Insurance: Mathematics and Economics 55 (2014) 18–29

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Multivariate negative binomial models for insurance claim counts

Peng Shi [a,*], Emiliano A. Valdez [b]

[a] Department of Actuarial Science, Risk Management, and Insurance, School of Business, University of Wisconsin - Madison, Madison, WI 53706, United States

[b] Department of Mathematics, Michigan State University, East Lansing, MI 48824, United States

HIGHLIGHTS

- We investigate multivariate count models using negative binomial distributions.
- We consider two class of modeling framework using different families of copulas.
- The first models the discrete count data directly using a mixture of max-id copulas.
- The second one employs elliptical copulas to join continuitized count data.
- We show the advantage of the copula approach over the common shock model.

ARTICLE INFO

Article history:
Received July 2013
Received in revised form
November 2013
Accepted 27 November 2013

Keywords:
Negative binomial distribution
Insurance claim count
Copula
Jitter
Multivariate model

ABSTRACT

It is no longer uncommon these days to find the need in actuarial practice to model claim counts from multiple types of coverage, such as the ratemaking process for bundled insurance contracts. Since different types of claims are conceivably correlated with each other, the multivariate count regression models that emphasize the dependency among claim types are more helpful for inference and prediction purposes. Motivated by the characteristics of an insurance dataset, we investigate alternative approaches to constructing multivariate count models based on the negative binomial distribution. A classical approach to induce correlation is to employ common shock variables. However, this formulation relies on the NB-I distribution which is restrictive for dispersion modeling. To address these issues, we consider two different methods of modeling multivariate claim counts using copulas. The first one works with the discrete count data directly using a mixture of max-id copulas that allows for flexible pair-wise association as well as tail and global dependence. The second one employs elliptical copulas to join continuitized data while preserving the dependence structure of the original counts. The empirical analysis examines a portfolio of auto insurance policies from a Singapore insurer where claim frequency of three types of claims (third party property damage, own damage, and third party bodily injury) are considered. The results demonstrate the superiority of the copula-based approaches over the common shock model. Finally, we implemented the various models in loss predictive applications.

1. Introduction

Modeling insurance claim counts is a critical component in the rakemaking process for property-casualty insurers. Typically insurance companies keep a comprehensive record of the claim history of their customers and have access to an additional set of personal information. The frequency of claim counts to a great extent reveals the riskiness of the insureds. Thus by examining the relation between claim counts and policyholders' characteristics, the insurer classifies the policyholders and determines the fair premium according to their risk level. For example, in the standard frequency-severity framework, one examines the number of claims and then the size of each claim given occurrence. In addition, the insurer could detect the presence of private information such as moral hazard and adverse selection through analyzing the claim behavior of policyholders, which is useful for the design of an insurance contract.

In practice, it is not uncommon for an insurer to observe claim counts of multiple types from a policyholder when various types of coverage are bundled into one single policy. For example, a homeowner's insurance could compensate losses from multiple perils, an automobile insurance could offer protection against third party and own damages, and so on. Our goal is to develop multivariate count regression models that accommodate dependency among different claim types.

Following the seminal contribution of Jorgenson (1961), count data regression techniques have been greatly extended and applied in various fields of studies. In general, three classes of

* Corresponding author. Tel.: +1 6083201366.
E-mail addresses: pshi@bus.wisc.edu (P. Shi), valdezea@math.msu.edu (E.A. Valdez).

0167-6687/$ – see front matter
http://dx.doi.org/10.1016/j.insmatheco.2013.11.011

approaches are discussed in the literature: the semi-parametric approach based on the pseudolikelihood method (Nelder and Wedderburn, 1972; Gourieroux et al., 1984), the parametric count regression models (Hausman et al., 1984), and the quantile regression that is a non-parametric approach (Machado and Silva, 2005). Comprehensive reviews for count regression can be found in Cameron and Trivedi (1998) and Winkelmann (2008). One straightforward approach of introducing correlation amongst multivariate count outcomes is through a common additive error. Kocherlakota and Kocherlakota (1992) and Johnson et al. (1997) provided a detailed discussion for the one-factor multivariate Poisson model. Along this line of study, Winkelmann (2000) proposed a multivariate negative binomial regression to account for overdispersion, Karlis and Meligkotsidou (2005) considered a multivariate Poisson model with a combination of common shocks to allow for a more flexible covariance structure, and Bermúdez and Karlis (2011) examined zero-inflated versions of the Poisson model. An alternative way to incorporate correlation among count data is the mixture model with a multiplicative error that captures unobserved individual-specific heterogeneity, for example, see Hausman et al. (1984) and Dey and Chung (1992). A common limitation of the above models is that the covariance structure is restricted to non-negative correlation. Such a limitation could be addressed through multi-factor models with examples that include the multivariate Poisson-log-normal model (Aitchison and Ho, 1989) and the latent Poisson-normal model (van Ophem, 1999). An emerging approach to constructing a general discrete multivariate distribution to support complex correlation structures is to use copulas. Despite of its popularity in dependence modeling, the application of copulas for count data is still in its infancy (Genest and Nešlehová, 2007). A relevant strand of studies of multivariate count data is regarding the longitudinal data. Unlike the genuine multivariate outcomes, longitudinal data usually has a large cross-sectional dimension but a small time dimension. See Boucher et al. (2008) for a recent survey on models of insurance claim counts with time dependence.

For purposes of risk classification and predictive modeling, we are more interested in the entire conditional distribution just as in many other applied research. Thus our study will limit to the parametric modeling framework that is based on probabilistic count distributions. Motivated by the characteristics of a claim data set from a Singapore automobile insurer, we are particularly interested in models based on negative binomial distributions.

A count variable N is known to follow a negative binomial (NB) distribution if its probability function could be expressed as

$$\Pr(N=n)=\frac{\Gamma(\eta+n)}{\Gamma(\eta)\Gamma(n+1)}\left(\frac{1}{1+\psi}\right)^{\eta}\left(\frac{\psi}{1+\psi}\right)^{n},$$

for $n=0,1,2,\ldots$

and is denoted as $N\sim \mathrm{NB}(\psi,\eta)$ for $\psi,\eta>0$. The mean and variance of N are $\mathrm{E}(N)=\eta\psi$ and $\mathrm{Var}(N)=\eta\psi(1+\psi)$, respectively. Compared with the Poisson distribution, the negative binomial accommodates overdispersion via parameter ψ. As $\psi\to 0$, overdispersion vanishes and the negative binomial converges to the Poisson distribution. For regression purposes, it is helpful to consider a mean parameterization that could be specified in terms of covariates as in $\lambda=\eta\psi=\exp(\boldsymbol{x}'\boldsymbol{\beta})$, where $\boldsymbol{x}$ denotes the vector of explanatory variables and $\boldsymbol{\beta}$ denotes the vector of regression coefficients. Then the negative binomial regression could come in two different parameterizations. The NB-I model is obtained for $\eta=\sigma^{-2}\exp(\boldsymbol{x}'\boldsymbol{\beta})$ and takes the form

$$f^{\text{NB-I}}(n|\boldsymbol{x};\boldsymbol{\beta},\sigma^2)=\frac{\Gamma(\sigma^{-2}\exp(\boldsymbol{x}'\boldsymbol{\beta})+n)}{\Gamma(\sigma^{-2}\exp(\boldsymbol{x}'\boldsymbol{\beta}))\Gamma(n+1)}\times\left(\frac{1}{1+\sigma^2}\right)^{\sigma^{-2}\exp(\boldsymbol{x}'\boldsymbol{\beta})}\left(\frac{\sigma^2}{1+\sigma^2}\right)^{n}.$$

The NB-II model is obtained for $\eta=\sigma^{-2}$ and takes the form

$$f^{\text{NB-II}}(n|\boldsymbol{x};\boldsymbol{\beta},\sigma^2)=\frac{\Gamma(\sigma^{-2}+n)}{\Gamma(\sigma^{-2})\Gamma(n+1)}\left(\frac{1}{1+\sigma^2\exp(\boldsymbol{x}'\boldsymbol{\beta})}\right)^{\sigma^{-2}}\times\left(\frac{\sigma^2\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\sigma^2\exp(\boldsymbol{x}'\boldsymbol{\beta})}\right)^{n}.$$

Though both models assume the same mean structure of the count variable, their difference could be characterized in terms of a dispersion function ϕ such that $\mathrm{Var}(N|\boldsymbol{x})=\phi\mathrm{E}(N|\boldsymbol{x})$. The NB-I model implies a constant dispersion $\phi=1+\sigma^2$, while the NB-II model allows for subject heterogeneity in the dispersion $\phi=1+\sigma^2\exp(\boldsymbol{x}'\boldsymbol{\beta})$ (see Winkelmann, 2008).

We examine three methods of constructing multivariate negative binomial models that allow for flexible pair-wise association and explore the possibility of using either NB-I or NB-II formulations. The first approach is to use common shock variables. Since this method relies on the additivity of the count distribution, only NB-I is suitable for this formulation. The other two approaches are based on parametric copulas: one working with the discrete count data directly with the mixture of max-id copulas that allows for flexible pair-wise association as well as global dependence, and the other employing elliptical copulas to join continuitized data while preserving the dependency among the original counts. In the empirical analysis, we look into an insurance portfolio from a Singapore auto insurer where claim frequency of three types of claims (third party property damage, own damage, and third party bodily injury) are considered, and we show that the copula-based approaches outperform the common shock model.

The rest of the paper focuses on the theory and applications of multivariate negative binomial regression models, and it is structured as follows: Section 2 describes the motivating dataset of insurance claim counts from a Singaporean automobile insurer. Section 3 briefly discusses the multivariate model using the combination of common shock variables. Section 4 explores the possibilities of using copulas to construct multivariate models with flexible dependence structures. The estimation and inference results are summarized in Section 5. Section 6 presents predictive applications and compares the performance of alternative models. Section 7 concludes the paper.

2. Data structure and characteristics

The motivating dataset of insurance claim counts is from a major automobile insurer in Singapore. According to the General Insurance Association of Singapore (GIA), automobile insurance is one of the largest lines of business underwritten by general insurers and the gross premium income accounts for over one third of the entire insurance market.

As in most developed countries, automobile insurance protects insureds from various types of financial losses. The protection in Singapore comes in hierarchies. The minimum level of protection, which is also a mandatory coverage for all car owners, covers death or bodily injury to third parties. Although not mandated by law, third party coverage often provides protection against the costs that may arise as a result of damage to third party properties. On top of third party benefits, fire and theft, the policy also covers damage from these respective hazards. The maximum protection is offered by a comprehensive policy, which additionally compensates for losses of the insured vehicle, and in many cases, the associated medical expenses for the insured.

To study the dependency among claim types and also to construct a homogeneous portfolio of policyholders, we limit the analysis to those individuals with comprehensive coverage. Our final sample includes one year of observation of 9739 individuals. For a given accident, there are three possible types of claims,