International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET207610

Generalized Linear Mixed Model Analysis of Acute Respiratory Infection Data on Children

Tiyas Yulita[1], Tika Widayanti[2]
[1]Politeknik Siber dan Sandi Negara, West Java, Indonesia
[2]Institut Teknologi Sumatera, Lampung, Indonesia

Article Info
Volume 7 Issue 6
Page Number : 110-115
Publication Issue :
November-December-2020

Article History
Accepted : 20 Nov 2020
Published : 10 Dec 2020

ABSTRACT

Statistical modeling often involves data which has a distribution of the exponential family. Generalized Linear Model (GLM) was developed to model these data by using a link function between the mean of the response variable and the linear form of the predictor variable. If the data of the response variable comes from several census blocks that are taken randomly, then the diversity between census blocks should not be ignored because it can increase bias. The Generalized Linear Mixed Model (GLMM) is a method that can capture a variety of random effects. However, it does not rule out if there are many predictor variables involved in the model and we use GLMMLasso as a combination method of GLMM and Lasso to shrink the parameter coefficients to zero, it is used to reduce the variance. In this study, a simulation was conducted to GLMMLasso use different numbers of predictor variables and different values of shrinkage coefficients to determine which shrinkage coefficient values have a minimum bias on parameter prediction. Acute Respiratory Infection (API) data on children in Jakarta is used to know the factors that could cause increased cases. The simulation result is the shrinkage coefficient which produces a minimum bias is 30, and the *R2* value of data analysis on the model is 99.24%

Keywords: Linear model, Mixed model, Lasso, Acute Respiratory Infection

I. INTRODUCTION

The regression model can explain the relationship between response and predictor variables to describe the relationship. The Classical regression model is developed with many assumptions, one of these is response variables was a normal distribution, but often this cannot be fulfilled due to several reasons. It is not uncommon for the data that is owned to spread out as an exponential family distribution, such as Gamma, Poisson, and Binomial, a classical regression with these conditions is not appropriate.

Generalized Linear Model (GLM) was developed to model data with response variables that are included in the Exponential family. GLM uses a link function that relates the mean response variable to the linear form of the predictor variable $g\left(E(\boldsymbol{Y})\right) = \boldsymbol{X}'\boldsymbol{\beta}$, g is

a monotonous and differentiable relationship function [5]. Since GLM is developed based on the distribution of the probability of the response variable, the link function depends on this distribution

In addition to the above, another problem arises if the responses we have come from several random sample cluster, or with random predictor variables, then the diversity between clusters and predictors should not be ignored that is then used a linear mixed model to capture this diversity. The combination of GLM and mixed models is the development of modeling that has a significant role in research.

The method involving the two is known as a Generalized Linear Mixed Model (GLMM). But it does not stop here, does not rule out if there are quite a lot of predictor variables so that if we keep using many predictor variables in modeling, the results obtained will vary with high variance. So Hastie [4] developed a method of shrinkage (shrinkage) the regression coefficient to zero or close to zero to increase the accuracy of the prediction which causes bias but the variety of predictions is reduced. GLMMLasso then emerged as a combination of these various methods.

ARI (Acute Respiratory Infection) is an event that can be considered rare but has a quite serious impact, especially in children. ARI is said to be at risk if it reaches the lungs which can become pneumonia. Pneumonia is an infectious disease that causes death, especially in children under five. In general, ARI disease occurs mostly in children. It is estimated that toddlers in Indonesia on average experience coughs and colds 3 to 6 times per year. WHO estimates the incidence of pneumonia in children under five in Indonesia with a percentage of 10-20% per year [1]. This study used data on ARI in children in Jakarta Province, because in that year the ARI incidence rate in Jakarta was the highest in Indonesia.

The selection of shrinkage coefficient in GLMMLasso is a problem on itself, so this study aims to obtain a shrinkage coefficient that produces a minimum bias by simulating several conditions, and the results will be used to selecting the right shrinkage coefficient, and then we observe the performance of GLMMLasso on child ARI data use 23 predictor variables.

II. METHODS AND MATERIAL

a. GLMMLasso

In estimation, GLMM is usually limited to many variables. When many predictors are involved in modeling, the results obtained are unstable. So that the variable selection procedure is very important in modeling. The classic way used to select predictor variables is by statistical testing which is usually problematic on the stability of the algorithm (forward-backward). An alternative approach that can be used to select predictor variables is the penalized regression technique. Lasso introduced by Tibshirani (1996), this method has become very popular as a regression approach that uses the $L_1 - penalty$ on its regression coefficient. This results in all coefficients shrinking towards zero or to zero so as to minimize the variance of the estimators obtained [4]. The basic idea is to maximize the *log-likelihood* $l(\boldsymbol{\beta})$ of the model by limiting the $L_1 - norm$ of the parameter vector β. So as to produce the Lasso parameter estimation as follows

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\beta} l(\beta), \text{ based on } \|\beta\|_1 \leq s \quad \ldots(1)$$

where $s \geq 0$ and $\|\cdot\|_1$ are $L_1 - norm$. The $\widehat{\boldsymbol{\beta}}$ as Lasso estimator can be obtained by optimizing the following equation

$$\widehat{\boldsymbol{\beta}} = arg \max_{\beta}[l(\boldsymbol{\beta}) - \lambda\|\boldsymbol{\beta}\|_1], \quad \ldots(2)$$

with $\lambda \geq 0$, while s and λ are the tuning parameters that must be determined, such as by using cross-validation. λ is often referred to as the shrinkage parameter (shinkage).

Suppose y_{it} is the observation t in cluster i, $i = 1, \dots, n$, $t = 1, \dots, T_i$ in $y_i^t = (y_{iT_1}, \dots, y_{iT_i})$, then $x_{it}^T = 1, x_{it_1}, \dots, x_{it_p}$ are covariate vectors, which are related to fixed effects and $z_{it}^T = 1, z_{it_1}, \dots, z_{it_p}$ are covariate vectors related to effects random. It is assumed that the conditional independent y_{it} with mean $\mu_{it} = E(y_{it}|\boldsymbol{b}_i, \boldsymbol{x}_{it}, \boldsymbol{z}_{it})$ and variance $var(y_{it}|\boldsymbol{b}_i) = \phi v(\mu_{it})$ where $v(.)$ is a known function of variance and ϕ is a scale parameter. The form of GLMM is as follows

$$g(\mu_{it}) = x_{it}^T\beta + z_{it}^T b_i = \eta_{it}^{par} + \eta_{it}^{rand} \quad \dots(3)$$

g is a monotone function and the link function can be derived continuously $\eta_{it}^{par} = x_{it}^T\beta$ is a linear parametric form with the parameter vector $\beta^T = (\beta_0, \beta_1, \dots, \beta_p)$ including the intercept and $\eta_{it}^{rand} = z_{it}^T b_i$ contains the random effect of certain groups where $b_i \sim N(0, Q)$ with Q covariance matrix of size qxq. So that the alternative form of GLMM becomes

$$\mu_{it} = h(\eta_{it}), \eta_{it} = \eta_{it}^{par} + \eta_{it}^{ran} \quad \dots(4)$$

where $h = g^{-1}$ is the inverse of the link function. In GLMM it is assumed that the conditional density function of y_it after the explanatory variable is given and the random effect b_i and is an exponential family

$$f(y_{it}|x_{it}, b_i) = exp\left\{\frac{(y_{it}\theta_{it} - k(\theta_{it}))}{\phi} + c(y_{it}, \phi)\right\} \quad \dots(5)$$

where $\theta_{it} = \theta(\mu_{it})$ is a natural parameter, $k(\theta_{it})$ is a specific function that depends on the type of exponential family, $c(.)$ is a constant of log normal and ϕ is a dispersion parameter. One method to maximize GLMM is the penalized quasi likelihood (PQL) (Breslow and Clayton (1993), Lin and Breslow (1996), and Breslow and Lin (1995)) [3]. Matrix of correlation $Q(\varrho)$ of random effect b_i depending on the unknown vector ϱ. In the basic concept of penalized, the combined likelihood function is defined by the parameter vector of the correlational structure ϱ together with the dispersion parameter ϕ in $\gamma^T = (\phi, \varrho^T)$ and the parameter vector $\delta^T = (\beta^T, b^T)$ with the log likelihood function:

$$l(\delta, \gamma) = \sum_{i=1}^n log(\int f(y_i|\delta, \gamma)p(b_i, \gamma)\, db_i) \quad \dots(6)$$

Where $p(b_i, \gamma)$ is the density of the random effect, Breslow and Clayton (1993) derive an approximation

$$l^{app}(\delta, \gamma) = \sum_{i=1}^n \log\big(f(y_i|\delta, \gamma)\big) - \frac{1}{2} b^T Q(\varrho)^{-1} b$$

With the penalty form $b^T Q(\varrho)^{-1} b$. Using the likelihood equation (6) it is developed by including the penalty $\lambda \sum_{i=1}^p |\beta_i|$ so that the form of the penelized likelihood of Breslow and Clayton (1993)

$$l^{pen}(\beta, b, \gamma) = l^{pen}(\delta, \gamma) = l^{app}(\delta, \gamma) - \lambda \sum_{i=1}^p |\beta_i|$$

Where $\hat{\gamma}$ is obtained from optimizing the function

$$\hat{\delta} = \arg\max_{\beta} l^{pen}(\delta, \hat{\gamma}) = \arg\max_{\beta} \left[l^{app}(\delta, \hat{\gamma}) - \lambda \sum_{i=1}^p |\beta_i|\right]$$

Penalty are used in two last equations considered a partial penalized approach if all the parameter vectors used $\delta^T = (\beta^T, b^T)$ are taken into account [3].

b. Gradient Ascent-GLMMLasso algorithm

The penalized log likelihood $l^{pen}(\delta, \gamma)$ cannot be derived, the derivative can be defined by the following equation:

$$l'_{pen}(\delta; v, \gamma) = \lim_{t \downarrow 0} \frac{1}{t}\big(l^{pen}(\delta + tv, \gamma) - l^{pen}(\delta, \gamma)\big)$$

The gradient Ascent algorithm uses the Taylor series approach and estimates each step of the penalized log likelihood l^{pen} for each $\hat{\delta}$ estimate of the second order of the Taylor series estimate:

$$l^{pen}\big(\hat{\delta} + ts^{pen}(\hat{\delta}, \gamma), y\big) \approx l^{pen}(\hat{\delta}, \gamma) + tl'_{pen}\big(\hat{\delta}; s^{pen}(\hat{\delta}, \gamma), y\big) + 0.5t^2 l''_{pen}\left(\hat{\delta}; s^{pen}(\hat{\delta}, \gamma)\right)$$

With $t > 0$ and $s^{pen}(.,.)$ $l''_{pen}(.;.)$ [3]

c. Methods

The method which used in this research is simulation and the use of data for the application of the method. In addition to knowing the performance of glmmLasso, the purpose of this simulation is to determine the shrinkage coefficient which results in a minimum bias value for the parameter β. The shrinkage coefficient (λ) used in this simulation is

10,15,20,30,40, 50,100. The model built is Poisson, here are the stages of the simulation method:

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j + b_i, i = 1, \ldots, 80$$

$$E[y_i] = \exp(\eta_i) := \lambda_i; \; y_i = \text{Poisson}(\lambda_i)$$

Determine the initial values for $\beta_0, \beta_1, \beta_2,$ and β_3, namely 0,1,-1,1,and $\beta_j = 0; j = 5,7, \ldots, 49,$ (odd) and $\beta_j = 1; j = 4,6, \ldots, 30$ (even)

1. Set the λ value of 10.20,30,40,50,60,70,80,100, 110,120.
2. Using p = 20,25,30, with a total of 80 observations.
3. Generating vector $x_{it}^T = (x_{it1}, \ldots, x_{it30})$ with a uniform distribution between the interval [0,1].
4. The random effect is generated with the normal distribution $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 0.4$.
5. Return to step 2 and the simulation is repeated 100 times for each p with all values of λ.
6. The result of the estimated parameter obtained is then calculated the average bias for β, which is the average of | β-b |.

In the simulation, the response variable is generated as much as 1 value for each cluster. To model using `glmmLasso` (). Then the depreciation value (λ) obtained from the simulation is used to model the ARI data for children in Jakarta Province in 2012. The measure of the goodness of the model is seen from the value of R^2, RMSE (Root Mean Square Error), and AIC (Akaike Information Criterion).

III.RESULTS AND DISCUSSION

The results of the average bias for the parameters are given in Table 1. The value of the depreciation coefficient (λ) which results in a minimum average parameter bias in the number of predictor variables (p) is 25, namely 30, this value will then be used in modeling data on child ARI in Jakarta Province. At p = 20, the depreciation value that produces the minimum bias is 70, while at p = 30, the depreciation value of 120 produces the minimum bias.

TABLE I. THE MEANS OF BIAS VALUE FOR $\boldsymbol{\beta}$

lambda (λ)	p		
	20	25	30
10	0,50956	0,40387	0,27389
20	0,49274	0,39986	0,27400
30	0,47994	0,39853	0,27529
40	0,47214	0,39975	0,27637
50	0,46582	0,40231	0,27687
60	0,46212	0,40506	0,27700
70	0,46117	0,40926	0,27741
80	0,46376	0,41321	0,27694
90	0,47278	0,41594	0,27627
100	0,47977	0,41894	0,27513
110	0,48957	0,42305	0,27441
120	0,49797	0,42669	0,27368

a. GLMMLasso in Child ARI Symptoms Cases in Jakarta

Given that Jakarta is the area with the highest cases of childhood ARI symptoms in Indonesia, modeling was carried out to describe the condition of symptoms of childhood ARI in Jakarta. The data used were obtained from Indonesian Demographic and Health Survey (IDHS). Information about the symptoms of childhood ARI was obtained from respondents, namely mothers who have children under five. Mothers were asked whether their children under five (0 to 59 months of age) had a cough accompanied by difficult or rapid breathing accompanied by inward chest pulls on breathing two weeks before the survey. These symptoms are consistent with ARI. It should be noted that the morbidity data collected is subjective based on the mother's perception of the disease without paramedic approval [2]. The prevalence of symptoms of ARI in children was considered based on the sex of the child, the mother's smoking status, the area of residence, the mother's education, and wealth. Other variables include nutritional information given to children, as well as whether or not vaccines are complete. ARI data can be considered as a Poisson distribution because it is a

count data with events that rarely occur at certain intervals or areas. The variables used in modeling are given in Table 2.

TABLE II VARIABLES ON MODEL

Variable	Explanation
Y	Number of children with ARI symptoms
X1	Average age of mothers (years)
X2	Average Length of Education of mothers (years)
X3	Average Number of Family Members
X4	Percentage of middle to lower economic families
X5	Percentage of mothers who smoke
X6	Percentage of children with male sex
X7	Average Age of Children (years)
X8	Percentage of children consuming breast milk
X9	Average birth weight in children (kg)
X10	Percentage of children consuming formula milk
X11	Percentage of children consuming additional food such as Cerelac
X12	Percentage of children consuming bread, pasta and other carbohydrate sources.
X13	Percentage of children consuming eggs
X14	Percentage of children consuming meat
X15	Percentage of children consuming brightly colored vegetables
X16	Percentage of children consuming green vegetables
X17	Percentage of children who consume vitamin A source fruits
X18	Percentage of children consuming other fruits
X19	Percentage of children consuming *Fe* source
X20	Percentage of children consuming fish.
X21	Percentage of children consuming nuts
X22	Percentage of children who consume dairy
X23	Percentage of children who received complete vaccines.
Census Block	87 census block (cluster)

GLMMLasso model:

$$g(\mu) = 0{,}69266 + 0{,}02157\ X5 - 0{,}01616\ X23 + b$$
$$\boldsymbol{b} \sim N(0, \sigma_b^2);\ \sigma_b^2 = 0{,}069$$

Where μ is the expected value of the number of children with symptoms of ARI in Jakarta. In the analysis results obtained the value π = 1, which means that there is no overdispersion. The value of the estimated model parameters can be seen in Table 3.

TABLE III ESTIMATED COEFFICIENT VALUE

Coefficient	GLMM	GLMMLasso
$\hat{\beta}_0$	-2.79771	0.69266
$\hat{\beta}_1$	0.04794	0
$\hat{\beta}_2$	0.00134	0
$\hat{\beta}_3$	0.26495	0
$\hat{\beta}_4$	-0.00284	0
$\hat{\beta}_5$	0.02831	0.02157
$\hat{\beta}_6$	0.00919	0
$\hat{\beta}_7$	-0.11265	0
$\hat{\beta}_8$	0.02467	0
$\hat{\beta}_9$	-0.65097	0
$\hat{\beta}_{10}$	-0.00634	0
$\hat{\beta}_{11}$	0.01955	0
$\hat{\beta}_{12}$	-0.00863	0
$\hat{\beta}_{13}$	-0.01255	0
$\hat{\beta}_{14}$	0.02355	0
$\hat{\beta}_{15}$	0.01596	0
$\hat{\beta}_{16}$	0.00644	0
$\hat{\beta}_{17}$	-0.02555	0
$\hat{\beta}_{18}$	0.01912	0
$\hat{\beta}_{19}$	-0.00964	0
$\hat{\beta}_{20}$	0.00467	0
$\hat{\beta}_{21}$	0.02058	0
$\hat{\beta}_{22}$	-0.02771	0
$\hat{\beta}_{23}$	-0.03165	-0.01616

From the results obtained on Table 3, it can be seen that none of the estimated values of the GLMM model is zero. Whereas in the GLMMLasso model a coefficient of zero is obtained. Shrinkage in the GLMMLasso model is considered more effective because we can immediately determine the predictor variables that affect the response variables. The goodness of the model can be seen from the AIC value obtained, GLMMLasso produces an AIC value that is smaller than GLMM which means that the GLMMLasso model is better than GLMM. The R^2 value obtained also shows that the predictor variables in the GLMMLasso model can better explain the diversity of the incidence of ARI symptoms better.

TABLE IV THE GOODNESS OF FITS

Model	R^2	RMSE	AIC
GLMM	55,05%	0,80819	230,4
GLMMLasso	99,24%	0,10486	155,3

From Table 4, we could see that the RMSE value in GLMMLasso is also smaller than GLMM. In modeling with GLMMLasso from data on ARI in children in Jakarta, it is found that the variable X5 has a positive effect on the symptoms of ARI, which means that the increasing percentage of smoking mothers will increase the number of children with ARI symptoms. Then the variable X23 has a negative effect on the incidence of ARI in children, which means that the higher the percentage of children who receive the complete vaccine, the lower the number of children with symptoms of ARI in Jakarta.

IV.CONCLUSION

From the results obtained, it can be concluded that simulation studies can assist in determining the shrinkage coefficient (λ) to model data with a minimum parameter bias in addition to the minimum variance due to the addition of lasso to GLMM. GLMMLasso modeling produces a better model based on a larger R^2 value and a smaller RSME than modeling results with GLMM. The AIC value in GLMMLasso is also smaller than GLMM. The addition of the depreciation coefficient (λ) makes the estimated coefficient for the predictor variable that has no effect to zero. This makes it easier for researchers to interpret the model because in a model without shrinkage, the interpretation becomes so complex.

REFERENCES

[1] Ministry of Health .2008. Profile of Indonesian Health 2007. Jakarta.

[2] Ministry of Health. 2013. Profile of Indonesian Health. 2012. Jakarta.

[3] Groll A, Tutz, 2012. Variable Selection for Generalized Linear Mixed Models by L1-penalized estimation. *Stat Comput* DOI 10.1007/s11222-0129359-z, Springer Science + Business Media. New York 2012.

[4] Hastie, T., Tibshirani, R., Wainwright, 2016, "Statistical Learning with Sparsity The Lasso and Generalizations," CRC Press. Taylor & Francis Group. 2016

[5] Muslim, A., et.al,. 2017. Pemodelan Data Curah Hujan Bulanan di Kecamatan Indramayu Tahun 1981-2014 menggunakan Generalized Linear Mixed Model Lasso (GLMMLasso), International Seminar on Science of Complex Natural System. Bogor, Indonesia.

Cite this article as :

Tiyas Yulita, Tika Widayanti, "Generalized Linear Mixed Model Analysis of Acute Respiratory Infection Data on Children", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 7 Issue 6, pp. 110-115, November-December 2020. Available at doi : https://doi.org/10.32628/IJSRSET207610
Journal URL : http://ijsrset.com/IJSRSET207610