JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 9, No. 1, January 2025, pp. 221-241

A Posteriori Premium Rate Calculation using Poisson-Gamma Hierarchical Generalized Linear Model for Vehicle Insurance

Fevi Novkaniza[1], Irene Devina Putri[1], Rahmat Al Kafi[1], Sindy Devila[1]
[1]Department of Matematics, Universitas Indonesia, Kampus FMIPA UI Depok, Indonesia
fevi.novkaniza@sci.ui.ac.id

Article History:
Received : 13-11-2024
Revised : 29-12-2024
Accepted : 02-01-2025
Online : 06-01-2025

Keywords:
Claim Frequency;
Conjugate Gradient;
Longitudinal Data;
Maximum Likelihood;
Random effects.

ABSTRACT

This study develops and applies the Poisson-Gamma Hierarchical Generalized Linear Model (PGHGLM) to address the challenge of determining accurate and fair premium rates in vehicle insurance. The PGHGLM models a mixture distribution for the response variable, influenced by random effects, and employs a logarithmic link function. Parameter estimation is conducted using the maximum likelihood method. However, since analytical estimation is not feasible, the numerical conjugate gradient method, specifically the Fletcher-Reeves algorithm, is utilized. The implementation of the PGHGLM uses the longitudinal Claimslong dataset, incorporating driver age as a covariate. The main contribution of this research lies in integrating a priori risk classification with a posteriori adjustment based on longitudinal claim frequency data. For datasets without covariates, trend parameters are incorporated into the model. For datasets with covariates, such as driver age, the average claim frequency is computed for each age category. Results show that posteriori premium rates increase with rising claim frequency from the previous year, with higher claim frequencies leading to larger rate adjustments in the subsequent year. Through the PGHGLM, a posteriori premium rate estimates are obtained for each age group of vehicle insurance policyholders. This study demonstrates the practical application of the PGHGLM in calculating precise premium rates. By analyzing a longitudinal vehicle insurance dataset, the model generates annual a posteriori premium rates tailored to age groups. These findings underscore the PGHGLM's robust methodological framework and its potential to enhance premium fairness, enable risk-adjusted pricing, and better tailor insurance products to diverse policyholder profiles.

https://doi.org/10.31764/jtam.v9i1.27837

A. INTRODUCTION

Risk and uncertainty are inherent aspects of human life, often resulting in financial losses. Insurance provides an effective mechanism to manage these risks by transferring them to an insurer in exchange for a premium (Lanfranchi & Grassi, 2022; Rejda, 2017; Rumson & Hallett, 2019a; Sheehan et al., 2023a). Among various insurance types, vehicle insurance plays a vital role in mitigating financial risks from accidents, theft, and damages, offering coverage such as liability, medical benefits, and accident compensation. This protection is particularly important for drivers facing risks from factors like poor road conditions, driver inexperience, and environmental hazards (Bagariang & Raharjanti, 2023a; Hsu et al., 2016a). To receive the protection, guarantee and risk transfer benefits offered, the policyholder must pay a premium to the insurance company. According to Indonesian Law No. 40 of 2014 on Insurance, a premium is the amount of money determined by the insurance company and agreed upon by

the policyholder to be paid based on the insurance agreement, or the amount determined based on the provisions of legislation that underlie compulsory insurance programs to obtain benefits. According to the law, there are four components of a premium: (i) basic premium, listed in the insurance policy with a fixed amount as long as there is no change in the protection guarantee, (ii) additional premium, which must be paid if there is a change in the policyholder's data resulting in expanded risk coverage, (iii) premium reduction, which represents a discount on the premium due to certain conditions, (iv) company tariff, which is the rate set by the insurance company association to prevent unhealthy competition among insurance companies. In an insurance agreement, there is also the total coverage amount, which is the amount of money that will be paid by the insurance company to the policyholder according to the agreed benefits. The most used premium calculation method is multiplying the premium rate by the total coverage amount (Niehaus, 2016; Lima Ramos, 2017).

An insurance pool, consisting of a collection of risks and premiums, is referred to as an insurance portfolio. An insurance portfolio can have several distinct risk classes based on the risk profile of each policyholder. When an insurance portfolio includes multiple risk classes and premiums, it is described as a heterogeneous portfolio (Boonen & Liu, 2022; Frostig, 2001; Frostig et al., 2007). Due to this heterogeneity, an insurance company cannot set a single premium rate for all risks in the portfolio, as it would be detrimental to both the company and the policyholders. For example, with a single premium rate, individuals with a low-risk profile would overpay, while those with a high-risk profile would underpay, benefitting disproportionately because the premium charged is too low (Antonio & Valdez, 2012). To prevent such inequities and potential losses, insurance companies must determine premium rates based on the policyholder's specific risk profile. In vehicle insurance, these risk profiles can be classified according to factors such as the driver's age, vehicle age, vehicle type, geographic location, and other attributes (Levitas et al., 2022). The policyholder's risk profile reflects the likelihood of filing a claim, defined as a request to the insurance company for compensation under the terms of the insurance policy (Levitas et al., 2022). A 2022 study by the Insurance Information Institute (III) in New York found that individuals with high-risk profiles are more likely to file claims. To address this, a risk classification scheme is employed to group risks into homogeneous categories, ensuring that policyholders with similar risk profiles pay premiums proportionate to their claim frequency and the benefits they receive (Lee et al., 2020).

In general insurance, premium rate calculation methods include judgment rating, class rating, and experience ratemaking (Rejda & McNamara, 2017). For motor vehicle insurance, experience ratemaking is commonly used, which involves two stages: a priori and a posteriori premium calculations. The a priori stage uses a risk classification scheme based on measurable factors associated with each policyholder's risk profile (Boucher & Denuit, 2006; Boucher & Inoussa, 2014). This stage is typically applied to new policyholders for whom the insurance company lacks historical claim data, relying instead on measurable attributes like age and gender at the time of observation (Tseung et al., 2022). However, unobservable or unmeasurable factors not considered in the a priori stage contribute to heterogeneity within the insurance portfolio and affect claim frequency (Antonio & Valdez, 2012; Boucher & Inoussa, 2014; Wolny-Dominiak & Sobiecki, 2014). Over time, as the insurer accumulates historical

claim data and detailed information about each policyholder's risk profile, the process moves to the a posteriori stage of premium rate calculation (Tseung et al., 2022). The data collected, including historical claims and policyholder risk profiles, is referred to as longitudinal data (Boucher & Inoussa, 2014). In the a posteriori stage, premium rate calculations refine the a priori assessments by incorporating heterogeneity factors previously unaccounted for (Antonio & Valdez, 2012; Boucher & Inoussa, 2014). For example, in vehicle insurance, risk factors considered during the a priori stage include driver age, gender, marital status, vehicle usage duration, and geographic location (Antonio & Valdez, 2012). Conversely, unmeasurable heterogeneity factors such as individual driving skills, road condition knowledge, emotional states, reflexes, and accident avoidance behavior are incorporated during the a posteriori stage, informed by longitudinal historical claim data (Lee et al., 2020).

To model the relationship between claim frequency and the influencing risk and heterogeneity factors in a posteriori premium calculation, statistical models such as the Generalized Linear Model (GLM), Linear Mixed Model (LMM), and Generalized Linear Mixed Model (GLMM) have been developed (Antonio & Beirlant, 2005; Tseung et al., 2022). Laird and Ware (1982) introduced the LMM, which extends the GLM used for a priori premium calculations by accommodating heterogeneity and correlation within insurance portfolios based on longitudinal claim data. However, the GLM is less suitable for longitudinal data as it cannot account for heterogeneity and correlation between responses at different times (Wolny-Dominiak & Sobiecki, 2014). To address these issues, the LMM was extended to the GLMM, which incorporates random effects for longitudinal data distributed within the exponential family (Antonio & Beirlant, 2005; Lee et al., 2020). These random effects not only model correlations between observations but also capture heterogeneity factors within the insurance portfolio, allowing for adjustments to the parameters of the response variable distribution (Gupta et al., 2004). Despite its advantages, the GLMM assumes normally distributed random effects, limiting its applicability when this assumption does not hold.

Gning et al. (2023) addressed this limitation by developing the Hierarchical Generalized Linear Model (HGLM) as an extension of the GLMM, enabling random effects to follow distributions other than normal. This study builds on these advancements by introducing the Poisson-Gamma HGLM to improve the accuracy and fairness of premium rate calculations in motor vehicle insurance. The model addresses heterogeneity in policyholders' risk profiles, particularly unobservable risk factors often overlooked in traditional methods, which can lead to inefficiencies in premium rate-setting. By accommodating non-normal random effects, the Poisson-Gamma HGLM allows for more accurate modeling of claim frequency and risk profiles in heterogeneous insurance portfolios (Antonio & Valdez, 2012; Gning et al., 2023). This ensures policyholders pay premiums proportional to their risk while maintaining the insurer's financial stability, effectively filling gaps in existing premium rate-setting methodologies and enhancing fairness and precision in motor vehicle insurance.

B. RESEARCH METHODS

1. Poisson-Gamma Mixture Distribution

As in (Li et al., 2024; Shirazi & Lord, 2019; Wu, 2022), let Y as a discrete random variable following a Poisson distribution with parameter λ denoted as $Y \sim Poisson\,(\lambda)$ with the conditional distribution of Y given λ is

$$\Pr(Y = y|\Lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} \qquad y = 0, 1,2, \ldots.$$

The parameter λ is a random variable following a Gamma distribution with parameters α and β, denoted as: $\Lambda \sim Gamma\,(\alpha, \beta)$ with the probability density function of Λ is

$$f(\lambda) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}\ \lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}} \qquad \alpha, \beta > 0.$$

Thus, the Pdf of Poisson-Gamma mixture distribution is

$$Pr(Y = y) = \frac{\Gamma(\alpha + y)}{y!\,\Gamma(\alpha)}\left(\frac{1}{1+\beta}\right)^{\alpha}\left(\frac{\beta}{1+\beta}\right)^{y}, y = 0,1,2, \ldots$$

The characteristics of Poisson-Gamma mixture distribution is as follows:

a. Mean

By the law of total expectation, the mean of poisson-gamma mixture distribution is

$$E(Y) = E[E(Y|\Lambda)] = \alpha\beta$$

b. Variance

By the law of total variance, the variance of poisson-gamma mixture distribution is

$$Var\,(Y) = \Big(Var(Y|\Lambda) + Var\big(E(Y|\Lambda)\big)\Big) = \alpha\beta + \alpha\beta^{2}$$

c. Moment Generating Function (MGF)

$$M(t) = E[e^{ty}] = \left(\frac{1}{1+\beta-\beta e^{t}}\right)^{\alpha}$$

d. Characteristic Function

$$\phi(t) = E\left[e^{ity}\right] = \left(1+\beta-\beta e^{it}\right)^{-\alpha}$$

The pdf plot of poisson-gamma mixture distribution is as shown in figure 1 and figure 2 below:

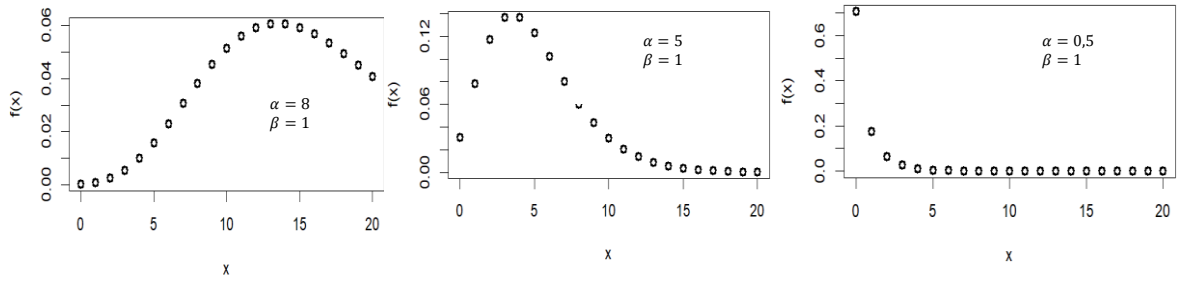

Figure 1. Poisson-Gamma Plot with varying values of α

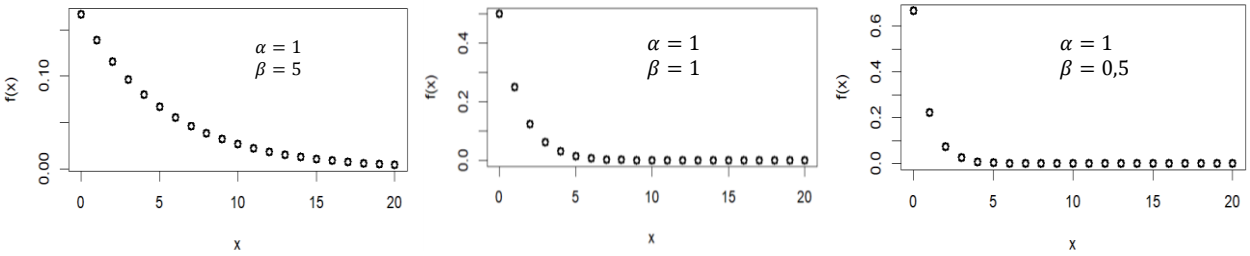

Figure 2. Poisson-Gamma Plot with varying values of β

2. Hierarchical Generalized Linear Model (HGLM)

The Hierarchical Generalized Linear Model (HGLM) is a statistical model particularly well-suited for modelling longitudinal data (Gning et al., 2023; Jin & Lee, 2024; Matsuyama, 2020). Longitudinal data involves repeated measurements on the same subjects over time, resulting in a complex structure where observations are correlated with one another. This correlation can arise from various factors, one of which is unobserved random effects. Given the presence of time correlations for each measurement, HGLM is a more appropriate model because it can accommodate these correlations by incorporating random effects into the model. For example, HGLM can be used to model a dataset of the number of claims from motor vehicle insurance. Suppose there is an insurance portfolio viewed as a single cluster. The number of claims occurring in this cluster at the initial time and subsequent times is likely correlated because policyholders within this portfolio cluster share similar risk characteristics. Let y be the response variable and u be the unobserved random component. The conditional log-likelihood for y given u has the form

$$l(\theta', \phi; y|u) = \frac{\{y\theta' - b(\theta')\}}{a(\phi)} + c(y, \phi)$$

where θ' denotes the canonical parameter and ϕ is the dispersion parameter. Let matrix $\boldsymbol{X}$ with $n \times J$ size denotes covariates and $\boldsymbol{\beta} = (\beta_1 \dots \beta_J)'$ denotes regression parameter with n represents the number of subject and J represents the number of covariates. Also, let μ be the conditional expectation of y given u and $\eta = g(\mu) = g(y|u)$ wheren $g(.)$ is link function in GLM with linear predictor given by:

$$\eta = g(\mu) = \eta + v = \boldsymbol{X\beta} + \boldsymbol{v}$$

where $v = v(u)$ is a strictly monotonic function of u. The distribution of u is assumed appropriately.

3. Poisson-Gamma Hierarchical Generalized Linear Model (PGHGLM)

In this section, there are two types of models that are discussed, which are model with covariates and without covariates. In PGHGLM with covariates, the representation of response variable Y_{kt} and covariates x_{kt} are explained, continued with the parameter estimation methods (Iqbal et al., 2023; Tawiah et al., 2020; Tzougas & Pignatelli di Cerchiara, 2021). Let's assume there are n individuals observed over a period T. The response variable used in the PGHGLM regression model is Y_{kt}, which is a count variable observed for individual k during period t, with $k = 1, \dots, n$ and $t = 1, \dots, T$. The representation of the response variable Y_{kt} is as follows:

$$\mathbf{Y_{kt}} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1T} \\ y_{21} \\ \vdots \\ y_{2T} \\ y_{n1} \\ \vdots \\ y_{nT} \end{bmatrix} \tag{1}$$

where y_{11} represents the observation of the first individual at the first period and y_{nT} represents the observation of the n-th individual at the T-th period. Suppose we want to determine the relationship between characteristics of k-th individual at the t-th period with response variable Y_{kt}, define covariate $x_{ktj}, j = 1, \dots, J$. The covariate matrix, denoted as $\mathbf{X} = (\boldsymbol{x_{kt1}}\ \boldsymbol{x_{kt2}}\ \dots\ \boldsymbol{x_{ktJ}})$ is of size $nT \times J$ as shown below

$$\mathbf{X} = \begin{bmatrix} x_{111} & x_{112} & \dots & x_{11J} \\ x_{121} & x_{122} & \dots & x_{12J} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1T1} & x_{1T2} & \dots & x_{1TJ} \\ x_{211} & x_{212} & \dots & x_{21J} \\ \vdots & \vdots & \ddots & \vdots \\ x_{nT1} & x_{nT2} & \dots & x_{nTJ} \end{bmatrix} \tag{2}$$

For k-th individual, it is assumed that there exist a real and positive random characteristic $\Theta_k, k = 1, \dots, n$ called random effect. Parameter of the regression model is expressed by vector $\boldsymbol{\beta} = \left(\beta_1 \dots \beta_J\right)'$. Poisson-Gamma HGLM requires the three following assumptions (Gning et al., 2023):

(A1) the random variable Θ_k is independent for $k = 1, \dots, n$ and $Y_{k1}, \dots, Y_{kT}$ also independen for $k = 1, \dots, n$; (A2) for each k and $\theta_k > 0$, the conditional distribusion of Y_{kt} given Θ_k is the Poisson distribution

$$Y_{kt}|\Theta_k \sim Poisson\ (\lambda_{kt}\Theta_{kt})$$

where the pdf form is as follows:

$$P(Y_{kt} = y_{kt} \mid \Theta_k = \theta) = \frac{e^{-\lambda_{kt}\theta}\lambda_{kt}\theta^{y_{kt}}}{y_{kt}!} \tag{3}$$

with $E(Y_{kt}|\Theta_k) = \lambda_{kt}\Theta_{kt}$, where $\lambda_{kt} = e^{\boldsymbol{x'_{kt}\beta}}$, $k = 1, \dots, n$; $t = 1, \dots, T$;

(A3) for each k the distribution of Θ_k is $gamma(a, a)$ with the pdf form is as follows:

$$g(\theta) = \frac{a^a}{\Gamma(a)}\theta_k^{a-1}e^{-a\theta_k} \tag{4}$$

To determine the non-conditional distribution of Y_{kt} that includes Θ_k as the random effect, mixing distribution technique is used. Thus, from (A2) and (A3), the pdf for Y_{kt} is:

$$P(Y_{kt} = y_{kt}) = \frac{\Gamma(a + y_{kt})}{y_{kt}!\,\Gamma(a)}\left(\frac{a}{(a + \lambda_{kt})}\right)^a\left(\frac{\lambda_{kt}}{(a + \lambda_{kt})}\right)^{y_{kt}} \tag{5}$$

for $y_{kt} = 0,1,2, \dots$ and 0 elsewhere. The cummulative distribution form of Y_{kt} is

$$\begin{aligned} F(y) &= \sum_{y_{kt}} P(Y_{kt} = y_{kt}),\ y \in \mathbb{R} \\ &= \sum_{y_{kt}} \frac{\Gamma(a + y_{kt})}{y_{kt}!\,\Gamma(a)}\left(\frac{a}{(a + \lambda_{kt})}\right)^a\left(\frac{\lambda_{kt}}{(a + \lambda_{kt})}\right)^{y_{kt}} \end{aligned}$$

Based on the law of total expectation, the mean of Y_{kt} is

$$E(Y_{kt}) = E[E(Y_{kt}|\Theta_k)] = \lambda_{kt} \tag{6}$$

and based on the law of total variance, the variance of Y_{kt} is

$$\begin{aligned} V(Y_{kt}) &= V\big(E(Y_{kt}|\Theta_{\mathrm{k}})\big) + E\big(V(Y_{kt}|\Theta_{\mathrm{k}})\big) \\ &= \frac{\lambda_{kt}^2}{a} + \lambda_{kt} \end{aligned} \tag{7}$$

The link function used of PGHGLM is the log function, expressed as $\eta = g(\mu) = \log(\mu)$ with the linear predictor form as follows:

$$g(\mu) = \ln[E(Y_{kt})] = \ln[\lambda_{kt}] = \boldsymbol{x}'_{\boldsymbol{kt}}\boldsymbol{\beta} \tag{8}$$

thus based on the property of natural logarithm, λ_{kt} is

$$\lambda_{kt} = \exp(\boldsymbol{x}'_{\boldsymbol{kt}}\boldsymbol{\beta}) \tag{9}$$

The representation of PGHGLM is:

$$\ln[E(Y_{kt})] = \boldsymbol{x}'_{\boldsymbol{kt}}\boldsymbol{\beta}_i\,, \qquad i = 1, \dots, j \tag{10}$$

with $\boldsymbol{\beta} = (\beta_1, \dots, \beta_j)'$ is the PGHGLM regression parameter that should be estimated. The parameter estimation method that is used is maximum likelihood estimation (MLE). The steps to obtain the likelihood function is as follows: First, for any $k = 1, \dots, n$, given Θ_k, the count variables $Yk_1, \dots, Yk_T$ are assumed independent. Therefore, their joint probability function is

$$\begin{aligned} & Pr(Y_{k1} = y_{k1}, \dots, Y_{kT} = y_{kT}) \\ &= \int_0^\infty P(Y_{k1} = y_{k1}, \dots, Y_{kT} = y_{kT}|\Theta_k) g(\theta_k)\, d\theta_k \\ &= \prod_{t=1}^{T} \frac{\lambda_{kt}^{y_{kt}}}{y_{kt}!} \frac{a^a}{\Gamma(a)} \frac{\Gamma(a + \sum_{t=1}^{T} y_{kt})}{(a + \sum_{t=1}^{T} \lambda_{kt})^{a+\sum_{t=1}^{T} y_{kt}}} \end{aligned} \tag{11}$$

Based on equation (11), the likelihood function can be constructed as

$$L(a, \boldsymbol{\beta}) = \prod_{k=1}^{n} \left[\frac{\Gamma(a + \sum_{t=1}^{T} y_{kt})}{(a + \sum_{t=1}^{T} \lambda_{kt})^{a+\sum_{t=1}^{T} y_{kt}}} \frac{a^a}{\Gamma(a)} \prod_{t=1}^{T} \left(\frac{\lambda_{kt}^{y_{kt}}}{y_{kt}!} \right) \right] \tag{12}$$

Based on equation (9), equation (12) can also be expressed as

$$\begin{aligned} & L(a, \boldsymbol{\beta}) \\ &= \prod_{k=1}^{n} \left(\frac{a^a}{\Gamma(a)} \frac{\Gamma(a + \sum_{t=1}^{T} y_{kt})}{(a + \sum_{t=1}^{T} \exp(\boldsymbol{x}'_{\boldsymbol{kt}}\boldsymbol{\beta}))^{a+\sum_{t=1}^{T} y_{kt}}} \prod_{t=1}^{T} \left(\frac{\exp(\boldsymbol{x}'_{\boldsymbol{kt}}\boldsymbol{\beta})^{y_{kt}}}{y_{kt}!} \right) \right) \end{aligned} \tag{13}$$

Let $s_k = \sum_{t=1}^{T} y_{kt}$ and $\mu_k = \sum_{t=1}^{T} \exp(\boldsymbol{x}'_{\boldsymbol{kt}}\boldsymbol{\beta})$, thus the likelihood function from equation (13) will be

$$L(a, \boldsymbol{\beta}) = \prod_{k=1}^{n} \left(\frac{a^a}{\Gamma(a)} \frac{\Gamma(a + s_k)}{(a + \mu_k)^{a+s_k}} \prod_{t=1}^{T} \left(\frac{\exp(\boldsymbol{x}'_{\boldsymbol{kt}}\boldsymbol{\beta})^{y_{kt}}}{y_{kt}!} \right) \right) \tag{14}$$

Second, to obtain the estimation for parameters a and $\boldsymbol{\beta}$ that maximize the likelihood function in (14), log-likelihood function is constructed as shown below:

$$\begin{aligned}\ell(a,\boldsymbol{\beta}) &= \ln[L(a,\boldsymbol{\beta})] \\ &= \ln\prod_{k=1}^{n}\left[\left[\prod_{t=1}^{T}\left(\frac{\exp(\boldsymbol{x_{kt}'\beta})^{y_{kt}}}{y_{kt}!}\right)\right]\frac{a^a}{\Gamma(a)}\frac{\Gamma(a+s_k)}{(a+\mu_k)^{a+s_k}}\right] \\ &= \boldsymbol{Y'X\beta} - \sum_{k=1}^{n} a+s_k \ln\left(a+\sum\nolimits_{t=1}^{T}\exp(\boldsymbol{x_{kt}'\beta})\right) + na\ln a \\ &\quad + \sum_{k=1}^{n}\ln\left(\frac{\Gamma(a+s_k)}{\Gamma(a)}\right) - \sum_{k=1}^{n}\sum_{t=1}^{T}\ln(y_{kt}!)\end{aligned} \tag{15}$$

In equation (15), the term $\ln\left(\frac{\Gamma(a+s_k)}{\Gamma(a)}\right)$ can be simplified based on the properties of gamma function as $\sum_{w=0}^{s_k-1}\ln(a+w)$ where $\sum_{w=0}^{s_k-1}\ln(a+w)=0$ if $s_k=0$. Therefore, substituting it to equation (15), we obtain the log-likelihood function for PGHGLM shown in equation (16).

$$\begin{aligned}\ell(a,\boldsymbol{\beta}) = \boldsymbol{Y'X\beta} - \textstyle\sum_{k=1}^{n}(a+s_k\ln(a+\sum_{t=1}^{T}\exp(\boldsymbol{x_{kt}'\beta}))) + \\ na\ln a + \textstyle\sum_{k=1}^{n}\sum_{w=0}^{s_k-1}\ln(a+w) - \sum_{k=1}^{n}\sum_{t=1}^{T}\ln(y_{kt}!)\end{aligned} \tag{16}$$

The partial derivatives for log-likelihood function in equation (16) to each of the parameters are:

a. Estimation of $\boldsymbol{\beta}$ parameter

$$\frac{\partial\ell(a,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = 0 \tag{17}$$

$$\begin{aligned}\Leftrightarrow \frac{\partial}{\partial\boldsymbol{\beta}}\Bigg[\boldsymbol{Y'X\beta} - \sum_{k=1}^{n} a+s_k\ln\left(a+\sum\nolimits_{t=1}^{T}\exp(\boldsymbol{x_{kt}'\beta})\right) + na\ln a + \sum_{k=1}^{n}\sum_{w=0}^{s_k-1}\ln(a+w) \\ - \sum_{k=1}^{n}\sum_{t=1}^{T}\ln(y_{kt}!)\Bigg] = 0 \\ \Leftrightarrow \frac{\partial}{\partial\boldsymbol{\beta}}\Bigg[\underbrace{\boldsymbol{Y'X\beta}}_{a} - \underbrace{\sum_{k=1}^{n} a+s_k\ln\left(a+\sum\nolimits_{t=1}^{T}\exp(\boldsymbol{x_{kt}'\beta})\right)}_{b}\Bigg] = 0\end{aligned}$$

For part (a), suppose $\boldsymbol{m} = \boldsymbol{Y'X\beta}$, where $\boldsymbol{Y'}$ and $\boldsymbol{X}$ are presented on equation (1) and (2). Then, m can be constructed as

$$\boldsymbol{m} = \underbrace{[Y_{11}\, Y_{12}\, Y_{13}\, \cdots Y_{1T}\, Y_{21}\, \cdots\, Y_{2T}\, Y_{n1}\, Y_{n2}\, \cdots\, Y_{nT}]}_{1\times nT} \underbrace{\begin{bmatrix} x_{111} & x_{112} & \cdots & x_{11J} \\ x_{121} & x_{122} & \cdots & x_{12J} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1T1} & x_{1T2} & \cdots & x_{1TJ} \\ x_{211} & x_{212} & \cdots & x_{21J} \\ \vdots & \vdots & \ddots & \vdots \\ x_{nT1} & x_{nT2} & \cdots & x_{nTJ} \end{bmatrix}}_{nT\times J} \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \vdots \\ \beta_J \end{bmatrix}}_{J\times 1}$$

In general, for any values of n and T, the derivative of m with respect to each β is:

$$\begin{aligned} \frac{\partial m}{\partial \beta_1} &= Y_{11}x_{111} + Y_{12}x_{121} + \cdots + Y_{nT}x_{nT1} \\ \frac{\partial m}{\partial \beta_2} &= Y_{11}x_{112} + Y_{12}x_{122} + \cdots + Y_{nT}x_{nT2} \\ &\vdots \\ \frac{\partial m}{\partial \beta_J} &= Y_{11}x_{11J} + Y_{12}x_{12J} + \cdots + Y_{nT}x_{nTJ} \end{aligned}$$

For part (b), suppose $z = \sum_{k=1}^{n} a + s_k \ \ln(a + \sum_{t=1}^{T} \exp(\boldsymbol{x'_{kt}\beta}))$ where

$\sum_{t=1}^{T} \exp(\boldsymbol{x'_{kt}\beta}) = \exp(\boldsymbol{x_{k1}\beta}) + \exp(\boldsymbol{x_{k2}\beta}) + \exp(\boldsymbol{x_{k3}\beta}) + \cdots + \exp(\boldsymbol{x_{kt}\beta})$ and the term $\exp(\boldsymbol{x_{kt}\beta}) = \exp\big(x_{kt1}\beta_1 + x_{kt2}\beta_2 + \cdots + x_{ktJ}\beta_J\big)$.

In general, for any values of n and T, the derivative of z with respect to each β is:

$$\begin{aligned} \frac{\partial z}{\partial \beta_1} &= \sum_{k=1}^{n} (a + s_k) \frac{x_{k11}\exp(\boldsymbol{x'_{k1}\beta}) + x_{k21}\exp(\boldsymbol{x'_{k2}\beta}) + \cdots + x_{kT1}\exp(\boldsymbol{x'_{kT}\beta})}{a + \exp(\boldsymbol{x'_{k1}\beta}) + \exp(\boldsymbol{x'_{k2}\beta}) + \ldots + \exp(\boldsymbol{x'_{kT}\beta})} \\ \frac{\partial z}{\partial \beta_2} &= \sum_{k=1}^{n} (a + s_k) \frac{x_{k12}\exp(\boldsymbol{x'_{k1}\beta}) + x_{k22}\exp(\boldsymbol{x'_{k2}\beta}) + \cdots + x_{kT2}\exp(\boldsymbol{x'_{kT}\beta})}{a + \exp(\boldsymbol{x'_{k1}\beta}) + \exp(\boldsymbol{x'_{k2}\beta}) + \ldots + \exp(\boldsymbol{x'_{kT}\beta})} \\ &\vdots \\ \frac{\partial z}{\partial \beta_J} &= \sum_{k=1}^{n} (a + s_k) \frac{x_{k1j}\exp(\boldsymbol{x'_{k1}\beta}) + x_{k2j}\exp(\boldsymbol{x'_{k2}\beta}) + \cdots + x_{kTj}\exp(\boldsymbol{x'_{kT}\beta})}{a + \exp(\boldsymbol{x'_{k1}\beta}) + \exp(\boldsymbol{x'_{k2}\beta}) + \ldots + \exp(\boldsymbol{x'_{kT}\beta})} \end{aligned}$$

Equation (17) can be expressed as:

$$\begin{aligned} &\frac{\partial \ell(a, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \\ \Leftrightarrow &\frac{\partial m}{\partial \boldsymbol{\beta}} + \frac{\partial z}{\partial \boldsymbol{\beta}} = 0 \end{aligned}$$

and the derivates for $\beta_j, j = 1, 2, \ldots, J$ is as follows:

$$\frac{\partial \ell(a,\boldsymbol{\beta})}{\partial \beta_j} = (Y_{11}x_{11j} + Y_{12}x_{12j} + \cdots + Y_{1T}x_{1Tj} + Y_{21}x_{21j} + \cdots + Y_{n1}x_{n1j} + Y_{n2}x_{n2j} + \cdots + Y_{nT}X_{nTj}) - \sum_{k=1}^{n}(a + s_k)\frac{x_{k1j}\exp(\boldsymbol{x'_{k1}\beta}) + x_{k2j}\exp(\boldsymbol{x'_{k2}\beta}) + \cdots + x_{kTj}\exp(\boldsymbol{x'_{kT}\beta})}{a + \exp(\boldsymbol{x'_{k1}\beta}) + \exp(\boldsymbol{x'_{k2}\beta}) + \ldots + \exp(\boldsymbol{x'_{kT}\beta})} = 0 \quad (18)$$

b. Estimation of a parameter

The derivatives for log-likelihood function in equation (16) respect to a is

$$\begin{aligned}\frac{\partial \ell(a,\boldsymbol{\beta})}{\partial a} &= \frac{\partial}{\partial a}\left(\sum_{k=1}^{n}(a+s_k)\ln\left(a+\sum_{t=1}^{T}\exp(\boldsymbol{x'_{kt}\beta})\right)\right. \\ &\qquad - a\ln a + \sum_{k=1}^{n}\sum_{w=0}^{s_k-1}\ln(a+w) \\ &\qquad \left. - \sum_{k=1}^{n}\sum_{t=1}^{T}\ln(y_{kt}!)\right) \\ &= \sum_{k=1}^{n}(a+s_k)\frac{\partial}{\partial a}\ln\left(a+\sum_{t=1}^{T}\exp(\boldsymbol{x'_{kt}\beta})\right) \\ &\qquad + \frac{\partial}{\partial a}(a+s_k)\ln\left(a+\sum_{t=1}^{T}\exp(\boldsymbol{x'_{kt}\beta})\right) \\ &\qquad - \left(a\frac{1}{a}+\ln a\right) + \sum_{k=1}^{n}\sum_{w=0}^{s_k-1}\frac{1}{a+w} \\ &= \sum_{k=1}^{n}\left(\frac{(a+s_k)}{a+\sum_{t=1}^{T}\exp(\boldsymbol{x'_{kt}\beta})} + \ln\left(a+\sum_{t=1}^{T}\exp(\boldsymbol{x'_{kt}\beta})\right)\right) \\ &\qquad - (1+\ln a) + \sum_{k=1}^{n}\sum_{w=0}^{s_k-1}\frac{1}{a+w} = 0\end{aligned} \quad (19)$$

It is challenging to obtain solutions to the homogeneous system of equations (18) and (19) analytically, so a numerical approach is required. For this purpose, the conjugate gradient method is used because it can solve large-scale optimization problems at a high convergence rate. The conjugate gradient method with the Fletcher-Reeves (FR) algorithm is used to find the solution to the log-likelihood function so that convergent values are obtained to serve as estimates for each parameter. The Fletcher-Reeves (FR) method is employed because it can solve optimization problems for nonlinear systems of equations and requires only first-order derivatives without needing the Hessian matrix or its approximation. The conjugate gradient method uses the following recursive formula:

$$V_{i+1} = V_i + \delta_i d_i, \quad (20)$$

where δ is a positive real number called the step size, and d_i is a non-zero number called the search direction. The numerical procedures to maximize the log-likelihood function is as follows:

Step 0. Initializing vector $V_0 = (a_0, \beta_{01}, \beta_{02}, \dots, \beta_{0J})$ and define

$$h(a, \beta_{01}, \beta_{02}, \dots, \beta_{0J}) = -l(a, \beta_1, \beta_2, \dots, \beta_J)$$

Step 1. Calculating gradient g_0

$$g_0 = \nabla h(a_0, \beta_{01}, \beta_{02}, \dots, \beta_{0J}) = \begin{pmatrix} \frac{\partial h(a_0, \beta_{01}, \beta_{02}, \dots, \beta_{0J})}{\partial a} \\ \frac{\partial h(a_0, \beta_{01}, \beta_{02}, \dots, \beta_{0J})}{\partial \beta_1} \\ \frac{\partial h(a_0, \beta_{01}, \beta_{02}, \dots, \beta_{0J})}{\partial \beta_2} \\ \vdots \\ \frac{\partial h(a_0, \beta_{01}, \beta_{02}, \dots, \beta_{0J})}{\partial \beta_J} \end{pmatrix}$$

If $\|\nabla h(a_0, \beta_{01}, \beta_{02}, \dots, \beta_{0J})\| \leq \epsilon$, iteration stops and $V_0 = (a_0, \beta_{01}, \beta_{02}, \dots, \beta_{0J})$ is a vector with optimal solution. Otherwise, continue to the next step.

Step 2. Calculate the search direction

If $i = 0$:

$$d_0 = -g_0 = -\nabla \left(h(a_0, \beta_{01}, \beta_{02}, \dots, \beta_{0J}) \right)$$

If $i > 0$:

$d_i = -\nabla \left(h(a_i, \beta_{i1}, \beta_{i2}, \dots, \beta_{iJ}) \right) + \omega_i d_{i-1}$. For Fletcher-Reeves method, the formula for ω_i is

$$\omega_i^{FR} = \frac{\|g_i\|^2}{\|g_{i-1}\|^2}$$

Step 3. Calculate the step size δ_0

Step size is calculated by exact line search method. By solving the problem $\underset{\delta_0 \in \mathbb{R}^+}{\operatorname{argmin}} \, l(V_0 + \delta_0 d_0)$, we obtain the exact value of δ_0.

Step 4. Construct vector $V_1 = (a_1, \beta_{11}, \beta_{12}, \dots, \beta_{1J})$.

With the iterative formula $V_{i+1} = V_i + \delta_i d_i$, we obtaion the formula for V_1 is $V_1 = V_0 + \delta_0 d_0$ dengan $d_0 = -\nabla h(a_0, \beta_{01}, \beta_{02}, \dots, \beta_{0J})$ and the value of δ_0 calculated from previous steps.

Step 5. Set $i = i + 1$ and go back to step 1.

This iteration runs until $\left\| \nabla h \left((a_i, \beta_{i1}, \beta_{i2}, \dots, \beta_{iJ}) \right) \right\| \leq \epsilon$. When $\left\| \nabla h \left((a_i, \beta_{i1}, \beta_{i2}, \dots, \beta_{iJ}) \right) \right\| \leq \epsilon$, iteration stop and $V_i = (a_i, \beta_{i1}, \beta_{i2}, \dots, \beta_{iJ})$ is the vector with optimal solution.

4. a-Posteriori Premium Rate Formulation with PGHGLM

The a-posteriori premium rate is the premium rate for the next period. In calculating the a-posteriori premium rate, the expected claim frequency for time $t+1$ is based on the historical claims from the previous year. In this case, the proposed premium rate for the insured $k$for time $t+1$ is given by the following formula:

$$E\left(Y_{k(t+1)}\middle|Y_{k1}=y_{k1},\dots,Y_{kt}=y_{kt}\right) \tag{21}$$

For $k=1,\dots,n$ and for each $t=1,\dots,T$ let

$$m_{k(t+1)}=E\left(Y_{k(t+1)}\middle|Y_{k1}=y_{k1},\dots,Y_{kt}=y_{kt}\right) \tag{22}$$

By substituting the pdf from assumptions (A1) until (A3) and based on the properties of conditional distribution, the a-posteriori premium rate formula is

$$m_{k(t+1)}=\frac{\lambda_{k(t+1)}\left(a+\sum_{s=1}^{t}y_{ks}\right)}{\left(a+\sum_{s=1}^{t}\lambda_{ks}\right)} \tag{23}$$

In the longitudinal data, which involves repeated measurements on the same subjects over time, the data has a complex structure, and observations are correlated over time. Therefore, it is necessary to calculate the correlation between the number of claims to understand the relationship between observation subjects or the response variable in the form of claim frequency of policyholders over time. For $k=1,\dots,n$ and $t,t_1,t_2=1,..,T$ and $t_1\neq t_2$, the covariance for Y_{kt_1} and Y_{kt_2} is

$$Cov\left(Y_{kt_1},Y_{kt_2}\right)=E\left(Y_{kt_1}Y_{kt_2}\right)-E(Y_{kt_1})E(Y_{kt_2})$$

Based on the law of total covariance and by the assumption of Y_{kt} independency, covariance for Y_{kt_1} and Y_{kt_2} is as follows:

$$Cov\left(Y_{kt_1},Y_{kt_2}\right)=\frac{\lambda_{kt1}\lambda_{kt2}}{a}.$$

Thus, the correlation coefficient between Y_{kt_1} and Y_{kt_2} is

$$\begin{aligned}\rho\left(Y_{kt_1},Y_{kt_2}\right)&=\frac{Cov\left(Y_{kt_1},Y_{kt_2}\right)}{\sqrt{Var\left(Y_{kt_1}\right)Var\left(Y_{kt_2}\right)}}\\&=\left(1-\frac{E(Y_{kt1})}{V(Y_{kt1})}\right)^{\frac{1}{2}}\left(1-\frac{E(Y_{kt2})}{V(Y_{kt2})}\right)^{\frac{1}{2}}\end{aligned} \tag{24}$$

C. RESULT AND DISCUSSION

We used the Claimslong dataset, a longitudinal claim frequency dataset obtained from the 'insuranceData' package in R. It contains data from 40,000 motor vehicle insurance policies observed over three years, with one covariate: the policyholder's age. The second dataset is from the motor vehicle insurance portfolio of the Automobile Common Statistics A.P.S.A.D for the years 1979–1981 in France, comprising claim frequency data for 1,044,454 motor vehicle insurance policyholders (C. Partrat and J. Besson, 1992), without covariates. We used the Claimslong dataset which is a longitudinal claim frequency dataset obtained from the 'insuranceData' package in R.

1. Data Description

The data with covariates used in this study was taken from the R Studio software and is titled Claims Longitudinal (Claimslong), available in the 'insuranceData' package. This data contains information on 40,000 motor vehicle insurance policyholders over three periods, consisting of policy ID numbers, driver ages, vehicle values, observation periods, and claim frequencies. From the available data, two variables considered in this thesis are given as follows:

a. NUMCLAIMS is a numerical variable that indicates the frequency of car insurance claims.
b. AGECAT is a categorical variable that indicates the driver's age category, with categories ranging from 1 (youngest) to 6 (oldest).

Table 1 below shows the vehicle insurance claim frequency data, including the observation year (t) and claim frequency (N_t) whch represented the number of claims in year t.

Table 1. Claimslong Frequency Data for Two Years

Claim	$\boldsymbol{N_0(2)}$	$\boldsymbol{N_1(2)}$	$\boldsymbol{N_2(2)}$	$\boldsymbol{N_3(2)}$	$\boldsymbol{N_4(2)}$	$\boldsymbol{N_{\geq 5}(2)}$	**Total**
$N_0(1)$	31397	2751	493	96	23	4	34764
$N_1(1)$	2505	784	266	89	35	25	3704
$N_2(1)$	360	242	131	79	39	24	875
$N_3(1)$	65	82	62	41	17	34	301
$N_4(1)$	16	28	24	16	22	23	129
$N_{\geq 5}(1)$	3	14	29	22	19	140	227
Total	34346	3901	1005	343	155	250	40000

Notes: $N_0(1)$ = the number of policyholders who filed 0 claims in the first year; $N_0(2)$ = the number of policyholders who filed 0 claims in the second year; $N_1(1)$ = the number of policyholders who filed 1 claim in the first year; $N_1(2)$ = the number of policyholders who filed 1 claim in the second year. and so on. For example, there were 31,397 policyholders who filed 0 claims in both the first and second years, and the total number of policyholders who had 0 claims in the second period was 34,346 policies. Figures 1 and Figures 2 display bar charts for claim frequencies in the first and second years, and the frequency distribution by age category, respectively.

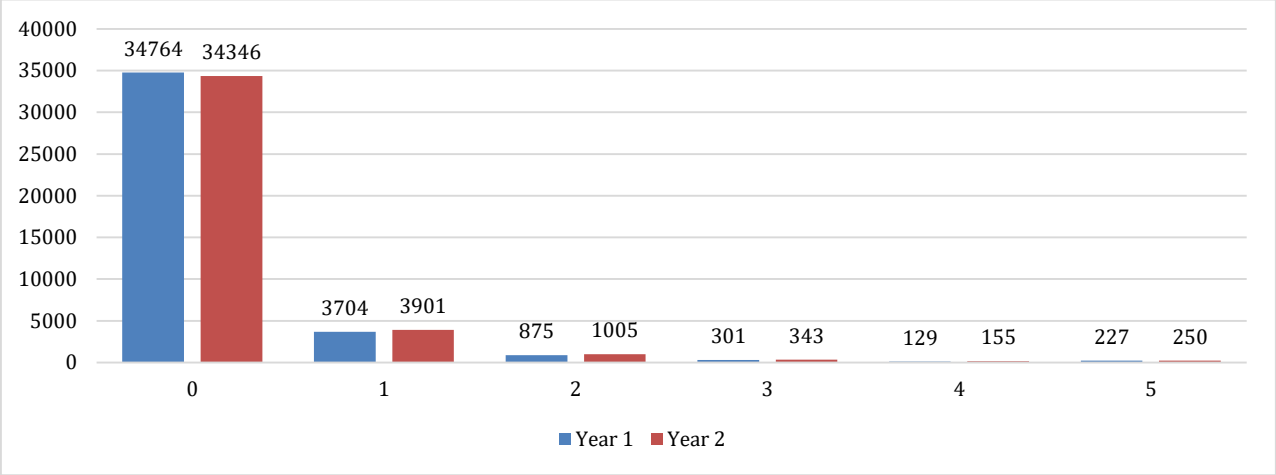

Figure 1. Claimslong Claim Frequency Data Bar Chart

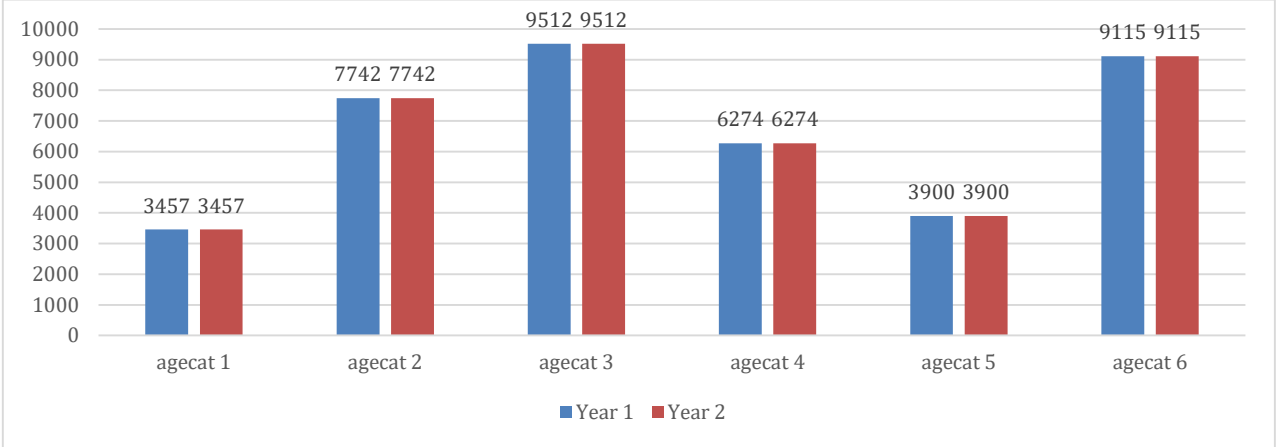

Figure 2. Frequency of Policyholder in Each Age Category Bar Chart

From Figure 2, the distribution of policyholders across age categories is consistent across both years. The largest group of policyholders is in category 3 (9,512 individuals), while the smallest group is in category 1 (3,457 individuals).

Table 2. Descriptive Statistics for Claimslong Data

Statistics	Year 1	Year 2
Mean	0.20020	0.22025
Variance	0.40122	0.44494
Min	0	0
Max	27	33

From Table 2 the average claim frequency in Year 1 is 0.2002, and in Year 2 it is 0.22025. The variance in Year 1 is 0.40122, and in Year 2 it is 0.44494. The claim frequencies range from 0 to 27 in Year 1 and from 0 to 33 in Year 2. The model uses Y_{kt} (claim frequency) as the dependent variable, and AGECAT as the categorical covariate. The age categories are encoded using dummy variables with category 1 as the base level.

- $x_{kt1} = \begin{cases} 1, & category\ 2 \\ 0, & elsewhere \end{cases}$
- $x_{kt2} = \begin{cases} 1, & category\ 3 \\ 0, & elsewhere \end{cases}$
- $x_{kt3} = \begin{cases} 1, & category\ 4 \\ 0, & elsewhere \end{cases}$
- $x_{kt4} = \begin{cases} 1, & category\ 5 \\ 0, & elsewhere \end{cases}$
- $x_{kt5} = \begin{cases} 1, & category\ 6 \\ 0, & elsewhere \end{cases}$

The PGHGLM model is employed to model claim frequency with age as the sole covariate. The model is expressed as:

$$\ln[\lambda_{kt}] = \beta_0 + \beta_1 x_{kt1} + \beta_2 x_{kt2} + \beta_3 x_{kt3} + \beta_4 x_{kt4} + \beta_5 x_{kt5}\ ; t = 1,2.$$

The Poisson-Gamma HGLM was selected due to its flexibility in handling overdispersed count data, such as insurance claims, where the variance exceeds the mean. The assumptions underlying this model include the independence of claims within the same year, a constant rate of claim occurrences across different age categories, and the use of a Gamma distribution for the random effects, which is appropriate for modelling insurance claim frequency where the variance is higher than the mean. Based on Section 2, parameter estimation can be performed using MLE with the help of conjugate gradient method. Therefore, the estimated values for each parameter are shown in Table 3.

Table 3. Estimated Value for Each Parameter

Parameter	Estimated Value
$\hat{a}$	1.1203526
$\hat{\beta}_0$	0.58352583
$\hat{\beta}_1$	0.1099651
$\hat{\beta}_2$	0.13942694
$\hat{\beta}_3$	0.09610736
$\hat{\beta}_4$	0.05906529
$\hat{\beta}_5$	0.13226208

Based on the estimated value for each parameter in Table 3 above, the average claim frequency for policyholders in each age category relative to category 1 (the youngest age category) is as follows: The average claim frequency for policyholders in age category 2 is $exp(0.1099651) = 1.116239113$ times higher than for policyholders in age category 1. The average claim frequency for policyholders in age category 3 is $exp(0.13942694) = 1.149614812$ times higher than for policyholders in age category 1. The average claim frequency for policyholders in age category 4 is $exp(0.09610736) = 1.100877248$ times higher than for policyholders in age category 1. The average claim frequency for policyholders in age category 5 is $exp(0.05906529) = 1.060844501$ times higher than for policyholders in age category 1. The average claim frequency for policyholders in age category 6 is

$exp(0.13226208) = 1.14140742$ times higher than for policyholders in age category 1. The average claim frequency for each age category also can be calculated, result shows in Table 4 below:

Table 4. Average Claim Frequency for Each Age Category

Age Category	Average Claim Frequency
Category 2	2.000687607
Category 3	2.06050845
Category 4	1.73153826
Category 5	1.901401266
Category 6	2.045797948

In Table 4, for base level category 1, the average claim frequency is $e^{\beta_0} = 1.792346813$.

2. A-Posteriori Premium Rate Calculation for Claimslong Data

Before calculating the a-posteriori premium rate, the correlation coefficient must be first calculated. Based on equation (41), the result for correlation coefficient is shown in Table 5 as below:

Table 4. Correlation of Claim Frequency for Each Age Category

Age Category	$\boldsymbol{\rho(Y_{k1}, Y_{kt_2})}$
Category 1	0.615355915
Category 2	0.641032308
Category 3	0.647783232
Category 4	0.637837313
Category 5	0.629237638
Category 6	0.646146769

If the correlation between claims at year 1 and year 2 is high, then the number of claims in subsequent years is closely related to the number of claims in the previous year. In other words, the number of claims in the previous year affects the number of claims in the coming year, resulting in a more differentiated a posteriori premium rate that considers the number of claims in previous years. Conversely, the a posteriori premium rate will not differ much from the previous year if the correlation coefficient is low. The a-Posteriori premium rate is calculated based on equation (40). To calculate the a-posteriori rate in year 3, where there is 0 claim in year 1 and 2, the formula is:

$$E(Y_3|Y_1 = 0, Y_2 = 0) = \frac{\hat{\lambda}_{k3}(\hat{a} + 0 + 0)}{(\hat{a} + \sum_{t=1}^{2} \lambda_{kt})}$$

Then, to find the premium rate difference between year 3 and year 2, the formula is

$$\%\, Diff = \frac{\left(E(Y_t|Y_{t-2}, Y_{t-1}) - E(Y_{t-1})\right)}{E(Y_{t-1})}$$

Based on those formula, the result of premium rate and difference is shown in Table 5.

Table 5. A-Posteriori Premium Rate Difference

Age Category	Claim Frequency	Premium Rate	%Rate Diff
Category 1	E(Y3|Y1=0, Y2=0)	0.426788668	-57.32%
	E(Y3|Y1=0, Y2=1)	0.80773005	-19.23%
	E(Y3|Y1=0, Y2=2)	1.188671433	18.87%
	E(Y3|Y1=0, Y2=3)	1.569612815	56.96%
	E(Y3|Y1=0, Y2=4)	1.950554197	95.06%
	E(Y3|Y1=0, Y2=5)	2.33149558	133.15%
Category 2	E(Y3|Y1=0, Y2=0)	0.437640508	-56.24%
	E(Y3|Y1=0, Y2=1)	0.828267984	-17.17%
	E(Y3|Y1=0, Y2=2)	1.218895459	21.89%
	E(Y3|Y1=0, Y2=3)	1.609522935	60.95%
	E(Y3|Y1=0, Y2=4)	2.000150411	100.02%
	E(Y3|Y1=0, Y2=5)	2.390777886	139.08%
Category 3	E(Y3|Y1=0, Y2=0)	0.440437561	-55.96%
	E(Y3|Y1=0, Y2=1)	0.83356162	-16.64%
	E(Y3|Y1=0, Y2=2)	1.226685678	22.67%
	E(Y3|Y1=0, Y2=3)	1.619809737	61.98%
	E(Y3|Y1=0, Y2=4)	2.012933795	101.29%
	E(Y3|Y1=0, Y2=5)	2.406057854	140.61%
Category 4	E(Y3|Y1=0, Y2=0)	0.436308714	-56.37%
	E(Y3|Y1=0, Y2=1)	0.825747462	-17.43%
	E(Y3|Y1=0, Y2=2)	1.215186211	21.52%
	E(Y3|Y1=0, Y2=3)	1.604624959	60.46%
	E(Y3|Y1=0, Y2=4)	1.994063707	99.41%
	E(Y3|Y1=0, Y2=5)	2.383502455	138.35%
Category 5	E(Y3|Y1=0, Y2=0)	0.432698096	-56.73%
	E(Y3|Y1=0, Y2=1)	0.818914093	-18.11%
	E(Y3|Y1=0, Y2=2)	1.205130089	20.51%
	E(Y3|Y1=0, Y2=3)	1.591346086	59.13%
	E(Y3|Y1=0, Y2=4)	1.977562083	97.76%
	E(Y3|Y1=0, Y2=5)	2.363778079	136.38%
Category 6	E(Y3|Y1=0, Y2=0)	0.439761646	-56.02%
	E(Y3|Y1=0, Y2=1)	0.832282399	-16.77%
	E(Y3|Y1=0, Y2=2)	1.224803152	22.48%
	E(Y3|Y1=0, Y2=3)	1.617323905	61.73%
	E(Y3|Y1=0, Y2=4)	2.009844658	100.98%
	E(Y3|Y1=0, Y2=5)	2.402365411	140.24%

From the correlation results and Table 4, it can be concluded that the correlation between claims in the first and second years is quite high, at around 60%. This results in a significant difference in the premium rate for the subsequent year, the third year, compared to the premium rate in the second year. To obtain the a-posteriori premium rate for the third year, the average base premium for motor vehicle insurance in Indonesia, around IDR 3,588,000, is used. The a posteriori premium can be calculated by multiplying (1 + rate) with the base premium. Therefore, the a posteriori premium in 2024 for a policyholder in age category 6 who claimed 3 times in 2023 and 0 times in 2022 is:

$$\text{IDR } 3.588.000 \times (1 + 1.617323905) = \text{IDR } 9.390{,}958\ ,$$

Thus, the premium for the third year will increase by a factor of (1 + rate) times the base premium. However, the difference in the rate increase will vary from the previous year depending on the claim frequency in the previous year. If the claim frequency in the previous year is higher, the difference in the rate for the following year will also be greater. The model's validity can be assessed using deviance residuals and goodness-of-fit tests, such as the Akaike Information Criterion (AIC). While the Poisson-Gamma HGLM is suitable for this dataset, one limitation is that it assumes the number of claims is independent across years. If claims in consecutive years are correlated, this assumption could lead to biased estimates. To address this, the model could be extended to account for temporal dependencies, such as using a time-series model or including lagged variables for past claims.

D. CONCLUSION AND SUGGESTIONS

The results of this study have direct implications for determining premium rates and managing risks in insurance companies. By modeling the claim frequency for different age groups, insurers can more accurately adjust their premium rates based on the likelihood of future claims. For instance, the estimated claim frequencies show that policyholders in older age categories tend to have a higher claim frequency, suggesting that premium rates for these groups should be adjusted accordingly. Additionally, the a-posteriori premium rate calculations, which take into account past claims, allow insurers to better reflect the risk profile of policyholders and tailor premiums more precisely to individual risks. The Poisson Gamma Hierarchical Generalized Linear Model (PGHGLM) is constructed by determining a mixture distribution for the response variable, which is influenced by random effects. This model utilizes a logarithmic link function, with parameter estimation conducted using the maximum likelihood method and the conjugate gradient technique for numerical optimization. From the PGHGLM construction, a formula for calculating posterior premium rates can be derived. PGHGLM is well-suited for datasets with a longitudinal structure, such as motor vehicle insurance data. When dealing with data that lacks covariates, trend parameters are incorporated into the model, while for datasets with covariates, the average claim frequency is computed for each age category, with the highest average observed in category 3. Furthermore, an increase in claim frequency from the previous year is associated with a corresponding rise in posterior premium rates.

ACKNOWLEDGEMENT

This research is supported by a grant from Hibah Riset FMIPA UI for the year 2023, identified by the number PKS-01/UN2.F3.D/PPM.00.02/2023, provided by the Faculty of Mathematics and Natural Sciences (FMIPA) at the University of Indonesia (UI). The authors gratefully acknowledge the financial support and assistance provided by this grant, which made this research possible.

REFERENCES

Antonio, K., & Beirlant, J. (2005). Actuarial Statistics With Generalized Linear Mixed Models. *Insurance: Mathematics and Economics*, *40*, 58–76. https://doi.org/10.1016/j.insmatheco.2006.02.013

Antonio, K., & Valdez, E. A. (2012). Statistical concepts of a priori and a posteriori risk classification in insurance. *AStA Advances in Statistical Analysis*, *96*(2), 187–224. https://doi.org/10.1007/s10182-011-0152-7

Bagariang, E., & Raharjanti, A. (2023a). Calculation of Motor Vehicle Insurance Premiums Through Evaluation of Claim Frequency and Amount Data. *Operations Research: International Conference Series*, *4*, 157–162. https://doi.org/10.47194/orics.v4i4.270

Boonen, T. J., & Liu, F. (2022). Insurance with heterogeneous preferences. *Journal of Mathematical Economics*, *102*, 102742. https://doi.org/https://doi.org/10.1016/j.jmateco.2022.102742

Boucher, J.-P., & Denuit, M. (2006). Fixed versus Random Effects in Poisson Regression Models for Claim Counts: A Case Study with Motor Insurance. *ASTIN Bulletin*, *36*(1), 285–301. https://doi.org/DOI: 10.2143/AST.36.1.2014153

Boucher, J.-P., & Inoussa, R. (2014). A posteriori ratemaking with panel data. *ASTIN Bulletin*, *44*, 587–612. https://doi.org/10.1017/asb.2014.11

Brown, R. L., & Lennox, W. S. (2015). *Introduction to Ratemaking and Loss Reserving for Property and Casualty Insurance*. ACTEX Publications.

Dickson, D. C., Hardy, M. R., & Waters, H. R. (2009). *Actuarial Mathematics for Life Contingent Risks*. Cambridge University Press.

Frostig, E. (2001). A comparison between homogeneous and heterogeneous portfolios. *Insurance: Mathematics and Economics*, *29*, 59–71. https://doi.org/10.1016/S0167-6687(01)00073-7

Frostig, E., Zaks, Y., & Levikson, B. (2007a). Optimal pricing for a heterogeneous portfolio for a given risk factor and convex distance measure. *Insurance: Mathematics and Economics*, *40*(3), 459–467. https://doi.org/https://doi.org/10.1016/j.insmatheco.2006.07.001

Gning, L., Diagne, M., & Tchuenche, M. (2023). Hierarchical generalized linear models, correlation and a posteriori ratemaking. *Physica A: Statistical Mechanics and Its Applications*, *614*, 128534. https://doi.org/10.1016/j.physa.2023.128534

Gupta, P., Gupta, R., & Ong, S.-H. (2004). Modelling Count Data by Random Effect Poisson Model. *Sankhyā: The Indian Journal of Statistics (2003-2007)*, *66*, 548–565. https://doi.org/10.2307/25053380

Hsu, Y.-C., Chou, P.-L., & Shiu, Y.-M. (2016a). An examination of the relationship between vehicle insurance purchase and the frequency of accidents. *Asia Pacific Management Review*, *21*(4), 231–238. https://doi.org/https://doi.org/10.1016/j.apmrv.2016.08.001

Iqbal, A., Shad, M. Y., & Yassen, M. F. (2023). Empirical E-Bayesian estimation of hierarchical poisson and gamma model using scaled squared error loss function. *Alexandria Engineering Journal*, *69*, 289–301. https://doi.org/https://doi.org/10.1016/j.aej.2023.01.064

Jin, S., & Lee, Y. (2024). Standard error estimates in hierarchical generalized linear models. *Computational Statistics & Data Analysis*, *189*, 107852. https://doi.org/https://doi.org/10.1016/j.csda.2023.107852

Laird, N. M., & Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, *38*(4), 963–974. https://doi.org/10.2307/2529876

Lanfranchi, D., & Grassi, L. (2022a). Examining insurance companies' use of technology for innovation. *The Geneva Papers on Risk and Insurance - Issues and Practice*, *47*(3), 520–537. https://doi.org/10.1057/s41288-021-00258-y

Lee, W., Kim, J., & Ahn, J. Y. (2020a). The Poisson random effect model for experience ratemaking: Limitations and alternative solutions. *Insurance: Mathematics and Economics*, *91*, 26–36. https://doi.org/https://doi.org/10.1016/j.insmatheco.2019.12.004

Lee, W., Kim, J., & Ahn, J. Y. (2020b). The Poisson random effect model for experience ratemaking: Limitations and alternative solutions. *Insurance: Mathematics and Economics*, *91*, 26–36. https://doi.org/https://doi.org/10.1016/j.insmatheco.2019.12.004

Levitas, J., Yavilberg, K., Korol, O., & Man, G. (2022). Prediction of Auto Insurance Risk Based on t-SNE Dimensionality Reduction. *Adv. Artif. Intell. Mach. Learn.*, *2*, 567–579. https://api.semanticscholar.org/CorpusID:254854018

Li, S., Dyk, D., & Autenrieth, M. (2024). *Poisson and Gamma Model Marginalisation and Marginal Likelihood calculation using Moment-generating Functions*. https://doi.org/10.48550/arXiv.2409.11167

Matsuyama, Y. (2020). Hierarchical Linear Modeling (HLM). In M. D. Gellman (Ed.), *Encyclopedia of Behavioral Medicine* (pp. 1059–1061). Springer International Publishing. https://doi.org/10.1007/978-3-030-39903-0_407

Niehaus, G. (2016). *The Role of Insurance in Enterprise Risk Management* (pp. 161–173). https://doi.org/10.1016/B978-0-12-800633-7.00012-2

Rejda, G. E. , M. M. J. . (2017). *Principles of Risk Management and Insurance* (13th ed.). Pearson Education Ltd.

Rumson, A. G., & Hallett, S. H. (2019a). Innovations in the use of data facilitating insurance as a resilience mechanism for coastal flood risk. *Science of The Total Environment*, *661*, 598–612. https://doi.org/https://doi.org/10.1016/j.scitotenv.2019.01.114

Sheehan, B., Mullins, M., Shannon, D., & McCullagh, O. (2023a). On the benefits of insurance and disaster risk management integration for improved climate-related natural catastrophe resilience. *Environment Systems and Decisions*, *43*(4), 639–648. https://doi.org/10.1007/s10669-023-09929-8

Shirazi, M., & Lord, D. (2019). Characteristics-based heuristics to select a logical distribution between the Poisson-gamma and the Poisson-lognormal for crash data modelling. *Transportmetrica A: Transport Science*, *15*(2), 1791–1803. https://doi.org/10.1080/23249935.2019.1640313

Širá, E., & RADVANSKÁ, K. (2015). *Insurance as a Means of Risk Transfer* (pp. 211–226). https://doi.org/10.14505/cfs.2014.ch9

Tawiah, K., Iddi, S., & Lotsi, A. (2020). On Zero-Inflated Hierarchical Poisson Models with Application to Maternal Mortality Data. *International Journal of Mathematics and Mathematical Sciences*, *2020*, 1–8. https://doi.org/10.1155/2020/1407320

Tseung, S., Chan, I. W., Fung, T. C., Badescu, A., & Lin, X. (2022). *A Posteriori Risk Classification and Ratemaking with Random Effects in the Mixture-of-Experts Model*. https://doi.org/10.48550/arXiv.2209.15212

Tzougas, G., & Pignatelli di Cerchiara, A. (2021). The multivariate mixed Negative Binomial regression model with an application to insurance a posteriori ratemaking. *Insurance: Mathematics and Economics*, *101*, 602–625. https://doi.org/https://doi.org/10.1016/j.insmatheco.2021.10.001

Wolny-Dominiak, A., & Sobiecki, D. (2014, November). *The Poisson regression with fixed and random effects in non-life insurance ratemaking*.

Wu, S. (2022). Poisson-Gamma mixture processes and applications to premium calculation. *Communications in Statistics - Theory and Methods*, *51*(17), 5913–5936. https://doi.org/10.1080/03610926.2020.1850791