Communications for Statistical Applications and Methods
2023, Vol. 30, No. 5, 437–451

https://doi.org/10.29220/CSAM.2023.30.5.437
Print ISSN 2287-7843 / Online ISSN 2383-4757

# Predictive analysis in insurance: An application of generalized linear mixed models

Rosy Oh[1,a], Nayoung Woo[b], Jae Keun Yoo[b], Jae Youn Ahn[b]

[a]Department of Mathematics, Korea Military Academy, Korea;
[b]Department of Statistics, Ewha Womans University, Korea

## Abstract

Generalized linear models and generalized linear mixed models (GLMMs) are fundamental tools for predictive analyses. In insurance, GLMMs are particularly important, because they provide not only a tool for prediction but also a theoretical justification for setting premiums. Although thousands of resources are available for introducing GLMMs as a classical and fundamental tool in statistical analysis, few resources seem to be available for the insurance industry. This study targets insurance professionals already familiar with basic actuarial mathematics and explains GLMMs and their linkage with classical actuarial pricing tools, such as the Bühlmann premium method. Focus of the study is mainly on the modeling aspect of GLMMs and their application to pricing, while avoiding technical issues related to statistical estimation, which can be automatically handled by most statistical software.

Keywords: generalized linear model, generalized linear mixed model, predictive analysis, ratemaking, Buhlmann premium, premium, insurance

## 1. Introduction

Generalized linear models (GLMs) and generalized linear mixed models (GLMMs) are fundamental tools for predictive analysis. These statistical tools have motivated many modern insurance-pricing techniques (Nelder and Verrall, 1997; Antonio and Beirlant, 2007; Wuthrich, 2019). Meanwhile, the credibility theory is a field within actuarial science that focuses on calculating risk premiums. It originated from Mowbray (1914) and Whitney (1918), who aimed to determine a premium that balanced the experience of an individual policyholder with that of a group of similar policyholders (Nelder and Verrall, 1997). While GLMMs and credibility theory are both widely used in risk classification in non-life insurance sectors, such as auto insurance, practitioners in these sectors tend to be more familiar with credibility theory.

GLMMs have several advantages over credibility theory. They offer a wider range of models and can be easily fitted and tested using standard statistical software. As a result, GLMMs provide additional modeling options for various actuarial problems, such as premium ratings and claims reserves. Furthermore, it is well known that credibility theory is closely related to the framework of GLMMs (Nelder and Verrall, 1997). While the insurance literature has some initial discussion on credibility theory based on GLMMs (Ohlsson, 2008; Lai and Sun, 2012), most assume above-average knowledge of the statistical aspects of GLMM, which makes its application inaccessible to traditional insurance practitioners.

[1] Corresponding author: Department of Mathematics, Korea Military Academy, PO Box 77-1 Nowon-gu, Seoul 01805, Korea. E-mail: rosy.oh5@gmail.com

Published 30 September 2023 / journal homepage: http://csam.or.kr

This study aims to explain GLMs and GLMMs in plain actuarial language and demonstrate how to obtain a simple and fair premium. The remainder of the article is organized as follows. Section 2 introduces the notations used. Section 3 provides the general settings of risk classification procedures for insurance using the language of GLMs. Sections 4 and 5 explain the modeling and prediction methods for GLMs and GLMMs, respectively. Section 6 provides the theoretical link between GLMMs and credibility theory. Finally, the results are discussed in Section 7.

## 2. Definitions and notations

We denote $\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{R}$, and $\mathbb{R}^+$ as the set of natural numbers, the set of non-negative integers, the set of real numbers, and the set of positive real numbers, respectively. We define $N_{it}$ as the number of claims in the $t^{th}$ policy year for the $i^{th}$ policyholder. We denote the realization of $N_{it}$ as $n_{it}$. Actuarial science literature often refers to $N_{it}$ as the frequency of insurance claims.

We differentiate a fixed effect from a random effect. The former is based on the policyholders' observed risk characteristics, whereas the latter is based on the policyholders' unobserved risk characteristics after controlling for fixed effects. Let the data matrix $\boldsymbol{X}_i$ of the $i^{th}$ policyholder have a dimension of $\tau \times (m+1)$, where $\tau$ is the number of observed years and $m$ is the total number of explanatory variables. Denote

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{X}_{i1} \\ \boldsymbol{X}_{i2} \\ \vdots \\ \boldsymbol{X}_{i\tau} \end{pmatrix} = \begin{pmatrix} 1 & X_{i11} & X_{i12} & \cdots & X_{i1m} \\ 1 & X_{i21} & X_{i22} & \cdots & X_{i2m} \\ & & \vdots & & \\ 1 & X_{i\tau 1} & X_{i\tau 2} & \cdots & X_{i\tau m} \end{pmatrix},$$

where $\boldsymbol{X}_{it} := (1, X_{it1}, \ldots, X_{itm})$ is the $(m+1)$-dimensional row vector of the a priori risk characteristics of the $i^{th}$ policyholder in the $t^{th}$ year. Let $\gamma_i$ denote the policyholder-specific random effect of the $i^{th}$ policyholder. For simplicity, we assume that the random effect $\gamma_i$ does not have the subscript $t$, because we assume that the a priori risk characteristics remain constant over time. Under GLM and GLMM, the distribution within the exponential dispersion family (EDF) has been addressed by McCullagh and Nelder (1989) and Breslow and Clayton (1993). EDF can be viewed as an abstraction of many well-known distributions, including normal, gamma, and Poisson distributions. This provides a unified approach to statistical modeling. However, owing to its abstract form, it may be difficult to understand intuitively. Therefore, we do not explore the benefits of a generalized form of EDF in this study, nor do we attempt to identify the best parametric assumption for the distribution that fits the data. Instead, we focus primarily on Poisson distribution because the general idea can easily be extended to other distributions within the EDF.

We use Pois($\lambda$) to denote a Poisson distribution with mean $\lambda$. Using gamma($\mu, \psi$), we denote a gamma distribution whose mean and variance parameters are $\mu$ and $\mu^2\psi$. Gamma($\mu, \psi$) can also be interpreted as a gamma distribution with the shape parameter $1/\psi$ and scale parameter $\mu\psi$. We use N($\mu, \sigma^2$) to denote a normal distribution, whose mean and variance are given by $\mu$ and $\sigma^2$, respectively. For $Y \sim$ N($\mu, \sigma^2$), we define the distribution of exp($Y$) as lognormal($\mu, \sigma^2$).

## 3. General setting

In terms of data, for each policyholder $i = 1, \ldots, z$ in $\tau$ policy years, with $m$ explanatory variables, the dataset has the following format:

$$(n_{i1}, \ldots, n_{i\tau}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i\tau}), \tag{3.1}$$

Table 1: Typical data example

| ID | Calendar year | Salary ($X_{it1}$) | Home value ($X_{it2}$) | $\cdots$ | Age ($X_{itm}$) | Frequency ($N_{it}$) |
|---|---|---|---|---|---|---|
| 1 | 2017 | 1500 | 20500 | $\cdots$ | 37 | 0 |
| 1 | 2018 | 1470 | 20900 | $\cdots$ | 38 | 1 |
| 1 | 2019 | 1563 | 20500 | $\cdots$ | 39 | 0 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| 102 | 2017 | 1401 | 45000 | $\cdots$ | 52 | 2 |
| 102 | 2018 | 1573 | 47000 | $\cdots$ | 53 | 0 |
| 102 | 2019 | 1562 | 50000 | $\cdots$ | 54 | 0 |

where $z$ is the total number of policyholders and $\boldsymbol{x}_{it} = (x_{it1}, x_{it2}, \dots, x_{itm})$. An example of a typical dataset is listed in Table 1.

First, we are interested in a fair premium, defined as $\mathbb{E}[N_{i\tau^*}]$, for the $i^{th}$ policyholder in calendar year $\tau^* = \tau + 1$. Throughout this study, although we do not specify so in the mathematical notation, we assume that the explanatory variables are always known and conditioned. Hence, for example, the fair premium $\mathbb{E}[N_{i\tau^*}]$ should be understood as

$$\mathbb{E}[N_{i\tau^*}] = \mathbb{E}[N_{i\tau^*} \mid \boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{i\tau}, \boldsymbol{x}_{i\tau^*}].$$

The fair premium for each policy year is determined solely by the fixed effect. In insurance, a fair premium based on the fixed effect is called the a priori rate and the risk classification process based on the fixed effect is called the a priori risk classification procedure. The premium determined by the a priori risk classification procedure can be further updated based on the claim history as follows:

$$\mathbb{E}[N_{i\tau^*} \mid n_{i1}, \dots, n_{i\tau}], \tag{3.2}$$

which is called the a posteriori rate. This premium adjustment mechanism based on claim history is called the a posteriori risk classification procedure. Typically, in insurance, the risk characteristics for a year are known at the beginning of each year. Hence, in predicting the premium in the $\tau^{*th}$ year, the explanatory variables $\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{i\tau^*}$ are assumed to be known. In the following sections, we examine the a posteriori risk classification procedure in (3.2) under various model assumptions.

## 4. Generalized linear model

GLMs are a flexible method that extend the fixed-effect linear regression model to response variables, which are not normally distributed. The expected value of $N_{it}$ depends on the observed risk characteristics with the following relation:

$$\mathbb{E}[N_{it}] = \eta^{-1}(\boldsymbol{X}_{it}\boldsymbol{\beta}),$$

where $\boldsymbol{X}_{it}\boldsymbol{\beta}$ is the linear predictor and $\eta(\cdot)$ is the link function, which relates $\mathbb{E}[N_{it}]$ to a linear predictor.

### 4.1. Modelling and estimation of the GLM

**Model 1. (Poisson GLM)** *Consider $z$ different policyholders with the observation given in the form of* (3.1) *and assume mutual independence among policyholders. Furthermore, we assume the following joint distribution of* $(N_{i1}, \dots, N_{i\tau^*})$*:*

*i. The distribution of the $i^{th}$ policyholder's frequency $N_{it}$ in the $t^{th}$ year is specified as*

$$N_{it} \sim \mathrm{Pois}(\Lambda_{it}),$$

*where*

$$\begin{aligned}\Lambda_{it} &:= \exp\left(\boldsymbol{X}_{it}\boldsymbol{\beta}\right)\\ &= \exp\left(\beta_0 + \beta_1 X_{it1} + \cdots + \beta_m X_{itm}\right)\end{aligned} \tag{4.1}$$

*with the coefficient parameters (fixed effects)* $\boldsymbol{\beta} := (\beta_0, \beta_1, \ldots, \beta_m)'$*. Here, we assume the link function* $\eta$ *to be a log function that gives the exponential expression in* (4.1).

*ii.* $N_{i1}, \ldots, N_{i\tau^*}$ *are independent.*

In Model 1, the only unknown parameters are the coefficients $\boldsymbol{\beta}$ shared by each policyholder. Hence, once the coefficients $\boldsymbol{\beta}$ are known, the conditional distribution of frequency for each policyholder in the future calendar year $\tau^* = t + 1$ can be determined exactly as

$$N_{i\tau^*} \sim \text{Pois}\left(\Lambda_{i\tau^*}\right),$$

with the a priori rate for the $i^{th}$ policyholder in calendar year $\tau^*$ given by $\mathbb{E}\left[N_{i\tau^*}\right] = \Lambda_{i\tau^*}$.

We assume that we have a dataset in the form of (3.1) with $z$ policyholders and that each policyholder is observed for $\tau$ years. Then, the log-likelihood function for Model 1 is written as

$$\ell\left(\boldsymbol{\beta}\right) := \sum_{i=1}^{z}\sum_{t=1}^{\tau} \log\left(f_{\boldsymbol{\beta}}\left(n_{it}\right)\right),$$

where $f_{\boldsymbol{\beta}}\left(n_{it}\right)$ is the probability mass function of the Poisson distribution given by

$$f_{\boldsymbol{\beta}}\left(n_{it}\right) := \frac{\exp\left(-\Lambda_{it}\right)\left(\Lambda_{it}\right)^{n_{it}}}{n_{it}!},$$

where $\Lambda_{it} = \exp(\boldsymbol{X}_{it}\boldsymbol{\beta})$. For the estimation of unknown coefficients $\boldsymbol{\beta}$, typically, maximum likelihood estimation (MLE) is considered and the estimator of $\boldsymbol{\beta}$ is defined as

$$\widehat{\boldsymbol{\beta}} := \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^{m+1}}\ \ell\left(\boldsymbol{\beta}\right).$$

## 4.2. Prediction in the GLM

In this subsection, we predict the a posteriori rate in (3.2), which is the premium in the $\tau^{*th}$ year. given the claim history up to the $\tau^{th}$ year. Under the settings in Model 1, the a posteriori rate defined in (3.2) can be obtained as

$$\mathbb{E}\left[N_{i\tau^*} \mid n_{i1}, \ldots, n_{i\tau}\right] = \mathbb{E}\left[N_{i\tau^*}\right] = \Lambda_{i\tau^*} \tag{4.2}$$

based on the assumption of independence in Model 1. The a posteriori rate in (4.2) coincides with the a priori rate. As (4.2) shows, the a posteriori rate is not a function of the past claim history under the setting in Model 1 and there is no premium adjustment by the claim history $n_{i1}, \ldots, n_{i\tau}$.

## 4.3. Data example

Throughout this study, we use the dataset for collision coverage on vehicles collected by the Wisconsin Local Government Property Insurance Fund (LGPIF) from 2006 to 2011 (Frees *et al.*, 2016). Collision coverage covers the impact of a vehicle on an object, the impact of a vehicle on an attached vehicle,

Table 2: Observable policy characteristics used as covariates

| Categorical variable | Description | | | Proportions |
|---|---|---|---|---|
| | Type of local government entity | | | |
| | Miscellaneous | | | 5.03% |
| | City | | | 9.66% |
| Entity type | County | | | 11.47% |
| | School | | | 36.42% |
| | Town | | | 16.90% |
| | Village | | | 20.52% |
| Continuous variable | | Min | Mean | Max |
| avgCoverage | Collision coverage amount | 0.009 | 1.097 | 13.135 |

Table 3: Estimation results under Model 1

| Parameter | | Est | Std.dev | z | $p$-value | |
|---|---|---|---|---|---|---|
| **Fixed effect** | | | | | | |
| (Intercept) | $\hat{\beta}_0$ | −2.339 | 0.302 | −7.758 | <0.0001 | * |
| Type = City | $\hat{\beta}_1$ | 1.618 | 0.314 | 5.158 | <0.0001 | * |
| Type = County | $\hat{\beta}_2$ | 2.633 | 0.307 | 8.572 | <0.0001 | * |
| Type = School | $\hat{\beta}_3$ | 1.021 | 0.309 | 3.310 | <0.0001 | * |
| Type = Town | $\hat{\beta}_4$ | −0.730 | 0.387 | −1.888 | 0.0591 | |
| Type = Village | $\hat{\beta}_5$ | 0.835 | 0.317 | 2.632 | 0.0085 | * |
| avgCoverage | $\hat{\beta}_6$ | 0.195 | 0.010 | 19.731 | <0.0001 | * |

or the overturning of a vehicle. The datasets contain 497 government entities. Each policyholder has six levels of entity types and average coverage over time as explanatory variables as shown in Table 2. For further explanation of the dataset, refer to Oh *et al.* (2020).

To fit the GLM in Model 1, we can use the `glm` function in `R`, whose form is given by

```
> glm(formula, family=familytype(link=linkfunction), data=dataset,...).
```

The specific form of the function under the setting in Model 1 is given by

```
> glm(Frequency~factor(entityType)+avgCoverage,
            family=poisson(link=log), data=mydata).
```

If the categorical variable (entity type) is presented as a dummy variable (e.g., TypeCity, TypeCounty) in a dataset, the formula specifying the fixed effects part of the function is modified as follows:

```
> glm(Frequency~TypeCity+TypeCounty+···+avgCoverage,
            family=poisson(link=log), data=mydata).
```

The estimation results of this model are summarized in Table 3. The asterisk (*) signs indicate that the parameters are significant at the significance level of 0.05. Based on the table, the a posteriori rate, which is equivalent to the a priori rate, can be obtained as

$$\widehat{\mathbb{E}[N_{it}]} = \exp\left(\boldsymbol{X}_{it}\hat{\boldsymbol{\beta}}\right). \tag{4.3}$$

Hence, the estimated parameter $\hat{\beta}_1$, for example, can be interpreted as the effect of the entity type of the county compared to the baseline, miscellaneous entity type, in the logarithm of the estimated a posteriori rate for the same average coverage. The table indicates that the county entity type is expected to incur more accidents than any other entity type having the same amount of coverage.

Within the same entity type, a higher average coverage corresponds to a higher expected accident count.

Moreover, the third policyholder in the dataset, for example, is in a county and has an average coverage of 2.495, so that under Model 1, the estimated expectation is

$$\begin{aligned}\exp\left(\hat{\beta}_0 + \hat{\beta}_2 + \hat{\beta}_6\, 2.495\right) &= \exp\left(-2.339 + 2.633 + 0.195 \times 2.495\right)\\ &= 2.182.\end{aligned}$$

## 5. Generalized linear mixed model

GLMMs are a popular method to model correlated data; they extend GLMs by including random effects in the predictor. The conditional expected value of $N_{it}|\gamma_i$ depends on the observed risk characteristics using the following relation:

$$\mathbb{E}\left[N_{it} \mid \gamma_i\right] = \eta^{-1}\left(\boldsymbol{X}_{it}\boldsymbol{\beta} + \gamma_i\right),$$

where $\gamma_i$ is the random effect and $\eta(\cdot)$ is a link function that converts $\mathbb{E}\left[N_{it}|\gamma_i\right]$ into a linear predictor $\boldsymbol{X}_{it}\boldsymbol{\beta}$. The random effect $\gamma_i$ measures the difference between the average score of policyholder $i$ and the average score in the portfolio.

### 5.1. Modelling and estimation of the GLMM

In Model 1, we assume that the selected explanatory variable $\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{i\tau^*}$ can explain the distribution of the frequency. However, in reality, it is almost impossible to describe human behavior and contributors are often not observed or overlooked. These unobserved components can be used in a regression setting, as shown in the following example:

***Example 1.*** Consider Model 1 with three explanatory variables: $X_{it1}$, $X_{it2}$, and $X_{it3}$. If we assume that they are observed, then the a priori premium is

$$\Lambda_{it} = \exp\left(\beta_0 + \beta_1 X_{it1} + \beta_2 X_{it2} + \beta_3 X_{it3}\right).$$

Now, we assume that only the first explanatory variable $X_{it1}$ is observed and the other explanatory variables, $X_{it2}$ and $X_{it3}$, are not. Then, a reasonable description of the a priori premium for frequency under this assumption is

$$\Lambda_{it} := \exp\left(\beta_0 + \beta_1 X_{it1}\right).$$

If the unobserved explanatory variables are denoted as

$$\gamma_i := X_{it2} + X_{it3},$$

where the explanatory variables are assumed invariant over time, we have the following conditional model:

$$N_{it} \mid \gamma_i \sim \text{Pois}\left(\exp\left(\beta_0 + \beta_1 X_{it1} + \gamma_i\right)\right). \tag{5.1}$$

By further defining $R_i := \exp(\gamma_i)$, the conditional distribution in (5.1) can be represented as

$$N_{it} \mid R_i \sim \text{Pois}\left(R_i \Lambda_{it}\right).$$

Regardless of whether observed or unobserved, the explanatory variables can also be assumed to be random variables. By placing an additional distributional assumption on the unobserved explanatory variables, Example 5.1 can be formally represented as follows:

**Model 2. (Poisson GLMM)** *Consider z policyholders with the observation given in the form of* (3.1), *assuming mutual independence among policyholders. Furthermore, we assume the following conditional joint distribution of* $(N_{i1}, \dots, N_{i\tau^*})$*:*

*i. Conditional on the random effect* $R_i$*, the distribution of the* $i^{th}$ *policyholder's frequency* $N_{it}$ *in the* $t^{th}$ *year is specified as*

$$N_{it} \mid R_i \sim \text{Pois}\left(R_i \Lambda_{it}\right),$$

*where*

$$\begin{aligned}\Lambda_{it} &:= \exp\left(\boldsymbol{X}_{it}\boldsymbol{\beta}\right)\\ &= \exp\left(\beta_0 + \beta_1 X_{it1} + \cdots + \beta_m X_{itm}\right)\end{aligned} \tag{5.2}$$

*with the coefficient parameters (fixed effects)* $\boldsymbol{\beta} := (\beta_0, \beta_1, \dots, \beta_m)'$*. Here, we assume the link function* $\eta$ *to be a log function that gives the exponential expression in* (5.2).

*ii. The random effect* $R_i$ *follows distribution function G with a parameter* $\psi$*.*

*iii. Furthermore, we assume that* $N_{i1}, \dots, N_{i\tau^*}$ *are independent conditional on* $R_i$*.*

In the case of a normal random effect, that is, $\gamma_i \sim \text{N}(0, \sigma^2)$, we have $R_i \sim \text{lognormal}(0, \sigma^2)$. Otherwise, if the random effect distribution is non-normal, the mixed model is called a hierarchical generalized linear mixed model (HGLMM) (Lee and Nelder, 1996). However, we do not differentiate between GLMMs and HGLMMs in this study. Although it is possible to assume that some of the parameters in (5.2) are random effects, known as random slopes in the classical GLMM literature, we do not pursue this direction for the sake of simplicity. We refer the reader to Zimmer (2018) for information on the impact of random slopes on insurance pricing.

In Model 2, the mean of the random effect $R_i$ is often set to be one for the identifiability issue. For example, we may assume

$$R_i \sim \text{gamma}\left(1, \psi\right) \tag{5.3}$$

with the mean of 1 and the dispersion parameter $\psi$. We refer to Section 2 for the details of parameterization. Note that using the random effect $R_i$ introduces longitudinal dependence among $N_{it}$s. Specifically, in Model 2, the law of total covariance is derived as

$$\begin{aligned}\text{cov}\left[N_{it_1}, N_{i,t_2}\right] &= \text{cov}\left[\mathbb{E}\left[N_{it_1} \mid R_i\right],\ \mathbb{E}\left[N_{it_2} \mid R_i\right]\right] + \mathbb{E}\left[\text{cov}\left[N_{it_1}, N_{it_2} \mid R_i\right]\right]\\ &= \text{cov}\left[\Lambda_{i,t_1} R_i, \Lambda_{i,t_2} R_i\right] + 0\\ &= \Lambda_{it_1}\Lambda_{it_2}\text{Var}(R_i).\end{aligned} \tag{5.4}$$

However, in Model 1, we have

$$\text{cov}\left[N_{it_1}, N_{it_2}\right] = 0. \tag{5.5}$$

An important implication in the comparison of the covariance in (5.4) under Model 2 and the covariance in (5.5) under Model 1 is that the random effect induces dependence among the observations. Thus, the random effect not only explains the unobserved explanatory components but also introduces dependence among observations. Hence, because the observations of the $i^{th}$ policyholder,

$$n_{i1}, \ldots, n_{i\tau}$$

are not independent under Model 2, the corresponding likelihood function under Model 2 cannot be written as the product of the likelihood function. Instead, we have the following log-likelihood function for the observations:

$$\ell_i(\boldsymbol{\beta}, \psi) := \log\left(\int \prod_{t=1}^{\tau} f_{\boldsymbol{\beta}}(n_{it} \mid r_i)\, g_{\psi}(r_i)\mathrm{d}r_i\right), \tag{5.6}$$

which requires one-dimensional integration. Here, $f_{\boldsymbol{\beta}}(\cdot \mid r_i)$ is the probability mass function of the Poisson distribution defined by

$$f_{\boldsymbol{\beta}}(n_{it} \mid r_i) := \frac{\exp(-\Lambda_{it} r_i)(\Lambda_{it} r_i)^{n_{it}}}{n_{it}!},$$

and $g_{\psi}$ is the probability density function of the random effect in (5.3). From a purely mathematical perspective, the introduction of a random effect in Model 2 is for modeling dependence within the observations. In this sense, the random-effect model is linked to the copula model (Frees and Valdez, 1998; Embrechts, 2009; Krupskii and Joe, 2013).

Next, for data in the form of (3.1), with $z$ policyholders and each policyholder observed for $\tau$ years, the log-likelihood function for Model 2 is written as

$$\ell(\boldsymbol{\beta}, \psi) := \sum_{i=1}^{z} \ell_i(\boldsymbol{\beta}, \psi).$$

For the estimation of unknown coefficients $\boldsymbol{\beta}$ and parameter $\psi$, MLE is usually considered, and the estimators of $\boldsymbol{\beta}$ and $\psi$ are typically defined as

$$\left(\widehat{\boldsymbol{\beta}}, \widehat{\psi}\right) := \arg\max_{\boldsymbol{\beta}, \psi}\, \ell(\boldsymbol{\beta}, \psi). \tag{5.7}$$

## 5.2. Prediction in the GLMM

The goal of the a posteriori risk classification is to predict the net premium for each policyholder. In Model 2, by definition, we have the following conditional expectation:

$$\mathbb{E}\left[N_{i,\tau^*} \mid R_i\right] = R_i \Lambda_{i,\tau^*}. \tag{5.8}$$

Because the random effect for the $i^{th}$ policyholder, $R_i$, is an unknown random variable, the conditional expectation in (5.8) cannot be used directly to predict the a posteriori rate in (3.2). However, in combination with the prediction of a random effect,

$$\mathbb{E}\left[R_i \mid n_{i1}, \ldots, n_{i\tau}\right], \tag{5.9}$$

the conditional expectation in (5.8) can calculate the a posteriori rate in (3.2) as follows:

$$\begin{aligned}\mathbb{E}\left[N_{i,\tau^*} \mid n_{i1},\ldots,n_{i\tau}\right] &= \mathbb{E}\left[\mathbb{E}\left[N_{i,\tau^*} \mid R_i, n_{i1},\ldots,n_{i\tau}\right] \mid n_{i1},\ldots,n_{i\tau}\right] \\ &= \Lambda_{i,\tau^*}\mathbb{E}\left[R_i \mid n_{i1},\ldots,n_{i\tau}\right].\end{aligned} \tag{5.10}$$

Hence, the prediction of the a posteriori rate in Model 2 is reduced to the estimation of predicting the random effect in (5.9). Note that the prediction of the random effect in (5.9) allows for a closed-form expression depending on the choice of distributional assumption for the random effect. Otherwise, the random effect in (5.9) can be predicted numerically, either based on numerical integration or a Monte Carlo simulation. The next subsection explains a specific method for predicting the random effect in (5.9).

## 5.3. Parametric assumptions for the random effect in the generalized linear mixed model

**Model 3.** *Considering the settings in Model 2, we further assume that the parametric assumption for $R_i$ in* (5.3) *is*

$$R_i \sim \text{gamma}\,(1,\psi)\,.$$

In Model 3, the likelihood function in (5.6) can be analytically expressed as

$$\begin{aligned}\ell_i\,(\boldsymbol{\beta},\psi) &= \log\left(\int \prod_{t=1}^{\tau} \frac{\exp\left(-r\lambda_{it}\right)\left(r\lambda_{it}\right)^{n_{it}}}{n_{it}!} \frac{1}{\Gamma(1/\psi)\psi^{1/\psi}}\left(r^{1/\psi-1}\right)\exp\left(-r/\psi\right)\mathrm{d}r\right) \\ &= \log\left(\frac{1}{\Gamma(1/\psi)\psi^{1/\psi}}\Gamma\left(1/\psi+\sum_{t=1}^{\tau} n_{it}\right)\left(1/\psi+\sum_{t=1}^{\tau}\lambda_{it}\right)^{-\left(1/\psi+\sum\limits_{t=1}^{\tau} n_{it}\right)}\prod_{t=1}^{\tau}\frac{\lambda_{it}^{n_{it}}}{n_{it}!}\right),\end{aligned} \tag{5.11}$$

where $\lambda_{it}$ denotes the realization of $\Lambda_{it}$ in (4.1).

Hence, the statistical estimation procedure is based on the simple optimization of (5.11), which does not require complicated numerical integration. Furthermore, since the gamma distribution is a conjugate prior of the Poisson distribution, the posterior mean of the random effect in (5.9) can be analytical obtained as follows:

$$\mathbb{E}\left[R_i \mid n_{i1},\ldots,n_{i\tau}\right] = \frac{1/\psi + \sum\limits_{t=1}^{\tau} n_{it}}{1/\psi + \sum\limits_{t=1}^{\tau} \lambda_{it}} \tag{5.12}$$

from the fact that the conditional distribution $R_i|n_{i1},\ldots,n_{i\tau}$ follows a gamma distribution with shape and scale parameters given by

$$1/\psi + \sum_{t=1}^{\tau} n_{it} \quad \text{and} \quad \frac{1}{1/\psi + \sum\limits_{t=1}^{\tau}\lambda_{it}},$$

respectively. In the conjugate families, the posterior mean of the individual's random effect in (5.12) can be represented by a weighted average between the prior mean of $R_i$ (weighted by its precision) and the ratio of the total number of insurance claims across $\tau$ years for policyholder $i$ to the sum of the respective rate parameters, $(\sum_{t=1}^{\tau} n_{it})/(\sum_{t=1}^{\tau} \lambda_{it})$ (weighted by the sum of the rate parameters):

$$\mathbb{E}\left[R_i \mid n_{i1}, \ldots, n_{i\tau}\right] = \frac{1/\psi + \sum\limits_{t=1}^{\tau} n_{it}}{1/\psi + \sum\limits_{t=1}^{\tau} \lambda_{it}} = (1)\frac{1/\psi}{1/\psi + \sum\limits_{t=1}^{\tau} \lambda_{it}} + \left(\frac{\sum\limits_{t=1}^{\tau} n_{it}}{\sum\limits_{t=1}^{\tau} \lambda_{it}}\right) \frac{\sum\limits_{t=1}^{\tau} \lambda_{it}}{1/\psi + \sum\limits_{t=1}^{\tau} \lambda_{it}}.$$

Because $\mathbb{E}\left[R_i|n_{i1}, \ldots, n_{i\tau}\right]$ later acts as a multiplier of the expectation from the fixed effects alone (see the example at the end of Section 5, under Model 3), $\mathbb{E}\left[R_i|n_{i1}, \ldots, n_{i\tau}\right] = 1$ suggests no modification in the premium. Additionally, the ratio $(\sum_{t=1}^{\tau} n_{it})/(\sum_{t=1}^{\tau} \lambda_{it})$ indicates how the observed number of claims relates to the expected number of claims over time for policyholder $i$. For example, if $\sum_{t=1}^{\tau} n_{it} > \sum_{t=1}^{\tau} \lambda_{it}$, then this would lead to a ratio greater than one, resulting in $\mathbb{E}\left[R_i|n_{i1}, \ldots, n_{i\tau}\right] > 1$.

So far, we have analyzed the properties of GLMMs in Model 2, in which the random effect $R_i$ is assumed to follow a gamma distribution. This distributional assumption is convenient because it provides an analytical expression for the likelihood function in (5.11) and the analytical conditional distribution of $R_i$. However, inherited from the linear mixed model, in which both the response and random variables are assumed to follow a normal distribution, the more popular choice of distributional assumption for $R_i$ is lognormal distribution. The following model provides a GLMM model with lognormal assumption for $R_i$.

**Model 4.** *We now consider the settings in Model 2 and we further assume that the parametric assumption for $R_i$ is*

$$R_i \sim \text{lognormal}\,(0, \psi)\,.$$

We have

$$\mathbb{E}\left[R_i\right] = \exp(\psi/2) \quad \text{and} \quad \operatorname{Var} R_i = (\exp(\psi) - 1)\exp(\psi)$$

under the parameterization in Model 4. Unlike the gamma distributional assumption for the random effect as in Model 3, the lognormal distributional assumption for the random effect does not lead to a closed-form expression for the likelihood function in (5.6). Hence, for the statistical estimation procedure in (5.7), numerical integration is required in (5.6). Standard numerical integration methods, such as the Laplace approximation (Laplace, 1986) and Gaussian quadrature (Golub and Welsch, 1969), can efficiently approximate the likelihood function for GLMMs. We explain that the a posteriori rate in (5.10) can be approximated as an affine function of observations in Section 6.1. In the next subsection the use of statistical programming for GLMMs is explained.

## 5.4. Data example

To fit the GLMMs in Model 2, we consider the `hglm` function in the hglm package (Rönnegård *et al.*, 2010) in R. The function is for fitting HGLMs via extended quasi-likelihood (Lee *et al.*, 2006), which extends GLMs by relaxing the assumption that error components are independent. Its grammatical form is as follows:

```
> hglm(formula, family=familytype(link=linkfunction),
        random=var, rand.family=familytype(link=linkfunction),
        data=dataset,...).
```

Table 4: Estimation results for the parameters under Model 3 and Model 4

| Parameter | | Model 3 | | | | | Model 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Est | Std.dev | z | *p*-value | | Est | Std.dev | z | *p*-value | |
| **Fixed effect** | | | | | | | | | | | |
| (Intercept) | $\hat{\beta}_0$ | −2.356 | 0.319 | −7.373 | <0.0001 | * | −2.849 | 0.364 | −7.829 | <0.0001 | * |
| Type = City | $\hat{\beta}_1$ | 1.048 | 0.370 | 2.833 | 0.0047 | * | 1.323 | 0.417 | 3.177 | 0.0015 | * |
| Type = County | $\hat{\beta}_2$ | 1.940 | 0.379 | 5.121 | <0.0001 | * | 2.229 | 0.422 | 5.284 | <0.0001 | * |
| Type = School | $\hat{\beta}_3$ | 0.920 | 0.333 | 2.766 | 0.0057 | * | 0.917 | 0.379 | 2.418 | 0.0157 | * |
| Type = Town | $\hat{\beta}_4$ | −0.909 | 0.390 | −2.331 | 0.0199 | * | −0.713 | 0.442 | −1.612 | 0.1072 | |
| Type = Village | $\hat{\beta}_5$ | 0.434 | 0.349 | 1.245 | 0.2131 | | 0.613 | 0.396 | 1.549 | 0.1216 | |
| avgCoverage | $\hat{\beta}_6$ | 0.383 | 0.048 | 8.038 | <0.0001 | * | 0.337 | 0.040 | 8.430 | <0.0001 | * |
| **Random effect** | $\psi$ | 1.072 | | | | | 1.191 | | | | |

Table 5: Selected estimation results for random effects under Models 3 and Model 4

| | Model 3 | | Model 4 | |
|---|---|---|---|---|
| | Estimate | Std.err | Estimate | Std.err |
| $R_1$ | 0.938 | 0.299 | 0.234 | 0.304 |
| $R_2$ | 0.647 | 0.247 | 0.004 | 0.238 |
| $R_3$ | 2.074 | 0.280 | 0.995 | 0.295 |
| $R_4$ | 0.875 | 0.238 | 0.262 | 0.238 |
| $R_5$ | 1.180 | 0.276 | 0.467 | 0.285 |
| $R_6$ | 0.921 | 0.314 | 0.302 | 0.313 |
| $R_7$ | 3.132 | 0.181 | 1.584 | 0.177 |
| $R_8$ | 0.336 | 0.345 | −0.451 | 0.286 |
| $R_9$ | 0.960 | 0.229 | 0.438 | 0.212 |
| $R_{10}$ | 0.264 | 0.360 | −0.906 | 0.341 |

This is similar to the `glm` function in `R` but with additional terms for the random effect: `random` and `rand.family`. First, we specify the policyholder-specific random effect of interest as `random=~1|PolicyNum`. To easily specify the distribution of the random effect, we consider the conditional expectation of frequency as

$$\mathbb{E}\left[N_{it} \mid \gamma_i\right] = \eta^{-1}\left(\boldsymbol{X_{it}\beta} + \gamma_i\right) = \exp\left(\boldsymbol{X_{it}\beta} + \gamma_i\right),$$

and having a log link function. To use the gamma distribution for the random effect in the function, the random effect $\gamma_i$ is given by $\gamma_i = \log(R_i)$ and $R_i \sim \text{gamma}(1, \psi)$. Thus, we need a log link function to specify the distribution for the random effect as `rand.family=gamma(link=log)`. Similarly, for the log-normal distribution, we can use the code `rand.family=gaussian(link=identity)`, which implies that $\gamma_i \sim \mathrm{N}(0, \psi)$; that is, $R_i \sim \text{lognormal}(0, \psi)$. The specific form of the function under the setting of Model 3 is

```
> hglm(Frequency~TypeCity+···+avgCoverage,
            family=poisson(link=log),
            random=~1|PolicyNum,
            rand.family=gamma(link=log),
            data=mydata).
```

Under the settings in Model 4,

```
> hglm(Frequency~TypeCity+···+avgCoverage,
            family=poisson(link=log),
            random=~1|PolicyNum,
            rand.family=gaussian(link=identity),
             data=mydata).
```

Alternative functions to fit GLMM are the `glmer` function of the lme4 package and the `glmmPQL` function of the MASS package. The `glmer` function is commonly used, and it fits the model via maximum likelihood. Meanwhile, `glmmPQL` uses the penalized quasi-likelihood approach to evaluate the likelihood of GLMM. The functions in Model 4 are

```
> glmer(Frequency~TypeCity+ ···+avgCoverage+(1|PolicyNum),
             family=poisson(link=log),
             data=mydata),
```

and

```
> glmmPQL(Frequency~TypeCity+···+avgCoverage,
              family=poisson(link=log),
              random=~1|PolicyNum,
              data=mydata).
```

The estimation results of these models for the parameters are summarized in Table 4. The estimates for the fixed effects differ slightly depending on the distribution of the random effect; however, we have the same interpretation for the tendency in risk characteristics. In this example, the function provides an estimate of the dispersion parameters for the random effect. Because the dispersion parameter of the gamma distribution is the reciprocal of the shape parameter and has a mean of 1, $\psi$ is implied. Furthermore, using the code `print(summary(res), print.ranef = TRUE)`, we obtain estimates of random effects for all policyholders. Table 5 lists the estimates for selected policyholders in Models 3 and 4. This function provides a different form of estimates depending on the link function of the random effect. From the estimation results in Tables 4 and 5, we obtain the estimated conditional expectation as follows:

$$\widehat{\mathbb{E}}\left[N_{i\tau^*} \mid n_{i1}, \ldots, n_{i\tau}\right] = \exp\left(\boldsymbol{X}_{i\tau^*}\hat{\boldsymbol{\beta}} + \hat{\gamma}_i\right).$$

Here, $\hat{\gamma}_i$ changes over time, based on historical observations, whereas the corresponding expectation of GLMs in (4.3) remains invariant with respect to time and historical observations. For example, the third policyholder in the dataset is in a county and has an average coverage of 2.495; thus, under Model 3, the estimated conditional expectation is

$$\begin{aligned}\exp\left(\hat{\beta}_0 + \hat{\beta}_2 + 0.38256\hat{\beta}_6\right) \cdot \hat{R}_3 &= \exp\left(-2.356 + 1.940 + 0.383 \times 2.495\right) \times 2.074 \\ &= 3.555,\end{aligned}$$

and under Model 4, it is

$$\begin{aligned}\exp\left(\hat{\beta}_0 + \hat{\beta}_2 + 0.38256\hat{\beta}_6 + \hat{\gamma}_3\right) &= \exp\left(-2.849 + 2.229 + 0.337 \times 2.495 + 0.995\right) \\ &= 3.375.\end{aligned}$$

## 6. Bühlmann method and GLMM

Although GLMMs offer a unified approach for calculating the a posteriori insurance rate, they often lack closed-form expressions, except for a few distributional assumptions; for instance, see Model 3. In this section, we demonstrate the incorporation of the credibility theory into the framework of GLMMs by providing a closed-form approximation of the a posteriori rate.

For simplicity, we consider the settings in Model 2 with the further assumption that the risk characteristics are the same across years; that is, $\boldsymbol{X}_{i1} = \cdots = \boldsymbol{X}_{i\tau^*}$. While one of the main interests of insurance is the prediction of the net premium of each policyholder defined in (5.8) using the a posteriori rate defined in (3.2), the closed form of the a posteriori rate defined in (3.2) is generally not possible. Alternatively, we may consider approximating the a posteriori rate for the $i^{th}$ policyholder defined in (3.2) as the affine function of the historical observations, by the following:

$$\text{Prem}_B := \widehat{\alpha}_0 + \widehat{\alpha}_1 N_{i1} + \cdots + \widehat{\alpha}_\tau N_{i\tau}, \tag{6.1}$$

where

$$(\widehat{\alpha}_0, \ldots, \widehat{\alpha}_\tau) := \arg\min_{(\alpha_0,\ldots,\alpha_\tau)\in\mathbb{R}^\tau} \mathbb{E}\left[\left(\mathbb{E}\left[N_{i\tau^*} \mid R_i\right] - (\alpha_0 + \alpha_1 N_{i1} + \cdots + \alpha_\tau N_{i\tau})\right)^2\right].$$

Following the classical procedure in the Bühlmann premium (Bühlmann and Gisler, 2006), the optimal premium for the $i^{th}$ policyholder in (6.1) is obtained as

$$\widehat{\alpha}_0 = (1 - Z_\tau)\,\mathbb{E}\left[N_{i\tau^*}\right] \quad \text{and} \quad \widehat{\alpha}_1 = \cdots = \widehat{\alpha}_\tau = \frac{Z_\tau}{\tau},$$

where $Z_\tau$, the Bühlmann factor, is given by

$$Z_\tau = \frac{\tau\,\mathrm{V}ar\left(\mathbb{E}\left[N_{i\tau^*} \mid R_i\right]\right)}{\mathbb{E}\left[\mathrm{V}ar\left(N_{i\tau^*} \mid R_i\right)\right] + \tau\,\mathrm{V}ar\left(\mathbb{E}\left[N_{i\tau^*} \mid R_i\right]\right)}.$$

Furthermore, $\mathbb{E}\left[N_{i\tau^*}\right]$, $\mathbb{E}\left[\mathrm{V}ar\left(N_{i\tau^*}|R_i\right)\right]$, and $\mathrm{V}ar\left(\mathbb{E}\left[N_{i\tau^*}|R_i\right]\right)$ can be obtained as

$$\begin{cases} \mathbb{E}\left[N_{i\tau^*}\right] := \Lambda_i, \\ \mathbb{E}\left[\mathrm{V}ar\left(N_{i\tau^*} \mid R_i\right)\right] := \Lambda_i, \\ \mathrm{V}ar\left(\mathbb{E}\left[N_{i\tau^*} \mid R_i\right]\right) := (\Lambda_i)^2\,\psi \end{cases}$$

from equations

$$\begin{cases} \mathbb{E}\left[N_{i\tau^*}\right] := \mathbb{E}\left[u\left(R_i\right)\right]; \\ \mathbb{E}\left[\mathrm{V}ar\left(N_{i\tau^*} \mid R_i\right)\right] := \mathbb{E}\left[v\left(R_i\right)\right]; \\ \mathrm{V}ar\left(\mathbb{E}\left[N_{i\tau^*} \mid R_i\right]\right) := \mathrm{V}ar\left(u\left(R_i\right)\right), \end{cases}$$

where

$$u\left(R_i\right) := \mathbb{E}\left[N_{i\tau^*} \mid R_i\right] = R_i\Lambda_i \quad \text{and} \quad v\left(R_i\right) := \mathrm{V}ar\left(N_{i\tau^*} \mid R_i\right) = R_i\Lambda_i.$$

## 7. Discussion and conclusion

In this study, we present an intuitive explanation of the structure of GLMs and GLMMs and their relationship with insurance pricing methods; and we explain their connection with the Bühlmann method in insurance. With recent advances in computing tools, such as deep neural networks, insurance pricing can become more accurate through the use of complex predictive expressions. However, the traditional credibility method, which seeks an intuitive representation of insurance prices, seems at odds with modern statistical techniques. An interesting future topic is to explore how to adapt modern statistical computing tools to the credibility method in order to improve accuracy while preserving its intuitive and simple form.

Here, we mainly focus on the Poisson distribution in GLM and GLMMs to model the frequency of insurance claims, which are commonly used for count data. However, based on the assumption that the variance and mean are the same in the distribution, there are limitations, such as incorrect inference when overdispersion is present in the dataset. To address overdispersion in the Poisson model, overdispersion adjustments can be made using the quasi-likelihood method (quasi-Poisson model) or an alternative distribution method, such as negative binomial distribution. We note that, with some limitation (Lee *et al*., 2020), the problem of overdispersion can be reduced for Poisson distribution by using the GLMM framework.

## Acknowledgment

Rosy Oh was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) and was funded by the Ministry of Education (NRF-2020R1I1A1A01067376). This research was supported by the Korean Risk Management Society. Jae Youn Ahn was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (RS-2023-00217022) and an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (RS-2022-00155966). The authors thank Ruhuan Feng (University of Illinois Urbana-Champaign) and Emiliano Valdez (University of Connecticut) for their comments and advice on this research.

## References

Antonio K and Beirlant J (2007). Actuarial statistics with generalized linear mixed models, *Insurance: Mathematics and Economics*, **40**, 58–76.

Breslow NE and Clayton DG (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, **88**, 9–25.

Bühlmann H and Gisler A (2005). *A Course in Credibility Theory and Its Applications* (Vol. 317), Berlin, Springer.

Embrechts P (2009). Copulas: A personal view, *Journal of Risk and Insurance*, **76**, 639–650.

Frees EW, Lee G, and Yang L (2016). Multivariate frequency-severity regression models in insurance, *Risks*, **4**, 4.

Frees EW and Valdez EA (1998). Understanding relationships using copulas, *North American Actuarial Journal*, **2**, 1–25.

Golub GH and Welsch JH (1969). Calculation of gauss quadrature rules, *Mathematics of Computation*, **23**, 221–230.

Krupskii P and Joe H (2013). Factor copula models for multivariate data, *Journal of Multivariate Analysis*, **120**, 85–101.

Lai TL and Sun KH (2012). Evolutionary credibility theory: A generalized linear mixed modeling approach, *North American Actuarial Journal*, **16**, 273–284.

Laplace PS (1986). Memoir on the probability of the causes of events, *Statistical Science*, **1**, 364–378.

Lee W, Kim J, and Ahn JY (2020). The poisson random effect model for experience ratemaking: Limitations and alternative solutions, *Insurance: Mathematics and Economics*, **91**, 26–36.

Lee Y and Nelder JA (1996). Hierarchical generalized linear models, *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 619–656.

McCullagh P and Nelder JA (1989). *Generalized Linear Models* (2nd ed), Chapman and Hall, London.

Mowbray AH (1914). How extensive a payroll exposure is necessary to give a dependable pure premium, *Proceedings of the Casualty Actuarial Society*, **1**, 24–30.

Nelder JA and Verrall RJ (1997). Credibility theory and generalized linear models, *ASTIN Bulletin: The Journal of the IAA*, **27**, 71–82.

Oh R, Shi P, and Ahn JY (2020). Bonus-malus premiums under the dependent frequency-severity modeling, *Scandinavian Actuarial Journal*, **2020**, 172–195.

Ohlsson E (2008). Combining generalized linear models and credibility models in practice, *Scandinavian Actuarial Journal*, **2008**, 301–314.

Rönnegård L, Shen X, and Alam M (2010). hglm: A package for fitting hierarchical generalized linear models, *The R Journal*, **2**, 20–28.

Whitney A (1918). The theory of experience rating, *Proceedings of the Casualty Actuarial Society*, **4**, 274–292.

Wuthrich MV (2021). From Generalized Linear Models to Neural Networks, and Back. [S.l.]: SSRN, Available from: https://ssrn.com/abstract=3491790

Zimmer DM (2018). The heterogeneous impact of insurance on health care demand among young adults: A panel data analysis, *Journal of Applied Statistics*, **45**, 1277–1291.

*Received February 7, 2023; Revised June 7, 2023; Accepted July 11, 2023*