RACSAM
Rev. R. Acad. Cien. Serie A. Mat.
VOL. **103 (**2), 2009, pp. 277–294
Matemática Aplicada / Applied Mathematics
Artículo panorámico / Survey

A survey on models for panel count data with applications to insurance

Jean-Philippe Boucher and Montserrat Guillén

Abstract. In insurance the expected number of claims per year given the observed characteristics of the covered risk is the basis for setting the price of a policy. Companies accumulate information of clients along several years, but in practice the panel data structure is not exploited. We review panel count data models that are useful in this context and present a new alternative based on the generalization of a compound sum.

Una revisión de los modelos para paneles de datos de enumeración con aplicaciones a seguros

Resumen. En seguros, el número esperado de reclamaciones por año dadas las características del riesgo cubierto es la base para establecer el precio de una póliza. Las compañías de seguros acumulan información de clientes a lo largo de varias anualidades, pero en la práctica la estructura de panel de los datos no se explota. Revisamos los modelos para paneles de datos de enumeración que son útiles en este contexto y presentamos una nueva alternativa basada en la generalización de una suma compuesta.

1 Introduction

Modeling count data is an essential part of insurance pricing. Count regression analysis allows identification of risk factors and prediction of the expected frequency of claims given the characteristics of the policyholders. An insurance premium is the amount of money that a client will pay to receive insurance coverage. A usual way to calculate the premium is to obtain the conditional expectation of the number of claims given the observable risk characteristics and to combine it with the expected claim amount or economic loss.

The literature on count regression analysis has grown considerably in the past years. Readers can overview the more general context of models for discrete longitudinal data in the text by Molenberghs and Verbeke [24, (2005)] where many applications and computational issues are discussed. For instance, binary longitudinal data or even the way of handling incomplete longitudinal data, where the observational period is not the same for all individuals. Another example is the article on the analysis of longitudinal count data with serial correlation (Xu et al. [36, (2007)]) using a state space model in an application to medical data. A description of univariate time series count models can be found in the classical book of Cameron and Trivedi [10, (1998)].

Presentado por / Submitted by José Garrido.
Recibido / Received: 15 de enero de 2009. *Aceptado / Accepted*: 4 de marzo de 2009.
Palabras clave / Keywords: Panel data, random effects, conditional distribution, zero-inflated distribution, hurdle distribution, compound sum.
Mathematics Subject Classifications: 62P05, 60E05, 62J99.

In Denuit et al. [12, (2007)] the reader can find a comprehensive review of count data models for cross-sectional data and its applications to automobile insurance rate-making. In this paper we address panel count data models in the context of insurance, so that we can see the advantages of using the information on each policyholder along time for modeling the number of claims. The existing literature has mainly advocated the use of the classical Poisson regression approach, but little has been said on other model alternatives and on model selection. We argue that new panel data models presented here that allow for time dependence between observations are closer to the data generating process that one can find in practice. Moreover, closed form expressions for future premiums given the past observations and model selection will definitely help practitioners to find suitable alternatives for modeling insurance portfolios that have accumulated some years of history.

Consider an insurance policy i over T consecutive years for which the number of claims reported to the insurance company are observed. Most often there is no accident, so that the observed number of claims is zero and note also that accidents may not be reported to the insurer as discussed in Boucher et al. [7, (2009)]. Let us call the vector of random variables $(N_{i,1}, \ldots, N_{i,T})$, which is the random counts to be modeled. For each individual policy i and year t, $t = 1, \ldots, T$, some covariate information exists, because the insurer knowns a vector of observable characteristics $(\mathbf{x}_{i,t}$ related to the individual. In automobile insurance, this is information on the insured driver and the insured vehicle. There are also characteristics that cannot be observed but they influence the number of accidents and therefore the number of claims. Examples of unobservable variables are swiftness of reflexes or respect of the driving code. The unobservable part of the model is called the random effect. We assume that the random effect is different for each individual and constant over time, so we will denote it by θ_i. Given the individual-specific random effect term θ_i, the claims $N_{i,1}$, $N_{i,2}$, ..., $N_{i,T}$ for each time period are independent. We assume that the covariates are independent from the random effects (see Mundlak [27, (1978)] or Hsiao [20, (2003)] for a general review).

A wide selection of models can be used to model the dependence that can exist between contracts of the same insured, see Boucher et al. [6, (2008)] for example. To account for this dependence that can exist between all the contracts of the same insured, one of the most popular way is the use of a common individual term (Hausman et al. [17, (1984)]).

This modeling has some natural interpretation for insurance data. Indeed, insurance data exhibit some variability that may be caused by the lack of information on some important classification variables (swiftness of reflexes, aggressiveness behind the wheel, consumption of drugs or drinking habits). These hidden features are usually captured by the individual random heterogeneity term. Using the notation of Hausman et al. [17, (1984)], the joint probability function of $N_{i,1}$, ..., $N_{i,T}$ is thus given by:

$$\begin{aligned}\Pr[N_{i,1}=n_{i,1},\ldots,N_{i,T}=n_{i,T}] &= \int_0^\infty \Pr[N_{i,1}=n_{i,1},\ldots,N_{i,T}=n_{i,T}|\theta_i]\, g(\theta_i)\,\mathrm{d}\theta_i \\ &= \int_0^\infty \left(\prod_{t=1}^{T} \underbrace{\Pr[N_{i,t}=n_{i,t}|\theta_i]}_{\text{Conditional Distribution}}\right) \underbrace{g(\theta_i)}_{\text{Random Effects Distribution}} \mathrm{d}\theta_i, \end{aligned} \tag{1}$$

where $g(\theta_i)$ is the density of θ_i, i represents an individual unit and t is the t-th observation of this individual.

The first approach proposed by Hausman et al. [17, (1984)] assumed a conditional Poisson distribution, where $g(\theta_i)$ is assumed gamma. In the literature, even if gamma random effects is by far the most popular hypothesis, actuaries also used a conditional Poisson distribution with other random effects distributions. Examples are the inverse Gaussian distribution (Willmot [33, (1987)]) or the lognormal distribution (Hinde [18, (1982)]). Boucher and Denuit [1, (2006)] compared all the fitting of these distribution with real insurance data. Very few papers propose other distributions for random effects. Some papers address time dependence, such as in Pinquet et al. [28, (2001)] or Purcaru et al. [29, (2004)] who used dynamic lognormal distributions. Boucher et al. [6, (2008)] proposed various transformation of the gamma, lognormal and inverse-Gaussian distributions by some changes on their parameters. Boucher et al. [7, (2008)] proposed to use a degenerated random effects distribution, with a mass-point for a zero-value of the random effects. Boucher and Guillén [8, (2008)] used a nonparametric estimation for the heterogeneity density based on a squared Kth-order polynomial expansion.

The second approach proposed by Hausman et al. [17, (1984)] assumed a conditional Negative Binomial distribution with beta random effects. This model was used in Boucher et al. [6, (2008)] for insurance data and provided the best fit compared to the Poisson-gamma distribution. In the actuarial literature, Gómez-Déniz et al. [15, (2008)] proposed another mixing distribution based on an inverse-Gaussian distribution. Other papers modeled claim severity and claim frequency (see, Frees et al. [13, (2001)] and Frees and Valdez [14, (2009)]).

Many attempts have recently been made in selecting other conditional distributions for panel data. In this paper, we propose a review of these conditional distributions as well as their applicability to insurance data. In the second section of the paper, we review the standard modeling approach of panel count data models, proposed by Hausman et al. [17, (1984)]. We also discuss some modeling difficulties resulting from the dependence between regressors and random effects and we obtain the predictive distribution resulting from observations of past values of $N_{i,t}$.

In the third part of the paper, we review the standard Poisson distribution, while in section 4 we explore panel data model for the conditional Negative Binomial distribution. The zero-inflated models proposed by Boucher et al. [7, (2008)], where an extra parameter is added to the count distribution, is described in section 5. In the sixth section, the hurdle distributions for panel data, proposed by Boucher et al. [4, 5, (2008)] is used. In the following section, we review a conditional count distribution where the time between two claims is no longer an exponential distribution, as proposed by Boucher and Denuit [2, (2007)]. In section 8, we propose a generalization for panel data of the *Negative Binomial X* introduced in actuarial science by Boucher et al. [3, (2007)]. Finally, in the last section, a numerical illustration is presented and some methods to compare model fits are described.

2 Panel Data Distributions

2.1 Endogeneous Regressors and Fixed Effects

In linear regression, correlation between covariates and the error term leads to inconsistency of the estimated parameters (Mundlak [27, (1978)] or Hsiao [20, (2003)] present a review in case of longitudinal data). The same problem exists for the count data regression when $\mathrm{E}[\theta|x_i] \neq \mathrm{E}[\theta]$ (Mullahy [26, (1997)]) and it leads to biased parameter estimates. In insurance, correlation between regressors and the error term is often present (Boucher and Denuit [1, (2006)]) and it may be caused by omitted variables that are correlated with the included ones.

As noted in Winkelmann [35, (2003)], consistent estimates may be found if corrections are made to the standard estimation procedures. Boucher and Denuit [1, (2006)] compared fixed and random effects models. they showed that standard estimation methods, like classical maximum likelihood, can still be used on joint distribution based on random effects. Indeed, the resulting parameter estimates, while being biased, represent the apparent effect on the frequency of claims, which is exactly the interest when the correlated omitted variables cannot be used in classification.

2.2 *A Priori* and *A Posteriori* Ratemaking

Property and liability motor vehicle insurers use classification plans to create risk classes. The classification variables introduced to partition risks into cells are called *a priori* variables (as their values can be determined before the policyholder starts to drive). Premiums for motor liability coverage often vary by the territory in which the vehicle is garaged, the use of the vehicle (driving to and from work or business use) and individual characteristics (such as age, gender, occupation and marital status of the main driver of the vehicle, for instance, if not precluded by legislation or regulatory rules).

As stated in the introduction, many important factors cannot be taken into account in the *a priori* risk classification. Consequently, tariff cells are still quite heterogeneous despite the use of many classification variables. This heterogeneity can be modeled by a random effect in a statistical model. It seems reasonable

to believe that the hidden characteristics are partly revealed by the number of claims at fault reported by the policyholders. Several empirical studies have shown that, if insurers were allowed to use only one rating variable, it should be some form of merit rating: the best predictor of the number of claims incurred by a driver in the future is not age or vehicle type but past claims history. Hence the adjustment of the premium from the individual claims experience in order to restore fairness among policyholders. In that respect, the allowance of past claims in a rating model derives from an exogenous explanation of serial correlation for longitudinal data. In this case, correlation is only apparent and results from the revelation of hidden features in the risk characteristics.

As a consequence, for each t period, the heterogeneity terms (denoted later as θ_i or ϕ_i) can be updated from past experience. Even if the parameters of the distributions are evaluated using maximum likelihood estimation, at each successive time period, the random effects θ_i can be updated. Indeed, the joint distribution can be expressed as:

$$\begin{aligned}&\Pr(N_1 = n_1, N_2 = n_2, \ldots, N_t = n_t) = \\ &\qquad \Pr(N_1 = n_1) \times \Pr(N_2 = n_2 | N_1 = n_1) \times \cdots \times \Pr(N_t = n_t | N_1 = n_1, \ldots, N_{t-1} = n_{t-1})\end{aligned}$$

Formally, the predictive distribution at time $T+1$ can be computed as:

$$\begin{aligned}\Pr(N_{i,T+1} =& n_{i,T+1} | n_{i,1}, \ldots, n_{i,T}) = \\ &= \int \Pr(N_{i,T+1} = n_{i,T+1} | \theta_i) \left(\frac{\left(\prod_{t=1}^{T} \Pr(N_{i,t} = n_{i,t} | \theta_i) \right) g(\theta_i)}{\int \left(\prod_{t=1}^{T} \Pr(N_{i,t} = n_{i,\tau} | \theta_i) \right) g(\theta_i) \, \mathrm{d}\theta_i} \right) \mathrm{d}\theta_i \qquad (2) \\ &= \int \Pr(N_{i,T+1} = n_{i,T+1} | \theta_i) \, g(\theta_i | n_{i,1}, \ldots, n_{i,T}) \, \mathrm{d}\theta_i\end{aligned}$$

where $g(\theta_i | n_{i,1}, \ldots, n_{i,T})$ is called the *a posteriori* distribution of the random effects θ_i, reflecting the past claims experience of insured i. If this *a posteriori* distribution can be expressed in closed form, moments of the predictive distribution can be found easily conditional on the random effects θ_i.

In actuarial science, *a priori* ratemaking is the premium charged for an insured without individual experience, i.e. the premium calculated using the distribution of $N_{i,1}$, while an experienced insured is charged using *a posteriori* ratemaking, i.e. by using the conditional distribution given the previous observed claims.

Exact predictive and posterior distributions for the random effects can only be expressed in closed form for some specific distributions. For other models, these distributions cannot be evaluated analytically. Consequently, a possible approach for evaluating these predictive distributions is the use of numerical computations or simulations, such as Markov chain Monte Carlo (MCMC) simulations.

3 Poisson

The simplest random effects model for count data is based on the Poisson distribution with an individual heterogeneity term that follows a specified distribution. Formally, we can express the classical Poisson random effects model as:

$$N_{i,t} | \theta_i \sim \text{Poisson}(\theta_i \lambda_{i,t}), \qquad i = 1, \ldots, N \quad t = 1, \ldots, T,$$

where i represent an insured and t the period of coverage, $\lambda_{i,t}$ is a positive parameter that will usually be related to individual known characteristics. If the gamma distribution of mean 1 and variance α is used, the joint distribution is equal to (Hausman et al. [17, (1984)]):

$$\begin{aligned}&\Pr[N_{i,1} = n_{i,1}, \ldots, N_{i,T} = n_{i,T}] = \\ &\qquad \left[\prod_{t=1}^{T} \frac{(\lambda_{i,t})^{n_{i,t}}}{n_{i,t}!} \right] \frac{\Gamma(\sum_{t=1}^{T} n_{i,t} + 1/\alpha)}{\Gamma(1/\alpha)} \left(\frac{1/\alpha}{\sum_{t=1}^{T} \lambda_{i,t} + 1/\alpha} \right)^{1/\alpha} \left(\sum_{t=1}^{T} \lambda_{i,t} + 1/\alpha \right)^{-\sum_{t=1}^{T} n_{i,t}}.\end{aligned}$$

This distribution, which has often been applied in practice (see chapter 36 of Johnson et al. [21, (1996)] for an overview), is known as a Multivariate Negative Binomial (MVNB) or Negative Multinomial. Note that this distribution can also be seen as the generalization of the bivariate Negative Binomial of Marshall and Olkin [23, (1990)]. For this distribution, $\mathrm{E}[N_{i,t}] = \lambda_{i,t}$ and $\mathrm{Var}[N_{i,t}] = \lambda_{i,t} + \alpha\lambda_{i,t}^2$, so overdispersion can be accounted for. Maximum likelihood estimates of the parameters and their variance estimates are straightforward.

The Poisson-gamma distributions has the following moments:

$$\mathrm{E}[N_{i,t}] = \lambda_{i,t}, \qquad \mathrm{Var}[N_{i,t}] = \lambda_{i,t} + \alpha\lambda_{i,t}^2.$$

As it is well known, we found that the *a posteriori* distribution of the heterogeneity term for the Poisson model with gamma random effects is also gamma distributed with parameters $\sum_t \lambda_{i,t} + 1/\alpha$ and $\sum_t n_{i,t} + 1/\alpha$. In consequence, the future premium (frequency part), which is equal to the expected number of reported claims, is equal to:

$$\mathrm{E}[N_{i,t+1}|N_{i,1},\ldots,N_{i,t}] = \lambda_{i,t+1}\frac{\sum_t n_{i,t} + 1/\alpha}{\sum_t \lambda_{i,t} + 1/\alpha}.$$

4 Negative Binomial

Negative Binomial distribution can also be used with random effects, as shown by Hausman et al. [17, (1984)]. Conditionally on the random effects δ_i, the conditional distribution has the following moments:

$$\mathrm{E}[N_{i,t}|\delta_i] = \lambda_{i,t}/\delta_i, \qquad \mathrm{Var}[N_{i,t}|\delta_i] = \mathrm{E}[N_{i,t}|\delta_i](1+\delta_i)/\delta_i$$

Thus, this conditional distribution implies overdispersion. Under the construction of (1), Hausman et al. [17, (1984)] assumed that the expression $\delta_i/(1+\delta_i)$ follows a beta distribution with parameter (a,b), with mean $a/(a+b)$ and variance $ab/\left((a+b+1)(a+b)^2\right)$. Following the development of Hausman et al. [17, (1984)], the joint distribution can be expressed as:

$$\Pr[N_{i,1} = n_{i,1},\ldots,N_{i,T} = n_{i,T}] = \frac{\Gamma(a+b)\Gamma(a+\sum_t \lambda_{i,t})\Gamma(b+\sum_t n_{i,t})}{\Gamma(a)\Gamma(b)\Gamma(a+b+\sum_t \lambda_{i,t}+\sum_t n_{i,t})}\prod_t^T \frac{\Gamma(\lambda_{i,t}+n_{i,t})}{\Gamma(\lambda_{i,t})\Gamma(n_{i,t}+1)}.$$

The moments of the Negative Binomial-Beta (NB-Beta) distribution are as follow:

$$\mathrm{E}[N_{i,t}] = \lambda_{i,t}\frac{b}{a-1}, \qquad \mathrm{Var}[N_{i,t}] = \lambda_{i,t}\frac{(a+b-1)b}{(a-1)(a-2)} + \lambda_{i,t}^2\left[\frac{(b+1)b}{(a-1)(a-2)} - \frac{b^2}{(a-1)^2}\right],$$

The *a posteriori* density of the heterogeneity term of the Negative Binomial with beta random effects, proposed by Hausman et al. [17, (1984)], has also a closed form. Indeed, using equation (2), it can be shown that the ratio $\delta_i/(1+\delta_i)$ follows a beta distribution with parameters $\sum_t \lambda_{i,t} + a$ and $\sum_t n_{i,t} + b$. Consequently, for this model, the frequency part of the future premium can be expressed as:

$$\mathrm{E}[N_{i,t+1}|N_{i,1},\ldots,N_{i,t}] = \lambda_{i,t+1}\frac{\sum_t n_{i,t} + b}{\sum_t \lambda_{i,t} + a - 1},$$

which has the same form as the future premium with the Poisson-gamma model, but allows more flexibility since an additional parameter is used to calculate the premium.

5 Zero-Inflated Distribution

The zero-inflated Poisson model has been shown to be a useful alternative to the Poisson distribution for cross-section data. Indeed, it often provides a good fit for the data and can easily be interpreted. The model

is based on a finite mixture model of two distributions combining an indicator distribution for the zero case and a standard count distribution (Mullahy [25, (1986)], Lambert [22, (1992)]). The distribution has been shown to be a natural candidate to deal with the large frequency of zero-values, which is exactly what is observed in insurance data. The zero-inflated Poisson (ZIP) distribution has two parameters ϕ and λ and has the following probability function:

$$\Pr[N=n]=\begin{cases}\phi+(1-\phi)\mathrm{e}^{-\lambda} & \text{for } n=0\\ (1-\phi)\dfrac{\mathrm{e}^{-\lambda}\lambda^n}{n!} & \text{for } n=1,2,\ldots\end{cases}.$$

For panel data modelling, we can treat the zero-inflated component as an individual parameter, add random effects to the mean parameter of the Poisson distribution or even use them together. By conditioning on these two random effects, the joint distribution can be expressed as:

$$\Pr[N_{i,1}=n_{i,1},\ldots,N_{i,T}=n_{i,T}|\epsilon_i,\theta_i]=\prod_{t=1}^{T}\left(I_{(n_{i,t}=0)}\phi_{it}+(1-\phi_{it})\frac{\mathrm{e}^{-\lambda_{it}\theta_i}(\lambda_{it}\theta_i)^n}{n!}\right)$$

where $\lambda_{it}=\exp(\boldsymbol{x}'_{it}\boldsymbol{\beta})$ and $\phi_{it}=\Phi(\boldsymbol{x}'_{it}\boldsymbol{\gamma}+\epsilon_i)$. Transposition is denoted by $'$ and there are two vector parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Φ denotes the cumulative distribution function of a standard normal. However because this distribution cannot be expressed in a simple form, instead, Boucher et al. [7, (2008)] choose to use time independent covariates, that is to say that covariates do not change over all the period observations of an individual unit. In this situation, the authors show that the joint conditional distribution can be modeled as:

$$\begin{aligned}\Pr[N_{i,1}=n_{i,1},\ldots,N_{i,T}=n_{i,T}|\phi_i,\theta_i]&=\prod_{t=1}^{T}\left(I_{(n_{i,t}=0)}\phi_i+(1-\phi_i)\Pr[N_{i,t}=n_{i,t}]\right)\\&=\sum_{j=0}^{T_0}\binom{T_0}{j}V_j^{\mathrm{Poi}}(n_{i,1},\ldots,n_{i,T}|\theta_i)\phi_i^{T_0-j}(1-\phi_i)^{(T-T_0)+j},\end{aligned}$$

where T_0 is the number of insured periods without claim and $V^{\mathrm{Poi}}(\cdot)$ is a function having the following Poisson form:

$$V_j^{\mathrm{Poi}}(n_{i,1},\ldots,n_{i,T}|\theta_i)=\frac{(\lambda_i\theta_i)^{\sum_{t=1}^{T}n_{i,t}}\exp\left(-(T-T_0+j)\lambda_i\theta_i\right)}{\prod_{t=1}^{T}n_{i,t}!}.$$

By this parametrization, using one or both random effects, Boucher et al. [7, (2008)] show that the joint distribution can be expressed in closed form. Indeed, an individual term ϕ_i that is beta distributed with parameters a_i and b, and an heterogeneity term θ_i that follows a gamma distribution of mean 1 and variance α are added to the model. Consequently, it leads to a multivariate zero-inflated Poisson Beta Gamma model (MZIP-BetaGamma) that can be expressed as:

$$\Pr[N_{i,1}=n_{i,1},\ldots,N_{i,T}=n_{i,T}]=\sum_{j=0}^{T_0}\binom{T_0}{j}V_j^{\mathrm{NB}}(n_{i,1},\ldots,n_{i,T})\frac{\beta(a+T_0-j,b+(T-T_0)+j)}{\beta(a,b)},$$

where T_0 is the number of insured periods without claim and the function $V_j^{\mathrm{NB}}(\cdot)$ has the following multivariate Negative Binomial form:

$$V_j^{\mathrm{NB}}(n_{i,1},\ldots,n_{i,T})=\frac{\Gamma(\sum_t^T n_{i,t}+1/\alpha)}{\Gamma(1/\alpha)\prod_t^T n_{i,t}!}\left(\frac{1/\alpha}{(T-T_0+j)\lambda_i+1/\alpha}\right)^{1/\alpha}\left(\frac{\lambda_i}{(T-T_0+j)\lambda_i+1/\alpha}\right)^{\sum_t^T n_{i,t}}.$$

Under the assumption of independence between random effects ϕ and θ, moments can be found using the conditional calculation:

$$\mathrm{E}[N_{i,t}] = \lambda_i \left(1 - \frac{a_i}{a_i + b}\right) \qquad \mathrm{Var}[N_{i,t}] = \lambda_i \left(1 - \frac{a_i}{a_i + b}\right) + \lambda_i^2 \left(1 - \frac{a_i}{a_i + b}\right) \left(\frac{a_i}{a_i + b} + \alpha\right).$$

Parameters can be evaluated using maximum likelihood estimation. Programs such as the NLMIXED procedure of the SAS system allow for this type of estimations when the log-likelihood can be expressed in closed form.

Using both individual effects leads to the following predictive distribution:

$$\Pr[N_{i,T+1} = n_{i,T+1} | n_{i,1}, \ldots, n_{i,T}] = \frac{\sum_{j=0}^{T_0^*} \binom{T_0^*}{j} V_j^{\mathrm{NB}}(n_{i,1}, \ldots, n_{i,T}, n_{i,T+1}) \, \beta(a + T_0^* - j, b + (T + 1 - T_0^*) + j)}{\sum_{j=0}^{T_0} \binom{T_0}{j} V_j^{\mathrm{NB}}(n_{i,1}, \ldots, n_{i,T}) \, \beta(a + T_0 - j, b + (T - T_0) + j)}.$$

where T_0^* is the updated value of T_0, considering $n_{i,T+1}$. Using the following notation:

$$K(j) = \frac{\binom{T_0}{j} \Gamma(a + T_0 - j) \Gamma(b + T + 1 - T_0 + j) ((T - T_0 + j)\lambda_i + 1/\alpha)^{-(\sum_t^T n_{i,t} + 1/\alpha)}}{(a_i + b + T) \sum_{k=0}^{T_0} \binom{T_0}{k} \Gamma(a + T_0 - k) \Gamma(b + T - T_0 + k) ((T - T_0 + k)\lambda_i + 1/\alpha)^{-(\sum_t^T n_{i,t} + 1/\alpha)}},$$

the predictive distribution can be expressed as:

$$\Pr[N_{i,T+1} = n_{i,T+1} | n_{i,1}, \ldots, n_{i,T}] = \begin{cases} 1 - \sum_{j=0}^{T_0} K(j)(1 - p^r) & \text{for } n_{i,T+1} = 0 \\ \displaystyle\sum_{j=0}^{T_0} K(j) \Pr_{\mathrm{NB}}[N_{i,T+1} = n_{i,T+1}; r, p] & \text{for } n_{i,T+1} = 1, 2, \ldots \end{cases}$$

where:

$$\Pr_{\mathrm{NB}}[N_{i,T+1} = n_{i,T+1}; r, p] = \binom{n_{i,T+1} + r}{r} p^r q^{n_{i,T+1}}$$

is the probability function of a Negative Binomial distribution with parameters equal to:

$$r = \sum_{t=1}^{T} n_{i,t} + 1/\alpha, \qquad p = \frac{(T - T_0 + j)\lambda_i + 1/\alpha}{(T + 1 - T_0 + j)\lambda_i + 1/\alpha}$$

and the expected predictive value is the equal to:

$$\mathrm{E}[N_{i,T+1} | n_{i,1}, \ldots, n_{i,T}] = \lambda_i \sum_{j=0}^{T_0} \frac{\left(\sum_t^T n_{i,t} + 1/\alpha\right) K(j)}{(T + 1 - T_0 + j)\lambda_i + 1/\alpha}.$$

Note that unlike the standard Poisson-gamma models, the predictive mean not only depends on the sum of number of past claims, but also on the number of insured periods without a claim (T_0).

Boucher et al. [7, (2008)] found that the generalizations of the zero-inflated Poisson distribution has an interesting interpretation for insurance data, where the number of accidents can be compared to the number of claims. The zero-inflated distributions applied to the number of claims can be used to model the behaviour of the insureds, i.e. to model the probability to file a claim.

6 Hurdle Distribution

The hurdle model was introduced by Cragg [11, (1971)] and reviewed by Mullahy [25, (1986)]. This model is characterized by the processes below and above the hurdle. Obviously, the most widely used hurdle model sets the hurdle at zero, which leads to a distribution with the following two processes: firstly, a dichotomous distribution that allows the participation of the second process; secondly, another process that specifies, the count number, on the condition that the first process *succeeds*. The first part of the model is a binary outcome model, and the second part is a truncated count distribution. Formally, this hurdle model is expressed as follows. Let $f_1(\cdot|\theta_1)$ and $f_2(\cdot|\theta_2)$ be two probability mass functions with respective support $\{0,1\}$ and $\{0,1,\ldots\}$ depending on parameter vectors θ_1 and θ_2. The counting random variable N follows the hurdle distribution if:

$$\Pr(N=n|\theta_1,\theta_2)=\begin{cases} f_1(0|\theta_1) & \text{for } n=0 \\ \dfrac{1-f_1(0|\theta_1)}{1-f_2(0|\theta_2)} f_2(n|\theta_2)=\Psi f_2(n|\theta_2) & \text{for } n=1,2,\ldots \end{cases},$$

where $\Psi=\frac{1-f_1(0|\theta_1)}{1-f_2(0|\theta_2)}$. Boucher et al. [4, 5, (2008)] generalized the hurdle distribution for panel data. As in the zero-inflated model, since random effects can be added to the model for both the first and the second process, the joint distribution can be generalized for more than one θ_i. Indeed, the joint distribution can be expressed as:

$$\Pr(N_{i,1}=n_{i,1},\ldots,N_{i,T}=n_{i,T})= \iint \prod_{t=1}^{T} f_1(0|\theta_{i,1})^{I_{(n_{i,t}=0)}} \left(1-f_1(0|\theta_{i,1})\right)^{1-I_{(n_{i,t}=0)}} f_2(n^*_{i,t}|\theta_{i,2})^{1-I_{(n_{i,t}=0)}} g(\theta_{i,1},\theta_{i,2})\,\mathrm{d}\theta_{i,1}\,\mathrm{d}\theta_{i,2},$$

where the following specific transformation $n^*=n-1$ has been used by Boucher et al. [4, 5, (2008)] to avoid the use of a truncated distribution. The joint distribution of the random effects $g(\theta_{i,1},\theta_{i,2})$ can be expressed by a copula. To obtain interesting predictive distributions, Boucher et al. [5, (2008)] use time independent covariates. For the zero-part of the model, the authors used a Bernoulli($\theta_{i,1}$) distribution where the parameter $\theta_{i,1}$ is beta(a_i,b)-distributed to account for the individual specificities. Covariates have been included in the model as $a_i=\exp(x_i'\,\beta)$ to be sure that parameter a_i is greater than zero. The positive part means are fitted using standard Poisson($\gamma_i\,\theta_{i,2}$) random effects models, where the mean variable can be expressed as $\gamma_i=\exp(x_i'\,\delta)$. The random effects $\theta_{i,2}$ follow a gamma distribution of mean 1 and variance α (i.e. both parameters are equal to $1/\alpha$). Consequently, with these conditional distributions, the joint distribution for all contracts of the same insured is expressed as:

$$\Pr[N_{i,1}=n_{i,1},\ldots,N_{i,T}=n_{i,T}]= \iint \prod_{t=1}^{T} \theta_{i,1}^{I_{(n_{i,t}=0)}} (1-\theta_{i,1})^{1-I_{(n_{i,t}=0)}} \left(\mathrm{e}^{-\gamma_i\theta_{i,2}}\frac{(\gamma_i\theta_{i,2})^{n^*_{i,t}}}{n^*_{i,t}!}\right)^{1-I_{(n_{i,t}=0)}} g(\theta_{i,1},\theta_{i,2})\,\mathrm{d}\theta_{i,1}\,\mathrm{d}\theta_{i,2}$$

The first moments of the hurdle for panel count data can be expressed as:

$$\begin{aligned}\mathrm{E}[N_{i,t}] &= \mathrm{E}[\theta_{i,1}]+\gamma_i\,\mathrm{E}[\theta_{i,1}\theta_{i,2}] \\ \mathrm{Var}[N_{i,t}] &= \gamma_i^2\,\mathrm{E}[\theta_{i,1}\theta_{i,2}^2]+\mathrm{E}[\theta_{i,1}\theta_{i,2}]\,[3\gamma_i-2\gamma_i\,\mathrm{E}[\theta_{i,1}]]+\mathrm{E}[\theta_{i,1}]-\mathrm{E}[\theta_{i,1}]^2-\gamma_i^2\,\mathrm{E}[\theta_{i,1}\theta_{i,2}]^2\end{aligned}$$

Boucher et al. [5, (2008)] suppose independent random effects where the joint distribution of the random effects can be expressed as the product of the marginal density functions $g_1(\theta_{i,1})$ and $g_2(\theta_{i,2})$, i.e. $g(\theta_{i,1},\theta_{i,2})=g_1(\theta_{i,1})g_2(\theta_{i,2})$. Because of independence, each process of the hurdle model for panel data with independent random effects can be expressed separately. Consequently, the two processes can also be analyzed separately for the *a posteriori* analysis. For the first process of the model, composed with a

Bernoulli-beta combination, it is well-known that the posterior distribution of the random effects is still beta distributed. The second process $N_i^* = N_i - 1$ of the hurdle distribution is to follow a Poisson distribution with gamma random effects. Thus, the *a posteriori* distribution of the random effect term is also a gamma distribution, as seen in section 3. By combining these two processes of the hurdle model, the expected value of the distribution, having past experience $1, 2, \ldots, T$ is equal to:

$$\begin{aligned}\mathrm{E}[N_{i,T+1}|n_{i,1},\ldots,n_{i,T}] &= \frac{\sum_{t=1}^{T} k_{i,t} + a_i}{T+b+a_i}\left(1+\gamma_i\frac{\sum_{t=1}^{T-T_0} n_{i,t}^* + 1/\alpha}{(T-T_0)\gamma_i + 1/\alpha}\right)\\ &= \left(\frac{(T-T_0)\gamma_i + a_i\gamma_i}{(T-T_0)\gamma_i + 1/\alpha}\right)\left(\frac{\sum_{t=1}^{T} n_{i,t} + (1+\gamma_i)/\alpha}{T+b+a_i}\right)\end{aligned}$$

where T_0 is equal to the number of periods without claim. As for the zero-inflated conditional distribution, the predictive premium of the hurdle model for panel data depends on the sum of past claims and on the number of insured periods without claims.

Often, as stated in the beginning of the paper, the models with random effects can be interpreted as models where hidden individual characteristics are captured by this additional random term. Since we work with two random effects terms, a dependence between these effects might be supposed since the same omitted characteristics affect both process. Consequently, Boucher et al. [4, (2008)] proposed to model this dependence with a Gaussian copula that leads to the following expression of the joint distribution of the random effects:

$$g(\theta_{i,1},\theta_{i,2}) = c(G_1(\theta_{i,1}), G_2(\theta_{i,2}))\; g_1(\theta_{i,1})\; g_2(\theta_{i,2}),$$

where the Gaussian copula is expressed as:

$$c^{\mathrm{Ga}}(G_1(\theta_{i,1}), G_2(\theta_{i,2})) = \frac{1}{\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2}\left(\frac{\rho^2\Phi^{-1}(G_1(\theta_{i,1}))^2 + \rho^2\Phi^{-1}(G_2(\theta_{i,2}))^2 - 2\rho\Phi^{-1}(G_1(\theta_{i,1}))\Phi^{-1}(G_2(\theta_{i,2}))}{1-\rho^2}\right)\right)$$

where Φ is the standard Normal distribution function. As for independent random effects, marginal density functions $g_1(\theta_{i,1})$ and $g_2(\theta_{i,2})$ are beta and gamma distributions. Obviously, the authors did not find closed form expression from this last model. Consequently, an alternative modeling approach must be used, such as numerical integration techniques, MCMC methods or others. Similarly, exact predictive and posterior distributions cannot be evaluated analytically. Consequently, a possible approach for evaluating these predictive distributions is the use of numerical computations or simulations.

As mentioned in Boucher et al. [3], the hurdle model possesses a natural interpretation for the number of reported claims. Indeed, it is reasonable to believe that the behaviour of the insureds is not same when they already have reported a claim. This suggests that two processes govern the total number of claims, as with the hurdle model.

7 Duration Models

It is well-known that if the time between two claims is exponentially distributed over a specified period over time, the distribution of the number of claims will be Poisson. Boucher and Denuit [2, (2007)] tried to generalize this situation by choosing other time duration distributions. More generally, let τ_i be the waiting time between the $(i-1)^{\text{th}}$ event and the i^{th} event. The k^{th} event thus occurs at time

$$\nu(k) = \sum_{i=1}^{k} \tau_i. \tag{3}$$

In this situation, the τ_i's are assumed to be independent and identically distributed.

Now, let $N(t)$ be the number of events occurring during the interval $[0,t]$. From (3), we can state that the relationship between the arrival time $\tau(i)$ and the count process $N(t)$ is $\nu(k) \leq t \Leftrightarrow N(t) \geq k$. Hence,

$$\begin{aligned} \Pr(N(t)=k) &= \Pr(N(t)<k+1) - \Pr(N(t)<k) \\ &= \Pr(\nu(k+1)>t) - \Pr(\nu(k)>t) \\ &= F_k(t) - F_{k+1}(t) \end{aligned} \tag{4}$$

where $F_k(t)$ is the distribution function of $\nu(k)$.

If the τ_i's have a common Exponential distribution, that is, if their respective probability density function is

$$f(\tau;\lambda) = \lambda \, \mathrm{e}^{-\lambda\tau},$$

then, as stated above, we find the classical Poisson process for claim counts. This means that the number of claims occurring in the time interval $[0,t]$ is Poisson distributed with mean λt, that is,

$$\Pr(N(t)=k) = \mathrm{e}^{-\lambda t}\frac{(\lambda t)^k}{ku!}.$$

Winkelmann [34, (1995)] suggested to take gamma distributed τ_i's, with density

$$f(\tau;\varphi,\lambda) = \frac{\lambda^\varphi}{\Gamma(\varphi)}\tau^{\varphi-1}\mathrm{e}^{-\lambda\tau},$$

where $\Gamma(\cdot)$ is the gamma function, $\varphi > 0$ and $\lambda > 0$. The stability under convolution of the gamma distribution for fixed λ parameter implies that $\nu(k)$ is also gamma distributed with density

$$f(\nu;\varphi,\lambda) = \frac{\lambda^{k\varphi}}{\Gamma(k\varphi)}\nu^{k\varphi-1}\mathrm{e}^{-\lambda\nu}.$$

Hence,

$$F_k(t) = \frac{1}{\Gamma(k\varphi)}\int_0^t \lambda^{k\varphi}\nu^{k\varphi-1}\mathrm{e}^{-\lambda\nu}\,\mathrm{d}\nu = G(\varphi k,\lambda t)$$

where the integral is known as an incomplete gamma function. Winkelmann [34, (1995)] shows that the gamma count distribution (GCD) can be found using equation (4):

$$\Pr(N(t)=k) = G(\varphi k,\lambda t) - G(\varphi k+\varphi,\lambda t)$$

with $G(0,\varphi\lambda)=1$.

Bradlow et al. [9, (2008)] show that another common model used in the duration analysis is based on the Weibull distribution. In this case, the τ_i's have density

$$f(\tau;c,\lambda) = \lambda c\tau^{c-1}\exp(-\lambda\tau^c)$$

for $\lambda > 0$ and $c > 0$. By using a k-fold convolution of the interarrival time distribution with the help of a Taylor series approximation Bradlow et al. [9, (2008)] obtained the following expression for the Weibull count distribution (WCD):

$$\Pr(N(t)=k) = \sum_{j=k}^{\infty}\frac{(-1)^{j+k}(\lambda t^c)^j\varphi_j^p}{\Gamma(cj+1)}$$

where

$$\varphi_j^0 = \frac{\Gamma(cj+1)}{\Gamma(j+1)}$$

$$\varphi_j^{p+1} = \sum_{m=p}^{j-1} \varphi_m^p \frac{\Gamma(cj - cm + 1)}{\Gamma(j - m + 1)}, \qquad p = 0, 1, 2, \ldots \qquad j = p+1, p+2, \ldots$$

Boucher and Denuit [2, (2007)] generalized these distributions for panel data as for the zero-inflated or the hurdle models. A generalization of the MVNB distribution can be found if random effects θ_i are supposed to follow a gamma$(1/\alpha, 1/\alpha)$ distribution:

$$\Pr(N_{i,1} = n_{i,1}, \ldots, N_{i,T} = n_{i,T}) = \int_0^\infty \prod_{\tau=1}^{T} \Big(G(\varphi n_{i,\tau}, t_{i,\tau}\lambda_{i,\tau}\theta_i) - G(\varphi n_{i,\tau} + \varphi, t_{i,\tau}\lambda_{i,\tau}\theta_i) \Big) h(\theta_i)\, \mathrm{d}\theta_i.$$

Complex computations are needed to obtain closed form from this last equation of the Multivariate gamma Count Distribution (MGCD). For the WCD, Boucher and Denuit [2, (2007)] expressed the joint distribution for all contracts of the same insured as the one shown for the panel data model of the GCD:

$$\Pr(N_{i,1} = n_{i,1}, \ldots, N_{i,T} = n_{i,T}) = \int_0^\infty \prod_{\tau=1}^{T} \left(\sum_{j=n_{i,\tau}}^{\infty} \frac{(-1)^{j+n_{i,\tau}} (\lambda_{i,\tau}\theta_i t_{i,\tau}^c)^j \varphi_j^p}{\Gamma(cj+1)} \right) h(\theta_i)\, \mathrm{d}\theta_i.$$

The first moments of these two models must be evaluated numerically. Obviously, as for other complex panel data distributions, exact predictive and posterior distributions for the random effects of the gamma and the Weibull count distribution can only be evaluated using numerical computations or simulations.

Non-constant hazard models, like MGCD or the MWCD, implies duration dependence. Consequently, when applied to insurance data, it means that within each year of contract, those models suppose that the report of a claim decreases the expected time to report an other claim.

8 Negative Binomial X

In a recent paper, Boucher et al. [3, (2007)] introduced a new model in actuarial science called *Negative Binomial X*. The model is based on a compound sum (or stopped-sum distributions) correspond to counting variables of the form:

$$N = \sum_{j=1}^{M} X_j$$

where the X_j's are integer-valued, independent and identically distributed, and where M and the X_j's are independent. The authors supposed that M is Poisson with mean λ and X_j is Logarithmic with parameter η, which means that N is Negative Binomial (using the standard assumption that $\sum_{j=1}^{M} X_j = 0$ if $M = 0$). For cross-section data, Boucher et al. [3, (2007)] showed by numerical applications on real insurance data that this model exhibits the best fit compared to zero-inflated, hurdle or Poisson distribution.

The original model was proposed by Santos Silva and Windmeijer [31, (2001)] who defined theNegBin_x regression model as follows: the parameter η_i of the Logarithmic distribution is expressed in terms of the available covariates as

$$\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}) = \frac{\eta_i}{1-\eta_i}$$

and the Poisson parameter is taken to be $\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$. Consequently, N_i is Negative Binomial with parameter $\lambda_i / \log(1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}))$ and $\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})$. After some simplifications, the probability mass function is given by:

$$\Pr(N_i = n_i) = \frac{\Gamma\left(n_i + \frac{\lambda_i}{\log(1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}))}\right) \exp(-\lambda_i)}{\Gamma(n_i+1)\Gamma\left(\frac{\lambda_i}{\log(1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}))}\right) (1+\exp(-\boldsymbol{x}_i'\boldsymbol{\gamma}))^{n_i}}.$$

Many generalizations for panel data can be made for this model. Indeed, following some well-known ideas in the renewal theory, the dependence between contracts of the same insureds can be constructed with the assumption that all T observations have the same distribution. This means that all the contracts of the same insured can be expressed by the following vector:

$$\left\{ N_{i,1} = \sum_{j=1}^{M_i} X_{j,1},\ N_{i,2} = \sum_{j=1}^{M_i} X_{j,2}, \ldots,\ N_{i,T} = \sum_{j=1}^{M_i} X_{j,T} \right\}.$$

where all the $X_{j,t}$, $t = 1, \ldots, T$ are i.i.d.. This modeling has close similarities with the common shock model used in Boucher et al. [7, (2008)], where the dependence between contracts of the same insured comes from a common individual random variable that is added to each time period (Holgate [19, (1964)] for the bivariate case). This model can be interpreted as if an individual specificity of an insured affects all his contracts. As for the common shock model, this generalization of the NegBin_x distribution for panel data cannot be satisfactory for insurance data. Indeed, the model becomes interesting for a situation where positive counts can be observed for all contracts, while in practice, in all automobile insurance portfolios, a great proportion of insureds does not report a single claim. As mentioned in Winkelmann [35, (2003)], this kind of modelling is interesting because mixing and compounding are related concepts. Indeed, using the notation $N_t(M_i) = \sum_{j=1}^{M_i} X_{j,t}$ compounding can be seen as:

$$\Pr[N_{i,1} = n_{i,1}, \ldots, N_{i,T} = n_{i,T}] = \sum_{m=0}^{\infty} \Pr[N_1(M_i) = n_{i,1}, \ldots, N_T(M_i) = n_{i,T} | M_i = m] \Pr[M_i = m]$$

which can be seen as a discrete version of model (1).

Another related generalization is to suppose that the dependence between the contracts of the same insured comes from the random variable X_i. In this situation, all the contracts of the same insured can be expressed by the following vector:

$$\left\{ \sum_{j=1}^{M_{i,1}} X_j,\ \sum_{j=1}^{M_{i,2}} X_j, \ldots,\ \sum_{j=1}^{M_{i,T}} X_j \right\}.$$

where all the $M_{i,t}$, $t = 1, \ldots, T$ are i.i.d.. Using the standard assumption that $\sum_{j=1}^{M_{i,t}} X_j = 0$ if $M_{i,t} = 0$ allows greater flexibility in the modelling and avoids the problem cited above. This generalization of the NegBin_x distribution is much harder than the previous one and is currently under investigation.

Instead of trying to generalize the NegBin_x distribution of Santos Silva and Windmeijer [31, (2001)] by supposing a constant number of claims or a constant X_j for all the contracts of the same insured, we propose to add an heterogeneity term to the random variable M, as it was done for all the other models presented in this paper. Since M is supposed to be Poisson distributed, gamma random effects seems a natural choice. As mentioned earlier, generally, the number of claims is modeled with Poisson distribution, where a random effects variable is added to the count distribution, but this gives too much weight and importance on the heterogeneity. Indeed, this kind of models generate predictive premiums that over-penalises insureds with many claims (see Young and DeVylder [37, (2000)] for example). By adding only a random effects on a single part of the conditional count distribution, these penalties are softer since the impact of the heterogeneity is weaker.

Using the results of Santos Silva and Windmeijer [31, (2001)], it is possible to express the joint distribution of all the contracts of the same insured to obtain a multivariate NegBin_x distribution (MVNB_x):

$$\begin{aligned} &\Pr[N_{i,1} = n_{i,1}, \ldots, N_{i,T} = n_{i,T}] \\ &\quad = \int_0^{\infty} \left(\prod_{t=1}^{T} \Pr[N_{i,t} = n_{i,t} | \theta_i] \right) g(\theta_i)\, \mathrm{d}\theta_i \end{aligned}$$

$$= \int_0^\infty \left\{ \left(\prod_{t=1}^{T} \frac{\Gamma\left(n_{i,t} + \frac{\theta_i \lambda_{i,t}}{\log(1+\exp(\boldsymbol{x}'_{i,t}\boldsymbol{\gamma}))}\right) \exp(-\theta_i \lambda_{i,t})}{\Gamma(n_{i,t}+1)\Gamma\left(\frac{\theta_i \lambda_{i,t}}{\log(1+\exp(\boldsymbol{x}'_{i,t}\boldsymbol{\gamma}))}\right) (1+\exp(-\boldsymbol{x}'_{it}\boldsymbol{\gamma}))^{n_{it}}} \right) \right.$$
$$\left. \frac{(1/\alpha)^{(1/\alpha)}}{\Gamma((1/\alpha))} \theta_i^{(1/\alpha)-1} \exp(-\theta_i(1/\alpha))\, \mathrm{d}\theta_i \right\}$$
$$= \frac{(1/\alpha)^{(1/\alpha)}}{\Gamma((1/\alpha))} \left(\prod_{t=1}^{T} \frac{1}{\Gamma(n_{i,t}+1)(1+\exp(-\boldsymbol{x}'_{it}\boldsymbol{\gamma}))^{n_{it}}} \right)$$
$$\times \int_0^\infty \left(\prod_{t=1}^{T} \frac{\Gamma\left(n_{i,t} + \frac{\theta_i \lambda_{i,t}}{\log(1+\exp(\boldsymbol{x}'_{i,t}\boldsymbol{\gamma}))}\right)}{\Gamma\left(\frac{\theta_i \lambda_{i,t}}{\log(1+\exp(\boldsymbol{x}'_{i,t}\boldsymbol{\gamma}))}\right)} \right) \theta_i^{(1/\alpha)-1} \exp\left(-\theta_i((1/\alpha) + \sum_{t=1}^{T} \lambda_{i,t}) \right) \mathrm{d}\theta_i$$

where $\lambda_{it} = \exp(\boldsymbol{x}'_{it}\,\boldsymbol{\beta})$. The last integration must be done numerically. By conditioning, the first two moments of the distribution are:

$$\mathrm{E}[N_{i,t}] = \frac{\lambda_{i,t} \exp(\boldsymbol{x}'_{i,t}\boldsymbol{\gamma})}{\log(1+\exp(\boldsymbol{x}'_{i,t}\boldsymbol{\gamma}))}$$
$$\mathrm{Var}[N_{i,t}] = (1+\exp(\boldsymbol{x}'_{i}\boldsymbol{\gamma}))\,\mathrm{E}[N_{i,t}] + \alpha\,\mathrm{E}[N_{i,t}]^2$$

As stated in Boucher et al. [3, (2007)], the model can have some interesting interpretations for insurance data. This *two-part* model can be used for modeling the number of injured persons or the number of third parties involved in a given claimed accident. Indeed, in these situations, M represents the number of accident while X_j is used to model the number of victims involved or parties affected by a single accident.

Exact predictive and posterior distributions for the random effects cannot be evaluated analytically, so they require a numerical approach.

9 Numerical Application

We worked with a sample of the automobile portfolio of a major company operating in Spain, that was already used for the hurdle distributions in Boucher et al. [4, 5, (2008)]. Only cars for private use were considered in this sample. The panel data contains information for the period from 1991 to 1998. Our sample contains $15{,}179$ policyholders who remained with the company for seven complete periods, resulting in $106{,}253$ insurance contracts. We have exogeneous variables (sex and age of the driver, years with the company and power of the vehicle) that are kept in the panel plus the annual number of accidents. For every policy of a single insured we used the initial information contained in his first contract. The total number of at-fault claims that took place within each year-long period was used for analysis. More details can be found in Boucher et al. [4, (2008)].

All models presented in the paper can be simplified to a Poisson distribution. Empirical applications of the zero-inflated, hurdle and duration panel data models showed in previous papers that all models exhibit a better fit compared to the standard MVNB distribution. Comparison between all these panel data models has not been done. These models are non-nested. Consequently, the models cannot be compared directly and we cannot use specification tests to distinguish between two models.

A standard method of comparing non-nested models (and also nested models) is to use the information criteria, such as the Akaike Information Criteria $(\mathrm{AIC}) = -2\log(L) + 2k$ or the Bayesian Information Criteria $(\mathrm{BIC}) = -2\log(L) + 2\log(n)k$, where k represents the number of parameters of the model and n the total number of observations. Table 1 shows the results of the fit of the distribution on our insurance data. Hurdle and zero-inflated[1] distributions seem to offer the best fit using the information criteria.

[1] The beta random effects of the MZIP-BetaGamma distribution were not significant. We then removed these random effects and worked instead with the MZIP-Gamma model.

Models	Number of parameters	Loglikelihood	AIC	BIC
MVNB	7	26,702.98	53,419.96	53,567.99
NB-Beta	8	26,671.64	53,359.29	53,343.29
MZIP-Gamma	8	26,664.17	**53,344.35**	53,328.35
Hurdle (ind.)	11	26,688.70	53,399.40	53,377.40
Hurdle (Gauss.)	10	26,662.47	53,344.94	**53,324.94**
MGCD	8	26,675.74	53,367.49	53,351.49
MWCD	8	26,671.34	53,358.67	53,342.67
MVNB_x	8	26,673.44	53,362.88	53,346.88

Table 1. Comparison of models for the Spanish data set - Information Criteria

Models	Godd Profile		Average Profile		Bad Profile	
	Mean	Variance	Mean	Variance	Mean	Variance
MVNB	0.0567	0.0595	0.0651	0.0688	0.0902	0.0974
NB-Beta	0.0564	0.0620	0.0657	0.0728	0.0910	0.1025
MZIP-Gamma	0.0534	0.0604	0.0668	0.0727	0.0885	0.0975
Hurdle (ind.)	0.0570	0.0644	0.0659	0.0717	0.0911	0.0997
Hurdle (Gauss.)	0.0577	0.0657	0.0664	0.0721	0.0911	0.0989
MGCD	0.0567	0.0614	0.0651	0.0713	0.0897	0.1017
MWCD	0.0566	0.0617	0.0651	0.0717	0.0906	0.1035
MVNB_x	0.0565	0.0619	0.0655	0.0721	0.0906	0.1015

Table 2. A priori Premiums

Deeper analysis could be done to compare the fit of all models. An interesting possibility is to test if the differences in the log-likelihood or the information criteria between the models are statistically significant. For independent observations, a log-likelihood ratio test for non-nested models, developed by Vuong [32, (1989)] and generalized by Rivers and Vuong [30, (2002)] can be used. This test cannot be applied directly to our panel data models since some observations —all contracts of the same insured— are not independent. However, as proposed by Golden [16, (2003)], an adapted Vuong test should be performed on non-nested models test. This test can be applied on correlated observations, and on panel data as done for instance in Boucher et al. [7, (2008)] and Boucher et al. [6, (2008)], but need complex intermediary steps before using this statistical test (gradient evaluation, autocorrelation check, etc.).

For illustration, we show the differences between models through the mean and the variance of insured profiles. Three profiles were selected and were classified as good, average and bad drivers. The results are given in Table 2. This table shows that the expected values of all profiles are fairly similar for the models studied. The greatest differences between models can be found in the variance estimates.

Differences between predictive premiums are also interesting to analyse. To illustrate this, we kept the estimated parameters out of the *a priori* analysis and projected a loss experience of 10 years. This way of analyzing the *a posteriori* models is very common in actuarial science. One can then many different claim experience situations that can arise in insurance companies. Table 3 shows the predictive premiums for an average risk profile for the MVNB and the NB-Beta models. For those models, the predictive premium only depends on the sum of reported claims.

The predictive premiums of the zero-inflated and the hurdle models do not only depend on the number of reported claims but also on the number of insured periods without claims. Table 4 shows the premium that should be charged to insureds depending on his past insured records. Interesting conclusions can be drawn

Models	A priori	Sum of claims 0	1	2	3	4
MVNB	0.0651	0.0413	0.0778	0.1143	0.1509	0.1874
NB-Beta	0.0657	0.0441	0.0770	0.1099	0.1428	0.1758

Table 3. Predictive Premiums (Average Risk Profile) for the MVNB and the NB-Beta Models

Models	T_0	A priori	Sum of claims 0	1	2	3	4
MZIPG	10	0.0668	0.0443	·	·	·	·
	9	0.0668	·	0.0783	0.1130	0.1480	0.1833
	8	0.0668	·	·	0.1116	0.1462	0.1809
	7	0.0668	·	·	·	0.1444	0.1786
	6	0.0668	·	·	·	·	0.1763
Hurdle (ind.)	10	0.0659	0.0448	·	·	·	·
	9	0.0659	·	0.0833	0.0876	0.0920	0.0963
	8	0.0659	·	·	0.1246	0.1304	0.1363
	7	0.0659	·	·	·	0.1683	0.1755
	6	0.0659	·	·	·	·	0.2140
Hurdle (Gauss.)	10	0.0663	0.0441	·	·	·	·
	9	0.0663	·	0.0776	0.1063	0.1336	0.1598
	8	0.0663	·	·	0.1113	0.1403	0.1687
	7	0.0663	·	·	·	0.1444	0.1750
	6	0.0663	·	·	·	·	0.1774

Table 4. Predictive Premiums (Average Risk Profile) for the Zero-Inflated and the Hurdle Models

from the analysis of predictive premiums. Indeed, for the hurdle model with independent random effects, we can see that the number of insured periods without a claim has a greater impact on the premium for the following year than the total number of reported claims. However, we also observe that the dependence between the two random effects of the hurdle model has a great impact on the premium. The correlated random effects hurdle models show premium values that are closer to those of the Poisson-gamma rather than the independent random effects model, since the impact of the reporting pattern is reduced.

It is interesting to note that for a fixed number of reported claims, the relation between the predictive premiums and T_0 is different for the MZIP-Gamma model than for the hurdle models. Indeed, the lowest premium for the MZIP-Gamma model is for small T_0, while it is for high T_0 for the hurdle models.

Because of the complex analytic form of the densities of the duration models, we were not able to specify clearly the sufficient statistic of the predictive distribution. In fact, the complete pattern of reporting as an importance on the predictive premiums, as shown in Table 5. For illustration, we only compute the predictive premiums for two extreme situations: in the case where all the claims were reported on different insured periods (A) and in the case where all the claims were reported on the same insured period (B). Differences between the two situations are too small. Moreover, we can also see that these predictive premiums are very close the values obtained for the MVNB and the NB-Beta distributions.

As for the duration model, we were not able to express the predictive distribution of the $N_{i,t}$ as a function of a specific sufficient statistic. Table 6 also shows the predictive premiums depending on situations A and B. However, in this case, we see that large differences can be observed. The relation between the reporting pattern and the predictive premiums is very close to the one observed with the hurdle model.

Models	Situation	A priori	Sum of claims 0	1	2	3	4
MGCD	A	0.0651	0.0427	0.0792	0.1114	0.1402	0.1682
	B	0.0651	·	·	0.1097	0.1391	0.1669
MWCD	A	0.0651	0.0437	0.0761	0.1082	0.1415	0.1742
	B	0.0651	·	·	0.1092	0.1424	0.1764

Table 5. Predictive Premiums (Average Risk Profile) for the Duration Models

Model	Situation	A priori	Sum of claims 0	1	2	3	4
MVNB$_x$	A	0.0655	0.0442	0.0783	0.1113	0.1434	0.1780
	B	0.0655	·	·	0.0970	0.1153	0.1318

Table 6. Predictive Premiums (Average Risk Profile) for the MVNB$_x$ Model

This is not a coincidence since there are close similarities between the two models that we would like to investigate further. Indeed, both models can be expressed as a compound sum $\sum_{i=1}^{M} X_i$, where:

1. $M \sim \text{Poisson}$ and $X_i \sim \text{Logarithmic}$ for the NegBin_x distribution;

2. $M \sim \text{Bernoulli}$, $X_i^* = X_i - 1$ and $X_i^* \sim \text{Poisson}$ for the hurdle distribution.

10 Conclusion

Cost-based pricing of individual risks is a key actuarial ratemaking principle. The price charged to policyholders is an estimate of the future costs related to the insurance coverage. The pure premium approach defines the price of an insurance policy as the ratio of the estimated costs of all future claims against the coverage provided by the insurance policy while it is in effect to the risk exposure, plus expenses.

The property and casualty ratemaking is based on a claim frequency distribution and a loss distribution. The claim frequency is defined as the number of incurred claims per unit of earned exposure. The exposure is measured in car-year for motor third party liability insurance (the rate manual lists rates per car-year).

In a free market, insurance companies need to use a rating structure that matches the premiums for the risks as closely as possible, or at least as closely as the rating structures used by competitors. This entails using virtually every available classification variable correlated to the risks, since failing to do so would mean sacrificing the chance to select against competitors, and incurring the risk of suffering adverse selection by them.

We have shown how panel data models can be useful in insurance to derive the distribution of the number of claims for one period ahead, given information on the past.

Acknowledgement. The authors would like to thank the anonymous referees for their careful reading of the manuscript and the resulting comments that greatly helped to improve the paper. Jean-Philippe Boucher would like to thank the *Université du Québec à Montréal* for providing *Programme d'aide financière à la recherche et à la création - PAFARC* grant and also acknowledges the financial support from the Natural Sciences and Engineering Research Council of Canada. Montserrat Guillén would like to thank the Spanish Ministry of Education and Science / FEDER SEJ2007-63298.

References

[1] BOUCHER, J.-P. AND DENUIT, M., (2006). Fixed versus Random Effects in Poisson Regression Models for Claim Counts: Case Study with Motor Insurance, *ASTIN Bull.*, **36**, 285–301.

[2] BOUCHER, J.-P. AND DENUIT, M., (2007). Duration Dependence Models for Claim Counts, *Deutsche Gesellschaft fur Versicherungsmathematik (German Actuarial Bulletin)*, **28**, 29–45.

[3] BOUCHER, J.-P.,DENUIT, M. AND GUILLÉN, M., (2007). Risk Classification for Claim Counts: A Comparative Analysis of Various Zero-Inflated Mixed Poisson and Hurdle Models, *N. Am. Actuar. J.*, **11**-4, 110–131.

[4] BOUCHER, J.-P., DENUIT, M. AND GUILLÉN, M., (2008). Correlated Random Effects for Hurdle Models Applied to Panel Data Count, *Université du Québec à Montréal*, working paper.

[5] BOUCHER, J.-P., DENUIT, M. AND GUILLÉN, M., (2008). Modeling of Insurance Claim Count with Hurdle Distribution for Panel Data, in *Advances in mathematical and Statistical Modeling, Statistics for Industry and Technology (SIT)*, Birkhäuser Boston, Inc..

[6] BOUCHER, J.-P., DENUIT, M. AND GUILLÉN, M., (2008). Models of Insurance Claim Counts with Time Dependence Based on Generalisation of Poisson and Negative Binomial Distributions, *Variance*, **2**(1), 135–162.

[7] BOUCHER, J.-P., DENUIT, M. AND GUILLÉN, M., (2008). Number of Accidents or Number of Claims? An Approach with Zero-inflated Poisson Models for Panel Data, *Journal of Risk and Insurance*, to appear.

[8] BOUCHER, J.-P. AND GUILLÉN, M., (2008). A Semi-Nonparametric Approach to Model Panel Count Data, *Université du Québec à Montréal*, working paper.

[9] BRADLOW, E., FADER, P., ADRIAN, M. AND MCSHANE, B., (2008). Count Models Based on Weibull Interarrival Times, *J. Bus. Econom. Statist.*, **26**(3), 369–378.

[10] CAMERON, A. C. AND TRIVEDI, P. K., (1998). *Regression Analysis of Count Data*, New York, Cambridge University Press.

[11] CRAGG, J., (1971). Some statistical models for limited dependent variables with application to the demand for durable goods, *Econometrica*, **39**(5), 829–844.

[12] DENUIT, M., MARÉCHAL, X., PITREBOIS, S. AND WALHIN, J.-F., (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Scales*, Wiley, New York.

[13] FREES, E., V. R., Y. AND YU, L., (2001). Case Studies Using Panel Data Models, *N. Am. Actuar. J.*, **5**(4), 24–42.

[14] FREES, E. W. AND VALDEZ, E. A., (2008). Hierarchical Insurance Claims Modeling, *J. Amer. Statist. Assoc.*, **103**, 484, 1457–1469.

[15] GÓMEZ-DÉNIZ, E., SARABIA, J. AND CALDERÓN OJEDA, E., (2008). Univariate and multivariate versions of the negative binomial-inverse Gaussian distributions with applications, *Insurance Math. Econom.*, **42**(1), 39–49.

[16] GOLDEN, R., (2003). Discrepancy Risk Model Selection Test for Comparing Possibly Misspecified or Nonnested Models, *Psychometrika*, **68**, 229–249.

[17] HAUSMAN, J., HALL, B. AND GRILICHES, Z., (1984). Econometric Models for Count Data with Application to the Patents-R and D Relationship, *Econometrica*, **52**, 909–938.

[18] HINDE, J., (1982). Compound Poisson Regression Models, in R. Gilchrist, ed., *GLIM 82: Proceeding of the International Conference on Generalised Linear Models*, New York, Springer-Verlag.

[19] HOLGATE, P., (1964). Estimation for the Bivariate Poisson distribution, *Biometrika*, **51**, 241–245.

[20] HSIAO, C., (2003). *Analysis of Panel Data*, Cambridge, Cambridge University Press, 2nd ed.

[21] JOHNSON, N., KOTZ, S. AND BALAKRISHNAN, N., (1996). *Discrete Multivariate Distributions*, New York, Wiley, 2nd ed.

[22] LAMBERT, D., (1992). Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing, *Technometrics*, **34**, 1–14.

[23] MARSHALL, A. AND OLKIN, I., (1990). Multivariate Distributions Generated from Mixtures of Convolution and Product Families, in *Topics in Statistical Dependence*, H. W. Block, A. R. Sampson and T. H. Savits, eds. Lecture Notes-Monograph Series, Vol. **16**, 371–393.

[24] MOLENBERGHS, G. AND VERBEKE, G., (2005). *Models for Discrete Longitudinal Data*, Springer.

[25] MULLAHY, J., (1986). Specification and Testing in some Modified Count Data Models, *J. Econometrics*, **33**, 341–365.

[26] MULLAHY, J., (1997). Instrumental Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior, *Review of Economics and Statistics*, **79**, 586–593.

[27] MUNDLAK, Y., (1978). On the Pooling of Time Series and Cross Section Data, *Econometrica*, **46**, 69–85.

[28] PINQUET, J., GUILLÉN, M. AND BOLANCE, C., (2001). Allowance for the Age of Claims in Bonus-Malus Systems, *ASTIN Bull.*, **31**, 337–348.

[29] PURCARU, O., GUILLEN, M. AND DENUIT, M., (2004). Linear Credibility Models Based on Time Series for Claim Counts, *Belgian Actuarial Bulletin*, **4**, 62–74.

[30] RIVERS, D. AND VUONG, Q., (2002). Model Selection Tests for Nonlinear Dynamic Models, *Econom. J.*, **5**, 1–39.

[31] SANTOS SILVA, J. AND WINDMEIJER, F., (2001). Two-part Multiple Spell Models for Health Care Demand, *J. Econometrics*, **104**, 67–89.

[32] VUONG, Q., (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses, *Econometrica*, **57**, 307–333.

[33] WILLMOT, G., (1987). The Poisson-Inverse Gaussian Distribution as an Alternative to the Negative Binomial, *Scand. Actuar. J.*, **87**, 113–127.

[34] WINKELMANN, R., (1995). Duration Dependence and Dispersion in Count Data Models, *J. Bus. Econom. Statist.*, **13**, 467–474.

[35] WINKELMANN, R., (2008). *Econometric Analysis of Count Data*, Springer-Verlag, Berlin, 5th ed.

[36] XU, S., JONES, R. AND GRUNWALD, G., (2007). Analysis of Longitudinal Count Data with Serial Correlation, *Biom. J.*, **49**(3), 416–428.

[37] YOUNG, V. AND DE VYLDER, F., (2000). Credibility in Favor of Unlucky Insureds, *N. Am. Actuar. J.*, **4**, 107–113.

Jean-Philippe Boucher
Département de Mathématiques
Université du Québec à Montréal
Québec, Canada.

Montserrat Guillén
Department of Econometrics, RFA-IREA
University of Barcelona
Spain.