Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 14(2) 2021: 194-205
http://ejournal.undip.ac.id/index.php/media_statistika

A STUDY OF GENERALIZED LINEAR MIXED MODEL FOR COUNT DATA USING HIERARCHICAL BAYES METHOD

Etis Sunandi[1,2], Khairil Anwar Notodiputro[2], Bagus Sartono[2]
[1]Department of Mathematics, University of Bengkulu, Bengkulu, Indonesia
[2]Department of Statistics, IPB University, Bogor, Indonesia

e-mail: khairil@apps.ipb.ac.id

DOI: 10.14710/medstat.14.2.194-205

Article Info:
Received: 8 Januari 2021
Accepted: 12 December 2021
Available Online: 11 Januari 2022

Keywords:
Absolute bias, GLMM, illiteracy, MCMC, Poisson Log-Normal

Abstract: Poisson Log-Normal Model is one of the hierarchical mixed models that can be used for count data. Several estimation methods can be used to estimate the model parameters. The first objective of this study was to examine the performance of the parameter estimator and model built using the Hierarchical Bayes method via Markov Chain Monte Carlo (MCMC) with simulation. The second objective was applied the Poisson Log-Normal model to the West Java illiteracy Cases data which is sourced from the Susenas data on March 2019. In 2019, the incidence of illiteracy is a very rare occurrence in West Java Province. So that, it is suitable as an application case in this study. The simulation results showed that the Hierarchical Bayes parameter estimator through MCMC has the smallest Root Mean Squared Error of Prediction (RMSEP) value and the absolute bias is relatively mostly similar when compared to the Maximum Likelihood (ML) and Penalized Quasi-Likelihood (PQL) methods. Meanwhile, the empirical results showed that the fixed variable is the number of respondents who have a maximum education of elementary school have the greatest risk of illiteracy. Also, the diversity of census blocks significantly affects illiteracy cases in West Java 2019.

1. INTRODUCTION

In statistical modeling, the Generalized Linear Mixed Model (GLMM) is a model which the linear predictor contains both random effects and fixed effects. The GLMM also inherits the idea of extending the linear model to non-normal data from the GLM. One of them is count data which distribute Poisson, known as the Mixed Poisson Model (Bolker et al., 2009).

The Mixed Poisson Model can be a two-level hierarchical model with the Log-Normal conjugate. This model is called the Poisson Log-Normal Model. The main difficulty of the model that is the estimation of the parameters that are not easy to obtain analytic solutions to maximize the marginal likelihood of the data. Due to this fact, different estimation methods based on approximations or simulations have been developed in recent years. Several methods include Penalized Quasi-Likelihood (PQL) (Breslow & Clayton, 1993), Maximum Likelihood (ML) using the Laplace method (McCullagh & Nelder, 1989), and Hierarchical Bayes using Markov Chain Monte Carlo (MCMC)(Gelman et al., 2013).

From National Socio-Economic Survey (Susenas), the national illiteracy rate in 2020 is 1.78% (Kemdikbud, 2020). Illiteracy cases are rare nowadays, especially in West Java Province. According to Statistic Indonesia, illiteracy is a condition of being unable to read and write. Its broad meaning is being able to read and write simple letters or sentences or to be able to read and write Braille characters. People with disabilities who have been able to read and write are classified as able to read and write (BPS, 2020).

This research was conducted to examine the performance of parameter estimators and models built using the Hierarchical Bayes MCMC method. Meanwhile, the PQL and ML methods are used as a comparison model. One of Hierarchical Bayes advantage is to provide inferences that are conditional on the data and are exact, without reliance on asymptotic approximation. Small sample inference proceeds in the same manner as if one had a large sample. Furthermore, this study also applies the Poisson Log-Normal model to the West Java Illiteracy Case data which is sourced from the March 2019 Susenas data.

2. LITERATURE REVIEW

2.1. Generalized Linear Mixed Model (GLMM)

Generalized Linear Mixed Model (GLMM) is typically constructed by incorporating random effects into the linear predictors of the conditional independent exponential family model. The two key elements in GLMM are independent of random effects and the distribution of random variables is an exponential family (Jiang, 2007).

The definition of GLMM follows (Jiang, 2007; McCulloh & Searle, 2001):

$$Y_i|\boldsymbol{u} \sim indep.\, f_{Y_i|\boldsymbol{u}}(y_i|\boldsymbol{u}), \tag{1}$$

$$f_i(y_i|\boldsymbol{u}) = exp\left\{\frac{y_i\xi_i - b(\xi_i)}{a_i(\phi)} + c_i(y_i, \phi)\right\},$$

$$E[Y_i|\boldsymbol{u}] = \mu_i,$$

$$g(\mu_i) = \boldsymbol{x_i'\beta} + \boldsymbol{z_i'u_i},$$

$$\boldsymbol{u} \sim f_U(\boldsymbol{u})$$

Where Y_i of the exponential family is the distribution conditional on a random effect $\boldsymbol{u}$. Second, the link function, g(·) is applied to the conditional mean of $Y_i|\boldsymbol{u}$ to obtain the conditional linear predictor. Finally, the linear predictor is assumed to consist of two components: the fixed effect, $\boldsymbol{x_i'\beta}$, and effect random, $\boldsymbol{z_i'u_i}$. In addition, $b(.)$, $a_i(.)$, $c_i(.)$, a known function, ϕ is a dispersion parameter that cannot know. The parameter $\boldsymbol{\eta_i} = \boldsymbol{x_i'\beta} + \boldsymbol{z_i'u}$, $g(\mu_i) = \boldsymbol{\eta_i}$, $b'(\xi_i) = \mu_i$ and $\xi_i = \eta_i$ (Jiang, 2007).

2.2. Poisson Log-Normal Model

The Poisson model is a model for count data. The term "count" refers to data with a non-negative integer response variable. As the name implies, shredding arises from studies that track the number of occurrences, for example, the number of defects in a quality improvement study, the number of disease occurrences in a medical study, the number of insects or birds or weeds in an ecological or agricultural study (Stroup, 2013).

Let y_{ij} is the count taken from the j-th observation and the i-th cluster which are independent of each other. The Poisson Log-Normal Model is written as follows (McCulloh & Searle, 2001):

$$y_{ij}|\boldsymbol{u} \sim indep.Poisson(\mu_{ij}), i = 1, 2, \dots m; j = 1, 2, \dots, n_i \quad (2)$$

$$g(\mu_{ij}) = \log(\mu_{ij}) = \boldsymbol{x_i'\beta} + \boldsymbol{u_i},$$

$$u_i \sim i.i.d\ N(0, \sigma_u^2)$$

Estimating parameters in the Poisson Log-Normal Model can be done in various ways, i.e Maximum Likelihood, Penalized Quasi-Likelihood, and Hierarchical Bayes via MCMC i.e Gibbs sampling. Research on the Mixed Poisson model has been conducted by several researchers (Berliana et al., 2019; Bermúdez et al., 2020; Mallya et al., 2018).

2.3. Maximum Likelihood (ML) Method

The Maximum Likelihood (ML) estimation is an estimate method of the model parameters. The ML method selects the set of values of the model parameters that maximizes the likelihood function. In the context of the exponential family, we do this by maximizing the log-likelihood function, $\ell(\theta;\ \boldsymbol{y}, \phi)$, with respect to the canonical parameter θ given the observation, $\boldsymbol{y}$, and parameters scale, ϕ (Stroup, 2013).

From Equation (1) can be written with the likelihood function integration in the distribution of $\boldsymbol{u}$ the q-dimension is as follows (McCulloh & Searle, 2001):

$$L = \int \prod_i f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{Y}|\boldsymbol{u}) f_{\boldsymbol{U}}(\boldsymbol{u}) d\boldsymbol{u}$$

The log-likelihood function of the Poisson Log-Normal model can be written as follows:

$$l = log\ (\prod_{i=1}^{m} \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} \frac{\mu_{ij}^{y_{ij}} e^{-\mu_{ij}}}{y_{ij}!} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}u_i^2} du_i \quad (3)$$

$$= y'X\beta - \sum_{i,j} log\ y_{ij}$$

$$+ \sum_i log \int_{-\infty}^{\infty} exp\left\{ y_i u_i - \sum_j e^{x_{ij}'\beta + u_i} \right\} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}u_i^2} du_i$$

Based on Equation (3), according to McCulloh & Searle (2001) the equation for the log-likelihood function for the fixed effect parameter β is

$$l = log \int f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{Y}|\boldsymbol{u}) f_{\boldsymbol{U}}(\boldsymbol{u}) d\boldsymbol{u} = log f_{\boldsymbol{Y}}(\boldsymbol{y})$$

So that

$$\frac{\partial l}{\partial \beta} = \frac{\partial}{\partial \beta} \int \frac{f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{Y}|\boldsymbol{u}) f_{\boldsymbol{U}}(\boldsymbol{u}) d\boldsymbol{u}}{f_{\boldsymbol{Y}}(\boldsymbol{y})} = \int \frac{\left[\frac{\partial}{\partial \beta} f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{Y}|\boldsymbol{u})\right] f_{\boldsymbol{U}}(\boldsymbol{u}) d\boldsymbol{u}}{f_{\boldsymbol{Y}}(\boldsymbol{y})} \quad (4)$$

Because $f_{\boldsymbol{U}}(\boldsymbol{u})$ does not depend on β, so it can be written as follows:

$$\frac{\partial f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{Y}|\boldsymbol{u})}{\partial \beta} = \left(\frac{1}{f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{Y}|\boldsymbol{u})} \frac{\partial f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{Y}|\boldsymbol{u})}{\partial \beta} \right) f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{Y}|\boldsymbol{u}) \quad (5)$$

$$= \frac{\partial log f_{Y|u}(Y|u)}{\partial \beta} log f_{Y|u}(Y|u)$$

Substituting Equation (5) to Equation (4) is obtained:

$$\frac{\partial l}{\partial \beta} = \int \frac{\frac{\partial log f_{Y|u}(Y|u)}{\partial \beta} f_{Y|u}(Y|u) f_U(u) du}{f_Y(y)} \tag{6}$$

$$= \int \left[\frac{\partial}{\partial \beta} f_{Y|u}(Y|u)\right] f_{U|y}(u|y) du$$

$$= X' E[W^*|y] - X' E[W^* \mu|y]$$

where $W^* = \left\{ {}_d\left[a(\phi) v(\mu_i) g_\mu(\mu_i)\right]^{-1} \right\}$

So the probability equation for β at Equation (2) when $W^* = I$ is as follows:

$$X'y = X' E[\mu|y] \tag{7}$$

Analogous to Equation (6), the probability function equation for the random effect parameter in the distribution $f_U(u)$ can be written as follows: let φ is a parameter in the distribution $f_U(u)$ so that:

$$\frac{\partial l}{\partial \varphi} = \int \frac{\partial log f_U(u)}{\partial \varphi} f_{U|y}(u|y) du = E\left[\frac{\partial log f_U(u)}{\partial \varphi} \middle| y\right] \tag{8}$$

The Parameters of the model are estimated using Equation (7) and (8). Since the exact likelihood function is difficult to calculate, the approach method is one of the natural alternatives. A well-known method for estimating integrals is called Laplace. Suppose we want to approach the integral form (Jiang, 2007):

$$\int exp\{-q(x)\} dx \tag{9}$$

With the Taylor expansion, $q(x) = q(\tilde{x}) + \frac{1}{2} q''(\tilde{x})(x - \tilde{x})^2 + \cdots$, Equation (9) can be approximated by

$$\int exp\{-q(x)\} dx \approx \sqrt{\frac{2\pi}{q''(\tilde{x})}} exp\{-q(\tilde{x})\}$$

So that

$$\int exp\{-q(\alpha)\} d\alpha \approx c|q''(\tilde{\alpha})|^{-1/2} exp\{-q(\tilde{\alpha})\}$$

Where c is a constant that depends only on the dimension of the integral.

2.4. Penalized Quasi-Likelihood (PQL) Method

For more complex models, the marginal or quasi-likelihood estimation equations are not available in closed form. The simplest method for adjusting the model uses the Laplace approach and is called the Penalized Quasi-Likelihood (PQL; (Breslow & Clayton, 1993)). If we apply the Laplace approach to the Integrated Quasi-Likelihood, the estimator of $\theta = [\beta^T, u^T]^T$ to remain is obtained by maximizing the penalized quasi log-likelihood

$$-\frac{1}{2\varphi}\sum_{i=1}^{n} d_i\left(y_i;\mu_i^b\right)-\frac{1}{2}\boldsymbol{u}^T\boldsymbol{R}^{-1}\boldsymbol{u}$$

Or the equivalent by solving the following estimation equations

$$\boldsymbol{U_\beta}=\boldsymbol{X}^T\boldsymbol{W}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})=\boldsymbol{0} \tag{10}$$

$$\boldsymbol{U_u}=\boldsymbol{Z}^T\boldsymbol{W}^{-1}\left(\boldsymbol{Y}-\mu\ \right)-\boldsymbol{R}^{-1}\boldsymbol{u}=\boldsymbol{0}$$

Where $\boldsymbol{W}=diag\{\varphi a_i^{-1}v(\mu_i\ )[\ g\ (\mu_i)]^2\}$ and **R** are components of the error variance. In Normal GLMM, the solution for (10) equals is obtained by solving the linear system repeatedly:

$$\begin{bmatrix}\boldsymbol{X^TW^{-1}X} & \boldsymbol{X^TW^{-1}Z}\\ \boldsymbol{Z^TW^{-1}XR^{-1}} & \boldsymbol{R^{-1}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}=\begin{bmatrix}\boldsymbol{X^TW^{-1}Y^*}\\ \boldsymbol{Z^TW^{-1}Y^*}\end{bmatrix}$$

where $\boldsymbol{Y^*}=\boldsymbol{\eta}\ \ +\left(\boldsymbol{Y}-\mu\ \right)\left[\ g'\left(\mu_i\ \right)\right]_{nx1}$ is called the work vector (McCullagh & Nelder, 1989).

2.5. Hierarchical Bayes (HB) Method via Markov Chain Monte Carlo (MCMC)

In the Bayesian context, let λ is parameter that has a prior distribution of $R(\lambda)$ and a posterior distribution of $R(\theta|\boldsymbol{y})$ where θ is the parameter of interest, given $\boldsymbol{y}$ data will be obtained. In HB, the model will be given in stages, namely $R(\boldsymbol{y}|\theta,\lambda_1)$ and $R(\theta|\lambda_2)$. The two models are combined with the prior distribution in the model parameters $\boldsymbol{\lambda}=(\lambda_1,\lambda_2)$. Suppose the parameter you want to find is $\vartheta=k(\theta)$. In the Hierarchical Bayes method (HB), the parameter φ will be obtained by calculating the mean of the posterior distribution $R(\theta|\boldsymbol{y})$ as follows (Gelman et al., 2013):

$$\hat{\vartheta}^{HB}=E(h(\theta)|\boldsymbol{y})=\int h(\theta)\ R(\theta|\boldsymbol{y})\ d\theta \tag{11}$$

In Equation (11), the posterior distribution is required $R(\theta|\boldsymbol{y})$ before searching $\hat{\vartheta}^{HB}$. The posterior distribution $R(\theta|\boldsymbol{y})$ can be obtained by applying the Bayes theorem, so that the joint posterior distribution of the small area u parameter λ will be obtained and the model parameter is given **y** data as follows:

$$R(\theta,\lambda|\boldsymbol{y})=\frac{S(\boldsymbol{y},\theta|\lambda)R(\lambda)}{R_1(y)} \tag{12}$$

By integrating the posterior distribution function together in Equation (12) to the parameters λ, the desired posterior distribution is obtained as follows:

$$R(\theta|\boldsymbol{y})=\int R(\theta,\lambda|\boldsymbol{y})\ d\lambda\ =\int R(\theta|\boldsymbol{y},\lambda)R(\lambda|\boldsymbol{y})\ d\lambda \tag{13}$$

By substituting Equation (13) into Equation (11), the following values $\hat{\varphi}^{HB}$ are obtained:

$$\hat{\vartheta}^{HB}=\iint h(\theta)\ R(\theta,\lambda|\boldsymbol{y})\ d\theta\ d\lambda \tag{14}$$

From Equation (14), it can be seen that to obtain the posterior mean $R(\theta|\boldsymbol{y})$ complex integration techniques are required, so that the posterior mean will be difficult to obtain analytically. Therefore, a numerical approach will be carried out. A random sample $\theta^{(k)}$ will be generated first from the posterior distribution $R(\theta|\boldsymbol{y})$, then $\int k(\theta)\ R(\theta|\boldsymbol{y})\ d\mu$ estimated

from the previously obtained random sample mean $\frac{1}{K}\sum_{k=1}^{K} h\left(\theta^{(k)}\right)$. Generating random samples from the posterior distribution $R(\theta|\boldsymbol{y})$ is not easy, so it takes a special method to resolve the issue. One method that can be used is the Markov Chain Monte Carlo (MCMC) (J. D. Hadfield, 2015; Rao & Molina, 2015).

MCMC is a method commonly used when the sample θ cannot be generated directly from the posterior distribution $R(\theta|\boldsymbol{y})$. Instead, the sample is θ generated iteratively in such a way that each sampling is expected to come from a distribution that is close to the posterior distribution $R(\theta|\boldsymbol{y})$ (Gelman et al., 2013). In the application of MCMC, a Markov chain is built $\left\{\theta^{(k)}, k = 1,2, \ldots\right\}$ where $\theta^{(k)}$ the sample is generated in k-iterations and depends on previous sampling $\left(\theta^{(k-1)}\right)$. When $k \to \infty$, the Markov Chain will converge towards the posterior distribution $R(\theta|\boldsymbol{y})$. In MCMC, the commonly used algorithms are Gibbs Sampling and Metropolis-Hasting. Several studies using the Bayes method have been carried out by Maulina et al. (2019) and Yanuar et al. (2020).

2.6. The Goodness of Fit of the Model Measures

The measures of goodness of fit that used in this research are the Root Mean Squared Error of Prediction (RMSEP) and Absolute Bias (AB). The MSEP value is obtained from the average squared difference between the actual and predicted data. Then the RMSEP value is calculated by rooting the MSEP value. While the absolute bias parameter estimator is obtained from the difference between the parameters and the expected value. The RMSEP value is calculated based on the testing data while the absolute bias is calculated based on the training data. The RMSEP measure is used to assess the goodness of the model's predictions. Meanwhile, Absolute Bias is used to assess the goodness of the parameter estimators of the model being built. The formula for RMSEP and Absolute Bias (AB) is as follow (Sunandi et al., 2021) :

$$RMSEP = \sqrt{\sum_{i=1}^{t} \frac{(y_i - \hat{y}_i)^2}{t}} \tag{15}$$

$$AB\left(\hat{\theta}\right) = \left|\theta - E(\hat{\theta})\right| \tag{16}$$

Where t is the number of validation data, y_i is the response variable and $\hat{y}_i$ is the estimated value of the response variable, and $\hat{\theta}$ is the parameter estimator θ.

METHODOLOGY

3.1. Data

The data used in this study are simulation data and empirical data. The simulation data are generated based on the Poisson Log-Normal model in equation (2). Some parameters are set in the model, namely $a = 0.5, b = 0.0005, \sigma_e^2 = 0.3$, and $\beta_0 = 0.5$.

Meanwhile, secondary data used is Susenas data on March 2019. The data collected by the Statistics Indonesia. The unit of observation is the household and the census block as a cluster. The variables used in this study, as response variable, was illiterate cases with the age of the respondent at least 12 years old. Furthermore, some fixed and random effect variables can be seen in Table 1.

Based on previous research (Astuti et al., 2017) that the level of education and work or not a person can be an indicator of the risk of illiteracy. In addition, visual, hearing, emotional, communication and concentration problems can hinder language skills so that it can lead to illiteracy (Rohmani Nur Indah, 2017). In addition, age is also an indicator of illiteracy (Mariyono, 2016). Furthermore, of the currently illiterate population, about two-thirds are women (Wahyuni et al., 2017). This means that gender can be an indicator of illiteracy cases.

Table 1. List of Fixed and Random Variables

Code	Variable Name
Fixed factor	
$\boldsymbol{X_1}$	Number of respondents who have visual impairments
$\boldsymbol{X_2}$	Number of respondents who have hearing loss
$\boldsymbol{X_3}$	Average respondent's age (in years)
$\boldsymbol{X_4}$	Number of respondents who have communication problems
$\boldsymbol{X_5}$	Number of respondents who have emotional disorders
$\boldsymbol{X_6}$	Number of respondents who have impaired concentration
$\boldsymbol{X_7}$	Number of respondents who are male
$\boldsymbol{X_8}$	Number of respondents who work
$\boldsymbol{X_9}$	Number of respondents graduated up to elementary school
Random factor	
u	Census block of Susenas on March 2019

3.2. Method

Simulation data is used to assess the performance of the MCMC, PQL, and ML methods in the Poisson Log-Normal Model. The simulation data and algorithm are as follows:

1. Determined the number of simulation replications $s = 100$, the number of observations $n = 24000$, the number of clusters $= n/10$, and the number of parameters $p = 11$
2. Generated $X_k \sim N(0,1)$
3. Generated $\sigma_u^2 \sim IG(a, b)$
4. Calculated σ_u
5. Generated $u_i \sim N(0, \sigma_u^2)$
6. Determined σ_e
7. Generated $e \sim N(0, \sigma_e^2)$
8. Generated $\beta_k \sim u(0,1)$
9. Determined $\beta_0 = 0.5$
10. Calculated $mu = \exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k X_{kij} + u_i\right)$
11. Generated $y_{ij} \sim Poisson(mu)$
12. The data was divided into 2 parts with stratified random sampling method, i.e training data (80%) and testing data (20%).
13. GLMM modeling was carried out using the PQL, ML, and Hierarchical Bayes methods via MCMC (1300 iterations)
14. Measured the goodness of fit the model using equation (15) and (16)
15. Interpreted of results

Analogous to the simulation flow, the analysis procedure on empirical data was as follows:

1. Data exploration.
2. The data was divided based on stratified Random sampling into 80% training data and 20% testing data.
3. GLMM modeling was carried out using the PQL, ML, and Hierarchical Bayes methods via MCMC (1300 iterations).
4. Measured the goodness of the model.
5. Interpreted of results.

4. RESULTS

4.1. Simulation Results

The simulation is designed to assess the performance of the MCMC method in estimating the parameters of the Poisson Log-Normal Model with the PQL and ML methods as a comparison. In this study, a simulation was carried out with a scenario iteration of 1300 gibbs sampling MCMC which was repeated 100 times. Computing programming using the R program packages "MCMCglmm" (J. Hadfield, 2012), "glmmML" (Brostrom, 2020), and "nlme" (Pinheiro, 2020).

When viewed from the measure of the goodness of the simulation results model, the Hierarchical Bayes method through MCMC has a smaller RMSEP value than PQL and ML, i.e. 5.194. Meanwhile, if viewed from the absolute bias of each parameter estimator, the three methods have relative similar values. All estimators have very small absolute bias values tending towards zero. Except b1, it still has a relatively large absolute bias value (0.67).

Table 2. Summary of Measures of the Goodness of fit the Model

Measure	PQL	ML	MCMC
RMSEP	324.955	239.422	5.194
AB(b0)	0.002	0.001	0.007
AB(b1)	0.644	0.644	0.642
AB(b2)	0.001	0.002	0.001
AB(b3)	0.003	0.002	0.004
AB(b4)	0.001	0.000	0.001
AB(b5)	0.002	0.002	0.002
AB(b6)	0.000	0.001	0.003
AB(b7)	0.002	0.002	0.001
AB(b8)	0.001	0.000	0.003
AB(b9)	0.001	0.001	0.001
AB(b10)	0.000	0.000	0.001

Similar to the absolute bias of the fixed parameter estimator, the absolute bias of the random parameter estimator of the Poisson Log-Normal Model also has the same trend in the three estimation methods. The absolute bias of cluster standard deviation ($\hat{\sigma}_u$) ranges from 0.096-0.097. Whereas the absolute bias of the unit standard deviation ($\hat{\sigma}_e$) ranges from 0.444-0.449. Of the three methods, MCMC has a slightly smaller absolute bias random effect when compared to PQL and ML. Overall, based on the measures of the goodness of the model, it can be claimed that the method of estimating Hierarchical Bayes parameters via MCMC is the best method used with the smallest RMSEP value.

Table 3. Absolute Bias of Random Parameter Estimator

Measure	PQL	ML	MCMC
AB($\widehat{\boldsymbol{\sigma}}_{\boldsymbol{u}}$)	0.097	0.097	0.096
AB($\widehat{\boldsymbol{\sigma}}_{\boldsymbol{e}}$)	0.449	0.445	0.444

4.2. Empirical Application to Illiteracy Case Data

Application of Poisson Log-Normal Model using MCMC, PQL, and ML performed on the Susenas West Java data. There are as many as 23,738 households as the unit of observation is divided into 2,405 census blocks as a cluster. Figure 1 shows that the distribution of illiteracy cases in West Java is quite diverse, both in districts and cities. Indramayu District has the largest percentage of illiterate cases, i.e 7.73% (186 cases). Meanwhile, Tasikmalaya City had the smallest percentage of illiterate cases, i.e 0.15% (3 cases).

Meanwhile, when viewed from the measure of the goodness of the model summarized in Table 4, the MCMC method has the smallest RMSEP value, which is 0.004. PQL and ML methods have RMSEP values of 0.090 and 0.041, respectively. It means, the MCMC method is more accurate in modeling the case of illiteracy in West Java in 2019.

Table 4 Measures of Goodness of Fit of Illiteracy Case Model

	MCMC	PQL	ML
RMSEP	0.004	0.090	0.041

Based on the estimation results of MCMC (Table 5), the variables of the number of respondents who have visual impairment (X_1), the number of respondents who have hearing problems (X_2), the average age of the respondents (in years) (X_3), the number of respondents who have communication problems (X_4), the number of respondents who have impaired concentration (X_6), the number of respondents who work (X_8), and the number of respondents who graduated up to elementary school (X_9) have a significant effect on the risk of illiteracy.

Increment 1 the number of respondents who have a vision impairment (X_1), hearing loss (X_2), communication disorders (X_4), or impaired concentration (X_6) can increase the risk of cases of illiterate 12.26%, 24.70%, 25.23%, or 24.31%, respectively. In addition, increasing the average age of the respondents by one year would increase the risk of illiteracy by 6.67%. Whereas an increase in the number of respondents who work (X_8) 1 person will reduce the risk of illiteracy by 14.72%. Furthermore, the increase in the number of respondents who graduated up to elementary school (X_9), 1 person, will increase the risk of illiteracy by 128.96%.

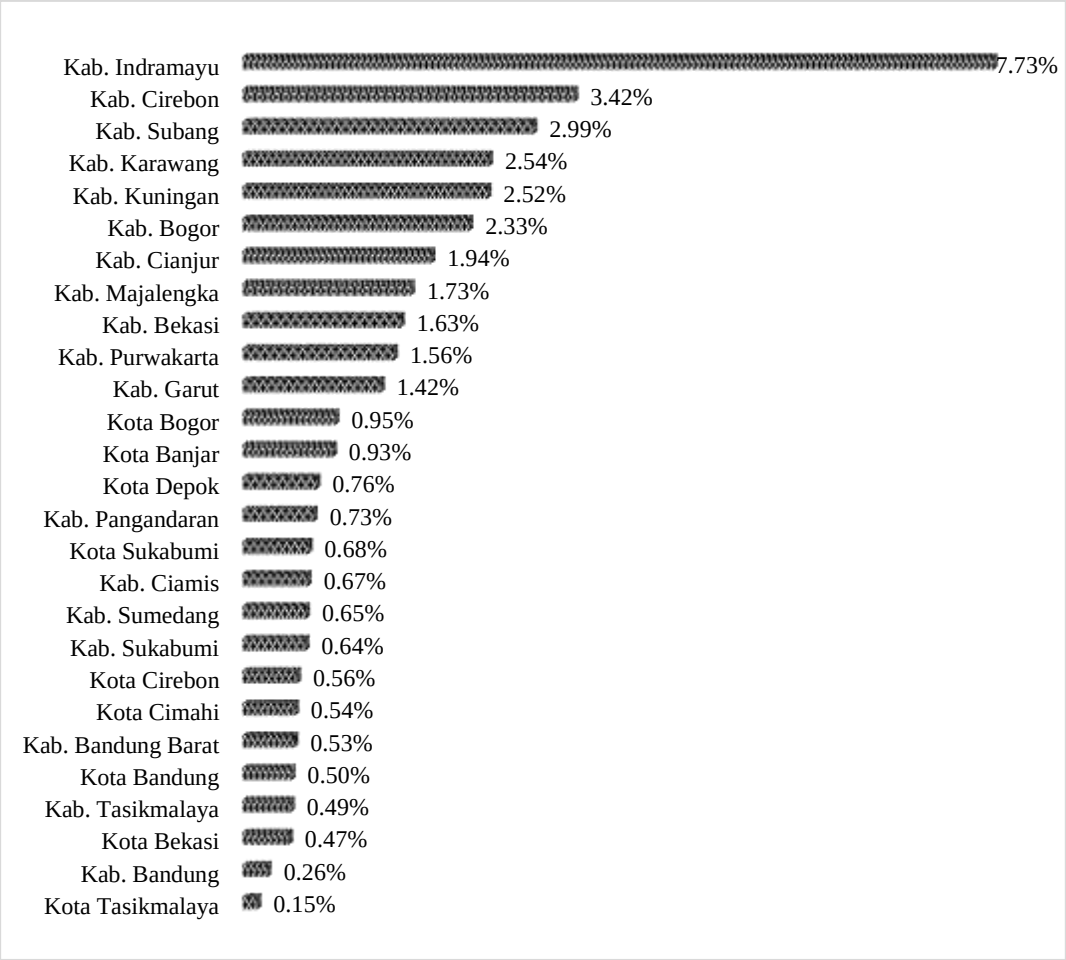

Figure 1. Distribution of Illiteracy Cases in West Java in 2019

In Table 6, there is information about the standard deviation estimator of the random variables used. The random effect of the census block as a cluster has an estimated standard deviation of 0.664 and a standard deviation of household as unit is 0.512. The two standard deviations of the random effect significantly affect the model. This means that there is a diversity of illiteracy cases in each Susenas census block in West Java 2019.

Table 5. MCMC Fixed Parameter Estimator of Poisson Log-Normal Model in Illiteracy Cases

	Estimate	**l-95%CI**	**U-95%CI**	**p-value**	**Relative Risk**
(Intercept)	-8.351	-8.753	-7.804	<0.001	
$\boldsymbol{X_1}$	0.116	0.023	0.192	0.014	1.123
$\boldsymbol{X_2}$	0.221	0.066	0.376	<0.001	1.247
$\boldsymbol{X_3}$	0.065	0.058	0.070	<0.001	1.067
$\boldsymbol{X_4}$	0.225	0.033	0.388	0.024	1.252
$\boldsymbol{X_5}$	0.148	-0.032	0.338	0.130	1.159
$\boldsymbol{X_6}$	0.218	0.017	0.397	0.026	1.243
$\boldsymbol{X_7}$	0.010	-0.048	0.089	0.872	1.010
$\boldsymbol{X_8}$	-0.159	-0.252	-0.064	<0.001	0.853
$\boldsymbol{X_9}$	0.828	0.777	0.875	<0.001	2.290

Table 6. MCMC Random Parameter Estimator Poisson Log-Normal Model in Illiteracy Cases

	Estimate	l-95% CI	U-95% CI
$\widehat{\boldsymbol{\sigma}}_{\boldsymbol{u}}$	0.664	0.295	0.9112
$\widehat{\boldsymbol{\sigma}}_{\boldsymbol{e}}$	0.512	0.3278	0.7719

5. CONCLUSION

The simulation results show that the Hierarchical Bayes method through MCMC has a smaller RMSEP value than PQL and ML, i.e. 5.94. Meanwhile, if viewed from the multiple bias of each model parameter estimator, the three methods have relatively the same value. Similar to the absolute bias of the fixed parameter estimator, the absolute bias of the random parameter estimator of the Poisson Log-Normal Model also has the same trend in the three estimation methods. From the three methods, MCMC has a slightly smaller relative bias random effect when compared to PQL and ML. Overall based on the measures of the goodness of fit of the model, it can be claimed that the Hierarchical Bayes parameters estimator through MCMC is the best method. Analogous to the simulation, empirical results show that the MCMC method has the smallest RMSEP value of 0.004. MCMC estimation based on the results of the variable and the number of respondent's graduate education up to elementary school (X9) has the biggest risk of illiteracy significantly.

REFERENCES

Astuti, N. K., Purhadi, P., & Andari, S. (2017). Pemodelan Angka Buta Huruf di Kabupaten/Kota se-Jawa Timur dengan Metode Geographically Weighted t Regression. *Jurnal Sains Dan Seni ITS*, *6*(2), 224–228. https://doi.org/10.12962/j23373520.v6i2.25005

Berliana, S. M., Purhadi, Sutikno, & Rahayu, S. P. (2019). Multivariate generalized Poisson regression model with exposure and correlation as a function of covariates: Parameter estimation and hypothesis testing. *AIP Conference Proceedings 2192, 090001*, 1–10. https://doi.org/10.1063/1.5139171

Bermúdez, L., Karlis, D., & Morillo, I. (2020). Modelling unobserved heterogeneity in claim counts using finite mixture models. *Risks*, *8*(10), 1–13. https://doi.org/10.3390/risks8010010

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. In *Trends in Ecology and Evolution* (Vol. 24, Issue 3, pp. 127–135). https://doi.org/10.1016/j.tree.2008.10.008

BPS. (2020). *Badan Pusat Statistik Provinsi Jawa Barat*. https://jabar.bps.go.id/Istilah/index?Istilah%5Bberawalan%5D=B

Breslow, N. E., & Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, *88*(421), 9–25. https://doi.org/10.2307/2290687

Brostrom, G. (2020). Package ' glmmML ': Generalized linear models with clustering. In *Cran*. http://cran.r-project.org/web/packages/glmmML/index.html

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013).

Bayesian data analysis, third edition. In *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC.

Hadfield, J. (2012). *MCMCglmm: MCMC generalised linear mixed models*.

Hadfield, J. D. (2015). MCMCglmm Course Notes. *Notes*, 1443.

Jiang, J. (2007). Linear and Generalized Linear Mixed Models and Their Applications. In *Linear and Generalized Linear Mixed Models and Their Applications*. Springer. https://doi.org/10.1007/978-0-387-47946-0

Kemdikbud. (2020). *Kementerian Pendidikan dan Kebudayaan » Republik Indonesia*. https://www.kemdikbud.go.id/main/blog/2020/09/pemerintah-terus-berkomitmen-dalam-mengentaskan-buta-aksara

Mallya, S., Sander, B., Roy-Gagnon, M. H., Taljaard, M., Jolly, A., & Kulkarni, M. A. (2018). Factors associated with human West Nile virus infection in Ontario: A generalized linear mixed modelling approach. *BMC Infectious Diseases*, *18*(1), 1–9. https://doi.org/10.1186/s12879-018-3052-6

Mariyono. (2016). Strategi Pemberantasan Buta Aksara Melalui Penggunaan Teknik Metastasis Berbasis Keluarga. *Pancaran*, *5*(1), 55–66.

Maulina, R. F., Djuraidah, A., & Kurnia, A. (2019). Pemodelan Kemiskinan Di Jawa Menggunakan Bayesian Spasial Probit Pendekatan Integrated Nested Laplace Approximation (INLA). *MEDIA STATISTIKA*, *12*(2), 140–151. https://doi.org/10.14710/medstat.12.2.140-151

McCullagh, P., & Nelder, J. A. (1989). Generalized Linear Models, Second Edition, Chapman & Hall/CRC, Boca Raton, FL. In *Chapman and Hall*.

McCulloh, C. E., & Searle, S. R. (2001). *Generalized, Linear, and Mixed Models* (1st ed.). John Wiley & Sons, Inc.

Pinheiro, J. (2020). *Title Linear and Nonlinear Mixed Effects Models*. https://bugs.r-project.org

Rao, J. N. K., & Molina, I. (2015). Small Area Estimation: Second Edition. In *Small Area Estimation: Second Edition*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118735855

Rohmani Nur Indah. (2017). Gangguan Berbahasa: Kajian Pengantar. In *UIN-Maliki Press*.

Stroup, W. W. (2013). Generalized Linear Mixed Models - Modern Concepts, Methods and Applications. In *International Statistical Review* (Vol. 81, Issue 3).

Sunandi, E., Notodoputro, K. A., & Sartono, B. (2021). A study on group lasso for grouped variable selection in regression model. *IOP Conference Series: Materials Science and Engineering*, *1115*(1). https://doi.org/10.1088/1757-899x/1115/1/012089

Wahyuni, S., Machfudz, M., & Badrih, M. (2017). Pemberdayaan Masyrakat Perempuan Melalui Pemberantasan Buta Aksara Guna Menumbuhkembangkan Usaha Kreatif Berbasis Literasi dan Potensi Lokal. *Jurnal Inovasi Pendidikan*, *1*(2), 48–71.

Yanuar, F., Sari, P. T., & Asdi, Y. (2020). Identification Of Rainfall Distribution In West Sumatera And Assessment Of Its Parameters Using Bayes Method. *MEDIA STATISTIKA*, *13*(2), 161–169. https://doi.org/10.14710/medstat.13.2.161-169