View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Jurnal Natural

Jurnal Natural

Vol. 21, (2) 2021
DOI 10.24815/jn.v21i2.19325

pISSN 1411-8513
eISSN 2541-4062

ORIGINAL RESEARCH

A generalized linear mixed model for understanding determinant factors of student's interest in pursuing bachelor's degree at Universitas Syiah Kuala

ASEP RUSYANA[1,2], KHAIRIL ANWAR NOTODIPUTRO[2*], BAGUS SARTONO[2]

[1]Department of Statistics, Universitas Syiah Kuala, Banda Aceh, Indonesia

[2]Department of Statistics, IPB University, Bogor, Indonesia

Abstract. Generalized Linear Mixed Model (GLMM) is a framework that has a response variable, fixed effects, and random effects. The response variable comes from an exponential family, whereas random effects have a normal distribution. Estimating parameters can be calculated using the maximum likelihood method using the Laplace approach or the Gauss-Hermite Quadrature (GHQ) approach. The purpose of this study was to identify factors that trigger student's interest to continue studying at Universitas Syiah Kuala (USK) using both techniques. The GLMM is suitable for the data because the variable response has a Bernoulli distribution, and the random effects are assumed to be having a normal distribution. Also, the model helps identify the relationship between the dependent variable and the predictors. This study utilizes data from six high schools in Banda Aceh city drawn using a two-stage sampling technique. Stage 1, we randomly chose six out of sixteen public senior high schools in Banda Aceh. Stage 2, we selected students from each school from four different major classes. The GLMM model includes one binary response variable, five numerical fixed-effects, and two random effects. The response variable is the interest of high school students to continue study at USK (yes or no). The five fixed effects in the model including scores of collaboration (C), Action (A), Emotion (E), Purposes (P), and Hope (H). Finally, the random effects are schools (S) and majors (M). In this study, both Laplace and GHQ techniques produce identical results. The predictors that can explain student interest are A, E, and H. These predictors have a positive effect. The random effects of schools and majors are not significantly different from zero. The model with three significant predictors is better than the complete predictor model.

Keywords: Gauss-Hermite Quadrature, GLMM, Laplace, student interest, Universitas Syiah Kuala

INTRODUCTION

The General Linear Mixed Model (GLMM) is a valuable framework for comparing how several variables affect different continuous variables [1]. The Generalized Linear Mixed Model (GLMM) was developed as a combination of Linear Mixed Model (LMM) and Generalized Linear Model (GLM) [2]. LMM has a Gaussian response variable with predictors of random effect, whereas GLM has a response variable having an exponential family. As a result, GLMM is a model that contains a response variable from the exponential family, fixed effects component, and Gaussian random effects [3]. Examples of exponential family distribution are Bernoulli, Binomial, Poisson, Normal and Exponential distribution [4]. Therefore, the GLMM can use a response variable that is more flexible than the linear model, which can only use a gaussian response variable.

Estimation of parameters for the GLMM can be done using the Pseudo-Likelihood (PL) [5], hierarchical likelihood (h-likelihood) [3], the maximum likelihood method based on the Laplace approximation [6], and maximum-likelihood based on Adaptive Quadrature [7]. The PL technique utilizes Taylor Series to do the optimalization of the model. The h-likelihood is helpful in predict random effects, fixed effects, and dispersion parameters [8]. The integral of distribution for h-likelihood sometimes has no analytical closed-form, so that we need a numerical approximation. One of the numerical methods is Laplace approximation [3]. Some studies reveal that Laplace method might be implemented using

*Corresponding Author:
khairil@apps.ipb.ac.id

Received: January 2021| Revised: May 2021 | Accepted: June 2021

smaller sample than the Adaptive Quadrature. Moreover, the Laplace can solve the interaction of random effects, but it cannot solve the quadrature case. Laplace is more efficient than the quadrature algorithm with one node.

GLMM has been applied to real data since 1989. GLMM model of Salamander mating experiment with three fixed effects and two random effects was created [9]. The model was known as logistic linear with random effect. The response variable scores are 0 and 1 where 0 means the salamander mates and 1 means the salamander did not mate, so that the variable respons has distribution Bernoulli. The fixed effects are Whiteside Female (WS_f), Whiteside Male (WS_m), and interaction of WS_f and WS_m. The random effects are the male and the female effects. The h-likelihood ran well to estimate the parameters of the model. Next, the fitting of the model for seizure patients with three covariates has been carried out. The model was known as Poisson log-linear mixed model at that time [10]. The response variable was the seizure-patient count, which has Poisson distribution. The fixed effect was the treatment, and the random effects were the *kth* visiting and age. Further, there was a problem of overdispersion in the model. The model coefficient parameter was estimated by h-likelihood, Laplace and Quadrature with one node and twenty five nodes [3].

GLMM has been implemented in several other specific cases in further development, including a model with time-series data in 2011 using natural cubic splines (NS) [11]. We can also find the application of GLMM to identify triggers for resettlement to Australia with repeated measurement in the data. In the research, Generalized Estimating Equations (GEEs) were used to estimate the parameters of the model [12]. The GLMM with the Penalized Quasi-Likelihood, the maximum likelihood with Laplace and Adaptive Gaussian Quadrature (AGQ) approximations for estimating model has been applied to simulation data with high dimension in 2017. Based on this research, the Laplace approach provided better result [13]. Then, GLMM without random effects was also applied to identify factors that explain student's interests. Variables of networks, goals and expectations were significant in the research [14].

Universitas Syiah Kuala (USK) is an A-accredited university since 2015. It is a public university ranked as Public Service Agency since 2018, and was awarded as the best university in Aceh by the Ministry of Education and Culture. Currently, USK has 1,558 lecturers with 39 percent having doctoral degrees, 60 percent having master's degrees, and 5 percent or 81 people professors. The number of students is around 33,000 people spread over 13 faculties. USK has a target to pass qualified graduates for working or job creators [15]. Based on [16], the higher the supervisor's education level of the students is the faster the students finish their thesis.

Many factors influenced student's interest to pursue their bachelor degree in USK. In this study, we included five fixed effects and two random effects as covariates. The future students usually hope to have many friends in the university, so collaboration or network development becomes a fixed effect. The other covariate is the ability to act. It is about how students collect information about USK, and it may influence the students' interest. The third fixed effect is the emotional factor that measures the willingness degree to study at USK. The fourth is the purpose factor which consists of several questions, including: (1) USK is my favorite campus; (2) USK can help me achieve my goals; (3) USK makes me get the job I want. The last fixed effect is the expectation factor which includes: (1) I hope USK has complete facilities; (2) I hope that tuition fees are affordable; (3) I hope USK has good quality lecturers; (4) I hope USK is a place comfortable study; (5) I hope USK offers many scholarships. The GLMM also involves two random effects, namely the school origin and the major. The School origin was assigned as a random effect because the six school origins were randomly selected from sixteen schools in Banda Aceh. Meanwhile, the major was chosen as a random effect because Science and Social Sciences were chosen from three majors in a senior high school.

USK currently applied a fair single tuition fee system to its students where the payment of the tuition fee was based on the student's financial capacity, but in fact many students or their parents felt that the costs were too high. The tuition fee was measured by the type of occupation of the parents, family assets owned, the amount of monthly electricity payments and the number of family dependents. Besides that, students through the Joint Entrance Examination were charged a different development fee at the beginning of their study in each study program. Therefore, several students did not do re-registration after they were accepted because of the high costs that must be bought by some students. There were students who choose to go to other universities, both public and private in Aceh.

Two reasons why this research was necessary are firstly, finding an accurate model that can

solve data solutions with binary response variables, some fixed effects, and some random effects, secondly, applying the model to identify the factors that influence high school student's interest in continuing their studies at USK. Therefore, USK was hoped to obtain information as a basis for implementing policies. The right policy is important to implement to reduce prospective students who do not re-register because of the tuition fee policies and they will satisfy when studying at USK.

Based on this background, the purposes of this study is to apply the GLMM model with the Laplace and one node Gauss-Hermite Quadrature approach in identifying factors that can trigger the interest of prospective students to continue their studies at USK. Futhermore the model is evaluated to find a model with only factors that significantly influence the response and higher model fit.

METHODOLOGY

Data

USK funds collecting of the data through Senior Lector Research Grant Program. The data were taken at the State Senior High Schools (Shortened to SMAN in Indonesian which stands for *Sekolah Menengah Atas Negeri*) in Banda Aceh City by the student enumerator from the Statistics Study Program - USK. The period of data collection was carried out from April until May 2019. This data collection was part of the research grant activity at USK in 2019.

The schools in Banda Aceh are categorized into three groups by the Education and Culture Office of Aceh Province, namely favorite, middle and ordinary. The number of schools in respective groups are five, five, and six, so that there are 16 schools in total with the number of students as many as 5,366 students (see Table 1).

The sample of the respondents was drawn in two stages. At the first stage, six schools were taken from the sixteen schools using stratified random sampling. Two selected favorite school samples are SMAN 3 and SMAN 4 Banda Aceh; the middle school samples are SMAN 5 Banda Aceh and SMAN 8 Banda Aceh; and the ordinary schools are SMAN 14 Banda Aceh and SMAN 16 Banda Aceh.

At the second stage, one class XI and one class XII of Natural Science, and one class XI and one class XII of Social Science were selected from each school randomly. The reason for selecting these classes is that several students consist of potential students to pursue their study at university. Finally, all students in the classes chosen are as respondents. The numbers of SMAN 3 and SMAN 4 students are 111 and 120 people, respectively, so the respondents of the favorite schools are 231 people. The number of students from SMAN 5 Banda Aceh and SMAN 8 Banda Aceh is 101 and 89 people, respectively, so that the number of middle school respondents is 190 students. The number of respondents of ordinary schools is 95. Therefore, the total of the respondent is 516 students, see Table 2.

Table 1. The public senior high school (SMAN) in Banda Aceh

Names of School	Rank	N
SMAN 1 Banda Aceh	F	418
SMAN 2 Banda Aceh	F	422
SMAN 3 Banda Aceh	F	583
SMAN 4 Banda Aceh	F	531
SMAN 10 Banda Aceh	F	272
Sub Total		2.226
SMAN 5 Banda Aceh	M	445
SMAN 7 Banda Aceh	M	508
SMAN 8 Banda Aceh	M	456
SMAN 9 Banda Aceh	M	387
SMAN 11 Banda Aceh	M	381
Sub Total		2.177
SMAN 6 Banda Aceh	O	327
SMAN 12 Banda Aceh	O	341
SMAN 13 Banda Aceh	O	40
SMAN 14 Banda Aceh	O	79
SMAN 15 Banda Aceh	O	62
SMAN 16 Banda Aceh	O	114
Sub Total		963
TOTAL		5.366

Notes: F = Favorite, M = Middle, O = Ordinary
N = size of student population

Table 2. Samples of the schools, the number of students for each major and school

Names of Schools	Major in The Schools		Total (Students)
	Natural Science (Students)	Social Science (Students)	
SMAN 3 Banda Aceh	55	56	111
SMAN 4 Banda Aceh	62	58	120
Sub Total			231
SMAN 5 Banda Aceh	56	45	101
SMAN 8 Banda Aceh	56	33	89
Sub Total			190
SMAN 14 Banda Aceh	29	30	59
SMAN 16 Banda Aceh	20	16	36
Sub Total			95
TOTAL			516

There are six variables in this research. Y is response variable whereas C, A, E, P and H are predictors. The predictors have scores which are arranged from three to six questions, see Table 3. The first predictor is affiliation with other people (*C*) which is composed of five questions, for the question example (1) I like activities that require collaboration with other people, (2) I am satisfied when my friends appreciating my efforts when working together. The answer choices given to respondents are Strongly Disagree (SDA), Disagree (DA), Agree (A) and Strongly Agree (SA). The options of the answers were provided in the questionary for all individual questions of the predictors. Then the second predictor is action (A) which is composed of three questions, parts of the questions are (1) I always study hard to pursue study in USK, (2) I look for information about USK from books, magazines, newspapers, and the internet. Furthermore, the emotional attitude predictor (*E*) is a set of four questions including (1) continuing to study at USK is my current dream, and (2) I am always interested if someone talks about USK. Then, the goal predictor (*P*) consists of five questions, including (1) I believe USK can help me achieve my goals, and (2) I believe that after graduating from USK I will get a good job. Finally, the fifth predictor is hope or expectation (*H*), including (1) I hope USK has complete facilities, and (2) I hope USK has good quality lecturers.

Method

Steps of the research are*:*

1. collecting data,
2. identifying outliers which have influence,
3. building the GLMM models,
4. estimating parameters, standard error (SE), t and p-value using the maximum likelihood method through the Laplace approach and the Gauss-Hermite Quadrature approach [11],
5. evaluating model fit through -2 Log Likelihood, AIC, AICC, BIC, HQIC and CAIC,
6. determining the factors that affect the interest of high school students to continue studying at USK,
7. based on result of the fifth step, building a model involving only influential predictors and its statistic fit.

Step two until step four was carried out with SAS 9.4 software.

Table 3. Information of response variables Y and fixed effects

Variable	Label	Scale
Y	Whether a student is interested to pursue in the USK (yes/no)	Binary
C	Collaboration	Numeric
A	Action	Numeric
E	Emotion	Numeric
P	Purposes	Numeric
H	Hope/ Expectation	Numeric

The data structure of response variables is presented in Table 4. The number of respondents from SMAN 3 Banda Aceh (n_1), SMAN 4 Banda Aceh (n_2), SMAN 5 Banda Aceh (n_3), SMAN 8 Banda Aceh (n_4), SMAN 14 Banda Aceh (n_5), and SMAN 16 Banda Aceh (n_6) are 111, 120, 101, 89, 59 and 36 students. The fixed effects consist of *C*, *A*, *E*, *P* and *H* while the random effects consist of *S* and *M*.

Table 4. Data structure for response variable *Y*, fixed effects and random effects *S* and *M*

No Resp.	Y_{ik}	Fixed Effects			Random Effects	
		C_{ik}	…	H_{ik}	S_{ik}	$MS_{ij(k)}$
1	Y_{11}	C_{11}	…	H_{11}	S_{11}	$MS_{11(1)}$
2	Y_{21}	C_{21}	…	H_{21}	S_{21}	$MS_{21(1)}$
…	…	…	…	…	…	…
n_1	$Y_{n_1 1}$	$C_{n_1 1}$	⋯	$H_{n_1 1}$	$S_{n_1 1}$	$MS_{n_1 2(1)}$
…	…	…	…	…	…	…
1	Y_{16}	C_{16}	…	H_{16}	S_{16}	$MS_{11(6)}$
2	Y_{26}	C_{26}	…	H_{26}	S_{26}	$MS_{21(6)}$
…	…	…	…	…	…	…
n_6	$Y_{n_6 6}$	$C_{n_6 6}$	⋯	$H_{n_6 6}$	$S_{n_6 6}$	$MS_{n_6 2(6)}$

Note:
$i = 1, 2, \ldots n_K$; $j = 1, 2$; $k = 1,2, \ldots, 6$;

n_k = the number of samples in the *k*th school,
Y_{ik} = Interest to study at USK for the *i*th respondent at the *k*th school (0 = not interested, 1 = interested),
C_{ik} = the score of cooperation of the *i*th respondent at the *k*th school,
H_{ik} = the expectation score of the *i*th respondent at the *k*th school,
S_{ik} = school of the *i*th respondent = *k*,
$MS_{ij(k)}$ = major of the *i*th respondent at the *k*th school ($m_{.1(.)}$ = Natural Science = 1, $m_{.2(.)}$ = Social Science = 2).

Generalized Linear Mixed Model (GLMM)

Two key elements in the GLMM [6]:

- observations are independent in random effects,
- the distribution of the random variable Y_i is an exponential family with probability density function:

$$f_{ik}(Y_{ik}|\boldsymbol{\alpha}) = exp\left\{\frac{Y_{ik}\xi_{ik} - b(\xi_{ik})}{a_{ik}(\phi)} + c_i(Y_{ik}, \phi)\right\} \quad (1)$$

Where: $Y_{ik}$is score of *i*th object in *k*th cluster, $b(.)$, $a_{ik}(.)$, $c_i(.)$ are functions, ϕ is the dispersion parameter which is or is not known. ξ_i is related with the conditional mean $\mu_i = E(Y_{ik}|\boldsymbol{\alpha})$. $\boldsymbol{\alpha}$ is a random effect vector. For example, Bernouli distribution is exponential family because its distribution can be written as equation (1). The process can be seen :

$f(y) = p^y(1-p)^{1-y}$ where: y = 0,1; 0 ≤ p ≤ 1

$$= exp[log(p^y(1-p)^{1-y})]$$
$$= exp[y\ log\ p + (1-y)\log(1-p)]$$
$$= exp\left[y\ log\frac{p}{1-p} + \log(1-p)\right]$$
$$= exp[y\xi - log\ (1-e^{\xi})]$$

It can be shown:

$Y_{ik} = y;\ \xi_{ik} = \xi = log\frac{p}{1-p}$;

$b(\xi_{ik}) = log\ (1-e^{\xi})$; $a_{ik}(\phi) = 1$;

$c_i(Y_{ik}, \phi) = 0$

The GLMM model can be written as [10] and [17]:

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Z\alpha} + \boldsymbol{\epsilon} \quad (2)$$

Where: $\boldsymbol{\eta}$ is a link function, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices, $\boldsymbol{\beta}$ is a fixed effect vector, $\boldsymbol{\alpha}$ is a vector of random effect and $\boldsymbol{\epsilon}$ is error vector of the model.

If Y_{ik} is the *i*th object in the *k*th cluster, *i* = 1, 2, …, *n* and *k* = 1, …, *K*, $\boldsymbol{\mu}_{ik} = E(Y_{ik}|b_l, x_{ik}, z_{ik})$, b_l is estimator of β_l and $V(Y_{ik}|\alpha_{ik}) = \phi v(\boldsymbol{\mu}_{ik})$ then the GLMM model can be written as [18]:

$$g(\boldsymbol{\mu}_{ik}) = \boldsymbol{x}_{ik}^T\boldsymbol{\beta} + \boldsymbol{z}_{ik}^T\boldsymbol{\alpha}_{ik} \quad (3)$$

Laplace's approach

The GLMM model in equation (2) can be estimated by:

$$\widehat{\boldsymbol{\eta}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\widehat{\boldsymbol{\alpha}} \quad (4)$$

Where: $\widehat{\boldsymbol{\eta}}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ are estimators of $\boldsymbol{\eta}$, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, respectively. The parameter estimators are obtained by minimizing the objective function in (5). It is also known as equation of the Laplace approximation to the marginal log likelihood.

$$log\{L(\boldsymbol{\beta}, \boldsymbol{\theta};\ \widehat{\boldsymbol{\alpha}}, \boldsymbol{y})\} =$$

$$\sum_{i=1}^{m}\begin{Bmatrix} n_i f(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\theta}, \widehat{\boldsymbol{\alpha}}_i) + \frac{n_{\alpha i}}{2} log\{2\pi\} \\ -\frac{1}{2} log|-n_i f''(\boldsymbol{\beta}, \boldsymbol{\theta}; \widehat{\boldsymbol{\alpha}}_i)| \end{Bmatrix} \quad (5)$$

Note:

$\boldsymbol{\beta}$ = constant effect parameter vector,
$\boldsymbol{\theta}$ = covariance parameter vector,
$\boldsymbol{y} = [\boldsymbol{y}_1', \dots, \boldsymbol{y}_K']'$,
K = the number of clusters,
$\boldsymbol{y}_i$ = the vector n_i x 1,
$\boldsymbol{\alpha}$ = the random effect vector,
$\widehat{\boldsymbol{\alpha}}$ = estimator of the random effect vector,
n_i = the number of objects in *i*th cluster,
f'' = the second derivative of function f.

Equation (5) usually reaches the minimum after going through several iterations. The $\widehat{\boldsymbol{\alpha}}$ is obtained after the objective function is minimum. In other words, the change in the value of the objective function is as small as possible or reaches a convergent character.

Gauss-Hermite Quadrature (GHQ)

The parameter estimators with the GHQ approach is obtained by minimizing the following objective function [8]:

$$p(\boldsymbol{y}_i) = \int \cdots \int p(\boldsymbol{y}_i|\boldsymbol{\alpha}_i, \boldsymbol{\beta}, \phi)p(\boldsymbol{\alpha}_i|\boldsymbol{\theta}^*)d\boldsymbol{\alpha}_i$$
$$\approx 2^{\frac{r}{2}}|f''(\boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}; \widehat{\boldsymbol{\alpha}}_i|^{-\frac{1}{2}}$$
$$\sum_{j_1=1}^{N_q} \cdots \sum_{j_r=1}^{N_q}\left[p(\boldsymbol{y}_i|\boldsymbol{a}_j^*, \boldsymbol{\beta}, \phi)p(\boldsymbol{a}_j^*|\boldsymbol{\theta}^*)\prod_{k=1}^{r} w_{j_k} exp\ z_{j_k}^2\right] \quad (6)$$

Note:

$\boldsymbol{\theta}$ = the vector of covariance parameters,
$\boldsymbol{\theta}^*$ = G-Side parameter vector. G-side is random effect in α whereas R-side is also known as residual effect,
ϕ = scale parameter,
r = the number of random effects,
$z_{j_k}^2$ = square of element of vector $\boldsymbol{z}_j^*$,
$\boldsymbol{w}$ = $\left[w_1, \dots, w_{N_q}\right]$ is GHQ weight vector,
w_{j_k} = element of GHQ weight,
$p(\boldsymbol{y}_i|\boldsymbol{\alpha}_i, \boldsymbol{\beta}, \phi)$ and $p(\boldsymbol{\alpha}_i|\boldsymbol{\theta}^*)$ = conditional distribution of random effects $\boldsymbol{y}_i$ and $\boldsymbol{\alpha}_i$, respectively,
$\int \dots d\boldsymbol{\alpha}_i$ = symbol of integral.

If $\boldsymbol{z} = \left[z_1, \cdots, z_{N_q}\right]$ are the standard abscissas for GHQ and $\boldsymbol{z}_j^* = \left[z_{j1}, \cdots z_{j_r}\right]$ is point in grid quadrature dimension-r then centered abscissas are

$$\boldsymbol{a}_j^* = \widehat{\boldsymbol{\alpha}}_i + 2^{1/2} f''(\boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}, \widehat{\boldsymbol{\alpha}}_i)^{-1/2}\boldsymbol{z}_j^* \quad (7)$$

Hypothesis test

Hypothesis of fixed effects are [16] [19]

$$H_o: \beta_l = \beta_{l0}$$

$$H_1: \beta_l \neq \beta_{l0}$$

l = 1,2,…,p

where : p is the number of fixed effects

After determining hypothesis, estimation of coefficients, standard error of the coefficient, t value and p-value are calculated and presented in Table 5.

Table 5. Layout of Model Solution

Effect	Est.	SE	DF	t Value	Pr > |t|
Intercept	$\hat{\beta}_0$	$SE(\hat{\beta}_0)$	*dbI*	*t-val.(int)*	*p val.(int)*
X_1	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	dbX_1	*t-val.*(X_1)	*p val.*(X_1)
X_2	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	dbX_2	*t-val.*(X_2)	*p val.*(X_2)
…	…	…	…	…	…
X_p	$\hat{\beta}_p$	$SE(\hat{\beta}_p)$	dbX_p	*t-val.*(X_p)	*p val.*(X_p)

Notes:
Est. = estimator, SE = Standard Error, DF = Degree of Freedom,

$$t\ Value = \frac{\hat{\beta}_l - \beta_{l0}}{SE(\hat{\beta}_l)}$$

For decision criteria, H_0 is rejected if *p value* is less than α. The value of α used in this case is 0.05.

S_{ik} and $MS_{ij(k)}$ are random effects in this research. Null hypothesis of S_k : $\sigma_k^2 = 0$ versus alternative hypothesis : $\sigma_k^2 > 0$. If S_k are significant or H_0 is rejected then it means that students' interest in different school is significant difference. Hypothesis of $MS_{ij(k)}$, H_0: $\sigma_j^2 = 0$ versus H_1: $\sigma_j^2 > 0$. if H_0 is rejected then it means that students' interest in major of school is significant difference.

Statistic Fits

Model fit can be analyzed with several criteria, see Table 6. The smaller the criterion score of a model is, the better the model is.

Table 6. Criteria for model fit information [8]

Criteria	Formulae
AIC	$-2\ell + 2d$ [20]
AICC	$-2\ell + 2dn^*/(n^* - d - 1)$, [21], [22]
HQIC	$-2\ell + 2d \log\log n$
BIC	$-2\ell + d\ log\ n$
CAIC	$-2\ell + d\ (log\ n + 1)$

Note:
ℓ = maximum value of log likelihood,
d = dimension of model,
n, n^* = reflection of the data size,
n = f for Laplace and Quadrature approximation,
n^* = n, unless $n < d + 2$, in which case $n^* = b + 2$,
b = the number of covariance parameters,
f = data size,
AIC = Akaike Information Criterion,
AICC = Corrected Akaike Information Criterion,
BIC = Bayesian Information Criterion,
CAIC = Consistent AIC and
HQIC = Hannan-Quinn Information Criterion.

Modelling the triggers for interest of students to continue study in USK

Model of the research can be written as:

$$\eta_{ik}^{(I)} = \beta_0 + \beta_1 C_{ik} + \beta_2 A_{ik} + \beta_3 E_{ik} + \beta_4 P_{ik} + \beta_5 H_{ik} + \alpha_1 S_{ik} + \alpha_2 MS_{ij(k)} + e_{ik} \quad (8)$$

Note:
i=1, 2,…,n_k; j=1,2; k = 1, 2, 3
i = index of respondent,
j = index of major,
k = index of school (random effect),
$\eta_{ik}^{(I)}$ = the linear predictor and it is modelled by the inverse logistic link function [12].

$$\eta_{ik}^{(I)} = log\left[\frac{\pi_{ik}^{(I)}}{1-\pi_{ik}^{(I)}}\right] \Rightarrow \pi_{ik}^{(I)} = \frac{1}{1+e^{-\eta_{ik}^{(I)}}} \quad (9)$$

C_{ik} = the score of cooperation of the ith respondent at the kth school,
A_{ik} = action score of the ith respondent at the kth school,
E_{ik} = emotional attitude score of the ith respondent at the kth school,
P_{ik} = score of the objectives of the ith respondent at the kth school,
H_{ik} = the expectation score of the ith respondent at the kth school,
S_{ik} = the kth school of ith respondent,
$MS_{ij(k)}$ = major of the ith respondent at the kth school ($m_{.1(.)}$ = Natural Science = 1, $m_{.2(.)}$ = Social Science = 2),
e_{ik} = random error for the ith respondent in the kth school.

$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \alpha_1,$ and α_2 are intercepts of fixed effect, and model coefficients of fixed effects and random effect.

RESULTS AND DISCUSSION

Outliers are values that often affect the results of statistical analysis. Outliers can be detected by a boxplot diagram. But there are also outliers that don't affect the results. In this study, there were several outliers in the data and after being evaluated they did not affect the results of the analysis. The boxplot for the data which have been cleaned in this study can be seen in Figure 1. Symbols of X1, X2 , X3, X4 and X5 are equal to symbols of C, A, E, P and H, respectively.

The optimum parameter estimator was obtained in the 14th iteration. The value of the objective function has already had a very small change since the 11th iteration, see Table 7. It means that the value of the objective function is convergence at 256.492. This can also be seen from the third column of the table, which shows

the score around zero. In this case, the objective function for Laplace approximation is represented by equation (5). The maximum likelihood method through the Laplace and Gauss-Hermes Quadrature with one node approaches produce similar output. We can see it from the output of iteration history, IC, fixed effect, and random effect solution.

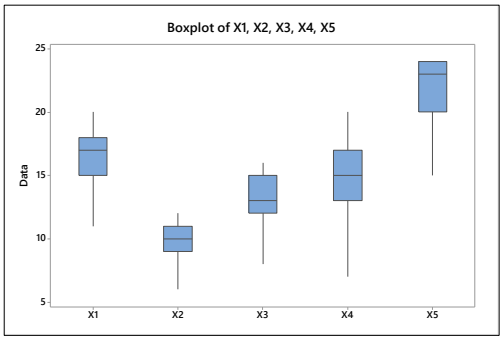

Figure 1. Boxplot of fixed effects

Table 7. Iteration history of Laplace and GHQ approximation

Iteration	Objective Function	Change
0	257.637	.
1	257.621	0.01595
2	257.441	0.18047
3	257.153	0.28725
4	256.915	0.23805
5	256.729	0.18684
6	256.612	0.11703
7	256.546	0.06589
8	256.522	0.02368
9	256.497	0.02474
10	256.494	0.00351
11	256.492	0.00158
12	256.492	0.00017
13	256.492	4E-06
14	256.492	7E-08

Table 8. Fit Statistics of Laplace approximate and GHQ with one node for complete model

-2 *Log Likelihood*	256.49
AIC (smaller is better)	272.49
AICC (smaller is better)	272.78
BIC (smaller is better)	270.83
CAIC (smaller is better)	278.83
HQIC (smaller is better)	265.82

Model fit can be seen from various criteria (Table 8). Information Criteria of Laplace approximation are the same as those of GHQ. Akaike Information Criterion (AIC) is one of the criteria that is often used to identify a suitable model from several models. The smaller the AIC is, the more suitable the model is.

The other model fit criteria are -2 Log Likelihood, Corrected Akaike Information Criterion (AICC), Bayesian Information Criterion (BIC), Consistent AIC (CAIC) and Hannan-Quinn Information Criterion (HQIC).

Table 9. Covariance Parameter Estimates of Laplace and GHQ approach

Cov Parm	Subject	Estimate	Standard Error
Intercept	*S*	0.144	0.2119
Intercept	*M(S)*	0.03474	0.1926

Where:
M(S) is levels of factor M nested under each level of S.

There is no influence of *S* and *M* on student interest (*Y*). The value of estimate *S* is 0.144 and *M(S)* is 0.03474 which can be called not significantly different from zero, see Table 9. This can also be interpreted that the places of schools in Banda Aceh do not show any difference in student interest in continuing school in USK. In addition, the major at SMAN also do not show any difference to interest in studying at USK.

Table 10. Solutions for Fixed Effects of Laplace and GHQ approach

Effect	Estimate	SE	DF	tValue	Pr > |t|
Intercept	-9.956	1.628	5	-6.12	0.002
C	-0.030	0.073	499	-0.41	0.680
A	0.300	0.127	499	2.36	0.019
E	0.515	0.133	499	3.87	<0.001
P	0.021	0.088	499	0.24	0.810
H	0.169	0.052	499	3.24	0.001

Where: SE = Standard Error,
DF = Degree of Freedom.

There are three fixed effects that affect the response variables, see Table 10. If the *p-value* is smaller than α, the variable can be said to have a significant effect on the response variable at the α level. Therefore, action (*A*), emotions (*E*), and expectations (*H*) have a significant effect at α = 0.05., see Table 10. Collaboration (*C*) and goals (*P*) have no significant effect. *A*, *E* and *H* have a positive influence, which means that the more actions are taken, the higher the emotions and the more expectations are, the more interested the students are to continue studying at USK.

Table 11 shows a further description of the random effects of *S* and *M* nested in S. The *p-value* (*Pr*> |*t*|) shows that all the level estimators of the random variables are not significantly different from zero. This may happen because USK provides study programs from less favorite to favorite and study programs in natural and social sciences.

Table 11. Solution for Random Effects with Laplace and GHQ approximation

Effect	School	Maj.	Est.	SE	t Value	Pr > |t|
S	1		0.186	0.350	0.53	0.596
S	2		0.187	0.343	0.54	0.586
S	3		0.276	0.408	0.68	0.498
S	4		-0.018	0.312	-0.06	0.954
S	5		-0.342	0.435	-0.79	0.433
S	6		-0.337	0.471	-0.72	0.475
M(S)	1	1	0.105	0.554	0.19	0.850
M(S)	1	2	-0.060	0.336	-0.18	0.858
M(S)	2	1	0.054	0.334	0.16	0.871
M(S)	2	2	-0.009	0.182	-0.05	0.960
M(S)	3	1	0.062	0.378	0.17	0.869
M(S)	3	2	0.004	0.184	0.02	0.982
M(S)	4	1	0.011	0.188	0.06	0.954
M(S)	4	2	-0.015	0.199	-0.08	0.939
M(S)	5	1	-0.031	0.254	-0.12	0.903
M(S)	5	2	-0.052	0.331	-0.16	0.876
M(S)	6	1	-0.010	0.194	-0.05	0.959
M(S)	6	2	-0.072	0.427	-0.17	0.867

Table 12. Fit statistics on model with fixed effects of Action, Emotions and Expectations through Laplace and GHQ approximation with one node

-2 *Log Likelihood*	258.44
AIC (*smaller is better*)	266.44
AICC (*smaller is better*)	266.51
BIC (*smaller is better*)	283.42
CAIC (*smaller is better*)	287.42
HQIC (*smaller is better*)	273.09

Table 12 shows the model fit values for the three influential fixed effects. The AIC value is 266.44, in other words it is smaller than the AIC of the complete model (AIC of complete model = 272.49). This shows that the model with three influential predictors is more suitable than the complete model. Besides AIC, the model fit value can be seen from -2 Log Likelihood, AICC, BIC, CAIC and HQIC. All the values of these three predictors are smaller than the values of the complete predictors, so this means that the model with three predictors is better than the model with five predictors. The analysis of the three-predictor model shows all the predictors that have a significant effect on student interest, see Table 13.

Handayani's research shows that Laplace is better than quadrature using several points [13]. This is different from this study which found that the results of the Laplace were the same as the results of the GHQ approach using one point.

Final model of student's interest is

$$\eta_{ik}^{(I)} = -10.186 + 0.293A_{ik} + 0.527E_{ik} + 0.166H_{ik} \quad (10)$$

Or

$$\eta_{ik}^{(I)} = log\left[\frac{\pi_{ik}^{(I)}}{1-\pi_{ik}^{(I)}}\right] = -10.186 + 0.293A_{ik} + 0.527E_{ik} + 0.166H_{ik} \quad (11)$$

Table 13. Solutions for a model with three significant effects

Effect	Est.	SE	DF	tValue	Pr > |t|
Int	-10.186	1.281	512	-7.95	<.001
A	0.293	0.124	512	2.38	0.018
E	0.527	0.097	512	5.43	<.001
H	0.166	0.048	512	3.45	0.001

Table 14. Odd ratio of significant fixed effect

Effect	Coef.	Odd Ratio
A	0.293	1.340
E	0.527	1.649
H	0.166	1.181

Odd ratio of a coefficient β_j can be counted by e^{β_j}. The odds ratio A is 1.340, see Table 14, it means that the difference in score on variable A of 1 will cause the difference in the interest score of 1.340 times. For example, a student who scores 9 for the variable A has an interest level of continuing to USK 1.340 times that of a student who scores 8 for the variable A. The odds ratio E and H are interpreted in the same way.

CONCLUSION

The GLMM model can be applied to identify triggers for student's interest in pursuing bachelor's degree in USK. The response has a Bernouli distribution, the fixed effects are cooperation (*C*), activities (*A*), emotions (*E*), purposes (*P*) and expectation (*H*), while the random effects are school origin (*S*) and majors (*M*). Results of the maximum likelihood estimation technique using the Laplace approach are as good as the Gauss-Hermite Quadrature approach with one node. This can be seen from the value of the information criteria which has the same value and even the other output results are the same. There are two models produced in this study, namely a full model consisting of seven predictors and a model containing three predictors that significantly affect the response variables, namely A, E and H. These factors have positive effects on the level of interest of students to continue studying at USK. The random effects of schools and majors do not have significantly different influences on the response variable. This implies that the origin of the schools and the majors in senior high school give the same level of students' interest in continuing their studies at USK. The values of AIC, AICC, BIC, CAIC and HQIC for the three predictor model have smaller scores than these values for

the five predictors. This shows that the model with three variables is better. In other words, *A*, *E* and *H* can explain the level of interest of senior high school students in Banda Aceh to study at USK.

ACKNOWLEDGMENT

Thanks to Mrs. Nurhasanah as a chief of Senior Lector Research Grant Program who has given a permission to be used the data.

REFERENCE

[1] Rutherford, A. 2001 *Introducing Anova and Ancova: A GLM Approach* (London: SAGE Publication)

[2] Rusyana, A.; Notodiputro, K. A.; Sartono, B. 2021. The lasso binary logistic regression method for selecting variables that affect the recovery of Covid-19 patients in China. *J. Phys.: Conf. Ser.* **1882** 012035**.**

[3] Lee, Y.; Ronnegard, L.; Noh, M. 2017 *Data Analysis using Hierarchical Generalized Linear Models with R* (Boca Raton: CRC Press)

[4] Casella, G.; Berger, R. L. 2002 *Statistical Inference* (Duxbury: Thomson Learning)

[5] Wolfinger, R. D. ; Tobias, R. D.; Sall, J. 1994. Computing Gaussian Likelihood and Their Derivatives for General Linear Mixed Models. *SIAM J. Sci. Comput.* **15** 1294–1310.

[6] Jiang, J. 2007 *Linear and Generalized Linear Mixed Models and Their Applications* (Davis: Springer)

[7] Evans, G. 1993 *Practical Numerical Integration* (New York: John Wiley & Sons)

[8] SAS Publishing 2008 *SAS/STAT® 9.2 User's Guide The GLIMMIX Procedure (Book Excerpt)* (North Carolina: SAS Institute)

[9] McCullagh, P.; Nelders, J. A. 1989 *Generalized Linear Models* (London: Chapman & Hall)

[10] Breslow, N. E.; Clayton, D. G. 1993. Approximate Inference in Generalized Linear Mixed Model. *J. Amer. Statist. Assoc.* **88** 9–25.

[11] Chuang, Y. H.; Mazumdar, S.; Park, T.; Tang, G.; Arena, V. C.; Nicolich, M. J. 2011. Generalised linear mixed models in time series studies of air pollution. *Atmos. Pollut. Res.* **2** 428–435.

[12] Namazi-Rad, M.-R.; Mokhtarian, P.; Shukla, N.; Munoz, A. 2016. A data-driven predictive model for residential mobility in Australia – A generalized linear mixed model for repeated measured binary data. *J. Choice Model.* **20** 49–60.

[13] Handayani, D.; Notodiputro, K. A.; Sadik, K.; Kurnia, K. 2017. A comparative study of approximation methods for maximum likelihood estimation in generalized linear mixed models (GLMM). in *AIP Conference Proceedings*. **1827** 020033.

[14] Nurhasanah; Rusyana, A.; Fitriana, AR. 2021. Binary logistic regression for identification of high school student interest in Banda Aceh city in continuing study at Universitas Syiah Kuala. *J. Phys.: Conf. Ser*. **1882** 012034.

[15] Rusyana, A.; Nurhasanah; Maulizasari. 2018. Description of the supporting factors of final project in Mathematics and Natural Sciences Faculty of Syiah Kuala University with multiple correspondence analysis. *IOP Conf. Ser. Mater. Sci. Eng.* **352** 012054.

[16] A.R., F.; Aida, J.; Salwa, N.; Rusyana, A. 2018. Classification of the length of study based on the student characteristics and academic performance in FMIPA Unsyiah. in *J. Phys.: Conf. Ser.* **1116** 022009.

[17] Stroup, W. W. 2013 *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications* (New York: CRC Press)

[18] Muslim, A.; Hayati, M.; Sartono, B.; Notodiputro, K. A. 2018. A Combined Modeling of Generalized Linear Mixed Model and LASSO Techniques for Analizing Monthly Rainfall Data. in *IOP Conf. Ser.: Earth Environ. Sci.* **187** 012044.

[19] Agresti, A. 2015 *Foundations of Linear and Generalized Linear Models* (USA: Wiley)

[20] Akaike, H. 1974 A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control AC-19*

[21] Hurvich, C. M.; Tsai, C. L. 1989. Regression and Time Series Model Selection in Small Samples. *Biometrika*. **76** 297–307.

[22] Burnham, K.P.; Anderson, D. R. 1998 *Model Selection and Inference: A Practical Information-Theoretic Approach* (New York: Springer-Verlag)