Generalized linear mixed models for dependent compound risk models

by Emiliano Valdez, Himchan Jeong, Jae Youn Ahn, and Sojung Park

Abstract

In ratemaking, calculation of a pure premium has traditionally been based on modeling frequency and severity in an aggregated claims model. For simplicity, it has been a standard practice to assume the independence of loss frequency and loss severity. In recent years, there is sporadic interest in the actuarial literature exploring models that departs from this independence. In this article, we extend the work of Garrido et al. (2016) which uses generalized linear models (GLMs) that account for dependence between frequency and severity and simultaneously incorporate rating factors to capture policyholder heterogeneity. In addition, we quantify and explain the contribution of the variability of claims among policyholders through the use of random effects using generalized linear mixed models (GLMMs). We calibrated our model using a portfolio of auto insurance contracts from a Singapore insurer where we observed claim counts and amounts from policyholders for a period of six years. We compared our results with the dependent GLM considered by Garrido et al. (2016), Tweedie models, and the case of independence. The dependent GLMM shows statistical evidence of positive dependence between frequency and severity. Using validation procedures, we find that the results demonstrate a more superior model when random effects are considered within a GLMM framework.

Keywords: Dependent frequency-severity models, random effects models, GLM, GLMM, ratemaking

1 Introduction and motivation

For many apparent reasons including ease of implementation, there has been an increase in popularity even among practitioners of the use of generalized linear models (GLMs) for insurance ratemaking, risk classification and many other actuarial applications. See, for example, Antonio and Valdez (2011), Frees et al. (2014) and Frees et al. (2016a). Originally synthesized by Nelder and Wedderburn (1972), GLMs extend the ordinary regression models to accommodate response variables that are not normally distributed and are rather members of the exponential family of distributions. As pointed out in Chapter 5 of Frees et al. (2014), the primary features of GLMs include a function that links the response variable to a set of predictor variables and a variance structure that is not necessarily constant across independent observations. It encompasses a wide variety of models that include the normal regression, Poisson regression, logistic regression, probit regression, to name a few. More importantly, it has been used by Garrido et al. (2016) to model the dependence between loss frequency and loss severity.

Following the formulation in Garrido et al. (2016), consider an insurer's portfolio with a class of policyholders where for a fixed time period, the number of claims incurred is N and the corresponding individual amount

of claims are denoted by $C_1, C_2, \ldots, C_N$. The total loss incurred can be written as the sum

$$S = C_1 + C_2 + \cdots + C_N$$

with the convention that when $N = 0$, the total loss S is also zero. In the case where $N > 0$, we define the average claim severity by $\overline{C} = S/N$ so that the aggregate loss can be expressed as $S = N\overline{C}$. Let $\mathbf{x} = (x_1, \ldots, x_p)$ be a set of p covariates within the class of policyholders. These p covariates may affect the means for loss frequency and loss severity differently. However, dependence is introduced in the model with the addition of N as a covariate affecting the mean for loss severity.

Both frequency N and average severity $\overline{C}$ are modeled as GLM that incorporates the effect of the various covariates. However, individual claims $C_1, C_2, \ldots, C_N$ are assumed to be independent with each C_i to have a reproductive Exponential Dispersion Family (EDF) structure, that is $C_i \sim EDF(\mu, \phi)$ conditional on knowing N. This is necessary so that conditionally on N, the average severity is also a member of EDF, that is, $\overline{C} \sim EDF(\mu, \phi/N)$. With link functions g_N and g_C for frequency and severity, respectively, it follows that

$$\nu = \mathbb{E}\left[N|\mathbf{x}\right] = g_N^{-1}(\mathbf{x}\alpha) \qquad \text{and} \qquad \mu_\theta = \mathbb{E}\left[\overline{C}|\mathbf{x}, N\right] = g_C^{-1}(\mathbf{x}\beta + \theta N)$$

With this specification, testing for independence is equivalent to testing whether $\theta = 0$ or not.

Using the tower property of expectation, the mean of aggregate loss can be expressed as

$$\mathbb{E}\left[S|\mathbf{x}\right] = \mathbb{E}\left[N\mathbb{E}\left[\overline{C}|\mathbf{x}, N\right]|\mathbf{x}\right] \neq \mathbb{E}\left[N|\mathbf{x}\right]\mathbb{E}\left[C|\mathbf{x}\right]$$

which clearly demonstrates the possibility of dependence. It has been shown in Garrido et al. (2016) that the variance of aggregate loss can be expressed as

$$Var(S|\mathbf{x}) = \phi_\theta \mathbb{E}\left[NV_C(\mu e^{\theta N})|\mathbf{x}\right] + \mu^2\left[\frac{1}{4}M_N^{''}(2\theta|\mathbf{x}) - (M_N^{'}(\theta|\mathbf{x}))^2\right].$$

In a simulation study done in Garrido et al. (2016), the presence of dependence can have serious biases on the value of the regression coefficients, their standard errors, as well as the mean and variance of the portfolio's aggregate loss. Using a real dataset, their work also found strong evidence of the presence of negative dependence between loss frequency and loss severity. This is an interesting result because it says that increasing level of number of claims decreases the amount of severity. An interesting observation was said in the paper that "claim counts and amounts are often negatively associated in collision automobile insurance because drivers who file several claims per year are typically involved in minor accidents."

The dependent GLM in Garrido et al. (2016) has a structure similar to that of the two-part frequency-severity model in Frees et al. (2011) where they considered that the amount of health care expenditures is affected by either the number of inpatient stays or outpatient visits. According to the result of their work, inpatient stays did not significantly affect the total annual health care expenditure but outpatient visits had a significantly negative impact on the total annual health care expenditure.

In Lee et al. (2016), a similar dependent frequency-severity model was considered that compared the effect of constant and varying dispersion parameters. The frequency is a Poisson regression model while the severity is a gamma regression model with a deterministic and assumed known function of frequency as additional

covariate. The model was calibrated using an auto insurance data set from the state of Massachusetts in year 2006. Interestingly, the result indicated a positive effect of frequency on severity, that is, high frequency led to increase in average severity.

Following an approach similar to that in Garrido et al. (2016), Park et al. (2017) calibrated the dependent two-part GLM with claims data from a Korean insurance company. Interestingly, the authors used the level of bonus-malus as an explanatory variable for predicting average severity. Bonus-malus is quite a common practice in Korea and other countries for penalizing bad drivers while rewarding good drivers. Their estimation results showed that for claims related to collisions, the dependency tends to change from positive to negative as the level of bonus-malus of a policyholder increases. However, there was lacking statistical evidence of dependency for liability claims.

Shi et al. (2015) applied a conditional two-part model with zero-truncated negative binomial distribution for frequency and a generalized gamma distribution for the average severity. To incorporate the dependency between frequency and average severity, they proposed the use of copulas but compared the results to the dependent two-part GLM used in Garrido et al. (2016). According to their study, in both the two-part model and the copula model, frequency had a positive correlation with average severity. Furthermore, they found using out-of-sample validation that the copula model outperforms the dependent two-part GLM.

The work of Shi et al. (2015) proposed the model which can accommodate heavy tail distribution for the claim severity as an extension of Czado et al. (2012) and Frees et al. (2011). The model estimation yielded a positive dependence between frequency and average severity. According to their interpretation, this is because the insured with high frequency may have a riskier driving habit, which leads to more severe claim when an accident happens.

There are certainly varied results regarding the relationship between frequency and average severity and this may be attributed to the differences in the type of model used and the characteristics inherent in the dataset used to calibrate the corresponding model.

For a typical portfolio of insurance policies, it is not uncommon to have observations of independent policyholders to come in a longitudinal format such as

$$(N_{it}, C_{itj}, \mathbf{x}_{it}, e_{it})' \tag{1.1}$$

for calendar year t, for $t = 1, \ldots, T_i$ where $T_i \leq T$ and for policyholder i, for $i = 1, \ldots, M$. There is a fixed number of calendar years T and we allow for unbalanced data. $\mathbf{x}_{it}$ refers to the vector of covariates describing policyholder characteristics and e_{it} refers to the length of exposure of the policyholder within calendar year t where $0 < e_{it} \leq 1$. The subscript j will become clear when we define our dependent compound risk model. Random effects models are typically used for such longitudinal observations with GLMM as a special case. Molenberghs and Verbeke (2005) provide an overview of statistical models for analyzing longitudinal data from clinical studies emphasizing the importance of "model formulation and parameter estimation."

In this article, we extend the GLM framework with dependent frequency and severity model suggested by Garrido et al. (2016) to the GLMM framework with dependent frequency and severity model. The primary contribution of our model is the addition of random, or subject-specific, effects in the linear predictor within

a dependent frequency and severity model. Here our subject is the policyholder or the insured in a portfolio of insurance contracts. Including random effects in the linear predictor reflects the idea that there is a natural heterogeneity across subjects (policyholders) and that the observations on the same subject may share common characteristics.

Longitudinal data models including GLMMs are not new in the actuarial literature. Frees et al. (1999) demonstrate the link between longitudinal data models and credibility models. Such link allows for a more convenient statistical analysis of credibility models as part of a ratemaking process; see Frees et al. (2001). On a similar note, Garrido and Zhou (2009) examined limited fluctuations credibility of GLMM estimators. Antonio and Beirlant (2007) provides motivations for using GLMMs for analyzing longitudinal data in actuarial science; in particular, claims datasets arising from a portfolio of workers' compensation insurance were used to calibrate GLMMs. The authors do not only provide the GLMM formulation but also give a discussion of estimation, inference and prediction useful for analyzing the data. In Antonio and Valdez (2011), GLMM has been discussed as a modern technique for risk classification, the process of "grouping of risks into various classes that share a homogeneous set of characteristics allowing the actuary to reasonably price discriminate." Finally, Boucher et al. (2008) suggested the use of random effects model for claim counts with time dependence. Indeed they compared such model with other types of models such as time-independent marginal models, models based on conditional distributions with artificial marginals, integer-valued autoregressive models, common shock models as well as copula models. Using an automobile insurance portfolio from a major company in Spain with data for years 1991 to 1998, they concluded that the random effects model is the most superior model for insurance claims data with time dependence.

Inspired by the work of Garrido et al. (2016) and the growing interest of the use of GLMM in insurance ratemaking and risk classification, we examine the usefulness of GLMM for modeling dependent compound risk. This appears to be a natural extension to make especially insurance data typically come in the form of (1.1). For example, auto insurance companies usually have data with policyholders observed for a period of consecutive years; such data provides a richer set of information for improved prediction, pricing and reserving. Preserving the GLM dependent frequency-severity structure proposed in Garrido et al. (2016), the extension is done with the extra addition of random effects to the linear predictor in a GLM framework. This addition of random effects allows us to capture the heterogeneity typically found when observations are not independent, but may be repeated. As said in Chapter 16, p. 400 in Frees et al. (2014), this "enables cluster-specific prediction ... and structure correlation within cluster."

The remainder of this paper is organized as follows. In Section 2, we give an introduction to generalized linear mixed models. The purpose of this is to make the paper self-contained and introduce the notation adopted throughout the paper. In Section 3, we describe the data used, some preliminary investigation of the data, and the specification of the GLMM used in model calibration. In Section 4, we give numerical results of our estimation. We also provide some methods we used in model validation. Finally, we conclude this paper with some remarks in Section 5. The appendices provide for details of the calculation of the likelihood equations and the derivation of the mean and variance of the aggregate loss.

2 GLMM and the dependent compound risk model

In this section, we introduce the concept of a GLMM by first briefly explaining GLM. This section also clarifies many of the notations adopted in this paper. Assume we have M independent observations with response variable y_i and known covariates $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$. GLMs extend ordinary linear models with normally distributed responses to include members from the exponential family with a density expressed in canonical form as

$$f(y_i|\beta, \tau) = \exp\left(\frac{y_i\gamma_i - \psi(\gamma_i)}{\tau} + c(y_i, \tau)\right), \quad \text{for } i = 1, \ldots, M \tag{2.1}$$

Within this exponential family, the relations $\mu_i = \mathbb{E}[y_i] = \psi^{'}(\gamma_i)$ and $Var[y_i] = \tau\psi^{''}(\gamma_i) = \tau V(\mu_i)$ hold for the mean and variance, respectively, where $g(\mu_i) = \mathbf{x}_i^{'}\beta$. Here, $V(\cdot)$ is the variance function while $g(\cdot)$ is called the link function that links the mean to the linear predictor. β denotes a $p \times 1$ fixed effects parameter vector in the linear predictor that is typically estimated from observed data.

Note that this model can be extended for longitudinal data by allowing for random, or subject-specific, effects in the linear predictor. Including random effects in the linear predictor reflects the idea that there is a natural heterogeneity across subjects (such as the insured) and that the observations on the same subject share common characteristics.

To further fix ideas, suppose we now have a dataset consisting of M independent subjects with response variable y_{it} and known covariates $\mathbf{x}_{it}'$. Here i refers to the subject, for $i = 1, \ldots, M$, observed during calendar year t for $t = 1, \ldots, T_i$ where $T_i \leq T$ and T is the overall total number of calendar years observed. Given the vector R_i describing the random effects for subject (or policyholder within the context of insurance applications) i, the response variable y_{it} has a density from the exponential family of the form

$$f(y_{it}|R_i, \beta, \tau) = \exp\left(\frac{y_{it}\gamma_{it} - \psi(\gamma_{it})}{\tau} + c(y_{it}, \tau)\right). \tag{2.2}$$

Within this framework, the following relations hold

$$\mu_{it} = \mathbb{E}[y_{it}|R_i] = \psi^{'}(\gamma_{it}) \qquad \text{and} \qquad Var(y_{it}|R_i) = \tau\psi^{''}(\gamma_{it}) = \tau V(\mu_{it}),$$

where $V(\cdot)$ is referred to as the variance function. Similar to the GLM construction, we model a transformation of the mean using the link function $g(\mu_{it}) = \mathbf{x}_{it}'\beta + z_{it}^{'}R_i$. Clearly, the link function $g(\cdot)$ provides a link between the mean and the linear form of the predictors including the random effects. Furthermore, β denotes a $p \times 1$ fixed effects parameter vector and R_i is a $q \times 1$ random effects vector. $\mathbf{x}_{it}(p \times 1)$ and $z_{it}(q \times 1)$ describe each subject i's covariate information for the fixed and random effects, respectively.

The specification of the GLMM is completed by assuming that the random effects, R_i, are mutually independent and identically distributed with density function $f(R_i|\sigma)$ where σ denotes the parameter vector in the density. For practical purposes, it is commonly assumed that the random effects follow a multivariate

normal distribution with zero mean vector and covariance matrix described by σ. This covariance structure allows us the flexibility to specify correlation within the subjects. It is clear, therefore, that within this framework, the dependence between observations on the same subject arises as a result of sharing the same random effects R_i.

Invoking the tower property of expectation, we find some general form of the unconditional mean

$$\mathbb{E}\left[y_{it}\right] = \mathbb{E}\left[\mathbb{E}\left[y_{it}|R_i\right]\right] = \mathbb{E}\left[g^{-1}(\mathbf{x}'_{it}\beta + z_{it}R_i)\right] \tag{2.3}$$

and based on the usual breakdown of the components of the variance, we also find some general form of the unconditional variance

$$\begin{aligned} Var[y_{it}] &= Var[\mathbb{E}\left[y_{it}|R_i\right]] + \mathbb{E}\left[Var[y_{it}|R_i]\right] \\ &= Var[\mu_{it}] + \mathbb{E}\left[\tau V(\mu_{it})\right]] \\ &= Var[g^{-1}(\mathbf{x}'_{it}\beta + z'_{it}R_i)] + \mathbb{E}\left[\tau V(\mu_{it})\right]. \end{aligned} \tag{2.4}$$

To better understand the consequences of adding random effects, as a simple illustration, consider the log link function $g(\mu) = \log\mu$, $V(x) = x^2$, $z_{it} = 1$ and $R_i \sim N(0, \sigma_R^2)$. It can easily be shown that the mean is

$$\mathbb{E}\left[y_{it}\right] = \mathbb{E}\left[\exp(\mathbf{x}'_{it}\beta + R_i)\right] = \exp(\mathbf{x}'_{it}\beta)\mathbb{E}\left[\exp(R_i)\right] = \exp(\mathbf{x}'_{it}\beta)\exp(\sigma_R^2/2)$$

while the variance is

$$Var(y_{it}) = Var(e^{\mathbf{x}'_{it}\beta + R_i}) + \tau\mathbb{E}\left[e^{(2\mathbf{x}'_{it}\beta + 2R_i)}\right] = e^{2\mathbf{x}'_{it}\beta + \sigma_R^2}\left[e^{\sigma_R^2}(1+\tau) - 1\right].$$

See Chapter 8 of McCulloch and Searle (2001).

For estimation purpose, the likelihood function for the unknown parameters β, σ and τ can be expressed as an integral as follows:

$$L(\beta, \sigma, \tau; \mathbf{y}) = \prod_{i=1}^{M} \int \prod_{t=1}^{T_i} f(y_{it}|\sigma, \beta, \tau) f(R_i|\sigma) dR_i \tag{2.5}$$

where $y = (y'_1, \cdots, y'_M)$ and the integration is done with respect to the q dimensional vector R_i. This integration presents some computational challenges for maximum likelihood estimation. To even have results that can give some kind of explicit expressions, we need to specify ostensibly referred to as conjugate distributions. For instance, when both the data and the random effects are normally distributed, the integral can be worked out analytically and explicit expressions can easily be derived. For more complex distributional assumptions, some approximations have been proposed in the literature. See McCulloch and Searle (2001). This is beyond the purpose of this paper. We will stick to the usual normal assumptions typically specified

for the random effects.

2.1 The dependent compound risk GLMM

As briefly stated in the introduction, a typical portfolio of insurance policies would contain observations of independent policyholders in a longitudinal format $(N_{it}, C_{itj}, \mathbf{x}_{it}, e_{it})'$ for calendar year t, for $t = 1, \ldots, T_i$ where $T_i \leq T$ and for policyholder i, for $i = 1, \ldots, M_t$. There is a fixed number of calendar years T and we allow for unbalanced data. $\mathbf{x}_{it}$ refers to the vector of covariates describing policyholder characteristics and e_{it} refers to the length of exposure of the policyholder within calendar year t where $0 < e_{it} \leq 1$.

For our purposes, N_{it} denotes the number of claims and C_{itj} denotes the claim size observed where j is an additional subscript needed to distinguish the many possible claims that may be incurred for each calendar year, hence, $j = 1, \ldots, N_{it}$. For each calendar year, we specify the distribution of claim severity applicable for the average claim in each calendar where we define

$$\overline{C}_{it} = \frac{1}{N_{it}} \sum_{j=1}^{N_{it}} C_{itj}.$$

The joint distribution for our dependent compound risk model can therefore be written as

$$f(n, \overline{c}|\alpha, \beta, b, u) = f_N(n|\alpha, b) \times f_{\overline{C}|N}(\overline{c}|\beta, u, n), \tag{2.6}$$

where the subscripts it have been suppressed for convenience and $N > 0$. The subscripts N and $\overline{C}|N$ are used to distinguish the density functions for frequency and severity, respectively. b is the random effects vector for the frequency while u is the random effects for the average severity. The construction in (2.6) is typically similar in structure to the two-part model of frequency and severity, with the exception of the addition of random effects and the addition of possible dependence between frequency and severity.

It is understood that when $N = 0$, the joint distribution simplifies to $f(0, 0) = f_N(0)$ and represents the probability of zero claims. In addition, when $N = 0$, the average severity will also be zero. Here we choose distributions for both frequency and average severity from the family of GLMM distributions. As a final specification of this dependent compound risk GLMM, we incorporate dependence between frequency and severity by adding the number of claims N as a linear predictor in the mean function for the average severity. While we assume that only the number of claims N of the current year can affect the mean of the severity of the current year just for simplicity of the model, this assumption can be generalized to assume that the number of claims N can affect the severity even though they are not in the same year. In particular, for the frequency component, we write the mean in terms of the linear predictor as

$$\mu_f = \mathbb{E}[N|\mathbf{x}] = g_N^{-1}(\mathbf{x}'\alpha + z_N'b),$$

where α is a parameter vector for the covariates associated with frequency. For the severity component, we

write

$$\mu_s = \mathbb{E}\left[\overline{C}|\mathbf{x}, N\right] = g_C^{-1}(\mathbf{x}'\beta + z_C'u + \theta N),$$

where β is a parameter vector for the covariates associated with average severity for which they may be different from that of frequency.

Finally, we can define the compound sum as

$$S_{it} = \sum_{j=1}^{N_{it}} C_{itj} = N_{it}\overline{C}_{it} \tag{2.7}$$

to represent the aggregate claims for policyholder i in calendar year t. Suppressing the subscripts, it is straightfoward to see that the unconditional mean of S can be expressed as

$$\mathbb{E}\left[S|\mathbf{x}\right] = \mathbb{E}\left[N\overline{C}|\mathbf{x}\right] = \mathbb{E}\left[N\mathbb{E}\left[\overline{C}|N,u\right]|\mathbf{x}\right] = \mathbb{E}\left[Ng_C^{-1}(\mathbf{x}'\beta + z'u + \theta N)|\mathbf{x}\right] \tag{2.8}$$

and the unconditional variance as

$$\begin{aligned} Var(S|\mathbf{x}) &= Var(\mathbb{E}\left[N\overline{C}|N,\mathbf{x}\right]) + \mathbb{E}\left[Var(N\overline{C}|N,\mathbf{x})\right] \\ &= Var(N\mathbb{E}\left[\overline{C}|N,\mathbf{x}\right]|\mathbf{x}) + \mathbb{E}\left[N^2 Var(\overline{C}|N,\mathbf{x})|\mathbf{x}\right] \\ &= Var(Ng_C^{-1}(\mathbf{x}'\beta + z'u + \theta N)|\mathbf{x}) + \mathbb{E}\left[N^2\tau V(g_C^{-1}(\mathbf{x}'\beta + z'u + \theta N))|\mathbf{x}\right] \end{aligned} \tag{2.9}$$

Note that if we are interested in the aggregate claims for calendar year t alone, we can easily define this as

$$S_t = \sum_{i=1}^{M} N_{it}\overline{C}_{it},$$

where M here is the total policies for each year. This should not preclude us from assuming that the number of policies may vary by year.

2.2 Negative binomial for frequency and gamma for average severity

To illustrate further the concepts, we now consider a negative binomial distribution for the frequency and a gamma distribution for the average severity.

Starting with the frequency component, we have the following model specification:

$$N|b \sim \text{ indep. NB}(\nu e^b, r) \text{ with density } f_N(N|b)$$

where

$$f_N(n|b) = \binom{n+r-1}{n}\left(\frac{r}{r+\nu e^b}\right)^r\left(\frac{\nu e^b}{r+\nu e^b}\right)^n. \tag{2.10}$$

Based on the log-link function, we have $\nu = \exp(\mathbf{x}'\alpha)$ where α is a $p \times 1$ vector of regression coefficients. The following relations hold:

$$\mathbb{E}\left[N|b\right] = e^{\mathbf{x}'\alpha+b} \qquad \text{and} \qquad Var[N|b] = e^{\mathbf{x}'\alpha+b}\left(1+e^{\mathbf{x}'\alpha+b}/r\right).$$

It is clear from these relations that unlike the Poisson distribution, overdispersion can easily be handled by the negative binomial distribution. An additional item to note is the accounting of the exposure e_{it} and this is accomplished through an adjustment to the mean parameter by multiplying it with e_{it}.

We assume that the random effect b has a normal distribution with mean zero and variance σ_b^2. We will need the density function of b, $f_b(b)$, and for our purposes, we will denote it by

$$f_b(b) = \frac{dF_b}{db} = \frac{1}{\sqrt{2\pi}\sigma_b}e^{-b^2/2\sigma_b^2} \text{ so that } \mathbb{E}\left[h(b)\right] = \int h(b)\frac{1}{\sqrt{2\pi\sigma_b^2}}e^{-b^2/2\sigma_b^2}db = \int h(b)dF_b.$$

For the average severity component, given $N > 0$, we have the following model specification:

$$\overline{C}|N,u \sim \text{ indep. Gamma}(\mu e^{u+\theta n}, \phi/n) \text{ with density } f_{\overline{C}|N}(\overline{c}|n,u)$$

where

$$f_{\overline{C}|N}(\overline{c}|n,u) = \frac{1}{\Gamma(n/\phi)}\left(\frac{n}{\mu e^{u+\theta n}\phi}\right)^{n/\phi}\overline{c}^{n/\phi-1}\exp\left(-\frac{n\overline{c}}{\mu e^{u+\theta n}\phi}\right). \tag{2.11}$$

Similar to the frequency component, we assume a log-link function so that $\mu = \exp(\mathbf{x}'\beta)$ where β is a $p \times 1$ vector of regression coefficients. The set of covariates selected for the severity component may be different from that of the frequency component. Any covariate that will not be considered significant may simply be ignored and naturally, this implies a corresponding beta coefficient of zero. Additionally, we will introduce the number of claims n as a covariate with coefficient θ. With this specification, the following relations hold:

$$\mathbb{E}\left[\overline{C}|N,u\right] = e^{\mathbf{x}'\beta+\theta n+u} \qquad \text{and} \qquad Var[\overline{C}|N,u] = e^{2(\mathbf{x}'\beta+\theta n+u)}\phi/n.$$

Note that both the negative binomial for frequency and gamma for average severity fall within the class of generalized linear mixed models. This is easy to verify.

We assume that the random effect u has a normal distribution with mean zero and variance σ_u^2 and its density function will similarly be denoted by f_u and defined as $f_u(u) = \frac{dF_u}{du} = \frac{1}{\sqrt{2\pi}\sigma_u}e^{-u^2/2\sigma_u^2}$.

In summary, our observed data consist of

$$(N_{it}, \overline{C}_{it}, \mathbf{x}_{it}, e_{it})' \tag{2.12}$$

for calendar year t, for $t = 1, \ldots, T_i$ where $T_i \leq T$ and for policyholder i, for $i = 1, \ldots, M$. The general form of the likelihood then can be expressed as

$$L = \int \int \prod_i \prod_t f(n_{it}, \overline{c}_{it}|b, u) dF_b dF_u. \tag{2.13}$$

One of the key property here is that the likelihood function L in equation (2.13) is divided into two product terms, where one term is only related with parameters for frequency while the other term is only related with parameters for severity. By using our proposed model, essentially we are optimizing two separate parts while still considering dependence between frequency and severity. Such decomposition of the likelihood function L enables us to optimize L using existing generalized linear mixed effect model packages in most statistical softwares.

To derive maximum likelihood estimates, we maximize the $\log L$ with respect to all the parameters by taking partial derivatives and setting each to zero. This leads us to the following set of $(2p+3)$ estimating equations that solve for $\hat{\alpha}$, $\hat{\beta}$, $\hat{\theta}$, $\hat{\sigma_b}$ and $\hat{\sigma_u}$, for $k = 1, \ldots, p$:

$$\sum_i \left[\frac{\int \sum_t \frac{\mathbf{x}'_{it}(k)}{r + e^{\mathbf{x}'_{it}\alpha + b}} (n_{it} - e^{\mathbf{x}'_{it}\alpha + b}) \prod_t f_N(n_{it}) dF_b}{\int \prod_t f_N(n_{it}) dF_b} \right] = 0$$

$$\sum_i \left[\frac{\int (b^2 - \sigma_b^2) \prod_t f_N(n_{it}) dF_b}{\int \prod_t f_N(n_{it}) dF_b} \right] = 0$$

$$\sum_i \left[\frac{\int \sum_t \frac{n_{it}\mathbf{x}'_{it}(k)}{e^{\mathbf{x}'_{it}\beta + \theta n_{it} + u}} (\overline{c}_{it} - e^{\mathbf{x}'_{it}\beta + \theta n_{it} + u}) \prod_t f_{\overline{C}|N}(\overline{c}_{it}|n_{it}) dF_u}{\int \prod_t f_{\overline{C}|N}(\overline{c}_{it}|n_{it}) dF_u} \right] = 0$$

$$\sum_i \left[\frac{\int \sum_t \frac{n_{it}^2}{e^{\mathbf{x}'_{it}\beta + \theta n_{it} + u}} (\overline{c}_{it} - e^{\mathbf{x}'_{it}\beta + \theta n_{it} + u}) \prod_t f_{\overline{C}|N}(\overline{c}_{it}|n_{it}) dF_u}{\int \prod_t f_{\overline{C}|N}(\overline{c}_{it}|n_{it}) dF_u} \right] = 0$$

$$\sum_i \left[\frac{\int (u^2 - \sigma_u^2) \prod_t f_{\overline{C}|N}(\overline{c}_{it}|n_{it}) dF_u}{\int \prod_t f_{\overline{C}|N}(\overline{c}_{it}|n_{it}) dF_u} \right] = 0$$

For the additional parameters r in the frequency model and ϕ in the average severity model, we find that there is no explicit form for the respective partial derivatives. We can estimate these parameters by using method of moments. Additional details of the derivation of the likelihood estimating equations developed above are in the appendix.

It is interesting to note that in this special case where we have negative binomial for frequency and gamma for average severity, we find that the unconditional mean is

$$\mathbb{E}\left[S|\mathbf{x}\right] = \mu\nu\mathbb{E}\left[[1-(\nu e^b/r)(e^\theta-1)]^{-r-1}\right] e^{\sigma_u^2/2+\theta} \tag{2.14}$$

and the unconditional variance is

$$\begin{aligned} Var(S|\mathbf{x}) \quad = \quad & \mu^2 e^{\sigma_u^2+2\theta} \times \Big(\phi\mathbb{E}\left[\nu e^b[1-(\nu e^b/r)(e^{2\theta}-1)]^{-r-1}\right] e^{\sigma_u^2} \\ & + \mathbb{E}\left[\nu^2 e^{2b+2\theta}(1+1/r)[1-(\nu e^b/r)(e^{2\theta}-1)]^{-r-2}\right] e^{\sigma_u^2} \\ & + \mathbb{E}\left[\nu e^b[1-(\nu e^b/r)(e^{2\theta}-1)]^{-r-1}\right] e^{\sigma_u^2} \\ & - \mathbb{E}\left[\nu e^b[1-(\nu e^b/r)(e^{\theta}-1)]^{-r-1}\right]^2 \Big), \end{aligned} \tag{2.15}$$

provided all the terms with expectation exist. Details of the derivation for equations (2.14) and (2.15) are provided in the appendix.

As a final remark, we broadly call the model constructed in this subsection to be the **dependent GLMM**. In the special case where we do not have random effects, that is, $\sigma_b = \sigma_u = 0$, we will call this the **dependent GLM** which is equivalent to that developed by Garrido et al. (2016). In the additional special case where $\theta = 0$, we have the **independent GLM**. In the validation section, we compared all these three models together with the **Tweedie GLM** and the **Tweedie GLMM**. The Tweedie models will be later described.

3 Data characteristics and preliminary investigation

For the empirical investigation in this article, we calibrated our proposed dependent GLMM frequency-severity model using the policy and claims experience data of a portfolio of automobile insurance policies from a general insurer in Singapore. The observed data is available for a period of six years, 1994–1999. This data has been obtained from General Insurance Association of Singapore, a trade association with representations from all the general insurance companies doing business in Singapore during the said 6-year period. This data or some extracted form of this data has been used in several articles including, but not limited to, Frees and Valdez (2008) and Shi and Valdez (2012).

Automobile insurance is one of the largest, if not the most important, lines of insurance offered by general insurers in the Singapore insurance market; its annual gross premium has historically been accounted for over a third of the entire insurance market. As with most other developed countries, auto insurance provides coverage at different layers, with the minimum layer protection which is mandatory, providing protection against death of bodily harm to third parties, regardless of who is at fault. In many countries, this is called third party liability coverage.

For estimation purpose, we used the policy and claims information for the first five years, 1994–1998, for which we call the training dataset. For validation, we used the observed information for year 1999. Our summary statistics in the rest of this section consists only of the training data. Here, we have a total number of $M = 41,831$ unique policyholders, each of which is observed T_i years and in essence, the maximum value of T_i is 5 years.

Table 1: Observable policy characteristics used as covariates

Categorical variables	Description		Proportions	
VehType	Type of insured vehicle:	Car	99.26%	
		MotorBike	0.54%	
		Others	0.2%	
Gender	Insured's sex:	Male = 1	80.74%	
		Female = 0	19.26%	
Cover Code	Type of insurance cover:	Comprehensive = 1	78.36%	
		Others = 0	21.64%	
Continuous variables		Minimum	Mean	Maximum
VehCapa	Insured vehicle's capacity in cc	10.00	1575.10	9996.00
VehAge	Age of vehicle in years	-1.00	6.65	46.00
Age	The policyholder's issue age	19.00	44.27	96.00
NCD	No Claim Discount in %	0.00	36.01	50.00

Observations in the portfolio consists of policies with comprehensive coverages for first party property damage and bodily injury as well as third party liability for property damage and bodily injury. We considered the total claims arising from all coverages.

From the portfolio, we have policy characteristics observed that we used as covariates in both the frequency and severity components and simple summary statistics for these covariates are provided in Table 1. We have a total of nine variables, including those classified as categorical or continuous and those transformed, in our policy file. Furthermore, it is typical for companies to classify automobile insurance policyholders according to both driver information and vehicle information; in our policy file, we have gender and issue age that relate to driver information, the rest are vehicle information. The table is indeed self-explanatory, but there are a few items worth noting about the insured portfolio used in this analysis. First, although in Singapore, motorbike is not uncommon, our dataset has predominantly car as the insured vehicle. Second, unlike in the United States, it is not uncommon to have fewer female drivers in the portfolio. Third, an examination of the age of vehicle measured in years indicate that the smallest number of age of vehicle insured is -1; it is possible to find vehicles purchased with models a year later. Also, the average age of vehicles is about 7 years

indicating there are generally newer and less than 10 years old vehicles insured. In Singapore, this is not at all surprising given the government restriction of allowing entitlement of vehicle ownership usually expiring in 10 years.

Table 2: Percentage and number of claims by count and year

	1994	1995	1996	1997	1998	Number	% of Total
0	92.2	92	91.7	91.9	90.2	98982	91.6
1	7.3	7.3	7.7	7.5	8.8	8288	7.7
2	0.5	0.6	0.6	0.6	0.9	700	0.6
3	0	0	0	0	0.1	42	0
4			0	0		4	0
5				0		1	0
Number	24690	22857	21437	20302	18731	108017	100

Table 2 provides a summary of the claim frequency distribution over the period from calendar year 1994–1998. For each year, this distribution shows that we have a significant percentage of zero claims. On the average, about 91.6% is the percentage of zero claims with this percentage ranging from 90.2% to 92.2%. This large percentage of zero claims is not uncommon for a portfolio of automobile insurance thereby allowing insurance companies the pooling of homogeneous risks through diversification. The aggregated total number of observations is 108,017 and when compared to the number of unique policyholders observed which is $M = 41,831$, this indicates that do have repeated observations for each unique policyholder for which the average number of years we observe each policyholder is a little more than 3 years. Moreover, notice that the number of policies observed varies each year, indicating the unbalanced nature of our data. For some years, we have policies that either lapse (leave the company) or newly join the company. In a calendar year, whenever a policyholder makes a claim, the most frequent number of times a policyholder claims is just once. We have very few policyholders with number of claims above 3 times in a calendar year.

Table 3: Average claim (AvgSev) by claim count and calendar year

	1994	1995	1996	1997	1998	AvgSev by count
1	4355	4028	4763	4219	4278	4329
2	8162	7545	8608	8325	7652	8025
3	6430	18564	12482	5018	11148	11395
4			9429	19363		16879
5				12149		12149
Annual AvgSev	4616	4378	5087	4558	4639	4655

Table 3 presents the average claim amount for different number of claims and for the different calendar years. The overall average claim amount is 4,655.09 and this varies within the range of 4,377.5 and 5,087.05. The more interesting observation we can deduce from this table is an examination of the last column: we see an increasing average amount of severity for increasing number of claims. For most calendar years, this possible positive correlation can also be observed. However, such a positive correlation cannot be deduced from Figure 1 below. This figure provides a further examination of the possible dependency between frequency and average severity. It is difficult to make a more conclusive relationship between frequency and average severity from this figure. However, this preliminary investigation of the data suggests that we can probably

make more conclusive statistical evidence of the relationship when we inject the relation of number of claims to average severity within either the GLM or GLMM framework.

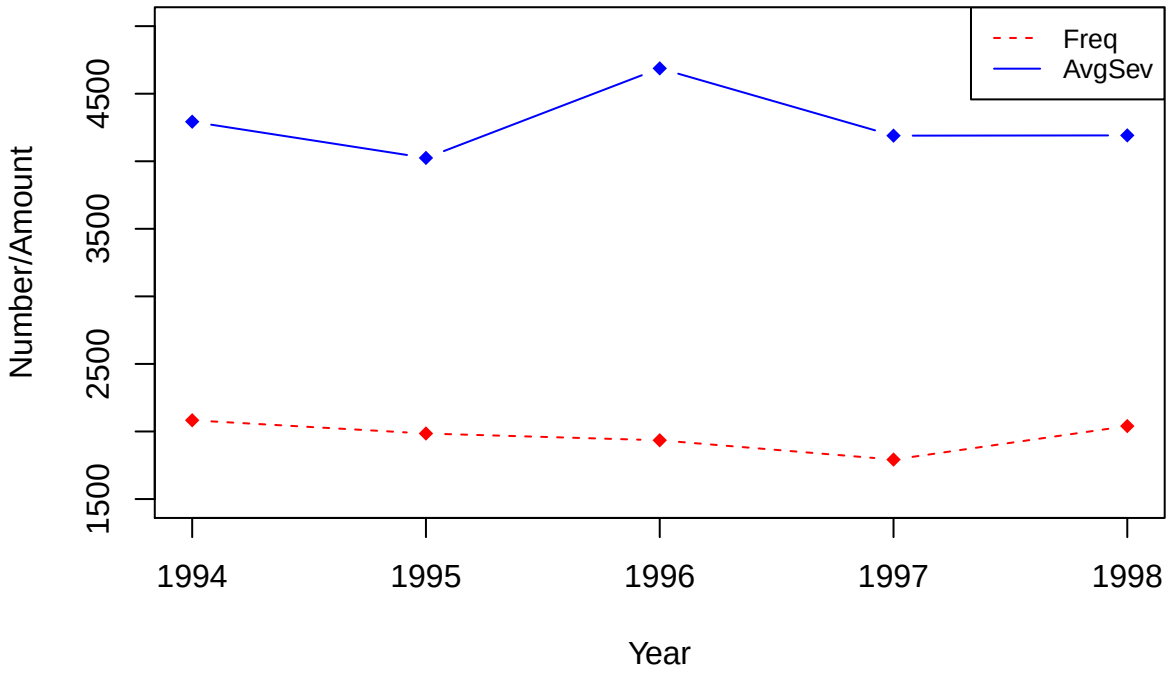

Figure 1: Frequency and average severity by calendar year

Figure 2 provides yet another preliminary observation about the possible relationship between frequency and average severity. In this figure, the x-axis is frequency and the y-axis is the average severity. Here we can see wider variation on the average severity for smaller number of claims. While this may indicate some evidence of possible relationship between frequency and average severity, it is difficult to draw a strong deduction of this relationship because it does not control for the effect of the heterogeneity of the policyholders.

In practice, the use of Poisson model for frequency is not uncommon. However, it is well known that the Poisson model cannot accommodate the large dispersion that is frequently observed with the number of claims. For our data, the overall average number of claims is 0.091 with a variance of 0.099. These values indicate a slight possibility of overdispersion. It is also widely accepted in the actuarial literature to use negative binomial to accommodate this possible overdispersion; see, for example, Frees and Valdez (2008).

In order to further motivate our choice of a negative binomial for frequency, we provide a goodness-of-fit test for the Poisson in comparison to the negative binomial. As shown in Table 4, the smaller test statistic observed for the negative binomial indicates that it is a better choice than the Poisson model.

To conclude this section on data, we provide the quality of the goodness of fit of choosing a gamma distribution model for average severity. Figure 3 provides log quantile-quantile (log-QQ) plots of fitting the gamma distribution for each calendar year for the period 1994–1998. We use the log-transformation here to summarize the quantiles because this helps us justify using a log-link function within the GLM framework later. It is generally well-known that the gamma model typically provides a good fit for lines of business with claims that have medium-sized tails. It is marginally clear from these log-QQ plots that we may have somewhat of a weak fit on both ends of the tail of the distribution for most calendar years. We wanted to restrict our

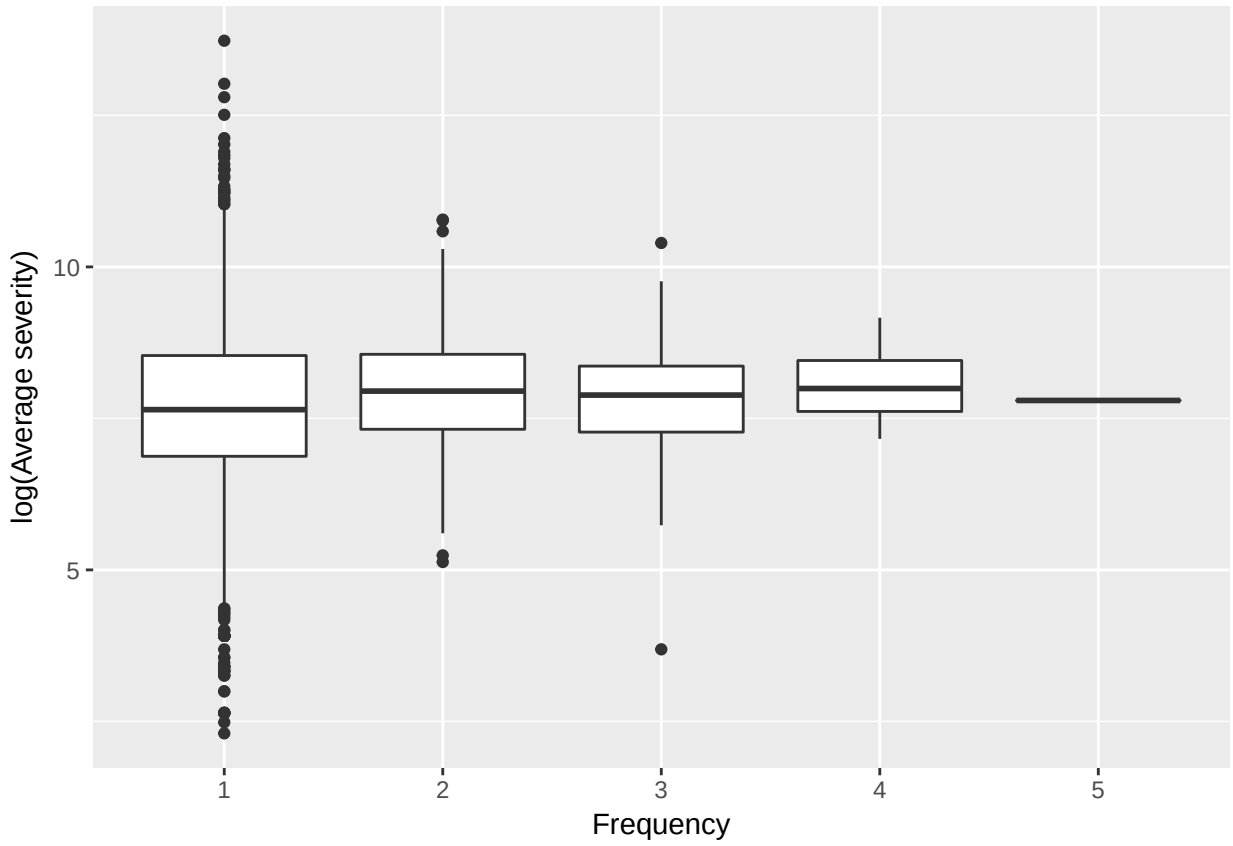

Figure 2: Graphical relationship of frequency and average severity, per policyholder

Table 4: Goodness-of-fit test for the frequency component

Count	Observed	Poisson	Negative Binomial
0	98982	98616.5	98978.9
1	8288	8979.1	8305.9
2	700	408.8	673.6
3	42	12.4	54
4	4	0.3	4.3
5	1	0	0.3
χ^2		67961.9	67674.6

marginal choice for the average severity to be within the class of exponential, or GLM, family. Furthermore, these plots are preliminary and do not consider the effect of policy characteristics for covariates or even the effect of the number of claims where we saw earlier that there may be some possible effect of varying average severity by frequency of claims. However, we feel that for future research, we need to address the quality of the fit on the tail of the severity distributions.

4 Results of estimation and model validation

This section provides details of the numerical results of calibrating the various models described in Section 2. Recall in that section we described the two-part frequency-severity model that is often used for fitting insurance claims data. The first component is the frequency model for which preliminary investigation of our data in Section 3 suggested the use of negative binomial. For the second component, we described the average severity model, given the number of claims observed. It was understood that the average severity is zero for zero number of claims. For positive number of claims, we described in Section 2 the use of a gamma distribution for average severity, one that falls within the class of GLM distributions. In addition, we provided some motivation for this choice in Section 3.

4.1 Numerical results

As earlier motivated and also widely known in the actuarial literature, the negative binomial model usually is a better choice for the frequency component than the Poisson model. The justifications in this case are the accommmodation of overdispersion that is typically present in insurance claims and the possible influence of unobserved random effects. Table 5 below summarizes the regression estimates for the negative binomial for both the GLM and the GLMM.

We now compare the two different negative binomial regression models for the frequency component. First, we note that the only difference between these two models is the presence of random effects in the GLMM. Second, we compare the regression coefficient estimates and their level of significance. Note that because historically we have known the non-linearity of age to claims, we added the square and cubic factors for age in our regression models. Broadly speaking, the effects of each of the explanatory variables summarized in Table 5 are not any much different between the two models. One can deduce that if random effect is added as in the GLMM, there is a slight decrease of the p-values of the corresponding coefficients. The signs of

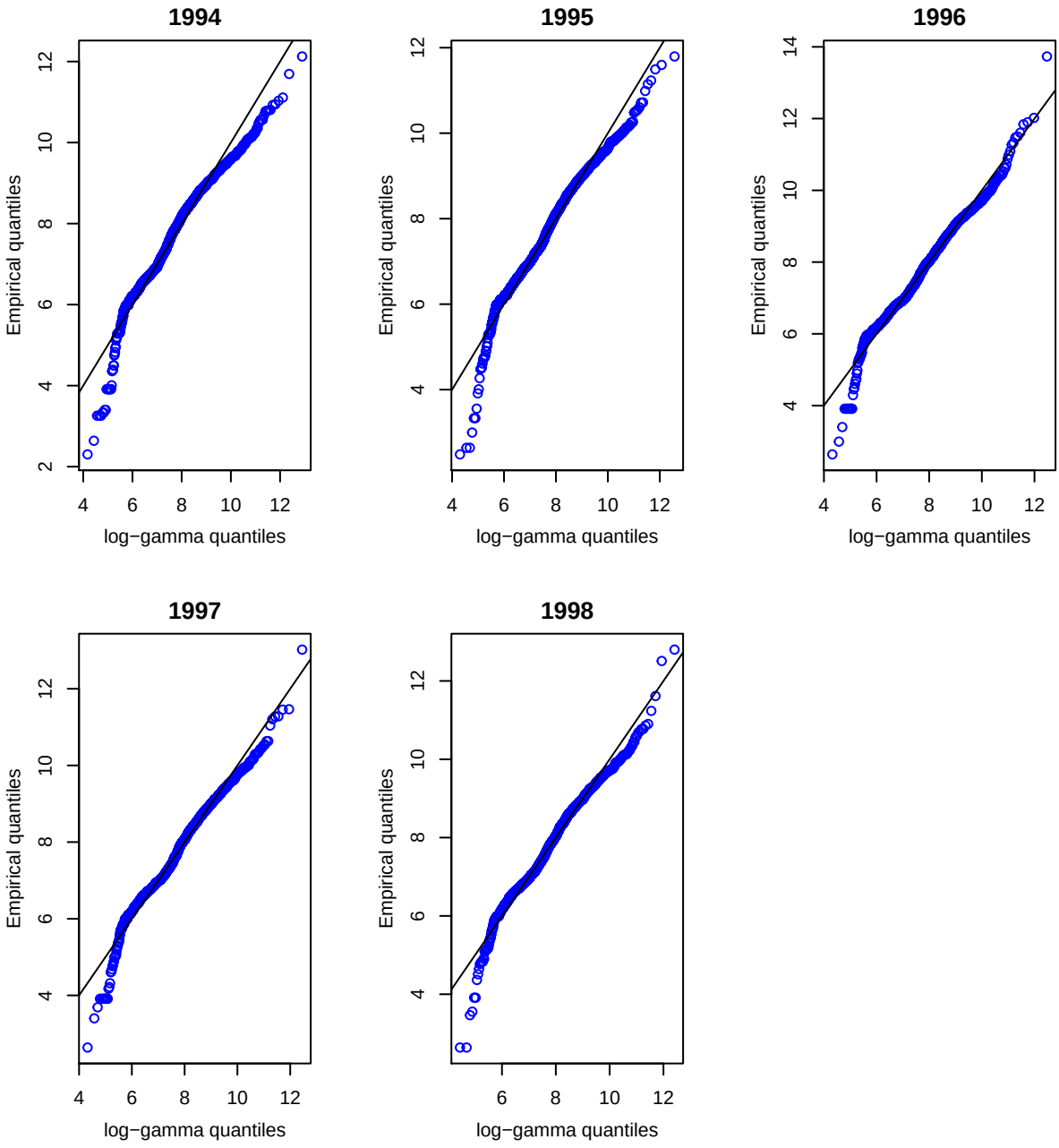

Figure 3: log-QQplots of fitting gamma to average severity for each calendar year

the coefficients are very consistent between the two models and appear to be intuitive. For example, when compared to cars, motorbikes are less likely to claim than cars and this might partly because cars are more frequently on the road. Older vehicles are less likely to claim and this may possibly be explained by fewer older vehicles insured. Male drivers are more likely to claim; historically, this has been intuitively explained by the more aggressive driving behavior of males than females. The effect of the policyholder's age does not seem to be significant in this portfolio. With respect to the effect of a premium discount, this is measured by the NCD score and the result is statistically significant and indeed quite intuitive. The negative sign indicates that policyholders with higher discounts are considered less risky driver and hence, tend to have lower likelihood of a claim.

Finally, we can examine the quality of the model fit by comparing the goodness-of-fit statistics that include the value of the log-likelihood, the AIC and BIC criteria. All three statistics are reported at the bottom of the table. Because we maximize the loglikelihood, we consider a greater value to be more superior. For either AIC or BIC, we consider a lower value to be more superior. An examination of all three model fit statistics suggests that the GLMM is slightly more superior model than the GLM.

We now turn to the average severity component, conditional on the event of at least one claim. There are three regression models we compared: the independent GLM, the dependent GLM and the dependent GLMM. Each of these was discussed in Section 2.

There are a few interesting observations we can draw from the regression estimates summarized in Table 6. First, for the most part, the regression estimates are consistent among all three models. To illustrate, consider the effect of vehicle type. The estimates suggest that claim amounts are larger for motorbikes than for any other types of vehicle. Although we found out in the frequency component, that motorbikes are less likely to claim, the result here suggests though that when motorbike policyholders claim, the severity impact is worse. This appears to be consistent with studies that suggest that the risk of fatal crash among motorbike drivers is far greater than for passenger cars. Moreover, these same studies suggest that the seriousness of the injuries are worse than those of passenger cars. Furthermore, not only do younger drivers have a higher likelihood of a claim, but the severity of claims is even far worse for younger drivers. Generally, younger drivers exhibit more aggressive behavior when it comes to driving and they have poorer judgment when intoxicated. Second, there are also some clear differences. For example, in the dependent GLMM, the regression estimates for the presence of a comprehensive coverage has a positive sign and the corresponding p-value suggests the significance of this variable. In each of the other two models, the effect of the presence of a comprehensive coverage is not statistically significant. The more interesting difference has to do with the coefficient estimate associated with 'count' between the dependent GLM and dependent GLMM. While both models provide significant evidence of dependence, the signs are opposite which suggests there is inconsistency of the possible relationship between frequency and claims. However, this inconsistency in sign is attributable to the presence of the random effects in the dependent GLMM. Controlling for the presence of the random effects account for the resulting differences in means among possible group of policyholders implied by the random effects. This is a favorable argument for using a dependent GLMM.

Finally, we examine the quality of the model fit by comparing the goodness-of-fit statistics that include the value of the log-likelihood, the AIC and BIC criteria. All three statistics are reported at the bottom of the table. An examination of all three model fit statistics suggests that the GLMM is far more superior model than the other two models.

Table 5: Regression estimates of the negative binomial model for frequency

	Negative binomial GLM			Negative binomial GLMM		
	Estimate	s.e.	Pr(>|t|)	Estimate	s.e.	Pr(>|t|)
(Intercept)	-4.2484	0.6098	0.0000	-4.2053	0.6217	0.0000
VTypeOthers	-0.4249	0.3076	0.1671	-0.4127	0.3081	0.1804
VTypeMBike	-1.3131	0.4575	0.0041	-1.2879	0.4591	0.0050
log(VehCapa)	0.3291	0.0442	0.0000	0.3309	0.0452	0.0000
VehAge	-0.0238	0.0038	0.0000	-0.0227	0.0038	0.0000
SexM	0.0966	0.0269	0.0003	0.0947	0.0274	0.0006
Comp	0.7252	0.0493	0.0000	0.7323	0.0496	0.0000
Age	-0.0232	0.0357	0.5164	-0.0281	0.0364	0.4400
Age ^{2}	0.0004	0.0008	0.6010	0.0005	0.0008	0.5373
Age ^{3}	0.0000	0.0000	0.6474	0.0000	0.0000	0.5961
NCD	-0.0120	0.0006	0.0000	-0.0114	0.0006	0.0000
σ_b^2				0.1992	0.4560	0.6484
Loglikelihood		-32345.3445			-32319.1436	
AIC		64714.6890			64664.2872	
BIC		64799.6805			64788.9577	

Table 6: Regression estimates of the gamma model for average severity

	Independent GLM			Dependent GLM			Dependent GLMM		
	Estimate	s.e.	Pr(>|t|)	Estimate	s.e.	Pr(>|t|)	Estimate	s.e.	Pr(>|t|)
(Intercept)	6.5138	1.4650	0.0000	6.6014	1.4472	0.0000	5.7330	0.7917	0.0000
VTypeOthers	0.9705	0.7135	0.1738	1.0329	0.7047	0.1428	-0.0040	0.3403	0.9905
VTypeMBike	3.3343	1.1948	0.0053	3.3210	1.1798	0.0049	2.1910	0.6041	0.0003
log(VehCapa)	0.5928	0.1042	0.0000	0.5911	0.1029	0.0000	0.3361	0.0573	0.0000
VehAge	-0.0285	0.0090	0.0015	-0.0289	0.0089	0.0011	-0.0153	0.0045	0.0008
SexM	0.0047	0.0621	0.9400	0.0014	0.0613	0.9823	-0.0404	0.0342	0.2372
Comp	0.0597	0.1155	0.6052	0.0653	0.1141	0.5672	0.1908	0.0557	0.0006
Age	-0.1540	0.0860	0.0733	-0.1514	0.0849	0.0746	-0.0323	0.0466	0.4892
Age ^{2}	0.0032	0.0018	0.0830	0.0031	0.0018	0.0830	0.0007	0.0010	0.4677
Age ^{3}	0.0000	0.0000	0.1202	0.0000	0.0000	0.1187	0.0000	0.0000	0.5082
NCD	-0.0050	0.0014	0.0002	-0.0052	0.0013	0.0001	-0.0043	0.0007	0.0000
Count				-0.1037	0.0537	0.0534	0.0668	0.0295	0.0235
σ_u^2							1.0814	0.4749	0.0228
Loglikelihood		-89874.3089			-89865.2618			-86071.9841	
AIC		179772.6178			179756.5235			172171.9683	
BIC		179857.6092			179848.5976			172271.1249	

In practice and in the actuarial literature, there is an increase in the interest of the use of Tweedie exponential dispersion family model to fit compound loss models. See, for example, Frees et al. (2016b). The Tweedie family of distributions belong to the exponential family with a variance function of the power form as $V(\mu) = \tau\mu^p$, for p not in $(0, 1)$. Special cases include the normal distribution when $p = 0$, the Poisson distribution when $p = 1$, and the gamma distribution when $p = 2$. However, when $1 < p < 2$, the Tweedie distribution can be derived as a compound Poisson-gamma distribution with a probability mass at zero. Although in this case, there is no explicit expression for the density function, the primary advantage of fitting such Tweedie models is to fit both the claim frequency and the claim severity simultaneously. It is interesting to note that Tweedie models have been applied in biostatistics and climatology.

Because of the increasing interest of this family of distributions for fitting loss models, we consider here the Tweedie regression models based on both the GLM, with only fixed effects, and the GLMM, with the addition of random effects. These Tweedie models explicitly ignore the possible dependence between frequency and severity. The results are summarized below in Table 7 where according to the goodness-of-fit statistics, the Tweedie GLM outperforms the Tweedie GLMM. In the next section, we will compare the performance of all the models described in this section.

Table 7: Regression estimates for the aggregate loss models based on Tweedie

	Tweedie GLM			Tweedie GLMM		
	Estimate	s.e.	Pr(>|t|)	Estimate	s.e.	Pr(>|t|)
(Intercept)	3.0667	1.6245	0.0591	3.0370	9.4923	0.7490
VTypeOthers	-0.1227	0.6909	0.8590	-0.1594	1.5039	0.9156
VTypeMBike	1.6520	0.5073	0.0011	1.6597	3.6119	0.6459
log(VehCapa)	0.8641	0.1163	0.0000	0.8713	0.8230	0.2898
VehAge	-0.0555	0.0099	0.0000	-0.0556	0.0350	0.1122
SexM	0.1186	0.0731	0.1047	0.1179	0.5238	0.8219
Comp	0.9371	0.1261	0.0000	0.9344	0.3059	0.0023
Age	-0.2044	0.0956	0.0326	-0.2059	0.5157	0.6897
Age ^{2}	0.0040	0.0020	0.0465	0.0040	0.0109	0.7115
Age ^{3}	0.0000	0.0000	0.0753	0.0000	0.0001	0.7424
NCD	-0.0166	0.0016	0.0000	-0.0166	0.0046	0.0003
σ_R^2				1632.6000	98.5019	0.0000
Loglikelihood		-112632.5098			-165017.4487	
AIC		225289.0196			330060.8975	
BIC		225374.0110			330185.5680	

4.2 Validation results

As we earlier alluded, we made the observations for year 1999 as our hold-out sample that we now use for model validation. The process of this model validation is to predict both the frequency (or count) and the average severity (or claim amount) to derive the total claim amount and compare the predicted with the actual (or observed) total claim amount for the policies in year 1999.

For our purposes, denote the actual value of the total claim amount by s_i and the corresponding predicted value, for the various models, by $\widehat{s}_i$. Here, $s_i = \sum_{k=1}^{N_i} C_{ik}$ for policyholder i in year 1999. Two validation measures were used to compare the models:

$$\text{Mean Squared Error:} \quad MSE = \frac{1}{M}\sum_{i=1}^{M}(\widehat{s}_i - s_i)^2$$

$$\text{Mean Absolute Error:} \quad MAE = \frac{1}{M}\sum_{i=1}^{M}\left|\widehat{s}_i - s_i\right|$$

Between two models, we would normally conclude that the one that produces: (1) a lower MSE is better; and (2) a lower MAE is better. These two validation measures provide the prediction accuracy of the various models. The difference between these two measures has to do with the norm used. Larger deviations relatively

have a greater effect on the MSE than smaller deviations. When compared to the MSE, the MAE is relatively less sensitive to large deviations. As such, the MSE is much more sensitive to observations that are considered outliers.

Note that we do have a total of five models being compared: the two-part independent GLM, the two-part dependent GLM, the two-part dependent GLMM, the Tweedie GLM and the Tweedie GLMM. Table 8 compares these validation measures produced by the two-part models and the Tweedie models, respectively. The values summarized in the table are self-explanatory.

For instance, the dependent GLMM outperforms the other models in terms of both validation measures for the prediction of total loss. This provides a strong argument for choosing the dependent GLMM.

Table 8: Validation measures for the five models

	MSE	MAE
Independent GLM	5957.589	774.9298
Dependent GLM	5962.351	815.6192
Dependent GLMM	5957.131	733.3369
Tweedie GLM	5957.777	767.0962
Tweedie GLMM	6226.111	782.1025

Besides the use of MSE and MAE, to simultaneously make a valid comparison among these five models, we applied the idea of Lorenz curve and calculated the Gini index for each model. The details of drawing the Lorenz curve and eventually computing the Gini index can be found in Chapter 2 of Frees et al. (2016a). The Gini index has also been used as model validation measure in Frees et al. (2016b).

Within the context of this paper, we proceed with the following three-step process to draw the Lorenz curve. We repeat this procedure for each model using the hold-out sample. From this Lorenz curve, we compute the corresponding Gini index.

1. From our hold-out sample, for $i = 1, 2, \ldots, M$, sort the observed loss Y_i according to the risk score S_i for which in our case, the predicted value from each model, in an ascending manner. That is, calculate the rank R_i of S_i between 1 and M with $R_1 = \text{argmin}(S_i)$.
2. Compute $F_{\text{score}}(m/M) = \frac{1}{M}\sum_{i=1}^{M} 1_{(R_i \leq m)}$, the cumulative percentage of exposures, and $F_{\text{loss}(m/M)} = \frac{\sum_{i=1}^{M} Y_i 1_{(R_i \leq m)}}{\sum_{i=1}^{M} Y_i}$, the cumulative percentage of loss, for each $m = 1, 2, \ldots, M$.
3. Plot $F_{\text{score}}(m/M)$ on the x-axis, and $F_{\text{loss}}(m/M)$ on the y-axis.

The predicted value as defined above is our pure premium according to the model being considered. The Gini index is equal to $2\times$ the area between the line of equality and the Lorenz curve drawn above. We then choose the model with the highest Gini index. According to Figure 4 below, the dependent GLMM is the best among all 5 models with a Gini index equal to 45.7%.

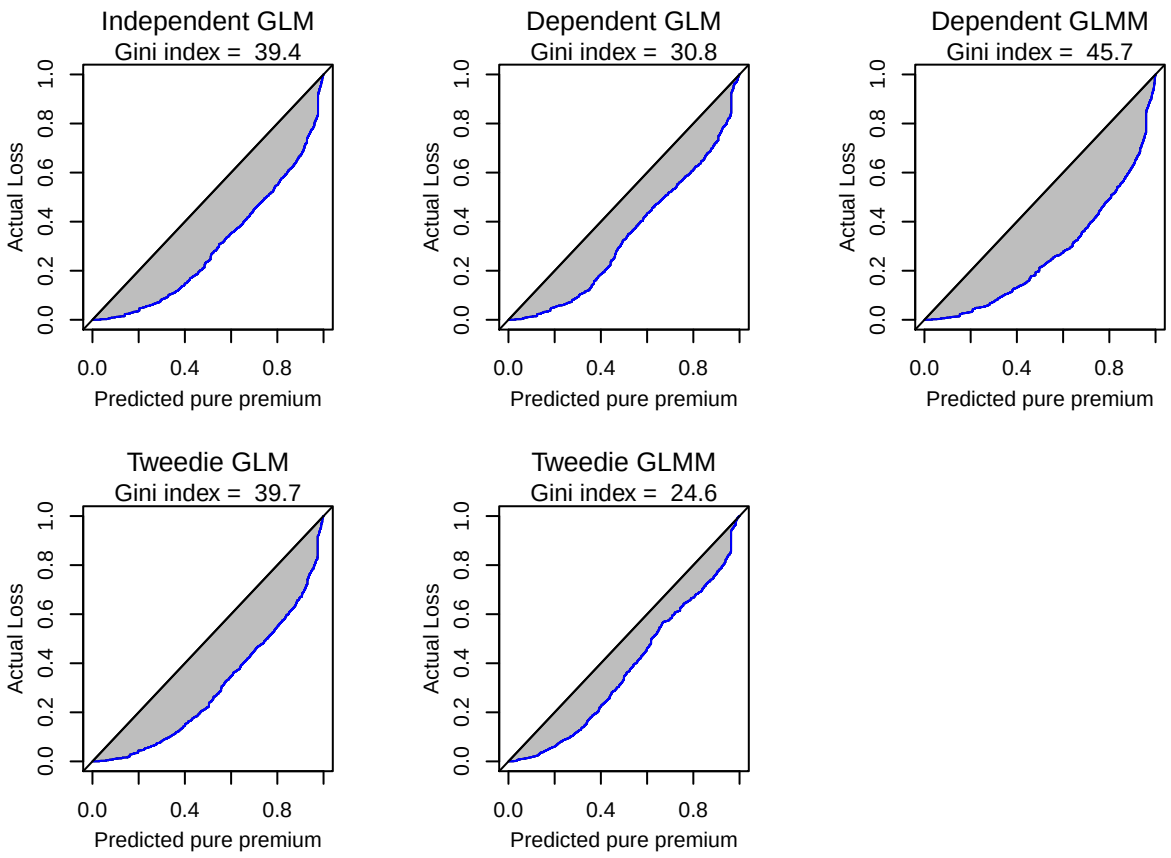

Figure 4: The Lorenz curve and the Gini index values for the five models

5 Concluding remarks

In this paper, we offered the use of the generalized linear mixed models (GLMMs) as alternative models with dependent claim frequency and claim severity. The concept is a natural extension of the dependent frequency-severity based on GLM introduced by Garrido et al. (2016). With dependent GLMM, we similarly introduced dependence by inducing frequency as an explanatory variable for the average severity but added random effects typically used to capture dependence because of the repeated observations. The GLMM is a natural extension to GLM when claims are observed for policyholders over a period of time. The GLMM has the advantage of being able to control for the random effects of groupings of policyholders, especially in terms of degrees of riskiness. This model is called the dependent GLMM throughout the paper.

To calibrate our model, we used the policy and claims experience data of a portfolio of automobile insurance policies from a general insurer in Singapore. The observed data is available for a period of six years, 1994–1999. We used the first 5 years of experience data as the training set for model estimation, and using the final year, 1999, as the hold-out sample for validation. To provide strong argument of the superiority of our proposed model, we considered four other models for which we called: the independent GLM, the dependent GLM (Garrido et al. (2016)), the Tweedie GLM, and the Tweedie GLMM. In most measures of model comparison, model validation and even the Gini index, the dependent GLMM outperforms all the four other models.

For future research, we need to examine a better model for severity. As indicated by our preliminary investigation, the gamma model did not do well to capture the long tailness of the loss distribution. For most insurance lines of business, this is usually the case and hence, we need further investigation. This paper is a very good start to justify the use of random effects within the class of generalized linear models.

Appendix A. The development of the log-likelihood equations

Based on our observed data, from equations (2.14) and (2.6), the likelihood can be expressed as

$$L = \prod_i \int \int \prod_t f(n_{it}, \overline{c}_{it}|b,u) dF_b dF_u = \prod_i \left(\int \prod_t f_N(n_{it}|b) dF_b \right) \left(\int \prod_t f_{\overline{C}|N}(\overline{c}_{it}|u, n_{it}) dF_u \right)$$

Thus, the log-likelihood can be expressed as

$$\begin{aligned} \ell &= \log L = \sum_i \left(\log \int \prod_t f_N(n_{it}|b) dF_b + \log \int \prod_t f_{\overline{C}|N}(\overline{c}_{it}|u, n_{it}) dF_u \right) \\ &= \sum_i \left(\log \int \prod_t f_N(n_{it}|b) dF_b \right) + \sum_i \left(\log \int \prod_t f_{\overline{C}|N}(\overline{c}_{it}|u, n_{it}) dF_u \right) \end{aligned}$$

and we can decompose this log-likelihood into $\ell = \ell_N + \ell_{\overline{C}|N}$ where

$$\ell_N = \sum_i \left(\log \int \prod_t f_N(n_{it}|b) dF_b \right)$$

and

$$\ell_{\overline{C}|N} = \sum_i \left(\log \int \prod_t f_{\overline{C}|N}(\overline{c}_{it}|u, n_{it}) dF_u \right)$$

We take the partial derivatives of the log-likelihood functions and set to zero:

$$\frac{\partial \ell_N}{\partial \alpha} = 0 \text{ for } k = 1, \ldots, p$$

$$\frac{\partial \ell_N}{\partial \sigma_b} = 0$$

$$\frac{\partial \ell_N}{\partial \theta} = 0$$

$$\frac{\partial \ell_N}{\partial \beta} = 0 \text{ for } k = 1, \ldots, p$$

$$\frac{\partial \ell_N}{\partial \sigma_u} = 0$$

The results yield to the $(2p + 3)$ estimating equations expressedly written in Section 2.

Appendix B. Details of the computation of the mean and variance of the compound sum

In this appendix, we provide the details of the derivation for the expression of the unconditional mean and variance of the aggregate claim as defined by $S = N\overline{C}$ according to our GLMM specification. For simplicity, here we drop all the subscripts. Using the notation that is conventional for the GLM framework, we define ν and μ so that $\nu = g_N^{-1}(\mathbf{x}'\alpha + z'b) = \mathbb{E}[N|\mathbf{x}]$ and $\mu = g_C^{-1}(\mathbf{x}'\beta + \theta n + z'u) = \mathbb{E}\left[\overline{C}|N, \mathbf{x}\right]$, using the link function $g_N(\cdot)$ for the frequency and $g_C(\cdot)$ for the average severity, respectively.

Therefore, in general, we can derive explicit formulas for the unconditional mean and variance of the aggregate claims as follows:

$$\begin{aligned}\mathbb{E}[S|\mathbf{x}] &= \mathbb{E}\left[N\overline{C}|\mathbf{x}\right] = \mathbb{E}\left[N\mathbb{E}\left[\overline{C}|N,u\right]|\mathbf{x}\right] \\ &= \mathbb{E}\left[Ng_C^{-1}(\mathbf{x}'\beta + \theta n + z'u)|\mathbf{x}\right]\end{aligned} \tag{5.1}$$

and

$$\begin{aligned}Var(S|\mathbf{x}) &= Var(\mathbb{E}\left[N\overline{C}|N,\mathbf{x}\right]) + \mathbb{E}\left[Var(N\overline{C}|N,\mathbf{x})\right] \\ &= Var(N\mathbb{E}\left[\overline{C}|N,\mathbf{x}\right]|\mathbf{x}) + \mathbb{E}\left[N^2Var(\overline{C}|N,\mathbf{x})|\mathbf{x}\right] \\ &= Var(Ng_C^{-1}(\mathbf{x}'\beta + \theta n + z'u)|\mathbf{x}) + \mathbb{E}\left[N^2\tau V(g_C^{-1}(\mathbf{x}'\beta + \theta n + z'u))|\mathbf{x}\right]\end{aligned} \tag{5.2}$$

Note that to simplify this, we can derive an expression for the unconditional mean and variance with our two-part depedent frequency severity GLMM:

$$N|b \sim \text{ indep. NB}(\nu e^b, r) \text{ with } b \sim N(0, \sigma_b^2)$$

and

$$\overline{C}|N,u \sim \text{ indep. gamma}(\mu e^u, \phi/n) \text{ with } u \sim N(0, \sigma_u^2)$$

We additionally assume log-link functions $g_N(\mu) = g_C(\mu) = \log\mu$ and that $z = 1$. For average severity, conditional on N, model specified above, we added θn in the linear predictor. Thus, we have $\nu = \exp(\mathbf{x}'\alpha)$ and $\mu = \exp(\mathbf{x}'\beta)$.

For the unconditional mean, we have

$$\begin{aligned}
\mathbb{E}\left[S|\mathbf{x}\right] &= \mathbb{E}\left[N\overline{C}|\mathbf{x}\right] = \mathbb{E}\left[N\mathbb{E}\left[\overline{C}|N,u\right]|\mathbf{x}\right] \\
&= \mathbb{E}\left[N\mu e^{n\theta+u}|\mathbf{x}\right] = \mu\mathbb{E}\left[Ne^{n\theta}|\mathbf{x}\right]\mathbb{E}\left[e^{u}|\mathbf{x}\right] \\
&= \mu\mathbb{E}\left[M'_{N|b,\mathbf{x}}(\theta)\right]e^{\sigma_u^2/2} \\
&= \mu\mathbb{E}\left[\nu[1-(\nu e^b/r)(e^\theta-1)]^{-r-1}\right]e^{\sigma_u^2/2+\theta} \\
&= \mu\nu\mathbb{E}\left[[1-(\nu e^b/r)(e^\theta-1)]^{-r-1}\right]e^{\sigma_u^2/2+\theta}
\end{aligned} \tag{5.3}$$

where we have used the following results which can be immediately deduced: $M_{N|b,\mathbf{x}}(t) = [1-(\nu e^b/r)(e^t-1)]^{-r}$ and $\mathbb{E}\left[Ne^{nt}|b,\mathbf{x}\right] = M'_{N|b,\mathbf{x}}(t) = \nu e^{b+t}[1-(\nu e^b/r)(e^t-1)]^{-r-1}$.

Note that the expectation in the final line above is with respect to the random effect b. If we therefore set $b=0$ and $u=0$, this leads us to the dependent GLM without random effects. In this case, we have $\mathbb{E}\left[S|\mathbf{x}\right] = \mu\mathbb{E}\left[M'_{N|\mathbf{x}}(\theta)\right]$, which gives us precisely that found in Garrido et al. (2016).

To derive the unconditional variance, we first note that $\mathbb{E}\left[S^2|\mathbf{x}\right] = \mathbb{E}\left[N^2\overline{C}^2|\mathbf{x}\right] = \mathbb{E}\left[N^2\mathbb{E}\left[\overline{C}^2|N,u\right]|\mathbf{x}\right]$ and because $\mathbb{E}\left[\overline{C}|N,u,\mathbf{x}\right] = \mu e^{n\theta+u}$ and $Var(\overline{C}|N,u,\mathbf{x}) = \phi\mu^2 e^{2(n\theta+u)}/n$, we can get

$$\mathbb{E}\left[\overline{C}^2|N,u,\mathbf{x}\right] = (\phi/n+1)\mu^2 e^{2(n\theta+u)},$$

$$\mathbb{E}\left[N^2\mathbb{E}\left[\overline{C}^2|N,u\right]|\mathbf{x}\right] = \mu^2\mathbb{E}\left[(\phi n+n^2)e^{2(n\theta+u)}|\mathbf{x}\right] = \mu^2(\phi\mathbb{E}\left[M'_{N|b,\mathbf{x}}(2\theta)\right] + \frac{1}{4}\mathbb{E}\left[M''_{N|b,\mathbf{x}}(2\theta)\right])e^{2\sigma_u^2}$$

and

$$\mathbb{E}\left[N^2e^{nt}|b,\mathbf{x}\right] = M''_{N|b,\mathbf{x}}(t) = \nu^2e^{2b+2t}(1+1/r)[1-(\nu e^b/r)(e^t-1)]^{-r-2} + M'_{N|b,\mathbf{x}}(t).$$

Finally, combining the expressions for $\mathbb{E}\left[S|\mathbf{x}\right]$ and $\mathbb{E}\left[S^2|\mathbf{x}\right]$, we have

$$\begin{aligned}
Var(S|\mathbf{x}) &= \mathbb{E}\left[S^2|\mathbf{x}\right] - \left(\mathbb{E}\left[S|\mathbf{x}\right]\right)^2 \\
&= \mu^2e^{\sigma_u^2}(\phi\mathbb{E}\left[M'_{N|b,\mathbf{x}}(2\theta)\right]e^{\sigma_u^2} + \frac{1}{4}\mathbb{E}\left[M''_{N|b,\mathbf{x}}(2\theta)\right]e^{\sigma_u^2} - \mathbb{E}\left[M'_{N|b,\mathbf{x}}(\theta)\right]^2) \\
&= \mu^2e^{\sigma_u^2+2\theta}(\phi\mathbb{E}\left[\nu e^b[1-(\nu e^b/r)(e^{2\theta}-1)]^{-r-1}\right]e^{\sigma_u^2} + \\
&\quad \mathbb{E}\left[\nu^2e^{2b+2\theta}(1+1/r)[1-(\nu e^b/r)(e^{2\theta}-1)]^{-r-2}\right]e^{\sigma_u^2} + \\
&\quad \mathbb{E}\left[\nu e^b[1-(\nu e^b/r)(e^{2\theta}-1)]^{-r-1}\right]e^{\sigma_u^2} - \\
&\quad \mathbb{E}\left[\nu e^b[1-(\nu e^b/r)(e^{\theta}-1)]^{-r-1}\right]^2)
\end{aligned} \tag{5.4}$$

Note that if we again set $b=0$ and $u=0$, we have the dependent GLM without random effects. In this case, we have

$$Var(S|\mathbf{x}) = \phi\mathbb{E}\left[NV_{C|\mathbf{x}}(\mu e^{\theta N})\right] + \mu^2\left\{\frac{1}{4}\mathbb{E}\left[M''_{N|\mathbf{x}}(2\theta)\right] - \mathbb{E}\left[M'_{N|\mathbf{x}}(\theta)\right]^2\right\},$$

which clearly corresponds to the one derived in Garrido et al. (2016).

It is also worth noting that in addition to removing the random effects, if we set $\theta = 0$, we end up with the unconditional mean and variance that corresponds to the case when frequency and average severity are independent.

References

Antonio, K. and Beirlant, J. (2007). Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics*, 40(1):58 – 76.

Antonio, K. and Valdez, E. A. (2011). Statistical concepts of a priori and a posteriori risk classification in insurance. *Advances in Statistical Analysis*, 96:187 – 224.

Boucher, J.-P., Denuit, M., and Guillén, M. (2008). Models of insurance claim counts with time dependence based on generalization of poisson and negative binomial distributions. *Variance*, 2(1):135–162.

Czado, C., Kastenmeier, R., Brechmann, E. C., and Min, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, 2012(4):278–305.

Frees, E. W., Derrig, R. A., and Meyers, G. (2014). *Predictive Modeling Applications in Actuarial Science, Volume 1: Predictive Modeling Techniques*. Cambridge University Press: Cambridge, U.K.

Frees, E. W., Derrig, R. A., and Meyers, G. (2016a). *Predictive Modeling Applications in Actuarial Science, Volume 2: Case Studies in Insurance*. Cambridge University Press, Cambridge, U.K.

Frees, E. W., Gao, J., and Rosenberg, M. (2011). Predicting the frequency and amount of health care expenditures. *North American Actuarial Journal*, 15(3):377 – 392.

Frees, E. W., Lee, G., and Yang, L. (2016b). Multivariate frequency-severity regression models in insurance. *Risks*, 4(4):1 – 36.

Frees, E. W. and Valdez, E. A. (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association*, 103(484):1457 – 1469.

Frees, E. W., Young, V., and Luo, Y. (1999). A longitudinal data analysis interpretation of credibility models. *Insurance: Mathematics and Economics*, 24(3):229 – 247.

Frees, E. W., Young, V., and Luo, Y. (2001). Case studies using panel data models. *North American Actuarial Journal*, 5(4):24 – 42.

Garrido, J., Genest, C., and Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, 70:205 – 215.

Garrido, J. and Zhou, J. (2009). Full credibility with generalized linear mixed models. *ASTIN Bulletin*, 39(1):61 – 80.

Lee, W., Park, S., and Ahn, J. (2016). Investigating dependence between frequency and severity via simple generalized linear models. working paper.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. John Wiley & Sons, Inc.: New York.

Molenberghs, G. and Verbeke, G. (2005). Longitudinal and incomplete clinical studies. *Metron - International Journal of Statistics*, 63(2):143–176.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370 – 384.

Park, S., Kim, J., and Ahn, J. (2017). Does hunger for bonus drive the dependence between claim frequency and severity? working paper.

Shi, P., Feng, X., and Ivantsova, A. (2015). Dependent frequency-severity modeling of insurance claims. *Insurance: Mathematics and Economics*, 64:417 – 428.

Shi, P. and Valdez, E. A. (2012). Longitudinal modeling of insurance claim counts using jitters. *Scandinavian Actuarial Journal*, 2012:1 – 21.