Generalized Linear Mixed Models by penalized Lasso in modelling the scores of Indonesian students

V M Santi1*, K A Notodiputro2, B Sartono2 and W Rahyu1

1 Statistics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Negeri Jakarta, Jakarta, 13220, Indonesia
2 Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, 16680, Indonesia

*vmsanti@unj.ac.id

Abstract. The Generalized linear mixed model (GLMM) is an extension of the generalized linear model by adding random effects to linear predictors to accommodate clustered or over dispersion. Severe computational problems in the GLMM modelling cause its use restricted for only a few predictors. When many predictors are available, the estimators become very unstable. Therefore, the procedure for selecting relevant variables is essential in modelling. The use of penalty techniques for selecting variables in mixed models is still rarely applied. In this article, the penalized Lasso approach proposed to handle these kinds of problems. The proposed methods select variables and estimate coefficients simultaneously in GLMM. Based on the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and standard error criteria, it was found that glmmLasso has a better performance than GLMM. For the factors affecting Indonesian’s student scores, where glmmLasso produces three significant covariates for the GLMM model while GLMM without penalized Lasso has five covariates, which means that the GLMM model is more complicated than glmmLasso. Gender, school quality based on National Examination (UN) scores and the opportunity for students to investigate to test their ideas are essential covariates as factors that influence the rating of Indonesian students.

1. Introduction
The high-dimensional linear regression model has widely studied in recent years. The most popular method for achieving sparse solutions is Lasso [1], which uses the penalty function ℓ1. Lasso is not only impressive in terms of its statistical properties but also because of its fast calculation in solving the problem of optimization convex. The Lasso proposed by Tibshirani has become a common approach to regression using the penalty function ℓ1 in the regression coefficient [1]. It results in the effect that all coefficients shrink to zero, and some set to zero. The basic idea is to maximize the log-likelihood l(β) of the model by limiting the norm-L1 from the β parameter vector. However, relatively few articles discussing high-dimensional regression problems involving non-convex loss functions, including Pan and Shen [2], Witten and Tibshirani [3] about clustering, [4,5] discussed the mixed Gaussian and Witten and Tibshirani [6] models of linear discriminant analysis.

For optimization problems in linear models, Lasso can solve by quadratic programming [1]. At the same time, Osborne [7] recommends an algorithm that simultaneously considers primal and dual problems, which are very efficient and can also apply to high-dimensional cases. In the last decade,
several improvements have designed for Lasso, such as the adaptive Lasso [8], SCAD [9], Elastic Net [10], Dantzig selector [11], Double Dantzig [12] and VISA [13].

Lasso has extended to more general models; for example, Tibshirani proposed a new method for selecting variables in the Cox model [14]. The principal works are to minimize partial-log likelihood in norm-L1 from parameters constrained by constants, which is done by an iterative two-step estimation scheme, using reweighted least square and adapting to constraints alternately through quadratic programming procedures. Gui and Li [15] enhanced this procedure, who suggested a repeated estimation approach based on the LARS algorithm, called the LARS-Cox method. But according to Segal and Goeman [16, 17], the two algorithms are computationally very demanding, that this procedure cannot be used very well for high dimensional data.

Generalized Linear Mixed Model (GLMM) is an extension of generalized linear models by adding random effects to linear predictors to accommodate clustered or overdispersion [18-21]. These models have received much attention and applied in many applications such as biology, ecology, medicine, pharmacy, and econometrics. Various estimation methods have proposed starting from numerical integration techniques [22] through “joint maximization methods” [19,23], where an estimation of parameters and random effects are assumed simultaneously to meet the Bayesian approach. An overview of this method can be found in McCulloch [20]. Severe computational problems in GLMM modeling are usually only limited to several predictor variables. The classic approach to selecting predictors based on test statistics with stability problems is usually a forward-backward selection procedure [24].

A flexible and efficient approach to generalized linear mixed models is an algorithm that follows the \( \ell_1 \)-regularized pathway by Park and Hastie [25], which extends the concept of the LARS algorithm [26] to generalized linear models. Another promising approach is to use component wise gradients, starting from the initial value of and then to go through a single coordinate \( \beta \), updating it according to the gradient of penalized likelihood [27–29]. Lately, Goeman [17] presented another approach based on a combination of gradient ascent optimization with the Newton-Raphson algorithm.

The use of penalty techniques for selecting variables in mixed models is still applied. For the mixed Gaussian model, Ni [30] proposed the SCAD penalty. At the same time, Wang [31] offers an adaptive mixed Lasso method, which can collect a large number of predictors and simultaneously contribute to population structure. Schelldorfer [32] developed \( \ell_1 \)-penalty estimation procedures that work for high-dimensional linear effect models based on maximum likelihood. Bondell [33] and Wang [34] consider joint selection cases for fixed and random effects in a linear model. Required in variable selection procedures for GLMM with a focus on successful longitudinal data analysis in Yang [35] and Groll and Tutz [36]. In this study, the Lasso method on GLMM applied to select variables and analyze the factors that influence the score of Indonesian students based on the 2015 PISA data.

2. Material and methods

2.1. Material

Based on the 2015 PISA data, there are 6,513 students in Indonesia between the ages of 12 and 15 who are measured using the questionnaire instrument. The quality of schools based on the results of the National Examination (UN) is assumed to be a fixed influence in the model. The number of independent variables used was 33 covariates in the scale of categorical and ratio based on previous research [37]. The limitation of this study is where the response variable used is the mean scores of science (IPA) of Indonesian students. It is because the 2015 PISA questionnaire focused more on science subjects.

2.2. Methods

In this study, the penalty Lasso function based on the gradient ascent algorithm used to select independent variables while estimating the parameters of the generalized linear mixed model regarding factors that influence students’ mean science scores in Indonesia based on the 2015 PISA data. Randomly selected schools used as a random intercept effect, which is 236 schools. Cross-Validation (CV) criteria and is used to determine the optimum lambda value. To see the goodness of the model, we use the
standard error criteria of the estimated parameter coefficients, Akaike Information Criterion (AIC), BIC, and standard deviation for random effects. The results of the selection obtained then compared with the GLMM without penalized Lasso. The generalized mixed linear model is based on covariates with schools as random effects and identity link function because the response variable assumed a Gaussian distribution and $b_j \sim N(0, \sigma^2_b)$ represents a school-specific random intercept.

3. Results and discussion

In this applied study, a total of 4,605 mean IPA scores of Indonesian students used as response variables. The student's science scores are data with a ratio scale. Fig. 1 below shows a scatter diagram of the standardized science scores of Indonesian students. The range of normalized response values ranges from -3 to +3.

![Scatter diagram of the mean scores of science of Indonesian student.](image)

The corresponding coefficient is buildups for glmmLasso, shown in Figure 2. The cutting off-line in Figure 2 shows the value of $s$, which is the opposite of the value of $\lambda$. Based on Figure 2, each category of covariates is changed as a dummy variable so that 86 independent variables are obtained as in the hardened linear mixture effect model above. The results of the parameter coefficients with the cutting off-limit based on the value $s$ as presented in Table 2.
Figure 2. The coefficient built-ups for glmmLasso on the mean science score of Indonesian student.

The penalty parameter $\lambda$ for glmmLasso is determined based on the Bayesian Information Criterion (BIC) at the interval $[0;200]$. The selected parameter $\lambda$ is quite large, which is $\lambda = 125$, which indicates that the penalty increases the fitting compared to the value of $\lambda = 0$. As a consequence, only a few variables enter the model. The BIC value based on several values of $\lambda$ can see in Figure 3 below.

Figure 3. BIC results for glmmLasso as a function of the penalty parameter $\lambda$; the optimal value of the parameter $\lambda$ is indicated by a vertical line.

Furthermore, the estimated value of the parameter coefficients of the generalized linear mixed-effect model is generated simultaneously with the penalized process of parameter coefficients whose values
are small close to zero. The values in parentheses are the standard error of each parameter coefficient, as presented in Table 1 below.

**Table 1. Estimated value parameter variables for the selection of parameters.**

| Variables | GLMM       | glmmLasso  |
|-----------|------------|------------|
| Intercept | 0.0434(0.043) | 0.079(0.043) |
| X₁(2)     | -          | -0.058 (0.039) |
| X₁(3)     | -          | -0.014 (0.037) |
| X₁(4)     | -          | -0.087 (0.071) |
| X₂(2)     | -0.095 (0.050) | -          |
| X₈        | -          | 0.024 (0.031) |
| X₁₃(2)    | 0.088 (0.041) | -          |
| X₁₄(2)    | 0.091 (0.048) | -          |
| X₁₅(2)    | -0.153 (0.043) | -          |
| X₁₅(3)    | -0.119 (0.057) | -          |
| X₂₉(2)    | 0.059 (0.032) | -          |
| X₃₁(2)    | -0.076 (0.032) | -          |
| X₃₁(3)    | -0.086 (0.049) | -          |
| X₂₉(2)    | -          | -0.109 (0.088) |
| X₂₉(3)    | -          | -0.022 (0.158) |
| b₀        | 0.604      | 0.601      |

Based on Table 1, glmmLasso produces three essential variables in the model. There are no variables selected in glmmLasso that also selected in GLMM. The GLMM model provides more variables so that it can say that the GLMM model is more complicated than the resulting model from glmmLasso. The parameter coefficient in the GLMM column in Table 1 is not a spare solution but is not significant for the model, so it is not write in Table. Gender, quality of the school, and students investigate to test the idea chosen as an essential variable in glmmLasso. It can see that the working principle of glmmLasso is like the Lasso group, where if one category is selected, then all categories of the covariate will also be selected as important variables for the model. While the standard deviation from random effects as presented in Table 1 above. It is seen that the standard deviation value of glmmLasso of 0.6012 is smaller than the standard deviation of GLMM.

Generalized linear mixed model (GLMM) results from glmmLasso based on the parameter coefficient values in Table 1 can be written as follows:

\[ g(\mu_{ij}) = 0.079 - 0.058 X₁(2) - 0.014 X₁(3) - 0.087 X₁(4) + 0.023 X₈ - 0.109 X₂₉(2) - 0.022 X₂₉(3) + 0.601 b_j. \]

Based on the above model, it can be interpreted that for the X₃ independent variable, that is, students conducting investigations with controls are categories (1 = all lessons). It can be explained that there is a decrease in students' mean science (IPA) scores (-0.014) if investigations are carried out by only a few learning and amounting to (-0.087) if very rarely students are allowed to conduct studies to test their ideas. For the sex variable (X₈), it found that there was an increase in the student's mean IPA scores of 0.024 for male students compared to female students. It is in line with the results of research conducted by Nuryoto [38], where the academic achievements of girls declined slightly in the Middle School (SMP) phase and returned to strength in high school (SMA) and higher education. Whereas for the school quality variable based on the results of the National Examination (UN), that is X₂₉ with category control (1 = high quality), it can see that there was a decrease in response value (-0.109) for medium-qualified school and (-0.022) for schools with low quality based on the national examination.
3.1. Model goodness
From Table 2, based on the Akaike Information Criterion (AIC) and BIC criteria, GlmmLasso has AIC and BIC values, which are both smaller than the AIC and BIC values of GLMM which indicate that GlmmLasso has better performance than GLMM without penalty Lasso. GLMM also produces more independent variables to be included in the model so that the model becomes more complicated compared to glmmLasso. Based on the standard error of the random effect, the results of glmmLasso also have a smaller value than GLMM, as seen in Table 1. The results obtained are following the results of Groll and Tutz [37]. They say that GlmmLasso has an excellent performance in selecting variables that can be applied to high dimensional data as well as in the most diverse linear mixed models.

Table 2. Model goodness test results.

|       | GLMM       | glmmLasso  |
|-------|------------|------------|
| BIC   | 610.4674   | -53.6011   |
| AIC   | -1249.38   | -1371.832  |

4. Conclusion
The factors that influence the mean scores of the science of Indonesian students have been studied based on the selection of variables that simultaneously form a generalized linear mixed model using penalized Lasso. GlmmLasso can produce spare solutions with simpler models and build a stable model. The GLMM model without penalized Lasso delivers a performance that is no better than glmmLasso based on the AIC and BIC criteria. The important variable chosen by GlmmLasso is gender, school quality, and the opportunity for students to investigate to test their ideas. The principle of glmmLasso is like the group Lasso.

Acknowledgement
Thank you to Prof. Dr. Khairil Anwar Notodiputro and Dr. Bagus Sartono for his guidance, direction, and input on this research. Thank you also to the Faculty of Mathematics and Natural Sciences, the State University of Jakarta, for their support.

References
[1] Tibshirani R 1996 Regression Shrinkage and Selection Via the Lasso J. R. Stat. Soc. Ser. B 58 267–88
[2] Pan W and Shen X 2007 Penalized model-based clustering with application to variable selection J. Mach. Learn. Res. 8 1145–64
[3] Witten D M and Tibshirani R 2010 A framework for feature selection in clustering J. Am. Stat. Assoc. 105 713–26
[4] Khalili A and Chen J 2007 Variable selection in finite mixture of regression models J. Am. Stat. Assoc. 102 1025–38
[5] Städler N, Bühlmann P and van de Geer S 2010 ℓ 1-Penalization for Mixture Regression Models Test 19 209–56
[6] Witten D M and Tibshirani R 2011 Penalized classification using Fisher’s linear discriminant J. R. Stat. Soc. Ser. B Stat. Methodol. 73 753–72
[7] Osborne M R, Presnell B and Turlach B A 2000 On the LASSO and its Dual J. Comput. Graph. Stat. 9 319–37
[8] Zou H 2006 The adaptive lasso and its oracle properties J. Am. Stat. Assoc. 101 1418–29
[9] Taylor P, Fan J and Li R Journal of the American Statistical Association Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties Variable Selection via Nonconcave Penalized 37–41
[10] Zou H and Hastie T 2005 Erratum: Regularization and variable selection via the elastic net
(Journal of the Royal Statistical Society. Series B: Statistical Methodology (2005) 67 (301-320)) J. R. Stat. Soc. Ser. B Stat. Methodol. 67 768
[11] Candes E and Tao T 2007 The Dantzig selector: Statistical estimation when p is much larger than n Ann. Stat. 35 2313–51
[12] James G M and Radchenko P 2009 A generalized Dantzig selector with shrinkage tuning Biometrika 96 323–37
[13] Radchenko P and James G M 2008 Variable inclusion and shrinkage algorithms J. Am. Stat. Assoc. 103 1304–15
[14] Tibshirani R 1997 The lasso method for variable selection in the cox model Stat. Med. 16 385–95
[15] Gui J and Li H 2005 Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data Bioinformatics 21 3001–8
[16] Segal M R 2006 Microarray gene expression data with linked survival phenotypes: Diffuse large-B-cell lymphoma revisited Biostatistics 7 268–85
[17] Goeman J J 2010 L1 penalized estimation in the Cox proportional hazards model Biometrical J. 52 70–84
[18] McCullagh P and Nelder J A 1989 Generalized Linear Models, Second Edition 532
[19] McCulloch C E and Searle S R 2001 Generalized Linear and Mixed Models Wiley Series I. Pdf (New York: John Wiley and Sons, Inc.)
[20] Molenberghs G and Verbeke G 2005 Models for Discrete Longitudinal Data (New York: Springer)
[21] Booth J G and Hobert J P 1999 Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm J. R. Stat. Soc. Ser. B Stat. Methodol. 61 265–85
[22] Schall R 1991 Estimation in generalized linear models with random effects Biometrika 78 719–27
[23] Breiman L 1996 Heuristics of instability and stabilization in model selection Ann. Stat. 24 2350–83
[24] Park M Y 2007 L 1 -regularization path algorithm for generalized linear models 659–77
[25] Efron B, Hastie T, Johnstone I and Tibshirani R 2004 Least Angle Regression Ann. Stat. 32 407–99
[26] Shevade S K and Keerthi S S 2003 A simple and efficient algorithm for gene selection using sparse logistic regression Bioinformatics 19 2246–53
[27] Kim Y and Kim J 2004 Gradient LASSO for feature selection Proceedings, Twenty-First Int. Conf. Mach. Learn. ICML 2004 473–80
[28] Genkin A, Lewis D D and Madigan D 2007 Large-scale bayesian logistic regression for text categorization Technometrics 49 291–304
[29] Ni X, Zhang D and Zhang H H 2010 Variable selection for semiparametric mixed models in longitudinal studies Biometrics 66 79–88
[30] Wang S and Song P X 2010 Doubly Regularized REML for Estimation and Selection of Fixed and Random Effects in Linear Mixed-Effects Models Biometrics 66 1069–77
[31] Hongmei Y 2007 Variable Selection Procedures for Generalized Linear Mixed Models in Longitudinal Data Analysis (North Carolina State University)
[32] Groll A and Tutz G 2011 Variable Selection for Generalized Linear Mixed Models by L1 -
Penalized Estimation Variable Selection for Generalized Linear Mixed Models by L 1 - Penalized Estimation

[37] Santi V M, Notodiputro K A and Sartono B 2019 Variable selection methods applied to the mathematics scores of Indonesian students based on convex penalized likelihood J. Phys. Conf. Ser. 1402

[38] Nuryoto S 1998 Perbedaan Prestasi Akademik Antara Laki-Laki Dan Perempuan Studi Di Wilayah Yogyakarta J. Psikol. 16–24