Comparison of GLM, GLMM and HGLM in Identifying Factors that Influence the District or City Poverty Level in Aceh Province

A Rusyana\textsuperscript{1,2}, A Kurnia\textsuperscript{1}, K Sadik\textsuperscript{1}, A H Wigena\textsuperscript{1}, I M Sumertajaya\textsuperscript{1} and B Sartono\textsuperscript{1}

\textsuperscript{1} Department of Statistics, IPB University, Bogor, 16680, Indonesia
\textsuperscript{2} Department of Statistics, Universitas Syiah Kuala, Banda Aceh, 23111, Indonesia

*E-mail: kusmans@apps.ipb.ac.id

Abstract. The purpose of research is to evaluate the GLM, GLMM and HGLM models to poverty data in Aceh Province and then identify the best model. The response variable is the percentage of district or city poverty while the fixed effect is population density, sex ratio, the number of populations, the number of industries, area types, percentage of PLN user, poverty line and percentage of productive age group. The random effect for the GLMM and HGLM models is the average monthly expenditure. The data in 2019 were taken from website of the Aceh Central Statistics Agency (BPS) on April 13, 2020. For aggregate, the response variable and the random effect met the normal and gamma distribution. The results showed that the population density has an influence on the percentage of poverty in the GLM and HGLM, while in the GLMM, there are no factors that affect it. The scores of determination coefficient (R\textsuperscript{2}) for GLM, GLMM and HGLM were 75.795\%, 68.441\% and 75.881\%, respectively whereas scores of RMSE of them were 0.121, 1.917 and 0.120, respectively. Because the HGLM model has the largest R\textsuperscript{2} and the smallest RMSE, the HGLM was said to be the best model for the case.

1. Background

The Linear Model (LM) has had a very broad development. LM has developed into the Generalized Linear Model (GLM). Furthermore, the combination of GLM and Linear Mixed Model (LMM) can form a Generalized Linear Mixed Model (GLMM). Then, the combination of GLMM and Joint GLM can form Hierarchical GLM (HGLM) \cite{1}. GLM is a linear model whose response variables have an exponential distribution of families such as binomial, poisson, gamma, and exponential. The linear model is GLM whose response variables have a normal distribution. Because GLMM is a combination of GLM and LMM, the GLMM component consists of the independent variable ($Y$), the independent variable coefficient ($\beta$), the independent variable ($X$), random effect ($\nu$), and model error ($\epsilon$). HGLM is GLMM in which the random effect on the independent variable does not spread Gaussian.

Aceh Province is an area located in the western part of Indonesia which has the provincial capital of Banda Aceh. The province has 18 districts and 5 cities with a population of nearly 54 million people. The land area is more than 5.5 million hectares \cite{2}.

Aceh Province is in the 6\textsuperscript{th} position with the highest percentage of poverty in Indonesia after Papua, West Papua, NTT, Maluku and Gorontalo in September 2019. The poor population in this region is 15 percent \cite{3}. Aceh’s high poverty ranking became big questions because Aceh has a large autonomy fund
every year, for example the autonomy fund for the last three years was IDR 8.27 trillion in 2017, IDR 2.4 trillion in 2018 and IDR 8.3 trillion in 2019. Meanwhile, the Aceh Revenue and Expenditure Budget is IDR 17.328 trillion in 2019. This poverty can be caused by natural disasters that often occur in Indonesia. Various natural disasters has been recorded 2,700 cases from 2016 to 2018 in Indonesia [4] and Aceh is an area that are often affected by these natural disasters.

The GLM and the GLMM has been researched or applied to several cases. Several approaches were carried out to get standardized logistic regression coefficient [5]. A logistic regression is known as a GLM model. Next, the GLM or Principal component regression, gamma-principal component and GLM has been compared to predict monthly rainfall and the GLM was better than the other models because of the non-negative result [6]. The other research, the GLMM has been combined with least absolute shrinkage and selection operator (LASSO) to get rainfall model and it was better than a the GLMM model or LASSO [7]. For gamma-pareto distributed data, it can be approached by GLM gamma [8].

Based on the description above, the purpose of this study is to build GLM, GLMM and HGLM models and then determine the best model of the three models. Next, identify the factors that influence the poverty level of districts / cities in Aceh Province.

2. Literature review

2.1. Generalized Linear Model (GLM)

GLM is developed from a linear model. GLM has three main components [9], namely:

1. A random component, namely $Y_i$ which is independent and has distribution with $E(Y_i) = \mu_i$ and $V(Y_i) = \sigma_i^2$.
2. The systematic component, namely $X_i$, $X_2$, ..., $X_p$, produces a linear estimator $\eta$, where $\eta = X\beta$.
3. The link function, namely $g(\mu)$.

Therefore, GLM has a model $g(E(Y_i|x)) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ whereas $Y_i$ is a function of the mean response value and $Var(Y_i) = \phi V(\mu_i)$. $Y$ is a random variable that has a distribution which is included in the exponential family. The forms of the exponential family distribution are:

$$f_Y(y, \theta, \phi) = \exp \left\{ \frac{\theta y - b(\theta)}{a(\theta)} + c(y, \phi) \right\}$$

(1)

2.2. Generalized Linear Mixed Model (GLMM)

GLMM is an extension of GLM where random effects are included in the model. GLMM is stated as [10]:

$$g(\mu_{ij}) = \alpha_i + \beta_j t_{ij} + \sum_{k=1}^{K} \gamma_k z_{i(k)}$$

(2)

Where:

- $\mu_{ij} = E(Y_{ij}|t_{ij})$ where $Y_{ij}$ is the j-th observation in the i-time place at $t_{ij}$
- $g(.)$ = link function
- $g(\mu_{ij}) = \text{linear predictor}$
- $\alpha_i$ and $\beta_i$ = random effect parameters
- $z_{i(k)}$ = variable $k$, $k = 1, 2, ..., K$;
- $\gamma_k = \text{the regression parameters K related to } z_{i(k)}$, $k = 1, 2, ..., K$.

When $Y_{ij}$ is binary then $g(\mu_{ij}) = log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right)$ and when $Y_{ij}$ has gamma distribution then $g(\mu_{ij}) = \mu^{-1}$. If $Y$ has a distribution that belongs to the exponential distribution family and the random effect $t_{ij}$ has a normal distribution then GLMM can be used but if the $t_{ij}$ is non-normal then the model is HGLM.

Let the GLMM model is written as:
\[ g(\mu_i) = x_i\beta + z_i b \]  

One of several techniques to estimate parameters is penalized quasi-likelihood (PQL). The solution can be solved with formulae [11]:

\[ \beta = (X' V^{-1} X)^{-1} X' V^{-1} \hat{y} \]  
\[ b = G' Z V^{-1} (\hat{y} - X\beta) \]

Where: \( V = W^{-1} + XGZ \), it is assumed that W is non singular matrix.

2.3. Hierarchical Generalized Linear Model (HGLM)

GLMM usually assumes a random effect in a model which has normal distribution. However, there are also cases where the random effect has not normal distribution so that the HGLM method develops [1].

2.4. Estimating Parameters

The parameters can be estimated by the maximum likelihood method using the glm and lmer functions in the lme4 and hglm packages published on April 4, 2019 in software R [12]. GLMM parameters can also be estimated with Gauss-Hermite Quadrature. After that, evaluating model after random effects were combined with response variable and fixed effects in a model is important [13]. The accuracy of the GLM, GLMM and HGLM models can be seen from root mean square error (RMSE) and the coefficient of determination [10].

\[ RMSE = \left( \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \right)^{1/2} \]  
\[ R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \]

3. Data and Methods

3.1. Data

The data used in this study came from the Central Statistics Board (BPS) on April 13, 2020 [2]. The response variable is the percentage of poor people in the district / city of Aceh Province (\( Y \)), while the independent variable is the population density of each district or city (\( X_1 \)), the ratio of the number of women and men (\( X_2 \)), the average expenditure (\( X_3 \)), the total population (\( X_4 \)), the number of industries (\( X_5 \)), type of area (\( X_6 \)), percentage of PLN users (\( X_7 \)), poverty line (\( X_8 \)) and percentage of productive population (\( X_9 \)).

3.2. Method

There are four (4) steps of the research. The first step determined or looked for percentage of district or city poverty (\( Y \)) and factors that were estimated to affect the district or city poverty level in Aceh Province. Then, the variables were explored using descriptive statistics and scatter plots. The third step developed the GLM, GLMM and HGLM models in this case including estimating parameters and testing their hypotheses, and the last step compared GLM, GLMM and HGLM models.

4. Results and Discussion

4.1. Descriptive Statistics

Aceh Province has twenty-three (23) districts / cities. The number of districts is eighteen (18) whereas the number of cities is five. The variables involved in this study were one response variable and 9 fixed effects and 1 random effect. The fixed effects consist of \( X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8 \) and \( X_9 \). The response variable is symbolized by \( Y \), which is the percentage of poor people in the district / city. The minimum percentage of poverty is 7.72 percent in the city of Banda Aceh while the maximum is 20.78 percent in
Aceh Singkil, see Table 1. The average percentage of poverty is 15.71 with a standard deviation of 3.49, the median is 15.60 percent.

The average population density \( (X_1) \) in Aceh Province is 392 people per kilometer with a standard deviation of 994 people per kilometer, see Table 1. The lowest population density is in Gayo Lues District while the most densely populated is in the provincial capital, Banda Aceh. Furthermore, the mean ratio of female to male ratio \( (X_2) \) was 100.76 with a standard deviation of 3.73. The minimum ratio is 93.91 in Kabupaten Pidie meaning that there are more males than females in Pidie while the maximum ratio is 109.39 in Aceh Jaya district.

Expenditures on food and non-food \( (X_3) \) are assumed to be random effects. The value of \( X_3 \) has been divided into four (4) parts, namely: (1) below quartile 1, (2) between quartile 1 and median, (3) between median and quartile 3 and (4) above quartile 3. The least expenditure is IDR 550,623 per capita per month in North Aceh Regency, while the highest is IDR 1,577,302 in Banda Aceh City. The average expenditure in Aceh Province is IDR 709,144. Furthermore, for the total population \( (X_4) \), the number of industries \( (X_5) \), the percentage of PLN \( (X_7) \), the poverty line \( (X_8) \) and the percentage of the productive age group \( (X_9) \) are 233,545 people, 2.78 respectively. companies, 98.88 percent, IDR 426,301 and 64.78 percent.

### Table 1. Descriptive statistics for \( Y, X_1 \) until \( X_9 \) variables.

| Variables | Minimum | Mean   | Median | Standard Error | Maximum |
|-----------|---------|--------|--------|----------------|---------|
| \( Y \)   | 7.22    | 15.71  | 15.60  | 3.49           | 20.78   |
| \( X_1 \) | 17      | 391.70 | 79     | 993.54         | 4734    |
| \( X_2 \) | 93.91   | 100.76 | 100.89 | 3.73           | 109.39  |
| \( X_3 \) | 550623  | 779380.22 | 709144 | 215688.79   | 1577302 |
| \( X_4 \) | 34874   | 233544.87 | 207202 | 150108.44   | 619407  |
| \( X_5 \) | 0       | 2.78   | 2      | 3.16           | 11      |
| \( X_6 \) | 0       | 0.22   | 0      | 0.42           | 1       |
| \( X_7 \) | 91.24   | 98.88  | 99.38  | 1.79           | 100     |
| \( X_8 \) | 351093  | 426301.30 | 417715 | 67277.88    | 628493  |
| \( X_9 \) | 57.75   | 64.78  | 64.52  | 3.01           | 73.24   |

### 4.2. Scatter plot of \( Y \) and \( X \)

Figure 1. Scatter plots of independent variables and dependent variable.
Scatter plot can be conducted to see relation of dependent variable and independent variables. The plot explains poor people and each of the independent variable, see Figure 1. From the figure, relation the response variable and the independent variables has not been clear, yet so we should see the relation in formal test in the models.

4.3. Model Analysis
The random response and effect variables have a Gamma distribution. The Kolmogorov Smirnov test was used to examine these two variables. The p value of the two distributions is more than the \( \alpha = 0.05 \), so the null hypothesis is accepted where \( H_0 \) is the population has gamma distribution, see Table 2.

Table 2. Kolmogorov Smirnov test for \( Y \) (percentage of poor population) and random effect \( X_3 \) (expenditure per person).

| Variables | \( P \) Value for Gamma Distribution |
|-----------|-------------------------------------|
| \( Y \)   | 0.784                               |
| \( X_3 \) | 0.284                               |

Population density influences percentage of poverty (Table 3). The \( p \) value of the variable for estimating the \( X \) coefficient is less than 0.05 so it means \( H_0 \) is rejected, in other words the coefficient of \( X_1 \) is significantly different from zero. The GLM model can be written as:

\[
\eta = \log(\mu) = 6.678 - 1.445 \times 10^{-4} X_1 - 9.570 X_2 - 2.138 \times 10^{-7} X_4 - 1.53 \times 10^{-3} X_6 - 0.1442 X_8 - 0.0166 X_7 + 1.402 \times 10^{-6} X_9 - 0.0277 X_3
\]  

(8)

Table 3. GLM model with the independent variables \( X_1, X_2, X_4, X_5, X_6, X_7, X_8 \) and \( X_9 \).

| Fixed Effect | Coefficient | Standard Error | \( t \) value | \( Pr(>|t|) | \( R^2 \) | RMSE |
|--------------|-------------|----------------|-------------|-----------------|--------|------|
| (Intercept)  | 6.678e+00   | 2.114e+00      | 3.158       | 0.00698**       |        |      |
| \( X_1 \)    | -1.445e-04  | 6.283e-05      | -2.300      | 0.03737*        |        |      |
| \( X_2 \)    | -9.570e-03  | 1.100e-02      | -0.870      | 0.39786         |        |      |
| \( X_4 \)    | -2.138e-07  | 3.148e-07      | -0.679      | 0.50806         |        |      |
| \( X_5 \)    | -1.530e-03  | 1.170e-02      | -0.131      | 0.89781         |        |      |
| \( X_6 \)    | -1.442e-01  | 1.203e-01      | -1.199      | 0.25055         |        |      |
| \( X_7 \)    | -1.660e-02  | 2.054e-02      | -0.808      | 0.43270         |        |      |
| \( X_8 \)    | 1.402e-06   | 6.603e-07      | 2.123       | 0.05206         |        |      |
| \( X_9 \)    | -2.770e-02  | 1.599e-02      | -1.732      | 0.10519         |        |      |

\( R^2 = 75.795\% \) \quad \text{RMSE} = 0.121

The fixed effects of this GLMM model are \( X_1, X_2, X_4, X_5, X_6, X_7, X_8 \) and \( X_9 \) while the random effect is \( X_3 \). Table 4 is the result of the lme4 package of R software. The results of the analysis of the model show that there is no fixed effect that affects the percentage of the poverty level, see Table 4. Random effect which is assumed has normal distribution also has no significant effect. The GLMM model can be written as:

\[
\eta = \log(\mu) = 84.2 - 1.377 \times 10^{-3} X_1 - 0.1674 X_2 - 4.418 \times 10^{-6} X_4 - 8.976 \times 10^{-3} X_5 - 2.505 X_6 - 0.3011 X_7 + 2.236 \times 10^{-5} X_8 - 0.4514 X_9
\]

(9)

Table 4. GLMM model with \( X_1, X_2, X_4, X_5, X_6, X_7, X_8, X_9 \) fixed effect and \( X_3 \) random effect.

| Fixed Effect | Coefficient | Standard Error | \( t \) value |
|--------------|-------------|----------------|--------------|
| (Intercept)  | 8.420e+01   | 3.377e+01      | 2.493        |
| \( X_1 \)    | -1.377e-03  | 1.003e-03      | -1.372       |
negative effect to that effects on poverty in a district or city is the level of population density for poverty data in Aceh. Based on the analysis of GLM, GLMM and HGLM, it is found that the HGLM model is the best model. GLM and GLMM are models that have a lower level of accuracy. The variable that effects on poverty in a district or city is the level of population density. The population density has negative effect to the percentage of poverty in the district / city.

The fixed effect of population density ($X_i$) has an effect to poverty level on alpha = 0.05 while the poverty line ($X_4$) has an effect to the percentage of poverty at alpha = 0.1. The results of this HGLM analysis are presented in Table 5. Random effect $X_1$ which is assumed has gamma distribution has no significant effect. The HGLM model can be written as:

$$\eta = \log(\mu) = 6.73 - 1.422\times10^{-4}X_1 - 9.897\times10^{-3}X_2 - 2.458\times10^{-7}X_3 - 1.358\times10^{-3}X_5 - 1.457\times10^{-1}X_7 - 1.696\times10^{-2}X_8 - 2.733\times10^{-2}X_9$$

(10)

The root mean square error (RMSE) of GLM, GLMM and HGLM has been calculated and the RMSE values of GLM, GLMM and HGLM are 0.121, 1.917 and 0.120. The RMSE GLM and HGLM values were not significantly different, while the GLM model could be said to be less accurate than the other models. The coefficient of determination of GLM, GLMM and HGLM was 75.795%, 68.441% and 75.881%, respectively.

### Table 5. HGLM model with $X_1$, $X_2$, $X_3$, $X_4$, $X_5$, $X_6$, $X_7$, $X_8$, $X_9$ fixed effect and $X_1$ random effect.

| Fixed Effect | Coefficient | Standard Error | t value | Pr(>|t|) |
|--------------|-------------|----------------|---------|----------|
| (Intercept)  | 6.730e+00   | 2.139e+00      | 3.146   | 0.0137   |
| $X_1$        | -1.422e-04  | 6.144e-05      | -2.315  | 0.0493   |
| $X_2$        | -9.897e-03  | 1.097e-02      | -0.902  | 0.3935   |
| $X_3$        | -2.458e-07  | 3.203e-07      | -0.767  | 0.4650   |
| $X_4$        | -1.358e-03  | 1.158e-02      | -0.117  | 0.9096   |
| $X_5$        | -1.457e-01  | 1.204e-01      | -1.211  | 0.2605   |
| $X_7$        | -1.696e-02  | 2.105e-02      | -0.806  | 0.4436   |
| $X_8$        | 1.389e-06   | 6.612e-07      | 2.102   | 0.0688   |
| $X_9$        | -2.733e-02  | 1.592e-02      | -1.716  | 0.1245   |

R$^2 = 75.881\%$  
RMSE = 0.120

The root mean square error (RMSE) of GLM, GLMM and HGLM has been calculated and the RMSE values of GLM, GLMM and HGLM are 0.121, 1.917 and 0.120. The RMSE GLM and HGLM values were not significantly different, while the GLMM model could be said to be less accurate than the other models. The coefficient of determination of GLM, GLMM and HGLM was 75.795%, 68.441% and 75.881%, respectively.

5. Conclusion
Based on the analysis of GLM, GLMM and HGLM, it is found that the HGLM model is the best model for poverty data in Aceh. GLM and GLMM are models that have a lower level of accuracy. The variable that effects on poverty in a district or city is the level of population density. The population density has negative effect to the percentage of poverty in the district / city.

References
[1] Lee Y Ronnegard L and Noh M 2017 Data Analysis using Hierarchical Generalized Linear Models with R (Boca Raton: CRC Press)

[2] Badan Pusat Statistika (BPS) Provinsi Aceh 2020 Statistika Dasar: Kemiskinan. https://aceh.bps.go.id/subject/12/kependudukan.html#subjekViewTab3. April 13, 2020
[3] Badan Pusat Statistika (BPS) 2020 Statistik Dasar: Kemiskinan dan Ketimpangan. https://www.bps.go.id/subject/23/kemiskinan-dan-ketimpangan.html#subjekViewTab3. April 14, 2020

[4] Yana M S Setiawan L Ulfa E M and Rusyana A 2018 Penerapan Metode K-Means dalam Pengelompokan Wilayah Menurut Intensitas Kejadian Bencana Alam di Indonesia Tahun 2013-2018 J. Data Anal 12 93-102

[5] Fitrianto A and Hanafi I 2014 On the Performance of Several Approaches to Obtain Standardized Logistic Regression Coefficients in the International Conference on Science, Technology and Social Sciences (ICSTSS) 533–538

[6] Soleh A M Wigena A H Djuraidah A and Saefuddin A 2015 Statistical Downscaling to Predict Monthly Rainfall Using Generalized Linear Model with Gamma Distribution Inform. Pertan. 24 2 215–222

[7] Muslim A Hayati M Sartono B and Notodiputro K A 2018 A Combined Modeling of Generalized Linear Mixed Model and LASSO Techniques for Analizing Monthly Rainfall Data in IOP Conference Series: Earth and Environmental Science 187 012044

[8] Hanum H Wigena A H Djuraidah A and Mangku I W 2016 Modeling Gamma-Pareto Distributed Data Using GLM Gamma Glob. J. Pure Appl. Math. 12 4 3569–3575

[9] Agresti A 2015 Foundations of Linear and Generalized Linear Models (USA: Wiley)

[10] Michel L. Brun F and Makowski D 2017 A framework based on generalised linear mixed models for analysing pest and disease surveys Crop Prot.

[11] Handayani D Notodiputro K A Kusman S and Kurnia A 2017 A comparative study of approximation methods for maximum likelihood estimation in generalized linear mixed models (GLMM) in AIP Conference Proceedings 1827 020033

[12] Alam M Ronnegrard L and Shen X, 2019, Package “hglm”: Hierarchical Generalized Linear Models, (Repository CRAN).

[13] Muslim A Kurnia A and Sadik K 2018 Hierarchical Generalized Linear Model Approach For Estimating Of Working Population In Kepulauan Riau Province in IOP Conference Series: Earth and Environmental Science 187