Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2020, 2020:67

https://doi.org/10.28919/cmbn/4960

ISSN: 2052-2541

TWEEDIE COMPOUND POISSON MODEL WITH FIRST ORDER AUTOREGRESSIVE TIME RANDOM EFFECT

FIA FRIDAYANTI ADAM[1,2], ANANG KURNIA[1,*], I. GUSTI PUTU PURNABA[3], I. WAYAN MANGKU[3], AGUS M. SOLEH[1]

[1]Department of Statistics, IPB University, Bogor 16680, Indonesia

[2]Vocational Program, Universitas Indonesia, Depok 16424, Indonesia

[3]Department of Mathematics, IPB University, Bogor 16680, Indonesia

Abstract: Modeling with the Tweedie compound Poisson distribution is mostly done based on the Generalized Linear Model (GLM). GLM can be expanded into the Generalized Linear Mixed Model (GLMM) if there are fixed effects and random effects. GLMM modeling with Tweedie compound Poisson response variables is still rarely done because it is not analytically tractable and the density function cannot be stated in closed-form. By using the h-likelihood method, GLMM modeling with Tweedie compound Poisson can be solved numerically. This research models the Tweedie compound Poisson response variable by using GLMM with two random effects, region and the time assumed to follow the first-order autoregressive process. A simulation study is carried out with an evaluation using the average relative bias and the average MSE. The simulation results show the greater the autoregressive coefficient results in the smaller value of the relative bias. MSE values that are close to zero indicate the model is very good in describing data. An application, which is conducted to model the total number of claims in a certain area and time based on the 2014

*Corresponding author

E-mail address: anangk@apps.ipb.ac.id

Received August 19, 2020

profile of risk and loss of motor vehicle insurance in Indonesia, shows model has small value of absolute bias and MSE.

Keywords: first-order autoregressive; GLMM; h-likelihood; Tweedie compound Poisson.

2010 AMS Subject Classification: 62J12, 97M30.

1. INTRODUCTION

The exponential dispersion model (EDM) is an exponential family distribution with additional dispersion parameters. EDM has an important role in modern data analysis because it is able to overcome problems where the response variable does not have a normal distribution. The density function of the random variable Y with the distribution including the EDM family is

$$f(y;\theta,\phi) = a(y;\phi)\exp\left(\frac{1}{\phi}\big(y\theta - k(\theta)\big)\right), \quad (1)$$

where k and a are known function, $\phi > 0$ is dispersion parameter, and θ is the natural parameter [2]. A characteristic of EDM is the mean-variance relation if the dispersion parameter is considered constant. In other words, if Y has EDM distribution with mean μ, variance function $Var()$, and dispersion parameter ϕ then $Var(Y) = \phi\, Var(\mu)$. If $Var(\mu) = \mu^p$, where p is index parameter, then Y is defined as Tweedie family distribution [2].

This research focuses on the Tweedie distribution family with a value of $p \in (1,2)$ which is the compound Poisson distribution. The random variable Y has compound Poisson distribution if $Y = \sum_{i=1}^{N} C_i . Y$ is built by two random variables namely N which has Poisson as the first distribution and C has gamma as the second distribution. This distribution has a probability of mass at zero and a skewed continuous distribution on the positive real line. As a result, this distribution can model data with a large zero. Henceforth the compound Poisson distribution is called the Tweedie compound Poisson distribution.

There are many applications of the Tweedie compound Poisson distribution. In the actuarial field, the distribution of Tweedie compound Poisson is used to model the total number of insurance claims [1], [5]. In the field of climatology, the distribution of Tweedie compound Poisson is used

to model rainfall [6], [8]. In the field of fisheries, Tweedie compound Poisson distribution models the amount of fish caught [12].

All those examples above are done based on a generalized linear model (GLM). GLM is an extension of the usual regression with the response variable not always coming from the normal distribution. GLM can be expanded into a generalized linear mixed model (GLMM) if there are fixed effects and random effects. GLMM with response variables have Tweedie's compound Poisson is still rarely performed. This is because the distribution itself is not analytically tractable and the density function cannot be stated in a closed-form [9], [10]. As a result, modeling involving the distribution of the Tweedie compound Poisson must be approximated numerically.

In general, the numerical methods used are mostly based on penalized quasi-likelihood (PQL) [7]. But the PQL method is only able to estimate the regression parameters. The PQL method is not yet equipped with the ability to estimate variance parameters needed in GLMM. Some methods had been done to overcome this problem such as Laplace approximation and the Gauss-Hermite quadrature adaptive method that are able to get variance parameter estimators in addition to estimating regression parameters [14].

The alternative method is hierarchical likelihood (h-likelihood) [13]. H-likelihood combines fixed and random effects in GLMM into an extended likelihood function. In this method the random effect does not have to be normal distribution like most in GLMM. For example, the random effect has gamma distribution while the response variable has Poisson distribution [3]. The h-likelihood method avoids the use of integrals in obtaining marginal likelihood. This method is also able to get the estimation of regression parameters and variance parameters.

Zhang [14] modeled Tweedie's compound Poisson using GLMM with one random effect. So the research question arises as to how the modeling involves the response variable that has Tweedie's compound Poisson distribution with two random effects? Supposed the random effect added is the time assumed to follow the first-order autoregressive process [11], the next research question is whether the h-likelihood method can be used to estimate the regression

parameters and variances in GLMM with a random effect of time following the first-order autoregressive process?

This study examines the development of a model with a Tweedie compound Poisson response variable with two random effects namely region and time which is assumed to follow the first order autoregressive process. The h-likelihood method is used to estimate regression parameters and variances. Details of model and method will be presented in section 2. To measure the goodness of the model, a simulation study is carried out in section 3 with the performance of the model based on the average relative bias and the Mean Squared Error (MSE). Section 4 is the application section which is carried out by modeling the total number of claims in some regions and time based on the 2014 risk and loss profile of a motor vehicle insurance insurance company in Indonesia. Finally section 5 contains conclusions.

2. MODEL AND METHOD

2.1 Basic Model

EDM with $Var(\mu) = \mu^p$ for $p \in (1,2)$ is a Tweedie compound Poisson distribution if $Y = \sum_{i=1}^{N} C_i$ where N has Poisson (λ) distribution and C_i has gamma $G(\alpha, \gamma)$ distribution. If $N = 0$ then $Y = 0$. If $N > 0$ then Y is the sum of C i.i.d gamma random variables. Tweedie compound Poisson linear mixed model is a mixed model in which the Y has the Tweedie compound Poisson distribution and if there is a random effect $\boldsymbol{v}$ and the relationship $\boldsymbol{\mu} = E(y|\boldsymbol{v})$ so

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zv} \tag{2}$$

through link function $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$, where $\boldsymbol{\beta}$ fixed effect vector, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the associated design matrix, and assumed $\boldsymbol{v} \sim N(\boldsymbol{0}, \boldsymbol{D})$. The variance component $\boldsymbol{D}$ is further expressed in terms of the relative covariance factor $\boldsymbol{\Lambda}$ such that $\boldsymbol{D} = \phi\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ [14]. As a result, the specification of equation (2) can be expressed as

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Z\Lambda v}^* = \boldsymbol{X\beta} + \boldsymbol{Z}^*\boldsymbol{v}^* \tag{3}$$

where $\boldsymbol{v}^* \sim N(\boldsymbol{0}, \phi\boldsymbol{I})$.

2.2 Proposed Model

Let y_{itj} be the jth observation that occurs in the region i and time t, where $i = 1,2,\dots,M$, $t = 1,2,\dots,T$, $j = 1,2,\dots,n_{it}$, and y_{itj} is also assumed to have Tweedie compound Poisson distribution and is related to a covariate variable x_{itj} through the following model

$$\begin{aligned} &E(y_{itj}|v_i,u_t) = \mu_{itj}, \\ &\log \mu_{itj} = \beta_0 + x_{itj}\beta_1 + v_i + u_t, \\ &u_t = \rho\, u_{t-1} + \varepsilon_t, |\rho| < 1, \end{aligned} \tag{4}$$

where β_0 and β_1 are fixed effect vectors, $v_i \sim iid\ N(0,\sigma_v^2)$ is the random effect of the i-th region, the u_t is the random effect of the time assumed to follow the first-order autoregressive process where ε_t is an error of the u_t assumed $\varepsilon_t \sim iid\ N(0,\sigma_\varepsilon^2)$, and ρ is the autoregressive coefficient. The random effects of v_i and u_t are assumed to be independent.

Suppose there is one covariate, then for each region i at time t, equation (4) can be denoted in the following matrix form

$$\begin{bmatrix} \log \mu_{it1} \\ \log \mu_{it2} \\ \vdots \\ \log \mu_{itn_{it}} \end{bmatrix} = \begin{bmatrix} 1 & x_{it1} \\ 1 & x_{it2} \\ \vdots & \vdots \\ 1 & x_{itn_{it}} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + v_i \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + u_t \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{5}$$

If $\log \mu_{itj} = \eta_{itj}$ then the above equation becomes

$$\begin{bmatrix} \eta_{it1} \\ \eta_{it2} \\ \vdots \\ \eta_{itn_{it}} \end{bmatrix} = \begin{bmatrix} 1 & x_{it1} \\ 1 & x_{it2} \\ \vdots & \vdots \\ 1 & x_{itn_{it}} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + v_i \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + u_t \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{6}$$

So for region i, equation (6) becomes

$$\begin{bmatrix} \boldsymbol{\eta}_{i1} \\ \boldsymbol{\eta}_{i2} \\ \vdots \\ \boldsymbol{\eta}_{it} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{i1} \\ \boldsymbol{X}_{i2} \\ \vdots \\ \boldsymbol{X}_{it} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + v_i \begin{bmatrix} \mathbf{1}_{n_{i1}} \\ \mathbf{1}_{n_{i2}} \\ \vdots \\ \mathbf{1}_{n_{iT}} \end{bmatrix} + u_t \begin{bmatrix} \mathbf{1}_{n_{1t}} \\ \mathbf{1}_{n_{2t}} \\ \vdots \\ \mathbf{1}_{n_{mt}} \end{bmatrix}. \tag{7}$$

Supposed that the number of observation is balanced and if defined $\boldsymbol{Z}_{1i} = \mathbf{1}_{n_i}$ and $\boldsymbol{Z}_{2t} = \mathbf{1}_{\boldsymbol{n_t}}$ the equation (7) above becomes

$$\boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_{1i}v_i + \boldsymbol{Z}_{2i}\boldsymbol{u}_i. \tag{8}$$

Model for all region is

$$\begin{bmatrix}\boldsymbol{\eta}_1\\ \boldsymbol{\eta}_2\\ \vdots\\ \boldsymbol{\eta}_m\end{bmatrix}=\begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\\ \vdots\\ \boldsymbol{X}_m\end{bmatrix}\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}+\begin{bmatrix}\boldsymbol{Z}_{11} & 0 & \cdots & 0\\ 0 & \boldsymbol{Z}_{12} & \cdots & 0\\ \vdots & \vdots & \ddots & \cdots\\ 0 & 0 & \cdots & \boldsymbol{Z}_{1m}\end{bmatrix}\begin{bmatrix}v_1\\ v_2\\ \vdots\\ v_m\end{bmatrix}+\begin{bmatrix}\boldsymbol{Z}_{21} & 0 & \cdots & 0\\ 0 & \boldsymbol{Z}_{22} & \cdots & 0\\ \vdots & \vdots & \ddots & \cdots\\ 0 & 0 & \cdots & \boldsymbol{Z}_{2T}\end{bmatrix}\begin{bmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\\ \vdots\\ \boldsymbol{u}_T\end{bmatrix}$$

or

$$\boldsymbol{\eta}=\boldsymbol{X\beta}+\boldsymbol{Z}_1\boldsymbol{v}+\boldsymbol{Z_2u} \tag{9}$$

where $\boldsymbol{Z}_1=\boldsymbol{I}_m\otimes\boldsymbol{Z}_{1i}$, $\boldsymbol{Z_2}=\boldsymbol{I}_T\otimes\boldsymbol{Z}_{2t}$, $\boldsymbol{\eta}=(\boldsymbol{\eta}_1',\boldsymbol{\eta}_2',\ldots\boldsymbol{\eta}_m')',\boldsymbol{X}=(\boldsymbol{X_1'},\boldsymbol{X_2'},\ldots,\boldsymbol{X_m'})'$, $\boldsymbol{v}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_m)'$, $\boldsymbol{u}=(\boldsymbol{u}_1',\boldsymbol{u}_2',\ldots\boldsymbol{u}_T')'$, $\boldsymbol{I}_m$ is identity matrix sized $m\times m$, and $\boldsymbol{I}_T$ is identity matrix sized $T\times T$.

Equation (4) assumed $v_i\sim iid\ N(0,\sigma_v^2)$, so the expectation value and covariance matrix of vector $\boldsymbol{v}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_m)'$ are

$$E(\boldsymbol{v})=\boldsymbol{0}\ \text{ and }\ Cov(\boldsymbol{v})=\boldsymbol{G_1}=\sigma_v^2\,\boldsymbol{I}_m. \tag{10}$$

On Equation (4) $\boldsymbol{u}_t$ is assumed independent and AR(1) so the expectation value and covariance matrix of vector $\boldsymbol{u}=(\boldsymbol{u}_1',\boldsymbol{u}_2',\ldots\boldsymbol{u}_T')'$ are

$$E(\boldsymbol{u})=\boldsymbol{0}\ \text{ dan }\ Cov(\boldsymbol{u})=\boldsymbol{G_2}=\frac{1}{(1-\rho^2)}\sigma_\varepsilon^2\,\boldsymbol{\Gamma}, \tag{11}$$

where $\boldsymbol{\Gamma}$ is symmetrical matrix sized $T\times T$ with element (t,t') is $\rho^{|t-t'|}$, $t=1,\ldots,T$ and $t'=1,\ldots,T$. Matrix $\boldsymbol{\Gamma}$ is

$$\boldsymbol{\Gamma}=\begin{bmatrix}1 & \rho & \cdots & \rho^{T-1}\\ \rho & 1 & \ddots & \vdots\\ \vdots & \vdots & \ddots & \vdots\\ \rho^{T-1} & \cdots & \rho & 1\end{bmatrix} \tag{12}$$

Equation (10) and equation (11) are rearranged in the form of relative covariance factor $\boldsymbol{\Lambda_1}$ and $\boldsymbol{\Lambda_2}$ so that $\boldsymbol{G_1}=\sigma_v^2\boldsymbol{\Lambda_1}\boldsymbol{\Lambda_1'}$ and $\boldsymbol{G_2}=\sigma_\varepsilon^2\boldsymbol{\Lambda_2}\boldsymbol{\Lambda_2'}$ where $\boldsymbol{\Lambda_1}\boldsymbol{\Lambda_1'}=\boldsymbol{I}_m$, $\boldsymbol{\Lambda_1}$ is a Cholesky decomposition matrix of d $\boldsymbol{I}_m$, and $\boldsymbol{\Lambda_2}\boldsymbol{\Lambda_2'}=\frac{\boldsymbol{\Gamma}}{(1-\rho^2)}$ where $\boldsymbol{\Lambda_2}$ is a Cholesky decomposition matrix of first-order autoregressive correlation matrix $\frac{\boldsymbol{\Gamma}}{(1-\rho^2)}$.

As a result, the proposed model on equation (4) becomes

$$\boldsymbol{\eta}=\boldsymbol{X\beta}+\boldsymbol{Z}_1\boldsymbol{\Lambda}_1\boldsymbol{v}^*+\boldsymbol{Z}_2\boldsymbol{\Lambda}_2\boldsymbol{u}^*=\boldsymbol{X\beta}+\boldsymbol{Z}_1^*\boldsymbol{v}^*+\boldsymbol{Z}_2^*\boldsymbol{u}^* \tag{13}$$

where $\boldsymbol{v}^* \sim N(\mathbf{0}, \sigma_v^2 \boldsymbol{I})$ and $\boldsymbol{u}^* \sim N\left(\mathbf{0}, \frac{1}{(1-\rho^2)}\sigma_\varepsilon^2 \boldsymbol{I}\right)$.

If link function is a logarithmic function then

$$\boldsymbol{\eta} = \log \boldsymbol{\mu} = \boldsymbol{X\beta} + \boldsymbol{Z_1^* v^*} + \boldsymbol{Z_2^* u^*} \text{ or} \tag{14}$$
$$\boldsymbol{\mu} = \exp \boldsymbol{\eta} = \exp (\boldsymbol{X\beta} + \boldsymbol{Z_1^* v^*} + \boldsymbol{Z_2^* u^*})$$

where $\boldsymbol{v}^* \sim N(\mathbf{0}, \sigma_v^2 \boldsymbol{I})$ and $\boldsymbol{u}^* \sim N\left(\mathbf{0}, \frac{1}{(1-\rho^2)}\sigma_\varepsilon^2 \boldsymbol{I}\right)$.

2.3 Parameter Estimation

In equation (14) above, it can be seen that the first parameters to be estimated are $\boldsymbol{\beta}$, $\boldsymbol{v}^*$, and $\boldsymbol{u}^*$. Those parameters are estimated by h-likelihood method. According to [3], h-likelihood is defined as

$$h = l_1 + l_2 + l_3 \tag{15}$$

where $l_1 = \log f(\boldsymbol{y} \mid \boldsymbol{v}^*, \boldsymbol{u}^*)$ is the log-density function for $\boldsymbol{y}$ given $\boldsymbol{v}^*$ and $\boldsymbol{u}^*$, l_2 is the log-density function for $\boldsymbol{v}^*$ with $\boldsymbol{v}^* \sim N(\mathbf{0}, \sigma_v^2 \boldsymbol{I})$, and l_3 is the log-density function for $\boldsymbol{u}^*$ with $\boldsymbol{u}^* \sim N\left(\mathbf{0}, \frac{\sigma_\varepsilon^2}{(1-\rho^2)} \boldsymbol{I}\right)$. Since $\boldsymbol{y}$ has Tweedie compound Poisson distribution, $\boldsymbol{y}$ is belong to the MDE family with $Var(\mu) = \mu^p$ and $1 < p < 2$. As a result, the density function of $\boldsymbol{y}$ is as defined in equation (1).

The parameter estimation solution can be done by maximizing the h-likelihood function above, by finding the solution of the equation $\frac{dh}{d\boldsymbol{\beta}} = 0$, $\frac{dh}{d\boldsymbol{v}^*} = 0$, and $\frac{dh}{d\boldsymbol{u}^*} = 0$. Since the Tweedie compound Poisson distribution is not a closed-form, then a numerical approximation is carried out. Parameter estimation can be solved by the Newton-Raphson method as follows

$$\begin{bmatrix} \boldsymbol{\beta}_{k+1} \\ \boldsymbol{v}_{k+1}^* \\ \boldsymbol{u}_{k+1}^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_k \\ \boldsymbol{v}_k^* \\ \boldsymbol{u}_k^* \end{bmatrix} + \boldsymbol{H}_k^{-1} \left(\begin{matrix} \frac{dh}{d\boldsymbol{\beta}} \\ \frac{dh}{d\boldsymbol{v}^*} \\ \frac{dh}{d\boldsymbol{u}^*} \end{matrix} \right) \Bigg| \begin{matrix} \boldsymbol{\beta} = \boldsymbol{\beta}_k \\ \boldsymbol{v}^* = \boldsymbol{v}_k^* \\ \boldsymbol{u}^* = \boldsymbol{u}_k^* \end{matrix} \tag{16}$$

where Hessian matrix $\boldsymbol{H}$ is

$$\boldsymbol{H} = \begin{bmatrix} -E\left(\frac{d^2h}{d\boldsymbol{\beta}^2}\right) & -E\left(\frac{d^2h}{d\boldsymbol{\beta} d\boldsymbol{v}^*}\right) & -E\left(\frac{d^2h}{d\boldsymbol{\beta} d\boldsymbol{u}^*}\right) \\ -E\left(\frac{d^2h}{d\boldsymbol{v}^* d\boldsymbol{\beta}}\right) & -E\left(\frac{d^2h}{d\boldsymbol{v}^{*2}}\right) & -E\left(\frac{d^2h}{d\boldsymbol{v}^* d\boldsymbol{u}^*}\right) \\ -E\left(\frac{d^2h}{d\boldsymbol{u}^* d\boldsymbol{\beta}}\right) & -E\left(\frac{d^2h}{d\boldsymbol{u}^* d\boldsymbol{v}^*}\right) & -E\left(\frac{d^2h}{d\boldsymbol{u}^{*2}}\right) \end{bmatrix}. \tag{17}$$

In every iteration, $(k+1)^{th}$ solutions satisfy the above equation. Iteration continue until solution converged.

After getting the parameter estimates of $\boldsymbol{\beta}$, $\boldsymbol{v}^*$, and $\boldsymbol{u}^*$, then the variance parameters will be estimated. To get the estimation of variance parameters, [13] defined *adjusted hierarchical likelihood* h_A as

$$h_A = h - \frac{1}{2}\ln\left\{\det\left(\frac{\boldsymbol{H}}{2\pi}\right)\right\} \tag{18}$$

where h is h-likelihood function from equation (15) and $\boldsymbol{H}$ is a Hessian matrix from equation (17).

The adjusted profile hierarchical likelihood is $h_p = h_A|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{v}^*=\widehat{\boldsymbol{v}}^*,\boldsymbol{u}^*=\widehat{\boldsymbol{u}}^*}$ where $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{v}}^*$, and $\widehat{\boldsymbol{u}}^*$ are estimated values from equation (16).

Variance parameters σ_v^2 and σ_ε^2 can be obtained by iteratively solving the equation

$$\begin{pmatrix}\sigma_v^2{}_{k+1}\\ \sigma_\varepsilon^2{}_{k+1}\end{pmatrix} = \begin{pmatrix}\sigma_v^2{}_{k}\\ \sigma_\varepsilon^2{}_{k}\end{pmatrix} + \boldsymbol{J}^{-1}\begin{pmatrix}\frac{dh_A}{d\sigma_v^2}\\ \frac{dh_A}{d\sigma_\varepsilon^2}\end{pmatrix}\Bigg|_{\substack{\sigma_v^2=\sigma_v^2{}_k\\ \sigma_\varepsilon^2=\sigma_\varepsilon^2{}_k}} \tag{19}$$

until convergent where $\boldsymbol{J}$ is Hessian matrix containing the second derivatives of adjusted hierarchical likelihood $\mathrm{h_A}$ function.

3. SIMULATION STUDY

3.1 Simulation Design

In this section a simulation study will be conducted to evaluate the goodness of the developed model. The determination of the parameter values in this simulation refers to [11] and [14]. The stages of the simulation are as follows:

1. Generating sample data with the following conditions
   a. The total areas are 35 regions (M = 35) and observation times are12 (T = 12). In each region and time there are 10 observations so that in total there are $S = 4200$ observations.
   b. Determine three categories of autoregressive coefficient values ρ = 0.2, ρ = 0.5, and ρ = 0.8.

c. Determine three categories of variance from the region random effect, $\sigma_v^2 = 0.1$, $\sigma_v^2 = 0.5$, and $\sigma_v^2 = 1$.

d. Set $\sigma_\varepsilon^2 = 0.5$, index parameter $p = 1.5$, and dispersion parameter $\phi = 1$.

e. Generate covariate variable $\boldsymbol{x}$~Normal (0,1).

f. Set the initial value $\boldsymbol{\beta_0} = (0,0)$, region $\boldsymbol{v}_0$, where $\boldsymbol{v}_0 \sim N(\boldsymbol{0}, \sigma_v^2 \boldsymbol{I}_M)$, and time $\boldsymbol{u}_0$, where $\boldsymbol{u}_0 \sim N\left(0, \frac{\sigma_\varepsilon^2}{(1-\rho^2)} \boldsymbol{I}_T\right)$. The values ρ, σ_v^2, and σ_ε^2 are taken from steps 1.b, 1.c, and 1.d.

g. Get matrices $\boldsymbol{Z_1^* = Z_1 \Lambda_1}$ and $\boldsymbol{Z_2^* = Z_2 \Lambda_2}$ where $\boldsymbol{Z_1 = I_M \otimes Z_{1i}}$, $\boldsymbol{Z_2 = I_T \otimes Z_{2t}}$, $\boldsymbol{Z_{1i} = 1_{n_i}}$ and $\boldsymbol{Z_{2t} = 1_{n_t}}$. $\boldsymbol{\Lambda_1}$ and $\boldsymbol{\Lambda_2}$ are Cholesky decomposition matrices so

$$\boldsymbol{\Lambda_1 \Lambda_1'} = \boldsymbol{I}_m \text{ and } \boldsymbol{\Lambda_2 \Lambda_2'} = \frac{\boldsymbol{\Gamma}}{(1-\rho^2)} \text{ where } \boldsymbol{\Gamma} = \begin{bmatrix} 1 & \rho & \cdots & \rho^{T-1} \\ \rho & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \rho \\ \rho^{T-1} & \cdots & \rho & 1 \end{bmatrix}$$ and ρ is specified on 1.b.

h. Generate response variable, $\boldsymbol{y}$, has Tweedie Compound Poisson distribution with parameter $\boldsymbol{\mu(y|v_0, u_0) = \mu} = \exp(\boldsymbol{X\beta_0 + Z_1^* v_0 + Z_2^* u_0})$, index parameter $p = 1.5$, and dispersion parameter $\phi = 1$.

2. Estimate the parameters of the fixed effect $\boldsymbol{\beta}$, the region random effect $\boldsymbol{v}^*$, and the time random effect $\boldsymbol{u}^*$ using the h-likelihood method until convergent.

3. Estimate variance component σ_v^2 and σ_ε^2 using *adjusted profile likelihood* h_A method until convergent.

4. Perform step 2 again by using the initial values $\boldsymbol{\beta}$, $\boldsymbol{v}^*$,$\boldsymbol{v}^*$, σ_v^2, and σ_ε^2 obtained from step 2 and 3 above.

5. Find the estimated value $\boldsymbol{\hat{\eta} = X\hat{\beta} + Z_1^* \hat{v}^* + Z_2^* \hat{u}^*}$ then $\boldsymbol{\hat{\mu}} = \exp \boldsymbol{\hat{\eta}}$ or $E(\hat{\boldsymbol{y}}|\hat{\boldsymbol{v}}, \hat{\boldsymbol{u}})$.

6. Repeat steps 1 to 6 above as many as $R = 100$ times.

7. Evaluate the model as in [14] by

   - Mean Relative Bias $= \frac{1}{R} \sum_{i=1}^{R} \frac{(\hat{\mu}_i - \mu_i)}{\mu_i}$.
   - $MSE = \frac{1}{R} \sum_{i=1}^{R} (\hat{\mu}_i - \mu_i)^2$.

3.2 Simulation Result

Modeling using directly generated data always produce not invertible Hessian matrix. It is necessary to scale the covariate and the response variables. The trial and error results by dividing the values of those two variables by 10000 produce a invertible Hessian matrix so that further modeling is done by scaling the data and based on the algorithm steps above. Model 1 is as in equation (4) then compared with the Model 2

$$E(y_{itj}|v_i, u_{it}) = \mu_{itj},$$
$$\log \mu_{itj} = \beta_0 + x_{itj}\boldsymbol{\beta_1} + v_i + u_t \quad (20)$$

that is, a model that does not have autoregressive assumptions on the time random effect. The initial values of the parameters are assumed similar to Model 1. Other assumption is σ_ε^2 becomes variance of u_t.

In this simulation study, estimating the parameters of both models was carried out using the h-likelihood method. The computational program was built using the R programming software. Much like [11] studies were conducted with known autoregressive coefficient values. In addition, the dispersion parameter values and Tweedie distribution index parameters were also assumed to be known referring to [14]. Simulation results with 100 replications can be seen in Table 1 and Figure 1

Table 1 Simulation result from 100 replications

Model 1 with time AR (1)

Rho	0.2			0.5			0.8		
Variance	0.1	0.5	1	0.1	0.5	1	0.1	0.5	1
Relative Bias	1.730	1.177	0.863	1.647	0.643	0.432	0.875	0.376	0.176
MSE	1.58E-06	3.39E-06	4.68E-06	1.62E-05	1.21E-05	5.65E-06	0.002	8.91E-05	8.91E-05

Model 2 without time AR (1)

Variance	0.1	0.5	1
Relative Bias	1.5118	0.7067	0.4602
MSE	0.0000	7.12E-07	1.73E-07

From Table 1 Model 1, for the same autoregressive coefficient, the greater the regions variances the smaller the relative bias. Then, the greater the autoregressive coefficient the smaller the relative bias produced. This shows that, the greater the autoregressive coefficient, the more unbiased the estimator can be. In addition, the greater the region variances the more unbiased the estimators produced. In other words, the autoregressive coefficient and regional variance influence the biasness of Model 1 estimators. Model 1 with a small autoregressive coefficient produce more bias than Model 2. For the largest autoregressive coefficient, Model 1 produces an unbiased estimator. On medium autoregressive coefficient, Model 2 produces an unbiased estimator only for a small variance. Generally, Model 1 produces more an unbiased estimator than Model 2. Figure 2 explains this with $rho = 0$ is for Model 2 and the remains are for Model 1. The simulation results in Table 1 also show the MSE values of the two models approaching zero. This shows that both models describe the data very well.

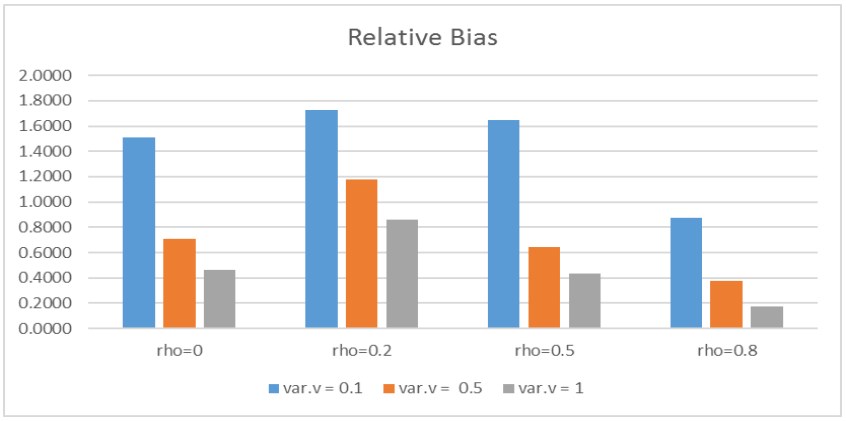

Figure 1 Relative Bias

4. APPLICATION STUDY

The data used comes from the Financial Services Authority (FSA), which is a report on the risk profile and loss of motor vehicle insurance of a general insurance company in Indonesia in 2014. The total claim becomes the observed response variable, the deductible becomes the fixed effect, and the region code as many as 35 regions as well as the month of occurrence to be random effects. The month of occurrence that are considered to follow the first order autoregressive process then.

For the purposes of this study the data were partly drawn through simple random sampling of 175,000 items of actual data. In each region and month, 10 policy numbers were taken.

Similar to the simulation study, the application study was also carried out on two models. Model 1 is using the autoregressive assumption on the time random effect as defined on equation (4) and Model 2 is without the autoregressive assumption on the time random effect as on equation (20).

To show that the response variable has Tweedie compound Poisson distribution, the index parameters of the Tweedie compound Poisson distribution must be between 1 and 2, or $1 < p < 2$. The Poisson compound distribution index of Tweedie p is obtained from the Tweedie package in R with the tweedie.profile () function. The program package also produces dispersion parameters ϕ. Figure 2 shows the highest likelihood profile value achieved by the index parameter p between 1.5 and 1.6.

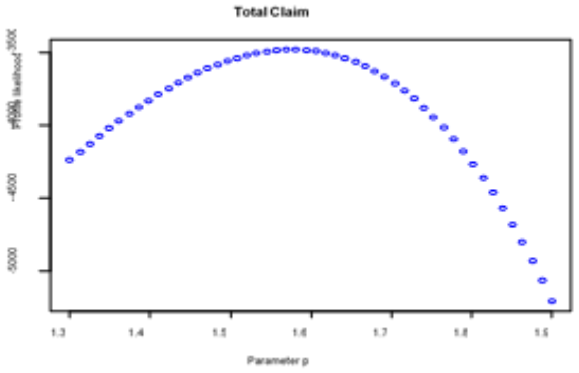

Figure 2 Likelihood profile of Total Claim

Table 2 shows total claim data has Tweedie compound Poisson distribution with index parameter $p = 1.58$ and dispersion parameter $\phi = 1.27$.

Table 2 Index and dispersion parameter

	Value
p	1.581633
ϕ	1.269497

The correlation coefficient between the total claims data at time t and $t-1$ is equal to ρ = 0.362 and the initial value $\sigma_\varepsilon^2 = 0.1$ is obtained by finding the variance of the difference in the total claims at time t reduced by multiplication between the correlation coefficient ρ with the total claim time t -1. The initial value of the area random effect $\sigma_\varepsilon^2 = 0.5$, and the regression parameters $\beta_0 = 0$, and $\beta_1 = 0$ refer to [4].

From Table 3 below it can be seen that the two models produce different estimates of regression parameters. The parameter β_0 in Model 1 is -2.357819 and β_1 is 13.291004. While in Model 2, the estimated value of the parameter β_0 is -2.378214 and β_1 is 13.2967676. As a result, the interpretation of the fixed effect on Model 1 is that if the deductible changes to one rupiah, then the expected total value of claims will change 56004.37 rupiahs, while in Model 2 if the deductible changes to one rupiah, then the expected total value of claims will change 55241.07 rupiahs.

Table 3 Estimate Parameter

	Model 1 With AR(1) assumption		Model 2 Without AR(1) assumption	
	Estimate	SE	Estimate	SE
Fixed Effects				
Intercept(β_0)	-2.357819	0.2347547	-2.378214	0.204398
Deductible(β_1)	13.291004	0.4012697	13.297676	0.3995944
Random Effect Variance				
Region	1.145744	0.1575271	1.104661	0.1504604
Time	0.133413	0.0048725	0.114967	0.0044087
Residual	0.678343		0.678252	

Table 4 Absolute Bias and MSE

	Mean Absolute Bias	MSE
Model 1 With AR(1) assumption	0.2283721	0.6691346
Model 2 Without AR(1) assumption	0.2376156	0.6783145

In the estimation of the random effects variance, Model 2 produces a relatively smaller variance than Model 1. In both models, the region's random effect is greater than the residual. This shows that there is diversity in the total value of claims between regions. Whereas the variance estimate of time random effect is smaller than the variance estimate of residuals. This shows there is a total diversity of claims in time but there is no variation in total claims between time

Evaluation of the model is done by calculating the mean absolute bias value and MSE of both models. Table 4 shows that the Model 1 produces a smaller mean absolute bias compared to the Model 2. In addition, the MSE value of Model 1 is smaller than the Model 1. This shows that in the application study, Model 1 is relatively better than the Model 2.

5. CONCLUSION

The h-likelihood method can be used to estimate the regression and variance parameters in GLMM with response variables has Tweedie compound Poisson distribution with two random effects namely region and time which are assumed to follow the first-order autoregressive process. The simulation study shows that for models with a time that are assumed to follow the first-order autoregressive process, the greater the autoregressive coefficient the relative bias produced is smaller and the greater the variance of regions the more unbiased the estimators produced. In other words, the autoregressive coefficient and regional variance influence the predictor's biasness.

MSE values close to zero indicate that the model describes the data well. The application study shows the absolute bias value and the MSE model with a time which is assumed to follow the first-order autoregressive process is smaller when compared to models with a time without first-order autoregressive assumptions. In general, models with a time that are assumed to follow the first-order autoregressive process are relatively better when compared to models with a time random effect without first-order autoregressive assumptions.

ACKNOWLEDGMENT

We would like to thank Kemenristek Dikti Republic of Indonesia for funding this research through 2020 Doctoral Grant Program.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] O.A. Quijano Xacur, J. Garrido, Generalised linear models for aggregate claims: to Tweedie or not?, Eur. Actuar. J. 5 (2015), 181–202.

[2] B. Jørgensen, Exponential dispersion models (with discussion), J. R. Stat. Soc. B. 49 (1987), 127–162.

[3] C. Lee, Y. Lee, Sire evaluation of count traits with a Poisson-Gamma hierarchical generalized linear model. Asian-Aust. J. Animal Sci. 11 (6) (1998), 642-647.

[4] H. Chandra, N. Salvati, R. Chambers, Small area estimation under a spatially non-linear model spatial statistics, Comput. Stat. Data Anal. 126 (2018), 19-38.

[5] J.S. Pai, K.J. Shand, X. Wang, Compound Poisson model with covariates, N. Amer. Actuar. J. 10 (4) (2006), 219-234.

[6] N.C. Dzupire, P. Ngare, L. Odongo, A Poisson gamma model for zero inflated rainfall data, J. Probab. Stat. 2018 (2018), 1012647.

[7] N.E. Breslow, D.G. Clayton, Approximate inference in Generalized Linear Mixed Models, J. Amer. Stat. Assoc.

88 (421) (1993), 9-25.

[8] P.K. Dunn, Occurrence and quantity of precipitation can be modelled simultaneously, Int. J. Climatol. 24 (2004), 1231–1239.

[9] P.K. Dunn, G.K. Smyth, Series evaluation of Tweedie exponential dispersion model densities, Stat. Comput. 15 (2005), 267-280.

[10] P.K. Dunn, G.K. Smyth, Evaluation of Tweedie exponential dispersion model densities by Fourier inversion, Stat. Comput. 18 (2008), 73–86.

[11] S. Muchlisoh, A. Kurnia, K.A. Notodiputro, IW. Mangku, Small Area Estimation of Unemployment Rate Based on Unit Level Model with First Order Autoregressive Time Effects, J. Appl. Probab. Stat. 12 (2) (2017), 51-63.

[12] S.G. Candy, Modelling catch and effort data using generalized linear models, the Tweedie distribution, random vessel effects and random stratum-by-year effects, CCAMLR Sci. 11 (2004), 59–80

[13] Y. Lee, J. A. Nelder, Hierarchical generalized linear model, J. R. Statis. Soc. B. 58 (1996), 619-678.

[14] Y. Zhang, Likelihood-based and Bayesian methods for Tweedie compound Poisson linear mixed models, Stat. Comput. 23 (2013), 747-757.