Selection of variables in generalized linear mixed model for smoker in Jambi Province

R Warti\textsuperscript{1,*}, K A Notodiputro\textsuperscript{2} and B Sartono\textsuperscript{2}

\textsuperscript{1}Department of Mathematics Education, Faculty of Tarbiyah and Teacher Training, UIN Sulthan Thaha Saifuddin, Jambi, 36363, Indonesia
\textsuperscript{2}Department of Statistics, Faculty of Mathematics and Natural Science, IPB University, Bogor, 16680, Indonesia

*riniwarti@uinjambi.ac.id

Abstract. Smoking is one of the health problems in Indonesia. Many factors cause a person to smoke, both originating from oneself and the environment. The statistical question that arises is how to choose the factors that are most significant in influencing people to smoke. These factors are the variables that will be used in modeling. This study aims to select the variables in the compressed linear mixed model using the Lasso penalty and the Boosting function, using the EM and REML algorithms. Respondents in this study were 160 smokers in Jambi Province. Based on the AIC value, the best model obtained from the selection of variables with the Boosting function and REML algorithm. The analysis shows that work, welfare level, and family members who smoke are the factors that influence people smoking in Jambi Province.

1. Introduction

Mixed models - both Generalized Linear Mixed Models (GLMM) and linear mixed models - have been used in various cases. The combined model combines both the same fixed effects for each observation or sample and random effects that apply to select samples or groups of samples. Through the use of random factors, the linear mixed model designed to handle repeated actions and other intricate study designs. Besides, GLMM tries to model data that does not follow the normal distribution. A direct relationship between the outcome and predictors redefined as a series of linear predictors and it to the expected value of the result via the "link" function. This link function, together with the variance of the expected return value, is chosen from the family members of the exponential distribution. [1]

Many methods can be used to select predictor variables in a linear mixed model [2]. One of them is the Lasso penalty (Least Absolute Shrinkage and Selection Operator or \(L_1\)) [1,3], which has the advantage of selecting variables by reducing some coefficients to zero. Variables whose factors reduced to zero can remove from the model without affecting the ability of the model to predict the desired results. Only Lasso chooses variables without having to select a threshold size coefficient to determine the relationship but has limitations when dealing with many correlating variables [1].

Many methods can be used to select predictor variables in a linear mixed model. One of them is the Lasso penalty (Least Absolute Shrinkage and Selection Operator or \(L_1\)) [1,3]. In its use in linear mixed models, several algorithms can be used to select variables with Lasso penalties including GLMMLasso, glmLasso, PIRLS, MM, and Penalized-EM [4-7]. Another method that can be used to select variables in linear mixed models is by Boosting [8,9], which works iteratively which can produce results that are
very similar to those given by Lasso. In its use in linear mixed models, several algorithms can be used to select variables with the Boosting function, including AdaBoost and Greedy [10-12].

The use of glmer algorithm, GLMMLasso, glmmlasso, bGLMM for selecting variables shows that the glmmlasso algorithm gives the best results [5]. However, this algorithm has not compared the results with the Boosting function. Based on the description above, the purpose of this study is to select the factors that influence people who smoke or not by the Lasso penalty method and the Boosting function in a linear mixed model with a case of smokers in Jambi Province. Next, determine the best model obtained from the process of selecting variables.

Based on the description above, the purpose of this study is to select the factors that influence people who smoke or not by the Lasso penalty method and the Boosting function in a linear mixed model with a case of smokers in Jambi Province. Next, determine the best model obtained from the process of selecting variables.

2. Methods

The research data sourced from 2017 Indonesian Demographic and Health Survey (IDHS), which contained 16363 respondents. It aged over 17 years, divided into adolescent, male, and female categories. The data used were 160 respondents from Jambi Province. Response variables in this study were smoking or not smoking, 14 predictor variables with the residence as random variables. Smoking or not categorization based on the number of cigarettes smoked per day. Response variables are binary scales and numerical predictor variables.

Furthermore, all variables will be selected using two methods, namely the lasso penalty and the Boosting function. By setting the X3 variable (residence) as a random variable, a general linear mixed model is formed [5]:

\[ g(\mu_{ij}) = \beta_0 + x_{1(2)}_{ij}\beta_1 + x_{1(3)}_{ij}\beta_2 + x_{1(4)}_{ij}\beta_3 + x_{2(2)}_{ij}\beta_4 + x_{2(3)}_{ij}\beta_5 + x_{3ij}\beta_6 + x_{4(2)}_{ij}\beta_7 + x_{4(3)}_{ij}\beta_8 + x_{4(4)}_{ij}\beta_9 + x_{5(2)}_{ij}\beta_{10} + x_{5(3)}_{ij}\beta_{11} + x_{5(4)}_{ij}\beta_{12} + x_{5(5)}_{ij}\beta_{13} + x_{5(6)}_{ij}\beta_{14} + x_{6(2)}_{ij}\beta_{15} + x_{7(2)}_{ij}\beta_{16} + x_{7(3)}_{ij}\beta_{17} + x_{7(4)}_{ij}\beta_{18} + x_{7(5)}_{ij}\beta_{19} + x_{8(2)}_{ij}\beta_{20} + x_{9(2)}_{ij}\beta_{21} + x_{10(2)}_{ij}\beta_{22} + x_{10(3)}_{ij}\beta_{23} + x_{11(2)}_{ij}\beta_{24} + x_{11(3)}_{ij}\beta_{25} + x_{12(2)}_{ij}\beta_{26} + x_{12(3)}_{ij}\beta_{27} + x_{13(2)}_{ij}\beta_{28} + x_{14(2)}_{ij}\beta_{29} + x_{14(3)}_{ij}\beta_{30} + b_j \]

where \( \mu_{ij} \) denoted as the expectation value from the mean smokers for the respondent \( i \) on the residence \( j \) and \( b_j \) represents a residence random intercept. The fitting of the model uses a logistic function because the response variable is assumed to be a bernoulli distribution.

In this study, the Lasso penalty function, which based on the glmmlasso algorithm and the glmmLasso function [13], is used to select predictor variables that estimated as parameters of the general linear mixed model which are factors affecting smoking people. The randomly chosen residence used as a random intercept effect with the Cross-Validation (CV) criteria as a determinant of the optimal lambda value. Then the variables are selected using the EM and REML algorithms. To see the goodness of the model, we use the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) of the two algorithms. The selection results obtained are then compared with the Boosting function based on the bGlmm algorithm and GMBoost function [14]. The variable selection also performed using the EM and REML algorithms.

3. Results and discussion

3.1. Sample description

The data of male respondents in 2017 Indonesian Demographic and Health Survey (IDHS) amounted to 16,363 people. This data collection is carried out individually in each of these households. The next sample of individuals selected from Jambi Province, totaling 160 people. Of the 160 people in this
sample, it turns out 68.95% are smokers, where 65.45% of these smokers live in rural areas, and 97.27% of smokers have family members who also smoke.

Characteristics of these smokers, 36.36% are early adult males aged 26-35 years, where 72.27% are heads of households with 52.73% having secondary education, and 46.36% are farmers on their land that are 95.45% earn income in cash and 29.09% are in the category of low-income families. This condition strengthened by the fact that 58.18% of smokers do not have health insurance, 69.09% smokers never read newspapers/magazines, 74.55% never listen to the radio, 95.45% at least once a week smokers watch TV and 68.18% never used the internet even though 87.27% smokers had cellphones.

3.2. Selection variables with Lasso penalty function
The sample used in this selection was 160 respondents from Jambi Province. The selection process uses two algorithms, EM and REML. Before selecting the variable, first, determine the tuning parameter value (lambda) for the Lasso penalty using the Cross-Validation (CV) criteria with a fold value = 10. The lambda value selection plot for the EM and REML algorithms can see in Figure 1.

![Figure 1](image1.png)

Figure 1. Tunning parameter for EM algorithm.

From Figure 1, it can see that for the value of $\lambda = 1$ to $\lambda = 13$, the IC value is still fluctuating (up and down), and the lambda value seen to be monotonically rising since $\lambda = 14$. That is, the optimum lambda value for these two algorithms is $\lambda = 13$. The results of selection on the EM algorithm show that for the amount of $\lambda = 13$ selected three variables, namely $X_5$ (work), $X_7$ (welfare level), and $X_9$ (family members who smoke) with the random variable coefficient $X_3$ is 0.1151 and intercept is 2.1045.

![Figure 2](image2.png)

Figure 2. Tunning parameter for REML algorithm.

From Figure 2, the selection of variables on the REML algorithm shows that for the value of $\lambda = 13$, selected three variables, namely $X_5$ (occupation), $X_7$ (welfare level), and $X_9$ (smoking family members)
with random variable coefficient X3 is 0.1320 and intercept is 2.1067. It shows that the variables selected with the Lasso penalty for the two algorithms are the same.

3.3. Variable selection with the boosting function
Selecting variables with the Boosting function uses two algorithms, EM and REML, and the GMMBoost function. The EM algorithm produces three selected variables, namely X5 (work), X7 (welfare level), and X9 (family members who smoke) with the random variable coefficient X3 being 0.2259 and interception is 2.3242. Selecting variables with the REML algorithm also produced three selected variables, namely X5 (occupation), X7 (welfare level), and X9 (smoking family members) with random variable coefficient X3 is 10.12 and interception is 2.2957. It shows that the variables selected with the Lasso penalty for the two algorithms are the same.

3.4. Selection of the best model
The indicator of the feasibility of the model can use two values, namely AIC and BIC. However, the Boosting function only issues AIC values. Based on the AIC value, the model produced by the Boosting function is better than the model generated by the Lasso penalty. The two AIC values for the Lasso penalty model are almost twice that of the Boosting function model. While the AIC value for the REML algorithm is smaller than the EM algorithm. The results followed the proposed of Buhlmann [15] that an AIC-based method for tuning parameter, namely, for choosing the number of boosting iterations more attractive since it is not required to run the algorithm multiple times for cross-validation as commonly used for Lasso.

| Methods            | Algorithm | AIC           | BIC           |
|--------------------|-----------|---------------|---------------|
| Lasso Penalty      | EM        | 151,4763      | 164,1122      |
|                    | REML      | 151,5479      | 164,3007      |
| Boosting Function  | EM        | 76,2830       |               |
|                    | REML      | 76,2686       |               |

From Table 1, it can see that the best model generated by selecting variables with Boosting using the REML algorithm. The results obtained are following the results of Andreas Myar et al. [16] which said that Boosting algorithms have triggered a lot of research during the last decade because it was offering various practical advantages like automated variable selection and implicit regularization of effect estimates.

4. Conclusions
The research showed that variable selection with the Boosting function is better than the Lasso penalty for the EM and REML algorithm, even though the AIC value generated by the REML algorithm is smaller than the EM algorithm. Although following different strategies for optimization and regularization, both methods imply similar constraints to the estimation problem leading to a comparable performance regarding prediction accuracy and variable selection in practice [17].

Acknowledgement
The authors would like to thank Prof. Dr. Khairil Anwar Notodiputro and Dr. Bagus Sartono for his guidance and direction on this research. Thanks also to Universitas Islam Negeri Sulthan Thaha Saifuddin Jambi for funding the publication of this study, as well as for researchers whose research is useful for this research. Hopefully, this publication is useful and a blessing.

References
[1] Tipton L, Cuenco K T, Huang L, Greenblatt R M, Kleerup E, Sciurba F, Duncan S R, Donahoe M P, Morris A and Ghedin E 2018 Measuring associations between the microbiota and repeated measures of continuous clinical variables using a lasso-penalized generalized linear
mixed model *BioData Min.* 11 12

[2] Tibshirani R 1996 Regression Shrinkage and Selection Via the Lasso *J. R. Stat. Soc. Ser. B* 58 267–88

[3] Schelldorfer J, Meier L and Bühlmann P 2014 Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using ℓ1-penalization *J. Comput. Graph. Stat.* 23 460–77

[4] Groll A and Tutz G 2014 Variable selection for generalized linear mixed models by L 1-penalized estimation *Stat. Comput.* 24 137–54

[5] Hu L, Lu W, Zhou J and Zhou H 2019 MM Algorithms for Variance Component Estimation and Selection in Logistic Linear Mixed Model *Stat. Sin.* 29 1585

[6] Pan L, Li Y, He K, Li Y and Li Y 2020 Generalized linear mixed models with Gaussian mixture random effects: Inference and application *J. Multivar. Anal.* 175 104555

[7] Hastie T 2007 Boosting algorithms: Regularization, prediction and model fitting *Stat. Sci.* 22 477–505

[8] Tutz G and Groll A 2010 Generalized linear mixed models based on boosting *Statistical Modelling and Regression Structures* (Springer) pp 197–215

[9] Bühlmann P and Hothorn T 2007 Boosting algorithms: Regularization, prediction and model fitting *Stat. Sci.* 22 477–505

[10] Groll A 2017 glmmLasso: Variable selection for generalized linear mixed models by L1-penalized estimation *R Packag. version 1* 25

[11] Groll A 2011 GMMBoost: Componentwise Likelihood-based Boosting Approaches to Generalized Mixed Models *R Packag. version 1*

[12] Bühlmann P 2006 Boosting for high-dimensional linear models *Ann. Stat.* 34 559–83

[13] Mayr A, Hofner B, Waldmann E, Hepp T, Meyer S and Gefeller O 2017 An update on statistical boosting in biomedicine *Comput. Math. Methods Med.* 2017

[14] Hepp T, Schmid M, Gefeller O, Waldmann E and Mayr A 2016 Approaches to regularized regression—a comparison between gradient boosting and the lasso *Methods Inf. Med.* 55 422–30

[15] Buja A, Mease D and Wyner A J 2007 Comment: Boosting algorithms: Regularization, prediction and model fitting *Stat. Sci.* 22 506–12

[16] Bühlmann P and Hothorn T 2007 Boosting algorithms: Regularization, prediction and model fitting *Stat. Sci.* 22 477–505

[17] Fan Y and Li R 2012 Variable selection in linear mixed effects models *Ann. Stat.* 40 2043