMA). Therefore, these are students in grade 7, grade 8, grade 9, grade 10, grade 11, and grade 12. The three Program of International Student Assessment (PISA) scores used are math literacy, scientific literacy, and reading literacy scores.

3.1. Analysis of the Relationship between Response Variables and Explanatory Variables

The relationship between the response variables used, i.e., the three PISA scores, must be examined to determine whether or not there is a significant relationship between the three response variables. If the relationship between the three response variables is not significant, it will result in the results of the multivariate analysis being relatively the same as the results obtained by univariate analysis.

Table 2. Descriptive Data and Relationships between Variables

Variables	Mean	Standard Deviation	Correlation coefficient			N
			Mathematics	Science	Reading	
Mathematics	409.70	75.28	1.00			1500
Science	396.90	75.64	0.87**	1.00		1500
Reading	421.15	67.42	0.84**	0.89**	1.00	1500

(**) Significant Correlation

Based on the results of the Pearson correlation test, it was found that the three response variables had a significant correlation coefficient and were classified as strong because the coefficient value was above 0.8 and the relationship was positive. Then, based on the scatter diagram between the response variables in pairs, it can be seen that the distribution of observations tends to spread and form a positive linear pattern. This indicates that there is a relatively strong relationship between the three response variables used and the relationship is positive, as shown in Figure 1 below. Therefore, the use of multivariate analysis can be carried out because of the strong correlation between the three response variables used.

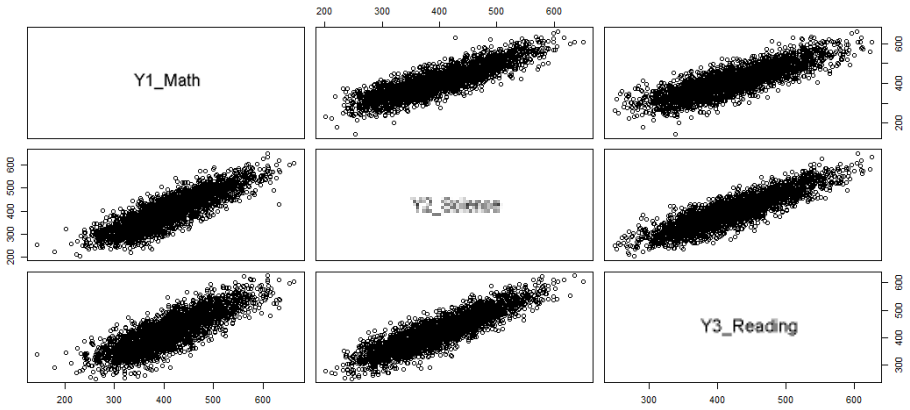

Figure 1. Scatter diagram between response variables

Based on checking the correlation between the explanatory variables, because all explanatory variables are of categorical type, the polychoric correlation is used. From 27 explanatory variables used in the model, it is found that there are explanatory variables that are strongly correlated with each other with a correlation coefficient value of more than 0.8. This indicates that there is multicollinearity in the explanatory variables that are strongly correlated, i.e., the existence of a computer (X_8) with educational software (X_9) of 0.80, the existence of a computer (X_8) with many computers (X_{16}) which has a correlation value of 0.86, repeating a class during elementary school (X_{24}) with repeating a class during junior high school (X_{25}) of 0.88, and repeating a class during high school (X_{26}) repeating a class during junior high school (X_{25}) of 0.83. Furthermore, two model specifications were formed based on the explanatory variables that were strongly correlated above, the first model, i.e., a model that did not involve the explanatory variables X_8 and X_{25}, and the second model, i.e., a model that did not involve X_8, X_{24}, and X_{26}. From the two models, the best model was selected based on the feasibility values of the AIC and BIC models as shown in Table 3 below.

Table 3. Selection of Independent Variables Based on Comparison of Models

Models	AIC	BIC
Model 1	– 1765.14	– 408.27
Model 2	– 1771.23	– 461.85

Based on the results of AIC and BIC obtained from both models, model 2 has smaller AIC and BIC values than model 1, and also model 2 has fewer model parameters because it uses 24 explanatory variables compared to the first model with 25 explanatory variables. Thus, model 2 is simpler than model 1. Therefore, the second model is the best model obtained based on the eligibility criteria of the model used.

3.2. Multivariate Linear Mixed Model

Before modeling the three PISA scores, i.e., the scores of mathematical literacy, scientific literacy, and reading literacy which are assumed to have a normal multivariate distribution, they are checked using the Doornik Hansen test as shown in Figure 2 below:

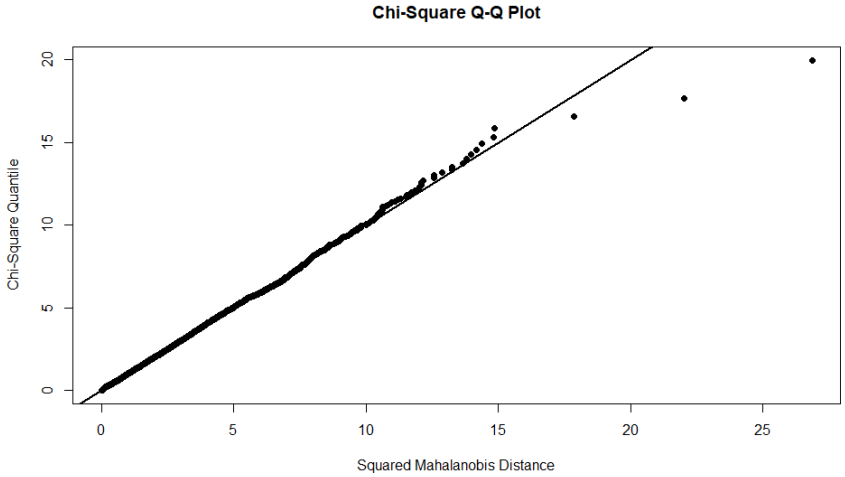

Figure 2. Quantile diagram of Response Variables

The results of testing the distribution of the three response variables obtained a calculated DH value of 11.56 with a p-value of 0.07, which means that the three response variables used have a normal multivariate distribution. Therefore, the specification of the model used is Multivariate Linear Mixed Models (MLMM) with schools that are used as a random effect on the model, which is assumed to be normally distributed.

The results of the simultaneous parameter significance test using the Wald chi-square test. The results of the chi-square count are 56791.8 with a p-value of 0.00 <0.05, so that it can be said that there is a significant effect of all model parameters simultaneously on the three response variables, including father's education, internet access, facilities at home, and the age of entering kindergarten (TK). These findings are in line with research conducted by Pakpahan [8], Santi et al. [9], and Santi et al. [10].

Table 4. Estimation of Varieties

	R (Residual Variance)			D (Random Variance)		
	Y1_Math	**Y2_ Science**	**Y3_ Reading**	**Y1_Mat**	**Y2_ Science**	**Y3_ Reading**
Y1_Math	2227.08			1548.12		
Y2_ Science	1610.15	2308.11		1424.41	1359.39	
Y3_ Reading	1296.81	1609.04	1886.37	1296.14	1209.15	1082.48

Based on Table 4 above, the estimation of the variance of random effects obtained from the three responses, the literacy scores of mathematics, science, and reading are 1548.12, 1359.39, and 1082.48, respectively, which involved a random effect in the form of 389 schools as well as the value of the variance of both the random effect and the residual variance that was not equal to zero indicating a relationship between the three response variables used. The variance value is relatively significant or not equal to zero, thus indicating that the variance between schools is significant. These results are in line with the previous assumption that the independence of the three response variables originating from the same school is a part that needs to be included in the model to avoid bias in the estimation of standard errors in parameter estimation. Based on the results of the estimation of the variance of random effects, it also explains that there is a significant difference in the effect of each school on students' mathematical literacy scores. Then, for reading literacy scores, the differences in schools have a small effect compared to the other two response variables. This is because reading literacy or students' reading ability is relatively a basic ability from within a student. Therefore, it does not depend on the origin of the student's school, whether the school is in the superior category or not, and vice versa for math and science abilities at each school.

4. CONCLUSIONS

The multivariate analysis model involving random effects, i.e., the Multivariate Linear Mixed Models (MLMM), can be said to be suitable for modeling PISA data, which involves three response variables, the scores of mathematical literacy, scientific literacy, and reading literacy, and involves a random effect in the form of students' schools. From testing the significance of the parameters for the three scores, it was simultaneously concluded that all the explanatory variables had a significant effect on the three PISA scores,

including father's education, internet access, facilities at home, and the age of entering kindergarten (TK). Based on the estimation of variance, it was found that school as a random effect has a significant influence on reading and science literacies.

AKNOWLEDGEMENT

The authors extend their gratitude to Mr. Khairil Anwar Notodiputro, Mrs. Indahwati, and Mr. Bagus Sartono for their guidance and direction, and to Irsyad Hasari for their assistance.

REFERENCES

[1] P. McCullagh and J. A. Nelder, "Generalized Linear Models, Second Edition." p. 532, 1989.
[2] T. J. Hastie and D. Pregibon, "Generalized linear models," *Statistical Models in S*. pp. 195–247, 2017, doi: 10.1201/9780203738535.
[3] M. A. Jaffa, M. Gebregziabher, L. M. Luttrell, and A. A. Jaffa, "Multivariate Generalized Linear Mixed Models With Random Intercept To Analyze Cardivascular Risk Markers In Type-1 Diabetic Patient," vol. 43(8), p. p.1447-1464, 2016, doi: 10.1186/s12967-015-0557-2.
[4] M. Gebregziabher, Y. Zhao, C. . Dismuke, N. Axon, J. K. Hunt, and L. E. Egede, "Joint modeling of multiple longitudinal cost outcomes using multivariate generalized linear mixed models," 2018.
[5] G. Oskrochi, E. Lesaffre, Y. Oskrochi, and D. Shamley, "An application of the multivariate linear mixed model to the analysis of shoulder complexity in breast cancer patients," *International Journal of Environmental Research and Public Health*, vol. 13, no. 3, 2016, doi: 10.3390/ijerph13030274.
[6] I. Pratiwi, "Efek Program Pisa Terhadap Kurikulum Di Indonesia," vol. Vol. 4(1), p. 51, 2019.
[7] OECD, "PISA 2018 Assessment and Analytical Framework," 2019.
[8] R. Pakpahan, "Faktor - Faktor Yang Memengaruhi Capaian Literasi Matematika Siswa Indonesia Dalam PISA 2012 Factors Affecting Literacy Mathematics Achievement Of Indonesian Student In PISA 2012," vol. 1, 2016.
[9] V. M. Santi, K. A. Notodiputro, and B. Sartono, "Variable selection methods applied to the mathematics scores of Indonesian students based on convex penalized likelihood," *J. Phys. Conf. Ser.*, vol. 1402, no. 7, 2019, doi: 10.1088/1742-6596/1402/7/077096.
[10] V. M. Santi, K. A. Notodiputro, and B. Sartono, "Generalized Linear Mixed Models by Penalized Lasso in Modelling The Score of Indonesian Students," *Journal of physics: Conference Series AASEC, August* 8, 2021.
[11] S. Breakspear, "How does PISA shape education policy making? Why how we measure learning determines what counts in education," *Cent. Strateg. Educ. Semin. Ser.*, no. 240, p. 16, 2014.
[12] A. G. Bluman, *Elementary Statistics: A Step by Step Approach*, Eighth Edi. New York, 2012.
[13] D. M. Berridge and R. Crouchley, *Multivariate generalized linear models using R*, vol. 39, no. 8. Taylor & Francis Grup, 2011.
[14] I. N. Latra, S. Linuwih, Purhadi, and Suhartono, "Estimation for Multivariate Linear Models," *Linear Model Theory*, vol. 10, no. 06, pp. 243–262, 2010, doi: 10.1002/9780470052143.ch12.
[15] C. E. McCulloch and S. R. Searle, *Generalized, Linear and, Mixed Models*. New York: John Wiley and Sons, Inc., 2001.
[16] J. Jiang and T. Nguyen, *Linear and Generalized Linear Mixed Models and Their Applications*, vol. 50, no. 1. Springer, 2021.
[17] OECD, "Sampling in PISA," no. March 2016, pp. 1–14, 2016.