Longitudinal Modeling of Insurance Claim Counts Using Jitters

Peng Shi
Division of Statistics
Northern Illinois University
DeKalb, IL 60115
Email: pshi@niu.edu

Emiliano A. Valdez
Department of Mathematics
University of Connecticut
Storrs, CT 06269
Email: emiliano.valdez@uconn.edu

September 8, 2011

Abstract

Modeling insurance claim counts is a critical component in the ratemaking process for property and casualty insurance. This article explores the usefulness of copulas to model the number of insurance claims for an individual policyholder within a longitudinal context. To address the limitations of copulas commonly attributed to multivariate discrete data, we adopt a "jittering" method to the claim counts which has the effect of continuitizing the data. Elliptical copulas are proposed to accommodate the intertemporal nature of the "jittered" claim counts and the unobservable subject-specific heterogeneity on the frequency of claims. Observable subject-specific effects are accounted in the model by using available covariate information through a regression model. The predictive distribution together with the corresponding credibility of claim frequency can be derived from the model for ratemaking and risk classification purposes. For empirical illustration, we analyze an unbalanced longitudinal dataset of claim counts observed from a portfolio of automobile insurance policies of a general insurer in Singapore. We further establish the validity of the calibrated copula model, and demonstrate that the copula with "jittering" method outperforms standard count regression models.

Keywords: Claim count, Copula, Jitter, Longitudinal data, Predictive distribution

1 Introduction

In property and casualty ratemaking, it is a common practice to account for the pattern of claims observed over a period of time. Past claims experience may provide invaluable insight into some of the risk characteristics of the policyholder that are typically considered unobservable and may be difficult to assess a priori. To illustrate, in the case of automobile insurance, peculiar driving abilities such as aggressiveness and judgement of the driver can be challenging to the actuary to evaluate at a time when a new policy is issued. However, as claims are observed over a length of time, the insurer is able to better assess the policyholder's potential risk characteristics and as a consequence, to make appropriate premium adjustments. In general terms, insured drivers with past claims are penalized while those with good claim records are rewarded. Experience rating is the term that is used to evaluate the past claim records of an insured policyholder in order to reach a fair and more equitable premium structure.

For pure premium calculations, the most customary approach to model the claims distribution is to use a two-part model which decomposes the total claims into claim frequency and claim severity. The claim frequency accounts for the number of times, usually in a period e.g. year, the policyholder makes a claim. The claim severity examines the amount of the claim, given a claim occurs. See, for example, Klugman et al. (2008). We focus our attention in this paper on the claim frequency or claim count component for which many believe to be the more important component. This is because the number of claims better reveals the inherent risk of a policyholder, and thus leads to a better risk classification. Besides, being unable to accurately predict the number of claims can have a devastating effect in an insurance portfolio. For each single claim made, a claim amount is often associated with it. Hence, each time a claim is not correctly predicted, the prediction bias is transmitted into the total insurance claim amount of the portfolio. To further emphasize the magnitude, according to the Department of Transportation and National Highway Traffic Safety Administration, in 2007, there were a total of approximately 10.5 million car crashes in the United States.

This paper considers modeling insurance claim counts observed over a period of time, or more commonly referred to as 'longitudinal insurance claim counts'. Assume that we observe claim counts denoted by N_{it} for the ith policyholder, $i = 1, \ldots, I$, in an insurance portfolio in year t, $t = 1, \ldots, T_i$. For each policyholder in the portfolio, the observable data is a vector of claim counts that can be expressed in the form $(N_{i1}, \ldots, N_{iT_i})$. The length of time T_i observed may differ among policyholders in the case when the data is unbalanced. Additionally, we have a set of observable covariates $\mathbf{x}_{it}$ frequently useful for the purpose of sub-dividing the portfolio into classes of risks with homogeneous characteristics. This process of risk classification is relevant for insurance pricing and valuation. It is a common practice in insurance ratemaking to account for risk classifying variables so that the insurer can optimally group homogeneous risks within an insurance portfolio. See Antonio and Valdez (2011).

Since the seminal work of Jorgenson (1961), count data regression has received extensive attention in the literature. Its applications range in various disciplines including but not limited

to actuarial science, biomedical statistics, economics, and sociology. We refer readers to the two excellent monograph, Cameron and Trivedi (1998) and Winkelmann (2008), and the vast literature therein for count data regression models. One strand of studies take a semiparametric approach, where the conditional mean of the count variable is estimated by a pseudolikelihood method. See, for example, Nelder and Wedderburn (1972) and Gourieroux et al. (1984). The semiparametric approach provides robust estimation of the expectation of the count of interest. However, little information could be derived for other aspects of the conditional distribution, which makes it very limited in many applied research. For example, the interest of our application is the predictive distribution that is used in the ratemaking process for property and casualty insurers. As an alternative, part of the literature, following the early work of Hausman et al. (1984), has been devoted to the development of fully parametric probabilistic count regression models. The most common count distributions used are Poisson and negative binomial where the covariates are linked through a function of the respective mean parameters. It is well-known for count data that sometimes we observe an excessive number of zeros more so than what these distributions can accommodate; hence, the use of generalized count distribution which includes zero-inflated and hurdle models, are increasing in popularity.

Various regression techniques have been proposed for longitudinal count data. In a longitudinal context, one hopes to capture the unobserved individual heterogeneity that helps improve the prediction of the count variable of interest. To capture the intertemporal dependence within subjects, the most popular approach is to introduce a common random effect, say α_i, to each observation. To fix ideas, if $(n_{i1}, \ldots, n_{iT_i})$ is the vector of observed counts, then the joint probability mass function for $(N_{i1}, \ldots, N_{iT_i})$ can be expressed as

$$\Pr(N_{i1} = n_{i1}, \ldots, N_{iT_i} = n_{iT_i}) = \int_0^\infty \Pr(N_{i1} = n_{i1}, \ldots, N_{iT_i} = n_{iT_i}|\alpha_i) f(\alpha_i) d\alpha_i$$

where $f(\alpha_i)$ is the pre-specified density function of the random effect. To evaluate the conditional probability in the first term of the above integral, the typical assumption is conditional independence as follows:

$$\Pr(N_{i1} = n_{i1}, \ldots, N_{iT_i} = n_{iT_i}|\alpha_i) = \Pr(N_{i1} = n_{i1}|\alpha_i) \times \cdots \times \Pr(N_{iT_i} = n_{iT_i}|\alpha_i).$$

Random effects models provide for an instinctive interpretation for insurance claims data. It captures unobservable individual risk characteristics that are not accounted for by the observable covariates. For example, in the case of automobile insurance, this may include unobservable behavior and judgmental abilities of the driver which could possibly be reflected in the pattern of claims observed over a period of time. The excellent survey paper in Boucher and Guillén (2009) explores various random effects models applied to automobile insurance data. Additional actuarial applications of these types of models are also discussed in the textbooks by Denuit et al. (2007) and Frees (2010).

In this paper, to account for within-subject serial correlation or intertemporal dependence, we

propose the use of indirect application of copula regression models to the vector of insurance claim counts $(N_{i1}, \ldots, N_{iT_i})$ for the ith policyholder. While copulas are undeniably popular partly because of their exceptional flexibility in modeling a variety of dependence structure, it has been highly debatable when they are directly applied to discrete data. This is because for multivariate discrete observations, there is the non-uniqueness of the copula and the resulting vague interpretation of the nature of dependence implied by the copula. See the arguments made by Genest and Nešlehová (2007). Despite this debate, there has been a number of applications of copula regression models to multivariate discrete data in various disciplines. For example, see Prieger (2002), Cameron et al. (2004), and Zimmer and Trivedi (2006) in economics, Song et al. (2009) in biostatistics, Purcaru and Denuit (2003) and Shi and Valdez (2011) in actuarial science.

To circumvent around the potential problems of direct applications of copulas, we apply the concept of jittering where we subtract an independent continuous random variable from each component of the multivariate discrete data. This has the effect of converting the discrete to continuous data. Henceforth, we are able to more comfortably specify copula functions to the resulting jittered multivariate continuous data. It has been demonstrated, in Denuit and Lambert (2005) where this approach has been suggested, that the concordance-based association measures, such as Kendall's *tau*, are preserved when jittering the discrete data. This allows us to naturally interpret the dependence implied by the copula for the jittered data. To our knowledge, this work is the first application of the jitter approach to longitudinal count data. Additionally, we do note that two studies employed copulas for the modeling of longitudinal insurance claims, Frees and Wang (2006) and Boucher et al. (2008). The primary difference of these studies compared with ours is that the former used copulas to construct the joint distribution of latent random variables, and the latter used copulas to construct the joint mass probability function of the claim counts directly. Additionally, we note that the jitter approach has recently been explored for multivariate random variables in the biostatistics literature by Madsen and Fang (2010) where our work is imminently related. For the type of copulas examined in this paper, we find that the class of elliptical copulas provides the advantage of accommodating flexible intertemporal dependence structures. In addition, the sub-copulas and marginals of elliptical copulas stay within the same copula family (see Landsman and Valdez (2003)), which is particularly useful for the unbalanced longitudinal nature with our application.

For empirical demonstration, we use claims experience data observed from an automobile insurance portfolio of a major general insurer in Singapore over a period of nine years: from 1993 to 2001. The data is considered unbalanced in the sense that we do not necessarily observe each policyholder during the entire period of observation. Some policyholders may have already been contracted after the start of the observation while others may have decided to leave the company before the end of the observation. In our dataset, we have a grand total of 4,006 observations, large enough to be able to comfortably perform model calibration and draw inference on the observed data.

The rest of this paper has been structured as follows. Section 2 provides for a discussion of

the various modeling framework suitable for longitudinal count data. Standard models examined include random effects count models. These models are discussed in some detail here because they are used to compare the performance of the jittered method. This section additionally describes the concept of applying copula models to jittered data. Whenever copula models are applied to data, some form of validation is necessary. We thus discuss the use of t-plots for validating the use of elliptical copulas. Section 3 summarizes the data used in our empirical investigation. Section 4 provides the calibration results together with model validation. Finally, we conclude in Section 5.

2 Modeling

Assume an insurance portfolio consisting of I policyholders is observed for T periods, and let $\mathbf{N}_i = (N_{i1}, \ldots, N_{iT_i})$ denote the vector of observable claim counts for the ith policyholder. Our goal is to be able to draw prediction on the number of claims in the subsequent period $T+1$ for this policyholder.

2.1 Review of Alternative Models

Traditional approach for modeling longitudinal count data is to incorporate random effects in the standard count regression model to account for individual heterogeneity. In this sub-section, we briefly discuss four standard formulations: random effects Poisson model, random effects negative binomial model, random effects zero-inflated Poisson model, and random effects zero-inflated negative binomial model. Because of the prediction purposes, we focus on fully parametric setup. The proposed copula-based count regression model will be compared against these four alternative models. We refer the reader to Denuit et al. (2007) and Frees (2010) for additional details on the actuarial modeling of claim counts.

2.1.1 Random Effects Poisson Model

The simplest form of random effects count regression assumes that the conditional number of claims follow a Poisson distribution, i.e.,

$$\Pr(N_{it} = n_{it}|\eta_i) = \frac{\tilde{\lambda}_{it}^{n_{it}} e^{-\tilde{\lambda}_{it}}}{n_{it}!} \tag{1}$$

with

$$\tilde{\lambda}_{it} = \eta_i \lambda_{it} = \eta_i \omega_{it} \exp(\mathbf{x}_{it}^{'} \boldsymbol{\beta}).$$

Here, ω_{it} is the weight parameter, representing the corresponding exposure, the length of time the insured keeps the policy inforce during the year. $\mathbf{x}_{it}$ denotes the vector of covariates in the tth period for policyholder i, and $\boldsymbol{\beta}$ represents the vector of corresponding regression coefficients. The subject-specific effect is captured by the random effect η_i. Many candidate distributions could be chosen for η_i. As shown in Hausman et al. (1984), when a gamma distribution with mean 1 and

variance $1/\psi$ is assumed for η_i, a closed form can be derived for the joint distribution of claim counts. It is easily shown that this random effect Poisson specification implies $\mathrm{E}[N_{it}] = \lambda_{it}$ and $\mathrm{Var}[N_{it}] = \lambda_{it} + \lambda_{it}^2/\psi$. The variance to mean ratio of $1 + \lambda_{it}/\psi$ indicates the overdispersion could be captured by the random effect.

Another way to think of Poisson model is to use generalized linear model framework, since Poisson distribution is a member of exponential family. A random effect model could be specified through the conditional link function. Using a canonical log link, one has

$$\tilde{\lambda}_{it} = \omega_{it} \exp(\alpha_i + \mathbf{x}_{it}^{'}\boldsymbol{\beta}).$$

A natural choice for the random intercept is $\alpha_i \sim N(0, \sigma^2)$. This specification is equivalent to a lognormal assumption for parameter η_i.

2.1.2 Random Effects Negative Binomial Model

Negative binomial regression is another popular choice to model count data. In a cross-sectional setup, it has the advantage of capturing overdispersion when compared with Poisson model. Different forms of negative binomial distribution are available for count data modeling, so with the incorporation of random effects. One well-known procedure to construct a negative binomial distribution is through a gamma-Poisson mixture formulation. Hausman et al. (1984) considered the random effect model:

$$\Pr(N_{it} = n_{it}|\nu_i) = \frac{\Gamma(n_{it} + \lambda_{it})}{\Gamma(\lambda_{it})\Gamma(n_{it} + 1)} \left(\frac{\nu_i}{1 + \nu_i}\right)^{\lambda_{it}} \left(\frac{1}{1 + \nu_i}\right)^{n_{it}}, \tag{2}$$

where $\lambda_{it} = \omega_{it} \exp(\mathbf{x}_{it}^{'}\boldsymbol{\beta})$, as defined above. This is known as NB1 distribution and could be derived by assuming a gamma distribution with parameters $(\lambda_{it}, \nu_i\lambda)$ for $\tilde{\lambda}_{it}$ in the Poisson distribution (Cameron and Trivedi (1998)). The conditional distribution has moments $\mathrm{E}[N_{ij}|\nu_i] = \lambda_{it}/\nu_i$ and $\mathrm{Var}[N_{it}|\nu_i] = \phi_i \mathrm{E}[N_{ij}|\nu_i]$ with $\phi_i = 1 + 1/\nu_i$. A convenient distributional assumption for the random effect is to assume that ϕ_i follows a beta distribution with parameters (a, b), under which, the joint distribution of claim counts has a closed form solution after integrating out the random effect. See Hausman et al. (1984).

As an alternative, when one assumes a gamma distribution with parameters (ψ, ψ) for the mean parameter $\tilde{\lambda}_{it}$ in the Poisson distribution, the gamma-Poisson mixture leads to another negative binomial distribution, known as NB2 distribution (see, for example, Greenwood and Yule (1920) and Dionne and Vanasse (1992) for its applications). The corresponding random effect model can be expressed as:

$$\Pr(N_{it} = n_{it}|\alpha_i) = \frac{\Gamma(n_{it} + \psi)}{\Gamma(\psi)\Gamma(n_{it} + 1)} \left(\frac{\psi}{\tilde{\lambda}_{it} + \psi}\right)^{\psi} \left(\frac{\tilde{\lambda}_{it}}{\tilde{\lambda}_{it} + \psi}\right)^{n_{it}}, \tag{3}$$

with

$$\mathrm{E}(N_{it}|\alpha_i) = \tilde{\lambda}_{it} = \omega_{it} \exp(\alpha_i + \mathbf{x}_{it}^{'}\boldsymbol{\beta}),$$

and

$$\mathrm{Var}(N_{it}|\alpha_i) = \tilde{\lambda}_{it} + \tilde{\lambda}_{it}^2/\psi.$$

Without loss of generality, a normal distribution is typically used for the random effect under the above specification.

2.1.3 Random Effects Zero-Inflated Models

Typically the number of claims contains a significant percentage of zeros, which cannot be captured by the variance function of Poisson and negative binomial distributions. One solution to address the issue of excess zeros is to use the zero-inflated count regression model introduced by Lambert (1992). Zero-inflated regression could be considered as a mixture of two stochastic processes, one only generating zeros and the other generating non-negative counts according to a standard count distribution. The probability that an individual count outcome is from the zero-generating process is usually determined by a binary regression model, for example, logit model with binomial distribution assumption. In a longitudinal context, subject-specific effects could be accommodated through random effects specification. In this application, we consider the following random effects formulation:

$$\Pr(N_{it} = n_{it}|\delta_i, \alpha_i) = \begin{cases} \pi_{it} + (1-\pi_{it})f(n_{it}|\alpha_i) & \text{if } n_{it} = 0 \\ (1-\pi_{it})f(n_{it}|\alpha_i) & \text{if } n_{it} > 0 \end{cases}. \tag{4}$$

In this model, π_{it} represents the probability that the observation n_{it} is from the zero-only process, and $f(n_{it}|\alpha_i)$ represents a standard count regression with subject-specific random effect α_i. In particular, we consider a random effects zero-inflated Poisson model (ZIP) with $f(n_{it}|\alpha_i)$ following (1), and a random effects zero-inflated negative binomial model (ZINB) with $f(n_{it}|\alpha_i)$ following (3). The probability π_{it} is determined by a logit regression of the form

$$\log\left(\frac{\pi_{it}}{1-\pi_{it}}\Big|\delta_i\right) = \delta_i + \mathbf{z}_{it}^{'}\boldsymbol{\gamma},$$

where random intercept δ_i captures individual heterogeneity, $\mathbf{z}_{it}$ is a set of covariates that could be different from the count data model and $\boldsymbol{\gamma}$ is the corresponding vector of regression coefficients.

2.2 Copula Model

In contrast to random effects, individual unobserved heterogeneity could be accommodated by the serial correlation among repeated observations in a longitudinal context. Towards this direction, we consider using a copula function to construct the joint distribution and thus capture the intertemporal dependence among insurance claims. Since claim counts are discrete, suppose there

exists a copula C, then the joint probability mass function of $(N_{i1}, \cdots, N_{iT})$ could be expressed as:

$$\Pr(N_{i1} = n_{i1}, \cdots, N_{iT} = n_{iT_i}) = \sum_{j_1=0}^{1} \cdots \sum_{j_{T_i}=0}^{1} (-1)^{j_1+\cdots+j_{T_i}} C(u_{1j_1}, \cdots, u_{Tj_{T_i}}), \tag{5}$$

where F_{it} denotes the distribution of N_{it}, $u_{t0} = F_{it}(n_{it})$ and $u_{t1} = F_{it}(n_{it} - 1)$ for $t = 1, \cdots, T_i$. See Song et al. (2009). This approach is subject to at least three computational difficulties. First, equation (5) contains 2^{T_i} terms, and thus become unmanageable to evaluate for T_i larger than four or five. Second, the calculation of the copula function C might involve high-dimensional non-tractable integration, for example, the family of elliptical copulas. Third, the non-uniqueness of the copula according to the Sklar's representation adds extra difficulty of applying copulas directly to discrete data (see, for example, Genest and Nešlehová (2007) for some discussion in the actuarial literature).

2.2.1 Continuous Extension with Jitters

To address the issues raised above, we adopt a "jitter" approach proposed by Madsen and Fang (2010). In this approach, the integer-valued claim count is converted to a continuous random variable. A copula is then used to model the dependency among the jittered continuous variables, while preserving the dependence relationship between the original discrete variables (see Denuit and Lambert (2005)).

Specifically, define $N_{it}^* = N_{it} - U_{it}$ where $U_{it} \sim \text{Uniform}(0, 1)$ and are independent for $t = 1, \cdots, T$ and $i = 1, \cdots, I$. Then given the distribution of the count variable N_{it} (we use F_{it} and f_{it} to denote its cumulative distribution function and probability function, respectively), one could derive the distribution of N_{it}^* as:

$$\begin{aligned} F_{it}^*(n) &= F_{it}([n]) + (n - [n]) f_{it}([n+1]) \\ &= \Pr(N_{it} \leq [n]) + (y - [y]) \Pr(N_{it} = [n+1]) \end{aligned} \tag{6}$$

and

$$f_{it}^*(n) = f_{it}([n+1]) = \Pr(N_{it} = [n+1]), \tag{7}$$

where F_{it}^* and f_{it}^* indicate the cumulative distribution function and density function of the jittered continuous random variable N_{it}^*, respectively. In this process, the claim count N_{it} could be retrieved from $N_{it} = [N_{it}^* + 1]$, and the parameters in F_{it}^* and f_{it}^* are exactly those of F_{it} and f_{it}, thus no information is lost in the "jitter" process.

According to Sklar's theorem, there exists a conditional copula $C(\cdot|\Omega)$ such that the joint distribution of the jittered claims for the ith policyholder $(N_{i1}^*, N_{i2}^*, \cdots, N_{iT}^*)$ could be expressed as:

$$F_i^*(n_{i1}^*, \cdots, n_{iT}^*|\Omega) = C(F_{i1}^*(n_{i1}^*|\Omega), \cdots, F_{iT}^*(n_{iT}^*|\Omega)|\Omega), \tag{8}$$

where Σ denotes available information, such as the set of covariates. Taking derivatives of equation

(8) leads to:

$$f_i^*(n_{i1}^*, \cdots, n_{iT}^*) = c(F_{i1}^*(n_{i1}^*), \cdots, F_{iT}^*(n_{iT}^*); \boldsymbol{\theta}) \prod_{t=1}^{T} f_{it}^*(n_{it}^*). \tag{9}$$

Here, $c(\cdot)$ represents the corresponding copula density, and $\boldsymbol{\theta}$ denotes the vector of association parameters in the copula. Note that the joint density in equation (9) represents a hypothetical case, since it is the discrete claim counts that one observes, rather than the jittered continuous variables. Based on the joint density of $(N_{i1}^*, \cdots, N_{iT}^*)$ in equation (9), one could derive the joint probability mass function of $(N_{i1}, \cdots, N_{iT})$ by averaging over the jitters $\boldsymbol{U}_i = (U_{i1}, \cdots, U_{iT})$:

$$f_i(n_{i1}, \cdots, n_{iT}) = \mathrm{E}_{\boldsymbol{U}_i} \left(c(F_{i1}^*(n_{i1} - U_{i1}), \cdots, F_{iT}^*(n_{iT} - U_{iT}); \boldsymbol{\theta}) \prod_{t=1}^{T} f_{it}^*(n_{it} - U_{it}) \right). \tag{10}$$

Furthermore, as shown in Denuit and Lambert (2005), the concordance-based association measures, such as Kendall's *tau*, of $(N_{i1}^*, \cdots, N_{iT}^*)$ are identical as those of $(N_{i1}, \cdots, N_{iT})$. This implies that we could interpret the dependency associated with the copula as that of the insurance claim counts.

2.2.2 Model Specification

To account for the potential overdispersion that is often observed with insurance claim counts data, we consider the negative binomial distribution for the marginal distribution of N_{it}:

$$f_{it}(n) = \Pr(N_{it} = n) = \frac{\Gamma(n + \psi)}{\Gamma(\psi)\Gamma(n + 1)} \left(\frac{\psi}{\lambda_{it} + \psi} \right)^{\psi} \left(\frac{\lambda_{it}}{\lambda_{it} + \psi} \right)^{n}, \tag{11}$$

with $\lambda_{it} = \omega_{it} \exp(\mathbf{x}_{it}^{'} \boldsymbol{\beta})$. Thus the distributions of the associated continuous variable N_{it}^* could be easily derived according to equations (6) and (7).

In longitudinal applications, elliptical copulas are commonly used to accommodate the within-subject serial correlation. Thus we consider members of this family for the jittered continuous variables, see, for example, Sun et al. (2008), Shi and Frees (2010), and Shi (2011). Elliptical copulas are extracted from multivariate elliptical distributions. A T-dimensional random vector $\boldsymbol{W} = (W_1, \cdots, W_T)$ follows a multivariate elliptical distribution, if for a $T \times 1$ vector $\boldsymbol{\mu}$ and a $T \times T$ positive-definite matrix $\boldsymbol{\Delta}$, its density is of form:

$$h(\boldsymbol{w}) = \frac{\kappa_T}{\sqrt{|\boldsymbol{\Delta}|}} g_T \left[\frac{1}{2} (\boldsymbol{w} - \boldsymbol{\mu})^{'} \boldsymbol{\Delta}^{-1} (\boldsymbol{w} - \boldsymbol{\mu}) \right],$$

where κ_T is a normalizing constant and $g_T(\cdot)$ is known as the density generator function. Using this notation, an elliptical distribution is often denoted by $\boldsymbol{W} \sim E_T(\boldsymbol{\mu}, \boldsymbol{\Delta}, g_T)$. Elliptical distributions are extensions of the multivariate normal distribution and thus share many of its tractable properties. This class of distributions was introduced by Kelker (1970) and was widely discussed in Fang et al. (1990). Some of their applications in actuarial science can be found in Landsman and Valdez (2003).

Let H be the cumulative distribution function of a T-dimensional multivariate elliptical random vector, H_t and h_t be the cumulative distribution function and density of the tth marginal, respectively, for $t = 1, \cdots, T$, then the corresponding elliptical copula is

$$C(\iota_1, \ldots, \iota_T) = H(H_1^{-1}(\iota_1), \ldots, H_T^{-1}(\iota_T)), \tag{12}$$

with its probability density function:

$$c(\iota_1, \ldots, \iota_T) = h(H_1^{-1}(\iota_1), \ldots, H_T^{-1}(\iota_T)) \prod_{t=1}^{T} \frac{1}{h_t(H_t^{-1}(\iota_t))}.$$

The elliptical copula remains invariant under an increasing transformation of the components of the vector $\boldsymbol{W}$. Thus we restrict our considerations to the copula associated with $E_T(\mathbf{0}, \boldsymbol{\Sigma}, g_T)$, where $\boldsymbol{\Sigma}$ is the dispersion matrix with diagonal entries equal to one.

We consider three commonly used dependence structure of $\boldsymbol{\Sigma}$: autoregressive of order 1 (AR(1)), exchangeable, banded Toeplitz with order 2. The AR(1) structure is a widely used specification for time series data, indicating that the correlation generally diminishes over time. In contrast, the exchangeable structure, also known as compound symmetry or uniform correlation in longitudinal data analysis, assumes a constant correlation over time. In a Toeplitz structure, the correlation between two observations depends only on their time difference, and the correlation becomes zero for far distant times. For demonstration purposes, the three dispersion matrix are as follows in the case where $T = 5$:

$$\Sigma_{AR} = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix} \quad \Sigma_{EX} = \begin{pmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{pmatrix} \quad \Sigma_{TOEP} = \begin{pmatrix} 1 & \rho_1 & \rho_2 & 0 & 0 \\ \rho_1 & 1 & \rho_1 & \rho_2 & 0 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ 0 & \rho_2 & \rho_1 & 1 & \rho_1 \\ 0 & 0 & \rho_2 & \rho_1 & 1 \end{pmatrix}.$$

2.2.3 Model Estimation

The parameters could be estimated using a likelihood-based estimation method. In the longitudinal application, we first consider the case of balanced data. Suppose all policyholders are observed for T periods, the discrete claim counts observations are denoted by $\mathbf{n}_i = (n_{i1}, \cdots, n_{iT})$ for $i = 1, \cdots, I$. Following the specification in Section 2.2.2, our goal is to estimate the vector of regression coefficients $\boldsymbol{\beta}$, the dispersion parameter ψ in the marginal distribution of N_{it}, as well as the vector of dependence parameters $\boldsymbol{\theta}$ in the elliptical copula.

Using an elliptical copula, the likelihood for the policyholder i could be shown as:

$$L_i(\boldsymbol{\beta}, \psi, \boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{U}_i}\left[\kappa_T|\boldsymbol{\Sigma}|^{-1/2} g_T\left(\frac{1}{2}\mathbf{V}_i^{*'}\boldsymbol{\Sigma}^{-1}\mathbf{V}_i^*\right)\left(\prod_{t=1}^{T}\kappa_1 g_1\left(\frac{1}{2}V_{it}^{*2}\right)\right)^{-1}\prod_{t=1}^{T}\Pr(N_{it}=n_{it})\right], \tag{13}$$

where $V_{it}^* = H_t^{-1}(F_{it}^*(n_{it}-U_{it}))$ and $\mathbf{V}_i^* = (V_{i1}^*, \cdots, V_{iT}^*)'$. This likelihood could be approximated by averaging over a large number of jitters:

$$\widetilde{L}_i(\boldsymbol{\beta}, \psi, \boldsymbol{\theta}) = \frac{1}{S}\sum_{s=1}^{S}\left[\kappa_T|\boldsymbol{\Sigma}|^{-1/2} g_T\left(\frac{1}{2}\mathbf{v}_{i,s}^{*'}\boldsymbol{\Sigma}^{-1}\mathbf{v}_{i,s}^*\right)\left(\prod_{t=1}^{T}\kappa_1 g_1\left(\frac{1}{2}v_{it,s}^{*2}\right)\right)^{-1}\prod_{t=1}^{T}\Pr(N_{it}=n_{it})\right], \tag{14}$$

where $\mathbf{v}_{is}^* = (v_{i1,s}^*, \cdots, v_{iT,s}^*)' = \left(H_1^{-1}(F_{i1}^*(n_{i1}-u_{i1,s}), \cdots, H_T^{-1}(F_{iT}^*(n_{iT}-u_{iT,s}))\right)'$ for $s = 1, \cdots, S$, and the $u_{it,s}$ for $t = 1, \cdots, T$ and $s = 1, \cdots, S$ are generated independently from uniform on (0,1). The total log likelihood function could be derived by summing up the logarithm of (14) over all policyholders. Note that though the above process assumes a balanced data set, the approach could be easily adapted for an unbalanced data where all policyholders are not observed during the complete sampling period. Assume policyholder i is observed for only T_i periods, a subset of the T periods. To account for the lack of balance, one could replace the copula in (12) with one of its subcopula, where each argument corresponds to each observation period. This can be naturally done in the case of elliptical copulas, where all subcopulas also belong to the same family. Specifically, one simply substitutes for the dispersion matrix $\boldsymbol{\Sigma}$ with $\boldsymbol{\Sigma}_i$, the corresponding submatrix of $\boldsymbol{\Sigma}$.

2.2.4 Copula Validation

The validation of copula specification has been widely discussed in the literature and many procedures have been proposed for general copulas, for example, see Fermanian (2005), Genest et al. (2006) and Scaillet (2007). A comprehensive review of goodness-of-fit test procedures for copulas can be found in Genest et al. (2009). To verify whether an elliptical copula is well specified, we adopt the t-plot method proposed by Sun et al. (2008). The t-plot exploits the properties of the elliptical distribution and is designed to evaluate the goodness-of-fit for the family of elliptical copulas. Shi (2011) modified the t-plot method for the case of unbalanced data and adapted the framework for both in-sample and out-of-sample validation purposes.

The null hypothesis of a t-plot is that a sample is from an elliptical multivariate distribution. In modeling the longitudinal insurance claim counts, we are interested in whether the copula in expression (10) is from the elliptical family. Suppose that policyholder i is observed for T_i ($\leq T$) periods. Following Shi (2011), such hypothesis could be tested according to the procedure below:

(i) Transform the claim counts to variables on $(0,1)$ by $\hat{\iota}_{it} = F^*_{it}(n_{it} - u_{it}; \hat{\boldsymbol{\beta}}, \hat{\psi})$ for $t = 1, \cdots, T_i$, where F^*_{it} is defined by equation (6) according to the negative binomial distribution in (11), and $\hat{\boldsymbol{\beta}}$ and $\hat{\psi}$ are maximum likelihood estimates of $\boldsymbol{\beta}$ and ψ, respectively. Under the null hypothesis, $\boldsymbol{\iota}_i = (\iota_{i1}, \cdots, \iota_{iT_i})$ is a realization of the hypothesized elliptical copula.

(ii) Compute the quantiles of $\hat{\iota}_{it}$ by $\hat{\zeta}_{it} = H_t^{-1}(\hat{\iota}_{it})$ for $t = 1, \cdots, T_i$, where H_t denotes the marginal distribution associated with the elliptical copula. Thus, if the copula is well-specified, $\hat{\boldsymbol{\zeta}}_i = (\hat{\zeta}_{i1}, \cdots, \hat{\zeta}_{iT_i})'$ follows the multivariate elliptical distribution of $E_{T_i}(\mathbf{0}, \boldsymbol{\Sigma}, g_{T_i})$.

(iii) Calculate vector $\hat{\boldsymbol{\zeta}}^*_i = (\hat{\zeta}^*_{i1}, \cdots, \hat{\zeta}^*_{iT_i})' = \widehat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\zeta}}_i$, and construct the t statistic for policyholder i:

$$t_i(\hat{\boldsymbol{\zeta}}^*_i) = \frac{\sqrt{T_i}\bar{\hat{\zeta}}^*_i}{\sqrt{(T_i-1)^{-1}\sum_{t=1}^{T_i}(\hat{\zeta}^*_{it} - \bar{\hat{\zeta}}^*_i)^2}},$$

with $\widehat{\boldsymbol{\Sigma}}$ the maximum likelihood estimator of $\boldsymbol{\Sigma}$ and $\bar{\hat{\zeta}}^*_i = T_i^{-1}\sum_{t=1}^{T_i}\hat{\zeta}^*_{it}$. Thus $t_i(\hat{\boldsymbol{\zeta}}^*_i)$ should be from a standard t distribution with $T_i - 1$ degrees of freedom.

(iv) Repeat steps (i) - (iii) for $i = 1, \cdots, I$, and calculate the t statistics $t_i(\hat{\boldsymbol{\zeta}}^*_i)$ for all policyholders in the sample. Define the transformed variable $\varsigma_i = G_{T_i-1}(t_i(\hat{\boldsymbol{\zeta}}^*_i))$, where G_{T_i-1} denotes the cumulative distribution function of a t distribution with $T_i - 1$ degrees of freedom. If the copula captures the dependence structure properly, $\boldsymbol{\varsigma} = (\varsigma_1, \cdots, \varsigma_n)'$ should be a random sample from a uniform random variable on (0,1). This can be easily verified using standard graphical tools or goodness-of-fit tests.

3 Data

In this paper, the longitudinal count regression models are calibrated using the number of claims in a portfolio of auto insurance from a general insurer in Singapore. Claim experience data for the insurance portfolio are collected over a period of nine years 1993-2001. Our data are obtained from the General Insurance Association of Singapore, a trade association representing general insurance companies in Singapore. Other studies that have examined Singapore auto insurance data include Frees and Valdez (2008), Frees et al. (2009), and Shi and Valdez (2011).

We restrict the analysis to "non-fleet" policies that provide coverage for a single policy only. Similar to several jurisdictions worldwide, Singapore requires drivers to have, at the minimum, a third party liability coverage to be able to drive a vehicle on the road, and at the same time, drivers have the liberty to choose beyond this minimum level of coverage. To construct a more homogeneous sample, we focus on policyholders with comprehensive coverage that account for the choice of most policyholders. Table 1 presents the claim frequency distribution over time. In each year, the claim counts distribution has a significant fraction of zeros. This is consistent with the insurance practice, where insures mange the risk pool through diversification effect. The number of policyholders changes over years, indicating the unbalanced nature of the data. The lack of balance could arise from many scenarios, for example, a customer might switch among different insurers, meaning an existing customer could leave the insurance or a new customer could join the insurer.

For this company, we note that the probability of accidents varies over time. Such observation could be related the insurer's underwriting standard. In practice, an insurance company could relax or restrict the underwriting benchmark in order to expand or shrink current business. Typically, a higher underwriting standard leads to more rejection of bad risks and thus a more favorable overall claim experience.

Table 1: Number and Percentage of Claims by Count and Year

	Percentage by Year									Overall	
Count	1993	1994	1995	1996	1997	1998	1999	2000	2001	Number	Percent
0	88.10	85.86	85.21	83.88	90.41	85.62	86.89	87.18	89.71	3480	86.9
1	10.07	12.15	13.13	14.29	8.22	13.73	11.59	11.54	9.71	468	11.7
2	1.47	2.00	1.25	1.83	0.00	0.65	1.37	0.92	0.57	50	1.25
3	0.37	0.00	0.21	0.00	1.37	0.00	0.15	0.18	0.00	6	0.15
4	0.00	0.00	0.21	0.00	0.00	0.00	0.00	0.18	0.00	2	0.05
Number	546	601	480	273	73	306	656	546	525	4006	100

A set of variables related to vehicle and policyholder characteristics comes along with the claim experience in the dataset, which helps explain and predict automobile accident frequency. These characteristics include the vehicle's age and brand, the driver's age, gender, and marital status, as well as an index of no claims discount (NCD). It is well-known that insurance company use experience rating to incorporate the policyholder's driving record into the ratemaking process, i.e., the occurrence of accident affect the policyholder's future premium. This design works as a mechanism to mitigate the moral hazard in a multi-period contract. A popular experience rating scheme is the bonus-malus system widely employed in Europe. Similar to the bonus-malus system, Singapore insurance industry adopts a NCD approach. The no claims discount scheme is introduced to encourage safe driving. An NCD increases 10% after a year without claims, and the maximum NCD level is 50%. If the NCD is at 50% or 40%, a claim during one year will reduce it to 30% or 20%, respectively. Two or more claims in one year can cause a complete lost of the policyholders current level of NCD. If the NCD is at 30% or lower levels, the entire discount is forfeited in case of accidents. Because clearly the NCD reveals the policyholders past driving records, we use it as a proxy to measure accident history.

Table 2 displays the effects of person level characteristics, age, gender, and NCD, on the claim frequency, suggesting the significant effects of these three variables. Though accounting for a small percentage in the insurance portfolio, young drivers has the higest accident rate, as shown in most automobile studies. In the meanwhile, older policyholders are found to be safer drivers. Table 2 also demonstrates the strong gender effects, with female drivers having lower accident rate, presumably due to their higher degree of risk aversion that is usually associated with defensive driving behavior. As expected, the drivers with lower NCD have more accidents, as predicted by the poor previous driving records reflected in NCD. Although not reported here, we also investigated the interaction effects among policyholder's characteristics.

Table 2: Number and Percentage of Claims by Age, Gender and NCD

	Percentage by Count					Overall	
	0	1	2	3	4	Number	Percent
Person Age (in years)							
25 and younger	73.33	23.33	3.33	0.00	0.00	30	0.75
26-35	87.49	11.12	1.19	0.10	0.10	1007	25.14
36-45	86.63	11.80	1.35	0.17	0.06	1780	44.43
46-60	86.85	11.92	1.05	0.18	0.00	1141	28.48
60 and over	91.67	6.25	2.08	0.00	0.00	48	1.2
Gender							
Female	91.49	7.98	0.53	0.00	0.00	188	4.69
Male	86.64	11.86	1.28	0.16	0.05	3818	95.31
No Claims Discount (NCD)							
0	84.83	13.17	1.61	0.26	0.13	1549	38.67
10	86.21	12.58	1.20	0.00	0.00	747	18.65
20	89.21	9.25	1.54	0.00	0.00	584	14.58
30	89.16	9.49	1.08	0.27	0.00	369	9.21
40	88.60	11.40	0.00	0.00	0.00	193	4.82
50	88.83	10.46	0.53	0.18	0.00	564	14.08
Number by Count	3480	468	50	6	2	4006	100

4 Inference

This section presents the results on model estimation and inference. The observations of years 1993-2000 are used to calibrate the model, and the observations in year 2001 are reserved for model validation purposes. We show that the copula model with jitters outperforms standard longitudinal count regression models.

4.1 Estimation Results

Both the random effects models and the proposed copula model are fitted with the longitudinal claim counts described in Section 3, and the parameter estimates are exhibited in Table 3 and Table 4, respectively. The preliminary analysis, in which the effects of both vehicle and driver characteristics as well as their interactions are examined, suggests a set of important covariates for the regression analysis: *young* is a binary variable indicating whether a driver is below 25 or not. The interaction term *midfemale* refers to mid-aged (in age range 30-50) female drivers. The NCD level is also reflected via a binary variable *zeroncd*, which equals one if the policyholder does not enjoy any discount in premium and zero otherwise. For simplicity, vehicle age *vage* is included as a continuous variable. In addition, we classify the brand of vehicles into various categories, and indicator variables *vbrand1* and *vbrand2* correspond to two of them.

Four random effects count regression models are considered in Table 3. As discussed above, many possible distributions for the random effect could be chosen. Here we use a gamma random effect for the Poisson model and a beta random effect for the negative binomial model, because these

formulations lead to a closed form solution for the joint distribution of insurance claim counts. In fact, we also estimated both models with a normal random effect and found no substantial difference in the estimated regression coefficients. A normal distributed random effect is employed in the zero-inflated (Poisson and negative binomial) models. The estimates in Table 3 shows the significant effects of covariates. As anticipated, young drivers are bad risks and have on average higher claim frequency. In contrast, mid-aged female category drives more safely and have a lower probability of accident. The policyholders with zero NCD, most possibly because of the poor driving history, have a higher accident rate. The effect of vehicle age is statistically significant, though small. The negative sign is consistent with the moral hazard interpretation that the driver of an older car has higher moral hazard risk than whom of a newer vehicle. The small size could be attributed to the potential nonlinear effect and the possible correlation between the vehicle age and driving experience. Note that the intercept in the negative binomial model is not comparable to other models. For comparison purposes, the intercepted must be adjusted to reflect the factor $b/(a-1)$. In fact, after this modification, the intercept is -1.7138, which is close to other random effects models.

Table 3: Estimates of standard longitudinal count regression models

	RE-Poisson		RE-NegBin		RE-ZIP		RE-ZINB	
Parameter	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
intercept	-1.7173	<.0001	1.6404	0.1030	-1.6780	<.0001	-1.7906	<.0001
young	0.6408	0.0790	0.6543	0.0690	0.6232	0.0902	0.6371	0.0853
midfemale	-0.7868	0.0310	-0.7692	0.0340	-0.7866	0.0316	-0.7844	0.0319
zeroncd	0.2573	0.0050	0.2547	0.0060	0.2617	0.0051	0.2630	0.0050
vage	-0.0438	0.0210	-0.0442	0.0210	-0.0436	0.0227	-0.0438	0.0224
vbrand1	0.5493	<.0001	0.5473	<.0001	0.5481	<.0001	0.5478	<.0001
vbrand2	0.1831	0.0740	0.1854	0.0710	0.1813	0.0777	0.1827	0.0755
LogLik	-1498.40		-1497.78		-1498.00		-1497.50	
AIC	3012.81		3013.57		3016.00		3017.00	
BIC	3056.41		3062.62		3070.50		3077.00	

To compare the performance of random effects models, we present the goodness-of-fit measures at the bottom of Table 3, including the log likelihood function evaluated at the estimated parameters, two commonly used model selection criteria, Akaike information criterion (AIC) and Bayesian information criterion (BIC). The goodness-of-fit statistics suggest negligible difference among these models. In particular, zero-inflated count data models do not demonstrate any advantage in modeling the claim counts for this particular insurer, implying that the excess zeros are fairly captured by the standard count distributions. The individual heterogeneity and overdispersion are accommodated by the random effect, which explains the comparable performance of the Poisson and negative binomial models.

Table 4 summarizes the results for the copula model with jitters. We report the parameter estimates and goodness-of-fit statistics for the Gaussian copula with three different dependence structures, AR(1), exchangeable, and banded Toeplitz. Both model selection criteria (a smaller

value) support the copula formulation compared with random effects models. Within the copula models, it seems that the model fit does not gain much from the extra parameter in the Toeplitz dependence. The best fit of the exchangeable dependence demonstrates the strong individual heterogeneity. The estimates of regression coefficients are consistent with economic hypothesis and comparable with Table 3.

It is worth stressing that the association parameter in the copula is about 0.12 and statistically significant. First, we cannot interpret this parameter directly because the copula is essentially used for the joint distribution of the jittered continuous variables. One solution is to translate it to concordance-based measures, for example, the corresponding Kendall's *tau* is 0.07. According to Denuit and Lambert (2005), the concordance-based association measures are reserved for the discrete variables. Thus one could interpret the Kendall's *tau* as the association measure among the discrete claim counts. Second, the significance of the association parameter implies strong unobserved individual effects. Such unobservable factors could be either the policyholder's inherent riskiness or the policyholder's driving attitude. On one hand, the serial association could arise from residual risk factors, though insurers use risk classification to identify the risk level of policyholder. On the other hand, the moral hazard effects could also cause serial dependence among claims, i.e., the coverage of insurance provides policyholders incentives for reckless driving behavior. Third, the small association agrees with our expectation, although one expects to observe correlation among claim counts mainly due to unobservable factors, be it selection or moral hazard effects. This is because when experience rating (i.e., the NCD scheme in Singapore) is used, the individual heterogeneity is significantly mitigated in a multi-period contract. Thus, it is not surprising to observe a weak association in the claim counts from an insurer with experience rating.

Table 4: Estimates of copula model with various dependence structures

	AR(1)		Exchangeable		Toeplitz(2)	
Parameter	Estimate	StdErr	Estimate	StdErr	Estimate	StdErr
intercept	-1.8028	0.0307	-1.8422	0.0353	-1.7630	0.0284
young	0.6529	0.0557	0.7130	0.0667	0.6526	0.0631
midfemale	-0.6956	0.0588	-0.6786	0.0670	-0.7132	0.0596
zeroncd	0.2584	0.0198	0.2214	0.0172	0.2358	0.0176
vage	-0.0411	0.0051	-0.0422	0.0056	-0.0453	0.0042
vbrand1	0.5286	0.0239	0.5407	0.0275	0.4962	0.0250
vbrand2	0.1603	0.0166	0.1752	0.0229	0.1318	0.0198
ϕ	2.9465	0.1024	2.9395	0.1130	2.9097	0.1346
ρ_1	0.1216	0.0028	0.1152	0.0027	0.1175	0.0025
ρ_2					0.0914	0.0052
LogLik	-1473.25		-1454.04		-1468.74	
AIC	2964.49		2926.08		2957.49	
BIC	3013.55		2975.13		3011.99	

4.2 Model Validation

As part of the model validation process, we evaluate the goodness-of-fit of the Gaussian copula. The modified t-plot method discussed in Section 2.2.4 is implemented and the uniform qq-plot is displayed in Figure 1. The linear trend along the 45 degree line provides evidence that the Gaussian copula is a suitable model for the dependency. A statistical test is performed for sample correlation. The correlation coefficient between the sample and theoretical quantiles is 0.937. Based on 5,000 simulation, the p-value for the correlation is 0.453, indicating the nonsignificant difference between the empirical and theoretical distributions.

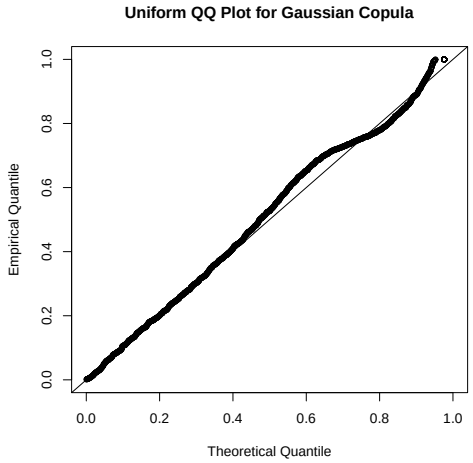

Figure 1: t-plot for the Gaussian copula with exchangeable dependence.

Finally, we compare different count regression models using hold-out observations in year 2001. Let N_{iT+1} denote the number of insurance claims in the testing period for the ith policyholder, $i = 1, \ldots, J$, with J representing the size of the hold-out sample. Given the realizations of N_{iT+1}, denoted by n_{iT+1}, three performance measures, log likelihood function (LogLik), mean squared error of prediction (MSPE), and mean absolute error of prediction (MSPE), are calculated according to:

- LogLik $= \sum_{i=1}^{J} \log\left(f_{iT+1}(n_{iT+1}|n_{i1}, \cdots, n_{iT_i})\right)$
- MSPE $= \sum_{i=1}^{J} \left[n_{iT+1} - \mathrm{E}(N_{iT+1}|N_{i1} = n_{i1}, \cdots, N_{iT_i} = n_{iT_i})\right]^2$
- MAPE $= \sum_{i=1}^{J} \left|n_{iT+1} - \mathrm{E}(N_{iT+1}|N_{i1} = n_{i1}, \cdots, N_{iT_i} = n_{iT_i})\right|$

Here, $f_{iT+1}(n_{iT+1}|n_{i1}, \cdots, n_{iT_i})$ represents the predictive distribution of claim counts given the past claim experiences. We calculate this quantity for the copula count model using the following relationship:

$$\begin{aligned} & f_{iT+1}(n_{iT+1}|n_{i1}, \cdots, n_{iT_i}) \\ =& \Pr(N_{iT+1} = n_{iT+1}|N_{i1} = n_{i1}, \cdots, N_{iT} = n_{iT_i}) \\ =& \frac{\mathrm{E}_{\boldsymbol{U}_i}\left(c(F^*_{i1}, \cdots, F^*_{iT_i}, F^*_{iT+1}; \boldsymbol{\theta}) f^*_{iT+1} \prod_{t=1}^{T_i} f^*_{it}\right)}{\mathrm{E}_{\boldsymbol{U}_i}\left(c(F^*_{i1}, \cdots, F^*_{iT_i}; \boldsymbol{\theta}) \prod_{t=1}^{T_i} f^*_{it}\right)}. \end{aligned} \tag{15}$$

where $F_{it}^* = F_{it}^*(n_{it} - U_{it})$ and $f_{it}^* = f_{it}^*(n_{it} - U_{it})$. Again, the evaluation of equation (15) involves averaging over the jitters. In the calculation of the performance measures, a large number of Monte Carlo simulations are used to ensure the convergence of the predictive distribution.

Table 5 displays the three performance measures for the copula models with different dependence structures. For comparison purposes, we also report the results for the random effects Poisson and negative binomial models. Consistent with the goodness-of-fit statistics presented in Section 4.1, the difference is insubstantial within each model category. However, noticeable difference is observed between the two model categories: the MSPE of copula models is comparable to that of random effects models, while the MAPE of copula models is less than random effect models. Further more, the copula models have larger likelihood for the hold-out observations than standard count regression models. The favorable results of copula models could be attributed to the more information, especially the dynamic associations, captured by a copula function.

Table 5: Model validation based on hold-out sample

	Standard Model		Copula Model		
	RE-Poisson	RE-NegBin	AR(1)	Exchangeable	Toeplitz(2)
LogLik	-177.786	-177.782	-168.037	-162.717	-165.932
MSPE	0.107	0.107	0.108	0.105	0.110
MAPE	0.213	0.213	0.197	0.186	0.192

5 Concluding Remarks

Modeling claim counts is a very important component in property and casualty insurance ratemaking. As claims are observed over a period of time, the insurer may be able to better evaluate the inherent risk characteristics of a policyholder and to adjust premiums as necessary. The type of data observed is longitudinal and with it comes the relevance of being able to frame a model that will help capture the intertemporal dependence in the pattern of claims. In this article, we investigated the usefulness of a different class of joint regression models for longitudinal count data that might provide better insight in understanding the possible association of insurance claim counts observed over time.

The most popular approach for modeling longitudinal count data is the class of random effects regression models. Within this class, the idea is to specify the class of distribution models for the count data and simultaneously, introduce random effects to capture the intertemporal dependence of the observations. The distribution of the random effects is typically pre-specified. Effectively, for insurance claims data, this has the natural interpretation of captivating the unobservable heterogeneity inherent for an individual policyholder. For comparison purposes, we examined random effects Poisson, negative binomial, and zero-inflated models.

Another approach for longitudinal count data, which has recently increased in popularity, involves the use of copula regression models. Copulas link the marginals in the observable longitudinal counts to its joint distribution function. This class of models provides for a rather flexible approach

to explicitly measure the possible presence of dependence in the count data observed over time. As we cited in the introduction, this approach has been used in disciplines such as economics, biostatistics, and actuarial science. Our approach falls within this framework.

Direct specification of a copula to a multivariate count data has been debatable because of non-uniqueness of the copula and the ambiguity of the implications of the dependence. The main difference of our approach is the indirect specification of copula regression models to the longitudinal insurance claim counts. Specifically, the proposed approach considers the application of copula regression models to the jittered multivariate claim count. The concept of jittering involves integration of a continuous random variable to each component in the multivariate claim count, thereby converting the discrete to continuous form. This technique avoids the debatable issue of calibrating copula models directly to multivariate discrete data. It has the additional advantage that with the continuous extension with jitters, the dependence structure of the original discrete data is preserved. This approach has been suggested by Denuit and Lambert (2005) and has been empirically explored in Madsen and Fang (2010).

To empirically demonstrate the usefulness of this jittered approach, we used a dataset with a grand total of 4,006 observations from a portfolio of automobile insurance obtained from a general insurer in Singapore. To account for the observable heterogeneity of the policyholder, we injected covariates through the marginal count distributions. Because of its flexibility to accommodate various form of dependence structure, we examined the class of elliptical copulas and in particular, we emphasized the usefulness of the Gaussian copula and investigated three different dependence structures: AR(1), exchangeable, and banded Toeplitz. Model parameters were estimated using the method of maximum likelihood. The model with exchangeable dependence structure appears to provide the best quality fit, although the results indicate only weak association. However, this is not atypical for insurance claim count data. In addition to, when it comes to prediction with hold-out samples, the continuous extension with jitters approach outperforms traditional random effects models. This indicates that the copula regression models with jitters proposed in this paper is not any less superior than conventional methods.

Finally, we would like to remark that there is a vast alternative models that may be suitable for longitudinal insurance claim counts data. One possible approach is to use a class of models for time series data. This class of models, although less popular for insurance claim counts, has been discussed in Cameron and Trivedi (1998) and it includes models with analogy to continuous time series data. Models widely discussed include for example, integer-valued ARMA models, autoregressive models, as well as serially correlated error models. The main appeal with time series models is its ability to capture trends and seasonality which otherwise would not be captured by models we discussed here. On the other hand, the primary disadvantage is the use of more complex methods to achieve efficient estimates of parameters. Several other classes of models are also explored in Molenberghs and Verbeke (2005). It will be interesting to examine these other classes of models and compare them with the continuous extension with jitters proposed in this paper. This is most notably suitable for further work.

References

Antonio, K. and E. A. Valdez (2011). Statistical concepts of a priori and a posteriori risk classification in insurance. *Advances in Statistical Analysis*. To appear.

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

Boucher, J.-P. and M. Guillén (2009). A survey on models for panel count data with applications to insurance. *RACSAM 103*(2), 277–294.

Cameron, A. C., L. Tong, P. K. Trivedi, and D. M. Zimmer (2004). Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. *Econometrics Journal 7*(2), 566–584.

Cameron, A. C. and P. K. Trivedi (1998). *Regression Analysis of Count Data*. Cambridge University Press: United Kingdom.

Denuit, M. and P. Lambert (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis 93*(1), 40–57.

Denuit, M., X. Maréchal, S. Pitrebois, and J.-F. Walhin (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. John Wiley & Sones Ltd.: West Sussex, England.

Dionne, G. and C. Vanasse (1992). Automobile insurance ratemaking in the presence of asymmetrical information. *Journal of Applied Econometrics 7*(2), 149–165.

Fang, K.-T., S. Kotz, and K. W. Ng (1990). *Symmetric Multivariate and Related Distributions*. Chapman & Hall/CRC.

Fermanian, J. (2005). Goodness-of-fit tests for copulas. *Journal of Multivariate Analysis 95*(1), 119–152.

Frees, E. and P. Wang (2006). Copula credibility for aggregate loss models. *Insurance: Mathematics and Economics 38*(2), 360–373.

Frees, E. W. (2010). *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press.

Frees, E. W., P. Shi, and E. A. Valdez (2009). Actuarial applications of a hierarchical insurance claims model. *ASTIN Bulletin 39*(1), 165–197.

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

Genest, C. and J. Nešlehová (2007). A primer on copulas for count data. *ASTIN Bulletin 37*(2), 475–515.

Genest, C., J. Quessy, and B. Rémillard (2006). Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scandinavian Journal of Statistics 33*(2), 337–366.

Genest, C., B. Rémillard, and D. Beaudoin (2009). Goodness-of-fit tests for copulas: a review and a power study. *Insurance: Mathematics and Economics 44*(2), 199–213.

Gourieroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood methods: Theory. *Econometrica 52*(3), 681–700.

Greenwood, M. and G. Yule (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society Series A 83*(2), 255–279.

Hausman, J., B. Hall, and Z. Griliches (1984). Econometric models for count data with an application to the patents-r&d relationship. *Econometrica 52*(4), 909–938.

Jorgenson, D. (1961). Multiple regression analysis of a poisson process. *Journal of the American Statistical Association 56*(294), 235–245.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā Series A 32*(4), 419–430.

Klugman, S. A., H. H. Panjer, and E. Willmot, Gordon (2008). *Loss Models: From Data to Decisions* (Third ed.). John Wiley & Sons, Inc.

Lambert, D. (1992). Zero-inflated poisson regression with an application to defects in manufacturing. *Technometrics 34*(1), 1–14.

Landsman, Z. M. and E. A. Valdez (2003). Tail conditional expectations for elliptical distributions. *North American Actuarial Journal 7*(4), 55–71.

Madsen, L. and Y. Fang (2010). Joint regression analysis for discrete longitudinal data. *Biometrics*. To appear.

Molenberghs, G. and G. Verbeke (2005). *Models for Discrete Longitudinal Data*. Springer.

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

Prieger, J. E. (2002). A flexible parametric selection model for non-normal data with application to health care usage. *Journal of Applied Econometrics 17*(4), 367–392.

Purcaru, O. and M. Denuit (2003). Dependence in dynamic claim frequency credibility models. *ASTIN Bulletin 33*(1), 23–40.

Scaillet, O. (2007). Kernel-based goodness-of-fit tests for copulas with fixed smoothing parameters. *Journal of Multivariate Analysis 98*(3), 533–543.

Shi, P. (2011). Multivariate longitudinal modeling of insurance company expenses. *Insurance: Mathematics and Economics*. To appear.

Shi, P. and E. Frees (2010). Long-tail longitudinal modeling of insurance company expenses. *Insurance: Mathematics and Economics 47*(3), 303–314.

Shi, P. and E. A. Valdez (2011). A copula approach to test asymmetric information with applications to predictive modeling. *Insurance: Mathematics and Economics 49*(2), 226–239.

Song, P. X.-K., M. Li, and Y. Yuan (2009). Joint regression analysis of correlated data using gaussian copulas. *Biometrics 65*(1), 60–68.

Sun, J., E. W. Frees, and M. A. Rosenberg (2008). Heavy-tailed longitudinal data modeling using copulas. *Insurance: Mathematics and Economics 42*(2), 817–830.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. Springer-Verlag: Berlin.

Zimmer, D. M. and P. K. Trivedi (2006). Using trivariate copulas to model sample selection and treatment effects: application to family health care demand. *Journal of Business and Economic Statistics 24*(1), 63–76.