Jeong, Himchan, Emiliano A. Valdez, Jae Youn Ahn, and Sojung Carol Park. 2021. "Generalized Linear Mixed Models for Dependent Compound Risk Models." *Variance* 14 (1).

Ratemaking and Product Information

Generalized Linear Mixed Models for Dependent Compound Risk Models

Himchan Jeong[1a], Emiliano A. Valdez[2b], Jae Youn Ahn[3c], Sojung Carol Park[4d]

[1] Department of Statistics and Actuarial Science, Simon Fraser University, [2] Department of Mathematics, University of Connecticut, [3] Department of Statistics, Ewha Womans University, [4] College of Business Administration and Institute for Industrial Systems Innovation, Seoul National University

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

Variance

Vol. 14, Issue 1, 2021

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 has been sporadic interest in the actuarial literature exploring models that depart from this independence. In this paper, the authors extend the work of Garrido, Genest, and Schulz (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, they quantify and explain the contribution of the variability of claims among policyholders through the use of random effects using generalized linear mixed models (GLMMs). The authors calibrated their model using a portfolio of auto insurance contracts from a Singapore insurer where they observed claim counts and amounts from policyholders for a period of six years. They compared their results with the dependent GLM considered by Garrido, Genest, and Schulz; Tweedie models; and the case of independence. The dependent GLMM shows statistical evidence of positive dependence between frequency and severity. Using validation procedures, the authors find that the results demonstrate a superior model when random effects are considered within a GLMM framework.

Himchan Jeong and Emiliano A. Valdez were supported by the CAE Research Grant on Applying Data Mining Techniques in Actuarial Science funded by the Society of Actuaries (SOA).

Sojung Carol Park acknowledges support from the Institute of Management Research at Seoul National University.

Address for Correspondence: himchan_jeong@sfu.ca

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, Derrig, and Meyers (2014); and Frees, Derrig, and Meyers (2016). Originally synthesized by Nelder

a Himchan Jeong is a Fellow of the Society of Actuaries (SOA) and holds a Ph.D. from the University of Connecticut. He has been actively involved in teaching and conducting research in actuarial science for several years. In recognition for his academic achievements and excellence, he has been awarded the James C. Hickman Scholarship from SOA recently in 2018-2020. His current research interest is predictive modeling for ratemaking and reserving of property and casualty insurance.

b Emiliano A. Valdez is a Fellow of the Society of Actuaries and has a Ph.D. from the University of Wisconsin in Madison. He joined the University of Connecticut in 2007 but had a brief stint (2013-2015) as professor and director of the actuarial science program at Michigan State University. In recent years, his research work has evolved around the applications of data science and statistical modeling to solve actuarial and insurance problems. The quality of his research has been recognized through awards that include the E. A. Lew Award, the Halmstad Memorial Prize, and the Hachemeister Prize.

c Jae Youn Ahn is an Associate Professor of the Statistics Department. He received his Ph.D. in Statistics from the University of Iowa. His research interests are in predictive analysis in insurance.

d Sojung Carol Park is a Professor of the Finance at College of Business Administration, Seoul National University. She received her Ph.D. in Insurance and Risk Management from the University of Pennsylvania. Her research interests are in empirical analysis on insurance market, financial consumer behavior, and Fintech.

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, Derrig, and Meyers (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. GLM comprises a wide variety of models including the normal regression, Poisson regression, logistic regression, and probit regression, to name a few. More important, it has been used by Garrido, Genest, and Schulz (2016) to model the dependence between loss frequency and loss severity.

Following the formulation in Garrido, Genest, and Schulz (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 amounts of claims are denoted by C_1, C_2, ..., C_N. The total loss incurred can be written as the sum

$$S = C_1 + C_2 + \ldots + C_N,$$

with the convention that when $N = 0$, the total loss S is also 0. 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 a GLM that incorporates the effect of the various covariates. However, individual claims C_1, C_2, ..., C_N are assumed to be independent with each C_i to have a reproductive exponential dispersion family 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 the exponential dispersion family, 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

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

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, Genest, and Schulz (2016) that the variance of aggregate loss can be expressed as

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

In a simulation study in Garrido, Genest, and Schulz (2016), the presence of dependence can have serious biases on the value of the regression coefficients, their standard errors, and the mean and variance of the portfolio's aggregate loss. Using a real data set, the authors' work also found strong evidence of the presence of negative dependence between loss frequency and loss severity. This is an interesting result because it indicates that an increasing number of claims decreases their severity. An interesting observation in the paper is 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" (Garrido, Genest, and Schulz 2016, 205).

The dependent GLM in Garrido, Genest, and Schulz (2016) has a structure similar to that of the two-part frequency-severity model in Frees, Gao, and Rosenberg (2011) wherein the authors considered that the amount of healthcare expenditures is affected by the number of either inpatient stays or outpatient visits. According to their results, inpatient stays did not significantly affect the total annual healthcare expenditure, but outpatient visits had a significantly negative impact on the total annual healthcare expenditure.

In Lee, Park, and Ahn (2019), 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 an additional covariate. The model was calibrated using an auto insurance data set from Massachusetts in 2006. Interesting to note, the result indicated a positive effect of frequency on severity, that is, high frequency led to an increase in average severity.

Following an approach similar to that in Garrido, Genest, and Schulz (2016), Park, Kim, and Ahn (2018) calibrated the dependent two-part GLM with claims data from a Korean insurance company. Interesting to note, the authors used the level of bonus-malus as an explanatory variable for predicting average severity. Bonus-malus is a quite 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, statistical evidence of dependency for liability claims was lacking.

Shi, Feng, and Ivantsova (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, Genest, and Schulz (2016). According to Shi, Feng, and Ivantsova's 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, Feng, and Ivantsova (2015) proposed a model that can accommodate heavy tail distribution for claim severity as an extension of Czado et al. (2012) and

Frees, Gao, and Rosenberg (2011). The model estimation yielded a positive dependence between frequency and average severity. According to Shi, Feng, and Ivantsova's interpretation, this is because insureds with high frequency may have riskier driving habits, leading to more severe claims when accidents happen.

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 data set used to calibrate the corresponding model.

For a typical portfolio of insurance policies, it is not uncommon for observations of independent policyholders to be 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, ... , T_i, where $T_i \leq T$ and for policyholder *i*, for *i* = 1, ... , *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 generalized linear mixed models (GLMMs) 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 paper, we extend the GLM framework with dependent frequency and severity model suggested by Garrido, Genest, and Schulz (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, Young, and Luo (1999) demonstrate the link between longitudinal data models and credibility models. Such a link allows for a more convenient statistical analysis of credibility models as part of a ratemaking process; see Frees, Young, and Luo (2001). On a similar note, Garrido and Zhou (2009) examine limited fluctuations credibility of GLMM estimators. Antonio and Beirlant (2007) provide motivations for using GLMMs for analyzing longitudinal data in actuarial science; in particular, claims data sets arising from a portfolio of workers' compensation insurance were used to calibrate GLMMs. The authors not only provide the GLMM formulation but also discuss estimation, inference, and prediction useful for analyzing the data. In Antonio and Valdez (2011), GLMM is 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" (187). Finally, Boucher, Denuit, and Guillén (2008) suggest the use of a random effects model for claim counts with time dependence. Indeed, they compare such a 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, and copula models. Using an automobile insurance portfolio from a major company in Spain with data from 1991 to 1998, they conclude that the random effects model is the superior model for insurance claims data with time dependence.

Inspired by the work of Garrido, Genest, and Schulz (2016) and the growing interest in 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, making insurance data typically take the form of equation (1.1). For example, auto insurance companies usually have data with policyholders observed for a period of consecutive years; such data provide a richer set of information for improved prediction, pricing, and reserving. Preserving the GLM dependent frequency-severity structure proposed in Garrido, Genest, and Schulz (2016), the extension is achieved with the 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 stated in Frees, Derrig, and Meyers (2014, Chapter 16, 400), this "enables cluster-specific prediction ... and structure correlation within cluster."

The remainder of this paper is organized as follows. In Section 2, we introduce GLMMs. The purpose of this is to make the paper self-contained and introduce the notations 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 data set consisting of M independent subjects with response variable y_{it} and known covariates $\mathbf{x}_i'$. 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:

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

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 a 0 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

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

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

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

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

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

while the variance is

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

See Chapter 8 of McCulloch and Searle (2001).

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

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

where $\mathbf{y} = (y_1', \cdots, y_M')$ 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 have results that can give some kind of explicit expressions, we need to specify what are 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 easily can 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 $i = 1, \ldots, M$. 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 year 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 therefore can 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 equation (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 also will be 0. 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

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

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 straightforward that the unconditional mean of S can be expressed as

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

and the unconditional variance as

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

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 number of 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

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

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:

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

It is clear from these relations that unlike the Poisson distribution, overdispersion can be handled easily by the negative binomial distribution. An additional item to note is the accounting of the exposure, 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 0 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

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

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

$$\begin{aligned}\overline{C}|N,u \sim \text{ indep. Gamma}&\left(\mu e^{u+\theta n}, \tfrac{\phi}{n}\right) \\ &\text{with density } f_{\overline{C}|N}\left(\overline{c}|n,u\right),\end{aligned}$$

where

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

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 simply may be ignored, and naturally this implies a corresponding beta coefficient of 0. In addition, we will introduce the number of claims n as a covariate with coefficient θ. With this specification, the following relations hold:

$$\mathbb{E}\left[\overline{C}|N,u\right] = e^{\mathbf{x}'\beta+\theta n+u}$$

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

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

We assume that the random effect u has a normal distribution with mean 0 and variance σ_u^2 and that 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 key property here is that the likelihood function L in equation (2.13) is divided into two product terms, where one term is related with parameters only for frequency while the other term is related with parameters only 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 software packages.

To derive maximum likelihood estimates, we maximize log L with respect to all the parameters by taking partial derivatives and setting each to 0. This leads us to the following set of $(2p + 3)$ estimating equations that solve for $\widehat{\alpha}$, $\widehat{\beta}$, $\hat{\theta}$, $\widehat{\sigma_b}$ and $\widehat{\sigma_u}$, for $k = 1, \ldots, p$:

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

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

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

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

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

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 the method of moments. Additional details about the derivation of the likelihood estimating equations developed above are in Appendix A.

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

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

and the unconditional variance is

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

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

As a final remark, we broadly call the model constructed in this subsection to be the **dependent GLMM.** In the special case wherein 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, Genest, and Schulz (2016). In the additional special case where $\theta = 0$, we have the **independent GLM.** In the validation section, we compared these three models together with the **Tweedie GLM** and the **Tweedie GLMM.** The Tweedie models will be described later.

3. DATA CHARACTERISTICS AND PRELIMINARY INVESTIGATION

For the empirical investigation in this paper, 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 are from a period of six years, 1994 to 1999. These data were obtained from the General Insurance Association of Singapore, a trade association with representations from all the general insurance companies doing business in Singapore during the said six-year period. All claim amounts are in Singapore dollar (SGD) currency. These data or some extracted form of these data have been used in several papers 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 historically has accounted for more than a third of the entire insurance market. As is the case in most developed countries, auto insurance provides coverage at different layers, with the minimum layer's being mandatory,

providing protection against death or bodily harm to third parties regardless of who is at fault. In many countries, this is called third-party liability coverage.

For estimation purposes, we used policy and claims information for the first 5 years, 1994 to 1998, which we call the training data set. For validation, we used the observed information for 1999. Our summary statistics in the rest of this section consist 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.

Observations in the portfolio consist of policies with comprehensive coverage 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 types of coverage.

We used observed policy characteristics from the portfolio as covariates in both the frequency and the 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, gender and issue age are driver information, and the rest of the information pertains to the vehicle. The table is indeed self-explanatory, but there are a few items worth noting about the insured portfolio used in this analysis. First, although motorbike insurance is not uncommon in Singapore, in our data set, usually cars are the insured vehicles. 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 indicates that vehicles can be age −1; it is possible to purchase a vehicle with a model year after the purchase year. Also, the average age of vehicles is about 7 years, indicating that insured vehicles are generally newer and less than 10 years old. In Singapore, this is not at all surprising given the government restriction on allowing entitlement of vehicle ownership that usually expires in 10 years.

Table 2 provides a summary of the claim frequency distribution from calendar years 1994 to 1998. For each year, there is a significant percentage of zero claims. On average, each year shows 91.6% 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; when compared to the number of unique policyholders observed, which is $M = 41{,}831$, this number indicates that there are repeated observations for each unique policyholder, who is observed, on average, for a little more than 3 years. Moreover, notice that the number of policies observed varies each year, indicating the unbalanced nature of our data. In some years, policies either lapse (leave the company) or newly join the company. In a calendar year, the most frequent number of times a policyholder makes a claim is just once. There are very few policyholders with more than three claims in a calendar year.

Table 3 presents the average claim severity amounts for numbers of claims and for calendar years. The overall average claim amount is 4,655.09, ranging from 4,377.50 to 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 numbers of claims. For most calendar years, this positive correlation can be observed. However, such a positive correlation cannot be deduced from Figure 1. This figure provides further examination of the possible dependency between frequency and average severity. It is difficult to find a more conclusive relationship between frequency and average severity in this figure. However, this preliminary investigation of the data suggests that we can probably find more conclusive statistical evidence of the relationship when we inject the relation of number of claims to average severity within either the GLM or the GLMM framework.

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

In practice, the use of the 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 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. Using a negative binomial to accommodate this possible overdispersion is also widely accepted in the actuarial literature; see, for example, Frees and Valdez (2008).

To further support our choice of a negative binomial for frequency, we provide a goodness-of-fit test for the Poisson model compared 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 from 1994 to 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 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 a somewhat weak fit on both ends of the tail of the distribution for most calendar years. We wanted to restrict our marginal choice for the average severity to be within the class of the exponential, or GLM, family. Furthermore,

Table 1. Observable policy characteristics used as covariates

Categorical Variable	Description		Proportion (%)	
VehType	Type of insured vehicle	Car	99.26	
		Motorbike	0.54	
		Other	0.20	
Gender	Insured's sex	Male = 1	80.74	
		Female = 0	19.26	
Cover Code	Type of insurance cover	Comprehensive = 1	78.36	
		Other = 0	21.64	
Continuous Variable		Minimum	Mean	Maximum
VehCapa	Insured vehicle's capacity in cc (cubic capacity)	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

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	98,982	91.6
1	7.3	7.3	7.7	7.5	8.8	8,288	7.7
2	0.5	0.6	0.6	0.6	0.9	700	0.6
3	0.0	0.0	0.0	0.0	0.1	42	0.0
4			0.0	0.0		4	0.0
5				0.0		1	0.0
Number	24,690	22,857	21,437	20,302	18,731	108,017	100

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

	1994	1995	1996	1997	1998	AvgSev by Count
1	4,355	4,028	4,763	4,219	4,278	4,329
2	8,162	7,545	8,608	8,325	7,652	8,025
3	6,430	18,564	12,482	5,018	11,148	11,395
4			9,429	19,363		16,879
5				12,149		12,149
Annual AvgSev	4,616	4,378	5,087	4,558	4,639	4,655

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 about the numerical results of calibrating the various models described in Section 2. Recall that 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 a 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 0 for zero claims. For a 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.

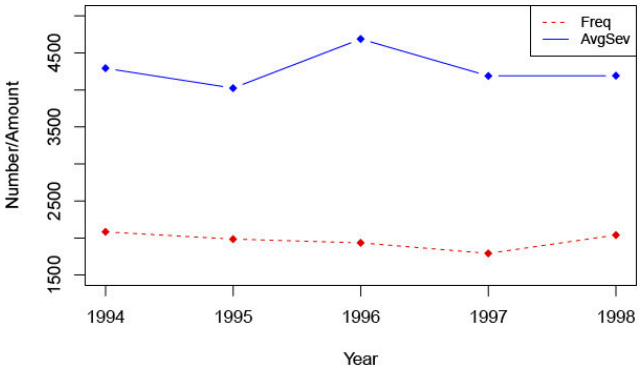

Figure 1. Frequency and average severity by calendar year

Note: Freq = frequency; AvgSev = average severity.

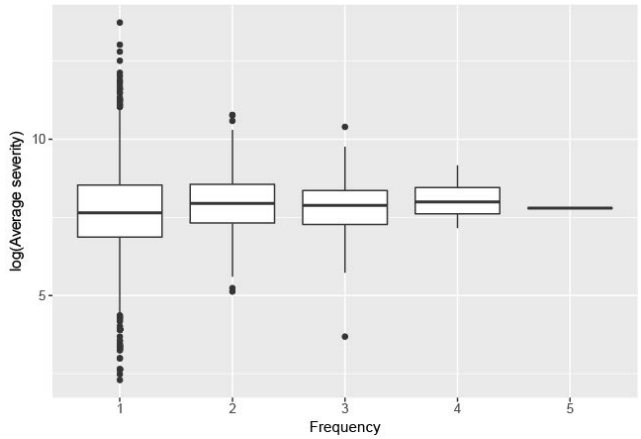

Figure 2. Graphical relationship of frequency and average severity, per policyholder

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

Count	Observed	Poisson	Negative Binomial
0	98,982	98,616.5	98,978.9
1	8,288	8,979.1	8,305.9
2	700	408.8	673.6
3	42	12.4	54.0
4	4	0.3	4.3
5	1	0.0	0.3
χ^2		67961.9	67674.6

4.1. NUMERICAL RESULTS

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

We now compare the two 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 levels of significance. Note that because historically we have known the nonlinearity of age to claims, we added the square and cubic factors for age in our regression models. Broadly speak-

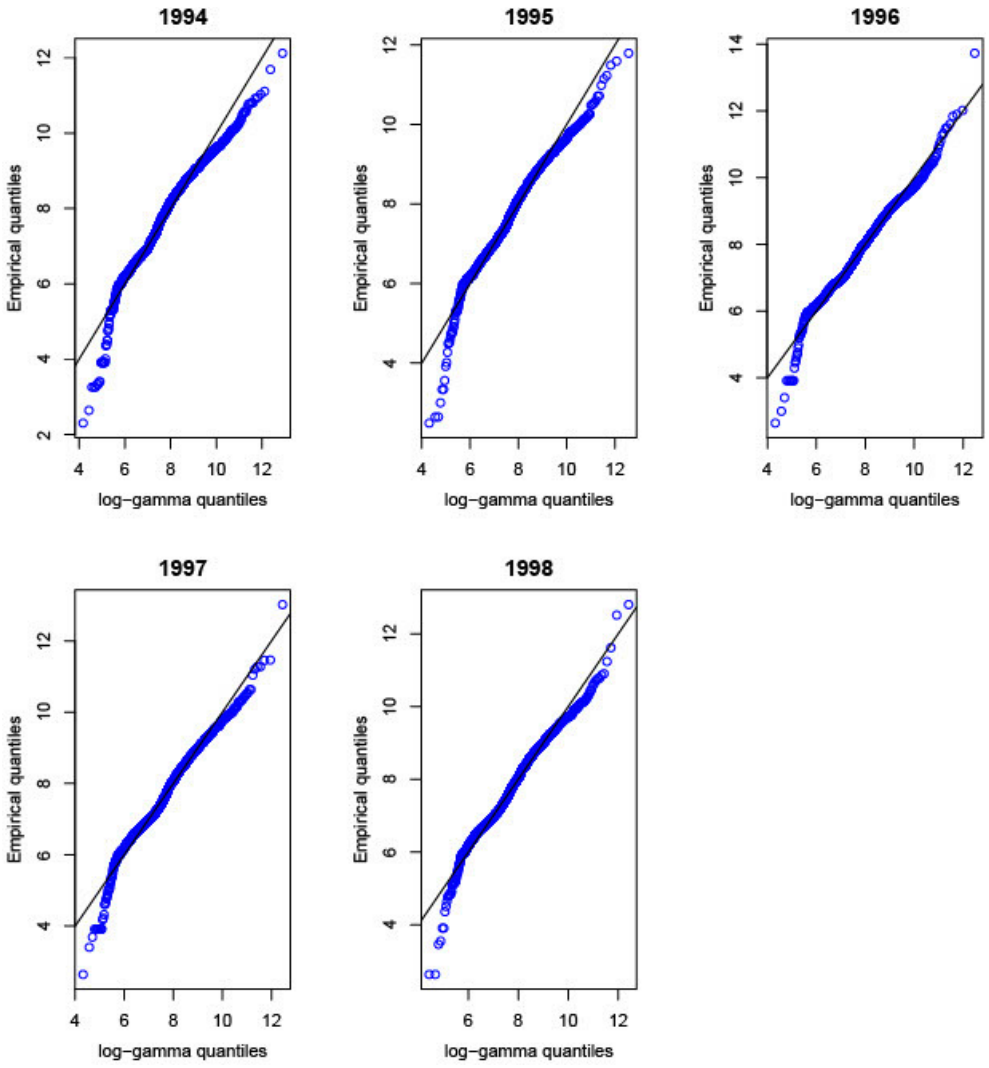

Figure 3. log-QQ plots of fitting gamma to average severity for each calendar year

ing, the effects of each of the explanatory variables summarized in Table 5 are not much different between the two models. One can deduce that if a random effect is added as in the GLMM, there is a slight decrease of the p-values of the corresponding coefficients. The signs of the coefficients are consistent between the two models and appear to be intuitive. For example, when compared to cars, motorbikes are less likely to claim, and this might partly be because cars are more frequently on the road. Older vehicles are less likely to claim, and this may be explained by fewer older vehicles' being insured. Male drivers are more likely to claim; historically, this has been intuitively explained by the more aggressive driving behavior of males than of 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 no claim discount score, and the result is statistically significant and indeed quite intuitive. The negative sign indicates that policyholders with higher discounts are considered less risky drivers and hence tend to have lower likelihoods of making claims.

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 Akaike information criterion (AIC) and Bayesian information criterion (BIC). All three statistics are reported at the bottom of Table 5. Because we maximize the log-likelihood, we consider a greater value to be superior. For either AIC or BIC, we consider a lower value to be superior. An examination of all three model fit statistics suggests that the GLMM is slightly superior to 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 vehicles. Although we found in the frequency component that motorbike drivers are less likely to make claims, the result here suggests that when motorbike policyholders make claims, the severity impact is worse. This is consistent with studies

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

	Negative Binomial GLM			Negative Binomial GLMM		
Variable	Estimate	*se*	Pr(>|t|)	Estimate	*se*	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
${\sigma_b}^2$				0.1992	0.4560	0.6484
Log-likelihood		−32345.3445			−32319.1436	
AIC		64714.6890			64664.2872	
BIC		64799.6805			64788.9577	

that suggest that the risk of fatal crashes among motorbike drivers is far greater than among passenger car drivers. Moreover, these same studies suggest that the seriousness of the injuries in motorbike crashes is worse than that in passenger car crashes. Furthermore, not only do younger drivers have a higher likelihood of claims, but the severity of claims is 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 also are some clear differences. For example, in the dependent GLMM, the regression estimate for the presence of 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 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, suggesting there is inconsistency in 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 accounts for the resulting differences in means among groups 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. All three statistics are reported at the bottom of Table 6. An examination of all three model fit statistics suggests that the GLMM is far superior to the other two models.

In practice and in the actuarial literature, there is increased interest in the use of the Tweedie exponential dispersion family model to fit compound loss models. See, for example, Frees, Lee, and Yang (2016). The Tweedie family of distributions belongs 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 0. 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 in 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 in Table 7 wherein 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.

4.2. VALIDATION RESULTS

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

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

	Independent GLM			Dependent GLM			Dependent GLMM		
Variable	Estimate	*se*	Pr(>|t|)	Estimate	*se*	Pr(>|t|)	Estimate	*se*	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
$\sigma_u{}^2$							1.0814	0.4749	0.0228
Log-likelihood		−89874.3089			−89865.2618			−86071.9841	
AIC		179772.6178			179756.5235			172171.9683	
BIC		179857.6092			179848.5976			172271.1249	

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

	Tweedie GLM			Tweedie GLMM		
Variable	Estimate	*se*	Pr(>|t|)	Estimate	*se*	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
$\sigma_R{}^2$				1632.6000	98.5019	0.0000
Log-likelihood		−112632.5098			−165017.4487	
AIC		225289.0196			330060.8975	
BIC		225374.0110			330185.5680	

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

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

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

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 MSE than do smaller deviations. When compared to MSE, MAE is relatively less sensitive to large deviations. As such, MSE is much more sensitive to observations that are considered outliers.

Note that we 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.

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

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 holdout sample. From this Lorenz curve, we compute the corresponding Gini index.

1. From our holdout sample, for i = 1, 2, ... , M, sort the observed loss Y_i according to the risk score S_i, which in our case is 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 = argmin(S_i).
2. Compute $F_{\text{score}}(m/M) = \frac{1}{M}\sum_{i=1}^{M} 1_{(R_i \le m)}$, the cumulative percentage of exposures, and $F_{\text{loss}(m/M)} = \frac{\sum_{i=1}^{M} Y_i 1_{(R_i \le m)}}{\sum_{i=1}^{M} Y_i}$, 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 two times the area between the line of equality and the Lorenz curve drawn above. We choose the model with the highest Gini index. According to Figure 4, the dependent GLMM is the best among all five models, with a Gini index equal to 45.7%.

5. CONCLUDING REMARKS

In this paper, we offered the use of the GLMMs as alternative models with dependent claim frequency and claim severity. The concept is a natural extension of the depen-

Table 8. Validation measures for the five models

Model	Mean Squared Error	Mean Absolute Error
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

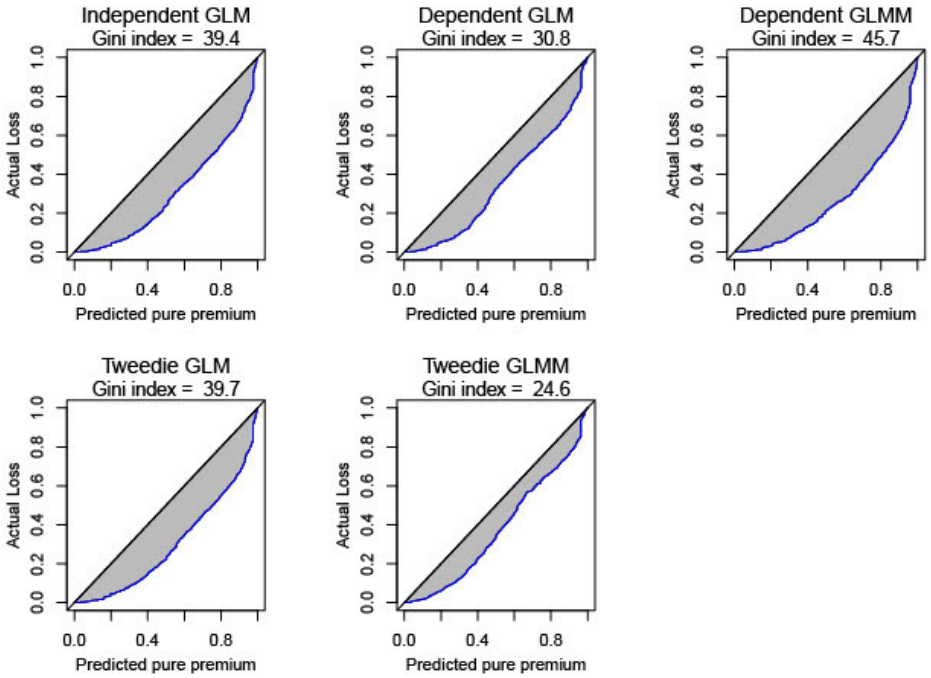

Figure 4. The Lorenz curve and the Gini index values for the five models

dent frequency-severity model based on GLM introduced by Garrido, Genest, and Schulz (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 of 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 are from a period of six years, 1994 to 1999. We used the first five years' experience data as the training set for model estimation and used the final year, 1999, as the holdout sample for validation. To provide a strong argument for the superiority of our proposed model, we considered four other models that we called the independent GLM, the dependent GLM (Garrido, Genest, and Schulz 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 the four other models.

In future research, we need to examine a better model for severity. As indicated by our preliminary investigation, the gamma model did not do well capturing the long tail 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 GLMs.

Submitted: September 22, 2017 EDT. Accepted: May 21, 2018 EDT. Published: September 29, 2021 EDT.

REFERENCES

Antonio, Katrien, and Jan Beirlant. 2007. "Actuarial Statistics with Generalized Linear Mixed Models." *Insurance: Mathematics and Economics* 40 (1): 58–76. https://doi.org/10.1016/j.insmatheco.2006.02.013.

Antonio, Katrien, and Emiliano A. Valdez. 2011. "Statistical Concepts of a Priori and a Posteriori Risk Classification in Insurance." *Advances in Statistical Analysis* 96 (2): 187–224. https://doi.org/10.1007/s10182-011-0152-7.

Boucher, J.-P., M. Denuit, and M. Guillén. 2008. "Models of Insurance Claim Counts with Time Dependence Based on Generalization of Poisson and Negative Binomial Distributions." *Variance* 2 (1): 135–62.

Czado, Claudia, Rainer Kastenmeier, Eike Christian Brechmann, and Aleksey Min. 2012. "A Mixed Copula Model for Insurance Claims and Claim Sizes." *Scandinavian Actuarial Journal* 2012 (4): 278–305. https://doi.org/10.1080/03461238.2010.546147.

Frees, Edward W., R. A. Derrig, and G. Meyers. 2014. *Predictive Modeling Applications in Actuarial Science*. Vol. 1, Predictive Modeling Technique. Cambridge, UK: Cambridge University Press.

———. 2016. *Predictive Modeling Applications in Actuarial Science*. Vol. 2, Case Studies in Insurance. Cambridge, UK: Cambridge University Press.

Frees, Edward W., Jie Gao, and Marjorie A. Rosenberg. 2011. "Predicting the Frequency and Amount of Health Care Expenditures." *North American Actuarial Journal* 15 (3): 377–92. https://doi.org/10.1080/10920277.2011.10597626.

Frees, Edward W., Gee Lee, and Lu Yang. 2016. "Multivariate Frequency-Severity Regression Models in Insurance." *Risks* 4 (4): 1–36. https://doi.org/10.3390/risks4010004.

Frees, Edward W., and Emiliano A. Valdez. 2008. "Hierarchical Insurance Claims Modeling." *Journal of the American Statistical Association* 103 (484): 1457–69. https://doi.org/10.1198/016214508000000823.

Frees, Edward W., Virginia R. Young, and Yu Luo. 1999. "A Longitudinal Data Analysis Interpretation of Credibility Models." *Insurance: Mathematics and Economics* 24 (3): 229–47. https://doi.org/10.1016/s0167-6687(98)00055-9.

———. 2001. "Case Studies Using Panel Data Models." *North American Actuarial Journal* 5 (4): 24–42. https://doi.org/10.1080/10920277.2001.10596010.

Garrido, José, C. Genest, and J. Schulz. 2016. "Generalized Linear Models for Dependent Frequency and Severity of Insurance Claims." *Insurance: Mathematics and Economics* 70 (September):205–15. https://doi.org/10.1016/j.insmatheco.2016.06.006.

Garrido, José, and Jun Zhou. 2009. "Full Credibility with Generalized Linear Mixed Models." *ASTIN Bulletin* 39 (1): 61–80. https://doi.org/10.2143/ast.39.1.2038056.

Lee, Woojoo, Sojung C. Park, and Jae Youn Ahn. 2019. "Investigating Dependence between Frequency and Severity via Simple Generalized Linear Models." *Journal of the Korean Statistical Society* 48 (1): 13–28. https://doi.org/10.1016/j.jkss.2018.07.003.

McCulloch, C. E., and S. R. Searle. 2001. *Generalized, Linear, and Mixed Models*. New York: Wiley.

Molenberghs, G., and G. Verbeke. 2005. "Longitudinal and Incomplete Clinical Studies." *Metron—International Journal of Statistics* 63 (2): 143–76.

Nelder, J. A., and R. W. M. Wedderburn. 1972. "Generalized Linear Models." *Journal of the Royal Statistical Society. Series A (General)* 135 (3): 370–84. https://doi.org/10.2307/2344614.

Park, Sojung C., Joseph H. T. Kim, and Jae Youn Ahn. 2018. "Does Hunger for Bonuses Drive the Dependence between Claim Frequency and Severity?" *Insurance: Mathematics and Economics* 83 (November):32–46. https://doi.org/10.1016/j.insmatheco.2018.09.002.

Shi, Peng, Xiaoping Feng, and Anastasia Ivantsova. 2015. "Dependent Frequency–Severity Modeling of Insurance Claims." *Insurance: Mathematics and Economics* 64 (September):417–28. https://doi.org/10.1016/j.insmatheco.2015.07.006.

Shi, Peng, and E. A. Valdez. 2012. "Longitudinal Modeling of Insurance Claim Counts Using Jitters." *Scandinavian Actuarial Journal* 2012:1–21.

APPENDICES

APPENDIX A. THE DEVELOPMENT OF THE LOG-LIKELIHOOD EQUATIONS

Based on our observed data, from equations (2.13) and (2.6), the likelihood can be expressed as

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

Thus, the log-likelihood can be expressed as

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

and

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

We take the partial derivatives of the log-likelihood functions and set to 0:

$$\begin{aligned} \frac{\partial l_N}{\partial \alpha} &= 0 \text{ for } k = 1, \ldots, p, \\ \frac{\partial l_N}{\partial \sigma_b} &= 0, \\ \frac{\partial l_N}{\partial \theta} &= 0, \\ \frac{\partial l_N}{\partial \beta} &= 0 \text{ for } k = 1, \ldots, p, \text{ and} \\ \frac{\partial l_N}{\partial \sigma_u} &= 0. \end{aligned}$$

The results yield to the (2*p* + 3) estimating equations 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:

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

and

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

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

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

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 assume 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

$$\begin{aligned} \mathbb{E}[S|\mathbf{x}] =& \mathbb{E}\left[N\overline{C}|\mathbf{x}\right] \\ =& \mathbb{E}\left[N\mathbb{E}\left[\overline{C}|N,u\right]|\mathbf{x}\right] \\ =& \mathbb{E}\left[N\mu e^{n\theta+u}|\mathbf{x}\right] \\ =& \mu\mathbb{E}\left[Ne^{n\theta}|\mathbf{x}\right]\mathbb{E}\left[e^u|\mathbf{x}\right] \\ =& \mu\mathbb{E}\left[M'_{N|b,\mathbf{x}}(\theta)\right] e^{\sigma_u^2/2} \\ =& \mu\mathbb{E}\left[\nu[1-(\nu e^b/r)(e^\theta-1)]^{-r-1}\right] \\ & \times e^{\sigma_u^2/2+\theta} \\ =& \mu\nu\mathbb{E}\left[[1-(\nu e^b/r)(e^\theta-1)]^{-r-1}\right] \\ & \times e^{\sigma_u^2/2+\theta}, \end{aligned} \tag{B.3}$$

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 $\mathbb{E}[S|\mathbf{x}] = \mu\mathbb{E}\left[M'_{N|\mathbf{x}}(\theta)\right]$, which gives us precisely what is found in Garrido, Genest, and Schulz (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

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

and

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

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

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

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

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

which clearly corresponds to what is derived in Garrido, Genest, and Schulz (2016).

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