NUMBER OF ACCIDENTS OR NUMBER OF CLAIMS? AN APPROACH WITH ZERO-INFLATED POISSON MODELS FOR PANEL DATA

Jean-Philippe Boucher
boucher.jean-philippe@uqam.ca
Département de Mathématiques
Université du Québec à Montréal
Québec, Canada
Michel Denuit
michel.denuit@uclouvain.be
Institut des Sciences Actuarielles
Université Catholique de Louvain
Belgium
Montserrat Guillen
mguillen@ub.edu
Department of Econometrics
University of Barcelona
Spain

Accepted for publication in the *Journal of Risk and Insurance*

Abstract

The *hunger for bonus* is a well-known phenomenon in insurance, meaning that the insured does not report all of his accidents to save bonus on his next year's premium. In this paper, we assume that the number of accidents is based on a Poisson distribution but that the number of claims is generated by censorship of this Poisson distribution. Then, we present new models for panel count data based on the zero-inflated Poisson distribution. From the claims distributions, we propose an approximation of the accident distribution, which can provide insight into the behavior of insureds. A numerical illustration based on the reported claims of a Spanish insurance company is included to support this discussion.

Claim Count, Accident Count, Panel Data, Random Effects, Zero-Inflated Model, Predictive Distribution, Model Comparison, Vuong Test.

Acknowledgements

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. Jean-Philippe Boucher and Michel Denuit gratefully acknowledge the financial support of the *Communauté française de Belgique* under the *Projet d'Action de Recherches Concertées / PARC 04/09-320*. Montserrat Guillen would like to thank the Spanish Ministry of Education and Science.

1 Introduction

In various applications involving count data, the data exhibit a high number of zero values. This led to the idea that a distribution with excess zeros can provide a good fit, such as the zero-inflated distribution. In insurance, the *hunger for bonus* (Philipson (1960), Lemaire (1977)) is a well-known phenomenon that represents the fact that insureds do not report all their accidents to save bonus on the following year's premium. However, actuaries and researchers continue to model the number of claims with standard count distributions, neglecting the bonus hunger phenomenon. In this paper, we assume that the number of accidents is based on a Poisson distribution but that the number of claims is generated by censorship of this Poisson distribution.

Risk classification techniques for claim counts have been the topic of many papers in the actuarial literature. Denuit et al. (2007) provides an exhaustive overview of count data models for insurance claims. For cross-sectional data, Boucher et al. (2007) studied zero-inflated models in motor insurance claim counts, and compared them to hurdle models. Boucher et al. (2008) worked on risk classification models for the number of claims, in the context of panel data, but only studied classical count data models, like Poisson and negative binomial. In this paper, we propose new zero-inflated models that generalize the distribution for longitudinal data by introducing random effects in the model.

We found that the generalizations of the zero-inflated Poisson distribution has interesting applications for insurance data, where the number of accidents can be compared to the number of claims. In the second section of the paper, we review the standard techniques used to consider the differences between the number of claims and the number of accidents. An overview of the zero-inflated model and the basis for the construction of panel data distributions are given in the third part of the paper. In section 4, we propose various types of parameterizations of zero-inflated models for panel data. Random effects are added to the zero-inflated term and the count distribution. Another kind of multivariate distribution, comprising a special degenerated random effects distribution, is shown to have interesting properties.

In section 5, we show that predictive distributions and their expected predictive value can be expressed in a closed form for all proposed distributions. Section 6 contains a numerical illustration performed on the reported claims in a sample from a Spanish insurance company to support the discussion, which allows us to analyze the hunger for bonus situation. We show that the expected value and the variance can differ significantly according to the model. Statistical tests to compare models are explained and a Vuong-Golden test is used to compare the fitting of non-nested models. Our results show that some of the models presented here with insurance data have a better fit than the commonly used Poisson distribution with gamma random effects.

2 Number of Claims versus Number of Accidents

In most of the bonus-malus schemes throughout the world (see Lemaire (1995) for an overview), a reported claim causes an increase in the premium for the following years. These types of system induce a hunger for bonus (Lemaire (1977)), where there is an incentive not to report all incurred claims since the resulting increase in subsequent premiums can be greater than the claim indemnity. Consequently, the number of claims and the number of accidents are different.

Traditionally, actuaries and researchers do not consider the differences between number of claims and accidents. They must use the number of claims in their analysis since the number of accidents is not directly available in insurance company databases. In their modelling of claims data, they tend to use Poisson distributions and add heterogeneity to improve the fit. However, the number of claims reveals two kinds of behavior: the driving behavior of the insured and the way the insured behaves when he has to report an accident or not. Adjusting two behaviors by adding a random effects variable that corrects all the misadjustments is far too general. In this section of the paper, by summarizing this area of research, we explore the standard way of linking the accidents and claim numbers.

2.1 Standard Approach and Lemaire's Model

A standard approach is based on a mix between a Poisson distribution that models the number of accidents and a Bernoulli variable that models the probability p that the accidents will be filed. Using the classic interpretation, this probability is set using a claiming threshold. If the cost of an accident is below this threshold, the accident is not reported, while it is reported in the opposite case. Thus, if we denote as M the number of accidents and as N the number of claims reported during a given period, we can express the model as:

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

where the B_j are i.i.d. Bernoulli(p) variables. Usually, the parameter p of the Bernoulli variable is estimated using the distribution of the claims amount. If we assume that the number of accidents M follows a Poisson(λ) distribution, using the compound frequency properties, we can easily prove that $N \sim Poisson(\lambda p)$.

Since the number of claims follows a Poisson distribution with parameters λp, it causes an identification problem when only claims data are considered. Indeed, with this model, it is impossible to derive the accident distribution from the claim distribution since the parameters can also be expressed as $\lambda p = \exp(\log(p) + \beta_0)$, where $\log(p)$ cannot be distinguished from the intercept β_0 of the model. Consequently, underreporting cannot be directly estimated. Moreover, under this model, since the number of claims is also following a Poisson distribution underreporting will not create an excess number of zeros with respect to the Poisson distribution for the claims distribution, as opposed to what can be seen in practice.

Based on this standard model, Lemaire (1977) developed an approach where policyholders report their accidents only if they can obtain some benefit, i.e. if the claim cost exceeds the discount obtained for a no-claim record. Lemaire's model assumes that insureds are completely rational: they know how a bonus-malus system works, they know its rules, and they can then calculate an optimal threshold from which it is profitable to claim. Consequently, when they have an accident, they have a certain likelihood of filing a claim with their insurer, associated with the cost of the accident and the number of claims already reported to their insurer in the same time period. This way of thinking is applied for each accident. Obviously, from Lemaire's point of view, the i.i.d. assumption of the B_j seems too strong here.

This modelling of the hunger for bonus phenomenon is appealing, but purely theoretical. Indeed, the model presumes a completely rational insured who has knowledge of the bonus-malus system, of

his annual claim frequency, discount factor, and the distribution of losses he faces. In practice, these assumptions do not hold. Consequently, for an insurance portfolio, the distribution of the number of claims may be different from what we would expect from that model. Other assumptions about the way insureds file their accidents may result in claims distribution different from the Poisson distribution.

3 Modelling Overview

3.1 Zero-Inflated Distribution

A high number of zero-values is often observed in the fitting of count data. This leads to the idea that a distribution with excess zeros can provide a good fit. A finite mixture models of two distributions combining an indicator distribution for the zero case and a standard count distribution (Mullahy (1986), Lambert (1992)) is a natural candidate to deal with the high number of zero-values. As a result, this distribution will account for the excess zeros of the empirical distribution. When applied to cross-section data, the density of this type of model, with $0 < \phi < 1$, can be expressed as:

$$Pr[N = n] = \begin{cases} \phi + (1-\phi)Pr[K = 0] & \text{for } n = 0 \\ (1-\phi)Pr[K = n] & \text{for } n = 1, 2, ... \end{cases}, \tag{3.1}$$

where the random variable K follows the standard distribution to be modified by an additional excess zero function. In the limiting case, where $\phi \to 0$, the zero-inflated model corresponds to the distribution of K.

Many distributions may be used with the zero-inflated models. Obviously, the classic distribution is the zero-inflated Poisson (ZIP) distribution. With the use of equation (3.1), the density of the ZIP model is :

$$Pr[N = n] = \begin{cases} \phi + (1-\phi)e^{-\lambda} & \text{for } n = 0 \\ (1-\phi)\frac{e^{-\lambda}\lambda^n}{n!} & \text{for } n = 1, 2, ... \end{cases}. \tag{3.2}$$

From this, we can determine the first two moments of the ZIP distribution $E[N] = (1-\phi)\lambda$ and $Var[N] = E[N] + E[N](\lambda - E[N])$. The ZIP model could be useful for modelling purposes because it accounts for overdispersion. Since the only difference between the ZIP model and the Poisson distribution is found when $N = 0$, it is easy to adjust the MLE equations of the Poisson distribution to find the parameters of the ZIP model. Other count distributions can also be used with the zero-inflated distributions, such as the negative binomial, the Poisson-inverse Gaussian or the Poisson-lognormal (Boucher et al. (2007)).

3.2 Cross-Section versus Panel Data

The data used for panel data analysis consist of N individual units, each having T observations. Data are usually subjected to cross-sectional analysis, where each observation of an individual unit is considered to be mutually independent. Thus, in this situation we worked with $N \times T$ independent observations. However, it is often the case that some data are a repetition for the same individual. Consequently, it can be useful to model the dependence between these observations.

Longitudinal data (or panel data) consist of repeated observations of individual units observed over time. Each individual is assumed to be independent, but correlation between observations of the same individual is allowed.

There are several models that include time dependence but the most popular way of dealing with these data is to use a common individual term (Hausman et al. (1984)) that affects all the observations of an individual unit. This random effect represents individual specificities that are constant over time. The hidden individual characteristics that are not captured by the covariates are also captured by this random effect.

Given the insured-specific random effect term θ_i, the annual claim numbers $N_{i,1}, N_{i,2}, \ldots, N_{i,T}$ are independent. and under the assumption that the covariates are independent from the random effects (see Hsiao (2003) for a general review), the joint probability function of $N_{i,1}, \ldots, N_{i,T}$ is thus given by:

$$\begin{aligned} Pr[N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T}] &= \int_0^\infty Pr[N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T}|\theta_i] g(\theta_i) d\theta_i \\ &= \int_0^\infty \left(\prod_{t=1}^{T} Pr[N_{i,t} = n_{i,t}|\theta_i] \right) g(\theta_i) d\theta_i, \end{aligned} \tag{3.3}$$

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. Many conditional distributions for the random variables $N_{i,t}$ can be chosen in conjunction with a distribution for the random effect θ_i. By conditioning, moments of the joint distribution can be found quite easily.

3.2.1 Multivariate Negative Binomial

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

$$N_{i,t}|\theta_i \sim Poisson(\theta_i \lambda_i), \quad i = 1, ..., N \quad t = 1, ..., T..$$

where $\lambda_i = exp(x_i'\beta)$, which is time independent, is used to model covariates in the distribution (Dionne and Vanasse (1989, 1992)). Many possible distributions for the random effects can be chosen. A gamma distribution can be used to express the joint distribution in closed form. Indeed, the joint distribution of all the observations of a singe unit, when the random effects follow a gamma distribution of mean 1 and variance α, is equal to (Hausman et al. (1984)) :

$$\begin{aligned} &Pr[N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T}] = \\ &\left[\prod_{t=1}^{T} \frac{(\lambda_i)^{n_{i,t}}}{n_{i,t}!} \right] \frac{\Gamma(\sum_{i=1}^{T} n_{i,t} + 1/\alpha)}{\Gamma(1/\alpha)} \left(\frac{1/\alpha}{T\lambda_i + 1/\alpha} \right)^{1/\alpha} (T\lambda_i + 1/\alpha)^{-\sum_{i=1}^{T} n_{i,t}}. \end{aligned} \tag{3.4}$$

This distribution, which has often been applied, is known as a multivariate negative binomial (MVNB) or negative multinomial. Using time invariant covariates, a sufficient statistic for this distribution is the sum of counts over time T, as we can see in the equation above. For the MVNB

distribution, $E[N_{i,t}] = \lambda_i$ and $Var[N_{i,t}] = \lambda_i + \alpha\lambda_i^2$, which accounts for overdispersion. Maximum likelihood and variance estimates of the parameters are straightforward.

The MVNB distributions have the following moments:

$$\begin{aligned} E[N_{i,t}] &= \lambda_i & (3.5)\\ Var[N_{i,t}] &= \lambda_i + \alpha\lambda_i^2 & (3.6)\\ Cov[N_{i,t}, N_{i,t+j}] &= \alpha\lambda_i^2, \quad j > 0. & (3.7) \end{aligned}$$

Other mixing distributions can be chosen to model the random effects, such as the inverse Gaussian (Holla (1966) or Shoukri et al. (2004)) or the lognormal (Hinde (1982)) distributions, which result in distributions with the same form for the two first moments, with distinctions found using higher moments. The MVNB distribution has often been used to model panel count data. Consequently, it can be used as a comparison with the zero-inflated models presented in this paper.

4 Multivariate Zero-Inflated Models

4.1 Generalizations of the Zero-Inflated Poisson 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 be interpreted quite easily. Obviously, the addition of a random effect to the model can generalize the model for panel data. However, if we look at equation (3.2), we will see that this generalization can be carried out in many ways. Indeed, 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 both random effects. By conditioning on these two random effects, the joint distribution can be expressed as:

$$\begin{aligned} Pr[N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T}|\phi_i, \theta_i] &= \prod_{t=1}^{T} \Big(I_{(n_{i,t}=0)}\phi_i + (1-\phi_i) Pr[N_{i,t} = n_{i,t}] \Big) \\ &= \sum_{j=0}^{T_0} \binom{T_0}{j} V_j^{Poi}(n_{i,1}, ..., n_{i,T}|\theta_i) \phi_i^{T_0-j} (1-\phi_i)^{(T-T_0)+j}, \end{aligned} \tag{4.1}$$

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

$$V_j^{Poi}(n_{i,1}, ..., n_{i,T}|\theta_i) = \frac{(\lambda_i\theta_i)^{\sum_{t=1}^{T} n_{i,t}} \exp\left(-(T - T_0 + j)\lambda_i\theta_i\right)}{\prod_{t=1}^{T} n_{i,t}!}. \tag{4.2}$$

Using one or both random effects, the joint distribution can be expressed in closed form. The conditional moments of this distribution are equal to:

$$
\begin{aligned}
E[N_{i,t}] &= \lambda_i \left(E[\theta_i] - E[\theta_i\phi_i]\right) && (4.3)\\
Var[N_{i,t}] &= \lambda_i \left(E[\theta_i] - E[\theta_i\phi_i]\right) \\
&+ \lambda_i^2 \Big(E[\phi_i\theta_i^2] - E[\phi_i^2\theta_i^2] + Var[\theta_i] \\
&\quad + E[\phi_i]^2 Var[\theta_i] + E[\theta_i]^2 Var[\phi_i] + Var[\phi_i]Var[\theta_i] - 2E[\phi_i]Var[\theta_i] \Big). && (4.4)
\end{aligned}
$$

We add random effects to the model with an individual term ϕ_i that is beta distributed with parameters a_i and b. Alternatively, we also assume that a heterogeneity term θ_i is added to the Poisson distribution and follows a gamma distribution of mean 1 and variance α. Consequently, when the two random effects ϕ and θ are assumed to be independent, as proposed in a GEE approach by Song (2005), it leads to a multivariate zero-inflated Poisson beta gamma model (MZIP-BetaGamma) that can be expressed as:

$$
Pr[N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T}] = \sum_{j=0}^{T_0} \binom{T_0}{j} V_j^{NB}(n_{i,1}, ..., n_{i,T}) \frac{\beta(a + T_0 - j, b + (T - T_0) + j)}{\beta(a, b)}, \tag{4.5}
$$

where T_0 is again the number of insured periods without claims and the function $V_j^{NB}(.)$ has the following multivariate negative binomial form:

$$
V_j^{NB}(n_{i,1}, ..., n_{i,T}) = \frac{\Gamma(\sum_t^T n_{i,t} + 1/\alpha)}{\Gamma(1/\alpha)\prod_t^T n_{i,t}!} \left(\frac{1/\alpha}{(T - T_0 + j)\lambda_i + 1/\alpha}\right)^{1/\alpha} \left(\frac{\lambda_i}{(T - T_0 + j)\lambda_i + 1/\alpha}\right)^{\sum_t^T n_{i,t}} \tag{4.6}
$$

For the purposes of illustration, let us also mention that apart from the gamma, the inverse-Gaussian can also be computed in a closed form, which leads to a multivariate zero-inflated Poisson inverse-Gaussian distribution (MZIP-BetaInverse Gaussian). Using an inverse-Gaussian distribution of unit mean and variance τ, the $V_j^{IG}(.)$ function is equal to:

$$
V_j^{IG}(n_{i,1}, ..., n_{i,T}) = \frac{\lambda_i^{\sum_t^T n_{i,t}}}{\prod_t^T n_{i,t}!} \left(\frac{2}{\pi\tau}\right)^{0.5} e^{1/\tau} (1 + 2\tau\lambda_i(T - T_0 + j))^{-s_i/2} K_{s_i}(z_i), \tag{4.7}
$$

where $s_i = \sum_t^T n_{i,t} - 0.5$, $z_i = (1 + 2\tau\lambda_i(T - T_0 + j))^{0.5}/\tau$ and $K_j(.)$ is the modified Bessel function of the second kind. Properties of the modified Bessel function of the second kind related to the Poisson-Inverse Gaussian distribution may be found in Shoukri et al. (2004). A lognormal distribution can also be used but the joint distribution cannot be expressed in closed form.

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

$$
\begin{aligned}
E[N_{i,t}] &= \lambda_i \left(1 - \frac{a_i}{a_i + b}\right) && (4.8)\\
Var[N_{i,t}] &= \lambda_i \left(1 - \frac{a_i}{a_i + b}\right) + \lambda_i^2 \left(1 - \frac{a_i}{a_i + b}\right)\left(\frac{a_i}{a_i + b} + \alpha\right) && (4.9)\\
Cov[N_{i,t}, N_{i,t+j}] &= \lambda_i^2 \left[\left(1 - \frac{a_i}{a_i + b}\right)^2 \alpha + \frac{a_i b}{(a_i + b)^2(a_i + b + 1)}(1 + \alpha)\right], \quad j > 0. && (4.10)
\end{aligned}
$$

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

4.1.1 Specification Tests

In the case where the random effects on the count distribution is removed, the joint distribution can be called multivariate zero-inflated Poisson beta (MZIP-Beta), while a multivariate zero-inflated Poisson gamma (MZIP-Gamma) can be used in the situation where ϕ is not randomly distributed. Compared to the MZIP-BetaGamma, these two models are obviously equivalent for some simple parameter restrictions. This can be tested using a classic hypothesis with two standard tests: the log-likelihood ratio and the Wald tests. These tests are asymptotically equivalent. In some cases, the parameter restriction concerns the boundary of the parameter space. One problem with the Wald or log-likelihood ratio tests arises when the null hypothesis is on the boundary of the parameter space. When a parameter is bounded by the H_0 hypothesis, the estimate is also bounded and the asymptotic normality of the MLE no longer holds under H_0. Consequently, a correction must be performed. Results from Chernoff (1954) for the log-likelihood ratio statistic and Moran (1971) for the Wald test showed that under the null hypothesis, the distribution of the LR statistic is a mixture of a probability mass of $\frac{1}{2}$ on the boundary and a half-$\chi^2(1)$ distribution above zero. Then, when testing at a level δ, one must reject the H_0 hypothesis if the test statistic exceeds $\frac{1}{2}\ \chi^2_{1-2\delta}(1)$, rather than $\chi^2_{1-\delta}(1)$. Consequently, in this situation, a one-sided test must be used. Analogous results stand for the Wald test since parameter distribution consists of a mass of one half at zero and a normal distribution for the positive values. In this case, the usual one-sided test critical value of $z_{1-\delta}$ is used (see Cameron and Trivedi (1998) for more details).

4.2 Generalization of the Poisson Distribution

Instead of considering the zero-inflated term and its standard count distribution as two different elements that can possess their own random effects, we propose to consider a Poisson distribution which has a special degenerated random effects distribution.

The motivation is taken from Boucher and Denuit (2006), who used fixed effects estimation to approximate the random effects of Poisson distribution for panel data. They observed a significant presence of zero-values for each individual because many insureds did not a report a single claim. This situation complicates the regression analysis given that a mass point at zero exists. As a result, the authors use a weighted regression for gamma, inverse-Gaussian and lognormal distributions to approximate the random effects distribution.

Instead of using a weighted regression, we propose to use a two-part distribution: the first part determines whether θ_i is different from zero while the second part, conditional on the success of the first part, determines its value. Formally, the distribution of the random effects could be modelled as follows:

$$g(\theta_i|\mu_i,\tau) = \begin{cases} \phi_i & \text{for } \theta_i = 0 \\ (1-\phi_i)\, f(\theta_i) & \text{for } \theta_i \geq 0 \end{cases}, \tag{4.11}$$

where ϕ_i as a mass point modelled as $\frac{\exp(x_i'\gamma)}{1+\exp(x_i'\gamma)} = \frac{\mu_i}{1+\mu_i}$ and $f(.)$ has a standard heterogeneity distribution of mean 1 (without loss of generality) and a dispersion parameter ν, such as gamma,

inverse-Gaussian or lognormal. When f follows a gamma distribution, the joint distribution of all claims reported by insured i can be expressed as:

$$\begin{aligned} Pr[N_{i,1}=n_{i,1},...,N_{i,T}=n_{i,T}] &= \int_0^\infty \prod_{t=1}^{T} \frac{e^{-\lambda_{i,t}\theta_i}(\lambda_{i,t}\theta_i)^{n_i}}{n_i!} g(\theta_i|\mu_i,\tau)d\theta_i \qquad (4.12) \\ &= I_{(T_0=T)}\phi_i + \\ &\quad (1-\phi_i)\frac{\Gamma(\sum_{i=1}^{T} n_{i,t}+1/\alpha)}{\Gamma(/\alpha)\prod_t^T n_{i,t}!}\left(\frac{1/\alpha}{T\lambda_i+1/\alpha}\right)^{1/\alpha}\left(\frac{\lambda_i}{T\lambda_i+1/\alpha}\right)^{\sum_{i=1}^{T} n_{i,t}} \qquad (4.13) \end{aligned}$$

This joint distribution can be called zero-inflated multivariate negative binomial (ZI-MVNB) since it has the form of a multivariate negative binomial distribution with a zero-inflated term equal to ϕ_i. For the MZIP-inverse Gaussian distribution, note that we can also use an inverse-Gaussian to model the degenerated random effects distribution, which would also lead to a closed form expression for the joint distribution. The model exhibits the following moments:

$$\begin{aligned} E[N_{i,t}] &= \lambda_i\left(1-\frac{\mu_i}{1+\mu_i}\right) \qquad (4.14) \\ Var[N_{i,t}] &= \lambda_i\left(1-\frac{\mu_i}{1+\mu_i}\right)+\lambda_i^2\alpha\left(1-\frac{\mu_i}{1+\mu_i}\right)^2 \qquad (4.15) \\ Cov[N_{i,t},N_{i,t+j}] &= \lambda_i^2\alpha\left(1-\frac{\mu_i}{1+\mu_i}\right)^2. \qquad (4.16) \end{aligned}$$

The model implies overdispersion, as do the other models constructed on a generalization of the zero-inflated Poisson distribution.

4.3 Potential Analysis based on the Zero-Inflated Models

The zero-inflated Poisson distribution for panel data is an appealing distribution that can be used in many areas where an excess of zero is detected. From a pure theoretical point of view, these distributions are appealing for biostatistics, econometrics and other statisticians. However, we found that these generalizations of the zero-inflated Poisson distributions have a particular application to insurance data.

As mentioned in Section 2, the uses of other distributions to model claim counts were motivated by the hunger for bonus situations that can occur in practice. The zero-inflated distributions applied to the number of claims can be used to model the behavior of insureds, i.e. to model the probability of reporting an accident. Indeed, because the models linked to a reporting decision at the period level, and not at the accident level as with Lemaire's model, we can conceive that each year, a number of insureds will not claim at all, whatever the case. However, in this situation, one might question why these insureds procure insurance. Some explanations refer to their fear of insurance, their having minimal protection (mandatory insurance), or their being insured only for major (with probability close to 0).

Another way of interpreting this model has some close connections with Lemaire's model, which also assumes that the number of accidents is Poisson distributed. In addition, it considers the probability of each accident's being reported. However, unlike the Lemaire's model, our models assumes

that the insureds do not really know how a bonus-malus system works and do not use any kind of algorithm when deciding whether to claim. More specifically, the first accident of each insured year indicates the way the insured will act for the rest of the year. Accordingly, if the first accident is reported, so will all the other accidents. If the first accident is not reported, nor will the other accidents. This is clearly an approximation, but seems realistic because insureds think that once they have *lost their bonus*, the other claims do not have an impact. Those that will not claim at first, because they made an effort to financially support their decision, tend to defend the way they act and will consequently not claim other accidents. In some highly uncommon situations where a major accident followed a non-reported accident, an insured would probably claim to his insurer. However, because the vast majority (more than 99.5% in our data) of the insureds reports less than two claims per year and given that major accidents are infrequent, this situation happens with a probability very close to 0. Nevertheless, this approximation error should always be kept in mind and be considered when the accident distribution is analyzed. As shown by Lemaire and Zi (1994), this interpretation seems fair with the Spanish data used in Section 6 since, at that time in Spain most insurers used harsh systems where all discounts were lost because of a single claim. However, we also think that this non-optimal strategy of deciding to report or not their first claim, followed by the same reporting behavior for every subsequent claims, can be applied to other jurisdictions. Indeed, these irrational behaviors of insureds can simply be explained by the fact that many of them do not understand the way insurers set the premiums.

Using a reporting decision at the period level allows us to distinguish the underreporting from the driving behavior. Consequently, using zero-inflated distributions on the number of claims, the idea is to *uncensor* these zero-inflated distributions to obtain an approximation of the accident frequency distribution. By removing all the effects of reporting that we modelled by the censorship parameter ϕ, the accident process is Poisson distributed (as in Lemaire's model), which is simple and easy to be understand. It seems intuitive to model the accident process by some classic count distribution such as the Poisson distribution because its interpretation is direct, as a limit of a Binomial distribution with the number of tries going to infinity and the accident probability tending to 0. Additionally, note that the zero-inflated models allow us to approximate the accident distribution, even without a deep understanding of the knowledge of the bonus-malus system.

For the modelling of claim counts, the ZI-MVNB model further generalizes the MZIP-BetaGamma models in the decision to claim the accident or not. Indeed, in the standard approach and in Lemaire's model, the decision to claim or not is made at each accident, while for the MZIP-BetaGamma model, the decision is made only for the first accident of each insured period, other accidents being filed similarly. In contrast, for the ZI-MVNB model, the decision is done only for the first accident of the first insured period, other accidents being reported similarly. This model seems appealing in the modelling of basic bonus-malus systems (like in Canada), where the bonus is lost if an accident has been claimed in the last 3 or 5 years.

Obviously, ideally, the ϕ parameter of the ZI-MVNB model should be modelled as dynamic, where it can decrease gradually each year, maybe because of the impact of a major accident. In short, we can interpret this model as a situation where some insureds will not claim at all for all insured periods. This kind of insured buys insurance only to obey the law, meaning that they will not report accident because their coverage is minimal or because an increase would make their premium too too expensive.

4.3.1 Distribution of the Number of Accidents

If the uncensored version of the zero-inflated distribution is used, all the ϕ parameters must be set to zero. In other words, the zero-inflated distributions are fitted to claims data to find the $\lambda_i = \exp(x_i'\beta)$ of the Poisson distribution of the number of accidents. We can also approximate the value of the heterogeneity parameter if we make the strong assumption that the heterogeneity of the claims distribution is the same as the one for the accident distribution, as done by Boucher and Denuit (2008), Walhin and Paris (2000) and even Lemaire (1977).

4.3.2 Other

Obviously, the analysis of the number of accidents distribution is interesting, but the analysis of the ϕ parameter, representing the censoring caused by the hunger for bonus that can be seen as an economic function, also has great potential. Indeed, it can be intriguing to determine what exogenous factors can affect this value. In the actuarial history, it is strongly believed that this kind of factor is highly influenced by the penalties assigned to drivers in case of a claim (Walhin and Paris (2000), Lemaire (1977)). However, we can also assume that the number of past claims or even the personality of the driver can have an effect.

Consequently, insureds that file many claims can be identified by some marketing studies to analyze the reasons that they exhibit this behavior. Introducing the new zero-inflated distributions with the analysis of the amount of claim distribution should enable actuaries to understand their insurance portfolio in greater depth. Consequently, deductibles and limits analysis can be done more directly.

Other analysis related to risk measurement can also be performed since profiles of insureds that do not file many claims can be seen as potentially riskier. Indeed, even if their driving conduct stays the same, the way they claim can change. Thus, they are affected by two processes, and the evolution of their claiming behavior can have more negative impacts on an insurer than would insureds that already claim almost all of their accidents. Fraud analysis can also be based on these zero-inflated models, where profiles of insureds, or characteristics of insureds who seem to claim more than expected can be revealed (see Artis et al. (1999) for an example of fraud analysis in insurance). As shown by Boucher and Denuit (2008), the analysis of the predictive premiums can also be seen with the *hunger for bonus* problem. Indeed, the authors use the MZIP-Gamma model for panel data and consider ϕ_i as a function that models the hunger for bonus as a function of the severity of the bonus-malus system.

5 Predictive Distributions

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).

Many important factors cannot be taken into account in the *a priori* risk classification. These include swiftness of reflexes, drinking habits or compliance with the highway code. Consequently, tariff cells are still quite heterogeneous despite the use of many classification variables. This heterogeneity can be modelled by a random effect in a statistical model. It is reasonable to believe that the hidden characteristics are partly revealed by the number of claims reported by the policyholders.

In consequence, for each period t, the heterogeneity terms (θ_i or ϕ_i) can be updated for past experience. Formally, the computation of the predictive distribution is performed using Bayesian analysis for the random effects:

$$\begin{aligned} & P(N_{i,T+1} = n_{i,T+1} | N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T}) \\ = & \int\int \Pr(N_{i,T+1} = n_{i,T+1} | \phi_i, \theta_i) \left(\frac{[\prod_t \Pr(N_{i,t} = n_{i,t} | \phi_i, \theta_i)]\, g(\phi_i, \theta_i) d\phi_i d\theta_i}{\int [\prod_t \Pr(N_{i,t} = n_{i,t} | \theta_{i,1}, \theta_i)]\, g(\phi_i, \theta_i)} \right) d\phi_i d\theta_i \\ = & \int\int \Pr(N_{i,T+1} = n_{i,T+1} | \phi_i, \theta_i) g(\phi_i, \theta_i | n_{i,1}, ..., n_{i,T}) d\phi_i d\theta_i, \end{aligned} \tag{5.1}$$

where $g(\phi_i, \theta_i)$ is the *a posteriori* distribution of the heterogeneity terms (θ_i or ϕ_i), reflecting the claims history of insured i. If this *a posteriori* distribution can be expressed in closed form, moments of the predictive distribution can easily be found by conditioning on the random effects.

5.1 MVNB distribution

As is well known, the posterior distribution of the random effect term for the Poisson model with gamma random effects is also gamma-distributed with parameters $T\lambda_i + 1/\alpha$ and $\sum_t^T n_{i,t} + 1/\alpha$. Using the posterior distribution, the predictive distribution can be computed directly as in section 3.2.1. For the purposes of illustration, we can express the expected predictive value of this model as:

$$E[N_{i,T+1} | N_{i,1}, ..., N_{i,t}] = \lambda_i \frac{\sum_t^T n_{i,t} + 1/\alpha}{T\lambda_i + 1/\alpha}. \tag{5.2}$$

This can be used for comparison with the other expected predictive values of the multivariate zero-inflated models. Note that the predictive distribution will depend only on the sum of past counts. Additionally, note that when the observed periods go to infinity, the expected predictive value of an individual unit will converge to its average number of past counts.

5.2 MZIP-BetaGamma Distribution

The use of both individual effects leads to the following predictive distribution:

$$\begin{aligned} & Pr[N_{i,T+1} = n_{i,T+1} | n_{i,1}, ..., n_{i,T}] \\ = & \frac{\sum_{j=0}^{T_0^*} \binom{T_0^*}{j} V_j^{NB}(n_{i,1}, ..., n_{i,T}, n_{i,T+1}) \beta(a + T_0^* - j, b + (T + 1 - T_0^*) + j)}{\sum_{j=0}^{T_0} \binom{T_0}{j} V_j^{NB}(n_{i,1}, ..., n_{i,T}) \beta(a + T_0 - j, b + (T - T_0) + j)}. \end{aligned} \tag{5.3}$$

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

$$K(j) = \frac{\binom{T_0}{j}\Gamma(a+T_0-j)\Gamma(b+T+1-T_0+j)((T-T_0+j)\lambda_i+1/\alpha)^{-(\sum_t^T n_{i,t}+1/\alpha)}}{(a_i+b+T)\sum_{k=0}^{T_0}\binom{T_0}{k}\Gamma(a+T_0-k)\Gamma(b+T-T_0+k)((T-T_0+k)\lambda_i+1/\alpha)^{-(\sum_t^T n_{i,t}+1/\alpha)}}, \tag{5.4}$$

the predictive distribution can be expressed as:

$$Pr[N_{i,T+1}=n_{i,T+1}|n_{i,1},...,n_{i,T}] = \begin{cases} 1-\sum_{j=0}^{T_0}K(j)(1-p^r) & \text{for } n_{i,T+1}=0 \\ \sum_{j=0}^{T_0}K(j)Pr_{NB}[N_{i,T+1}=n_{i,T+1};r,p] & \text{for } n_{i,T+1}=1,2,... \end{cases} \tag{5.5}$$

where:

$$Pr_{NB}[N_{i,T+1}=n_{i,T+1};r,p] = \binom{n_{i,T+1}+r}{r}p^r q^{n_{i,T+1}} \tag{5.6}$$

is the probability function of a negative binomial distribution with parameters equal to:

$$r = \sum_{t=1}^{T} n_{i,t}+1/\alpha, \quad p=\frac{(T-T_0+j)\lambda_i+1/\alpha}{(T+1-T_0+j)\lambda_i+1/\alpha} \tag{5.7}$$

and the expected predictive value is the equal to:

$$E[N_{i,T+1}|n_{i,1},...,n_{i,T}] = \lambda_i\sum_{j=0}^{T_0}\frac{\left(\sum_t^T n_{i,t}+1/\alpha\right)K(j)}{(T+1-T_0+j)\lambda_i+1/\alpha}. \tag{5.8}$$

5.3 ZI-MVNB Distribution

The predictive distribution of the degenerated random effects model can also be computed analytically:

$$Pr[N_{i,T+1}=n_{i,T+1}|n_{i,1},...,n_{i,T}] = \frac{I_{(n_{i,T+1}=0)}\phi_i+(1-\phi_i)L(n_{i,1},...,n_{i,T+1};T+1)}{I_{(T=T_0)}\phi_i+(1-\phi_i)L(n_{i,1},...,n_{i,T};T)}, \tag{5.9}$$

with:

$$L(n_{i,1},...,n_{i,T};T)=\frac{\Gamma(\sum_{i=1}^T n_{i,t}+1/\alpha)}{\Gamma(1/\alpha)\prod_t n_{i,t}!}\left(\frac{1/\alpha}{T\lambda_i+1/\alpha}\right)^{1/\alpha}\left(\frac{\lambda_i}{T\lambda_i+1/\alpha}\right)^{\sum_{i=1}^T n_{i,t}}. \tag{5.10}$$

This model can be analyzed for two kinds of insureds: those who reported at least one claim and the others who did not report any. For insureds who reported at least once, the predictive distribution returns to a standard multivariate negative binomial distribution since the first parts of the numerator and the denominator of equation (5.9) fall. Indeed, we found that the a posteriori distribution of the random effect term is a gamma distribution with parameters equal to $T\lambda_i+1/\alpha$

and $\sum_t^T n_{i,t} + 1/\alpha$, in keeping with the standard Poisson-Gamma combination. The first moments can be found easily where, for example, the expected value is equal to:

$$E[N_{i,t+1}|N_{i,1},...,N_{i,t},T_0 \neq T] \quad = \quad \lambda_i \frac{\sum_t^T n_{i,t} + 1/\alpha}{T\lambda_i + 1/\alpha}. \tag{5.11}$$

When the insured is claim-free, the predictive distribution has the following form:

$$Pr[n_{i,T+1}|n_{i,1},...,n_{i,T},T = T_0] \quad = \quad \frac{I_{(n_{i,T+1}=0)}\phi_i + (1-\phi_i)\left(\frac{1/\alpha}{T\lambda_i+1/\alpha}\right)^{1/\alpha} Pr_{NB}[n_{i,T+1};r,p]}{\phi_i + (1-\phi_i)\left(\frac{1/\alpha}{T\lambda_i+1/\alpha}\right)^{1/\alpha}} \tag{5.12}$$

where the notation $Pr_{NB}[n_{i,T+1};r,p]$ is defined in equation (5.6), with

$$r \quad = \quad n_{i,T+1}, \quad p = \frac{T\lambda_i + 1/\alpha}{(T+1)\lambda_i + 1/\alpha}. \tag{5.13}$$

Since the moments of this negative binomial are easily found, we can obtain the predictive moments quite directly. For example, the first moment of this predictive distribution can be expressed as:

$$E[N_{i,t+1}|N_{i,1},...,N_{i,t},T_0 = T] \quad = \quad \lambda_i \frac{(1-\phi_i)\left(\frac{1/\alpha}{T\lambda_i+1/\alpha}\right)^{1+1/\alpha}}{\phi_i + (1-\phi_i)\left(\frac{1/\alpha}{T\lambda_i+1/\alpha}\right)^{1/\alpha}}, \tag{5.14}$$

from which we can see that the premium of insured i goes to zero, which is his average number of reported claims, if the number of insured periods increases significantly.

5.4 Comparisons between Models

As mentioned earlier, generally, the number of claims is modelled with a Poisson distribution, where a random effects variable that corrects two kinds of behavior (driving behavior and the way an insured report accidents) is added to the count distribution. Regarding the Poisson distribution, the zero-inflated models add weight to the zero probability. As Mullahy (1997) asserts, the unique addition of an heterogeneity to the Poisson distribution can also correct the number of estimated insureds without claims. However, if used directly on claims data, this *correction* will put too much weight and importance on heterogeneity. Indeed, this kind of model generates predictive premiums that overpenalize insureds with many claims (see Section 6.5.1 or Young and De Vylder (2000) for example). All the zero-inflated models soften these penalties because the impact of the heterogeneity is weaker.

6 Numerical Application

We applied the presented models to the number of reported claim in third-liability. The reasons behind the good fit of the zero-inflated models for cross-section data can justify the generalizations

Variable	Description
v1	equals 1 for women and 0 for men
v2	equals 1 if the client has been with the company between 3 and 5 years
v3	equals 1 if the client has been with the company for more than 5 years
v4	equals 1 if the insured is 30 years old or younger
v5	equals 1 if engine power is larger to or equal than 5500 cc

Table 6.1: Exogenous variables

into panel data, where random effects are introduced in the zero-inflated term and the count distribution. Indeed, the hunger of bonus can be seen as individually specific, while the absence of some important classification variables (swiftness of reflexes, aggressiveness behind the wheel, consumption of drugs, etc.) can be used to justify the random effects on the count distribution.

6.1 Assumptions

As seen in the presentation of the zero-inflated models for panel data, covariates are time independent, that is to say that they do not change over all the observations of an individual unit. This assumption was also made by Gourieroux (1999) for insurance data. Even if this assumption causes additional heterogeneity in the data, and therefore a certain amount of anti-selection, it is not as restrictive as it might seem. Indeed, the majority of the time dependent covariates that are used in insurance involve the age of the insured or the length of stay with the company. These variables do not evolve randomly in time because the change can already be known in advance. Consequently, the estimated coefficient of a parameter can still be interpretable because the evolution of the parameter is known. Other covariates can also change in time, such as the city or the type of vehicle belonging to the driver, but this kind of major change often involves the issuance of another policy.

6.2 Data Used

In this paper, we worked with a sample of the automobile portfolio of a major company operating in Spain. Only cars for private use were considered in this sample. The panel data contain information for the period from 1991 to 1998. Our sample comprises 15,179 policyholders who remained with the company for seven complete periods. We used six complete years to estimate the parameters, which means that we are working with 91,074 contracts for the estimations, and reserved the last year to compare the predictions of the models with the observed results. Five exogenous variables (see Table 6.1) were kept in the panel plus the annual number of accidents. For every insured we used the initial information from his first policy. The total number of at fault claims that took place within each annual period was also used. The average claim frequency of the portfolio is 6.73%.

In this paper, the exogenous variables, shown in Table 6.1, were used to model the distribution parameters. The characteristics of the insureds were expressed through functions of the score statistic $\beta_0 + x_i'\beta$, where β_0 is the intercept and $\beta' = (\beta_1, ..., \beta_p)$ is a vector of regression parameters for explanatory variables $x_i = (x_{i,1}, ..., x_{i,p})$.

Parameter	MVNB (std. err)	MZIP-Gamma	ZI-MVNB
γ_0	.	−0.7299 (0.1755)	−1.9749 (0.4240)
γ_1	.	−0.6391 (0.2193)	−1.6402 (0.9953)
γ_3	.	0.6827 (0.1328)	1.2859 (0.3949)
γ_4	.	−0.2499 (0.1527)	−0.8115 (0.6014)
γ_5	.	−0.4193 (0.1183)	−0.9291 (0.2919)
β_0	−2.6811 (0.0378)	−2.3062 (0.0522)	−2.5514 (0.0344)
β_1	0.1145 (0.0438)	. .	. .
β_2	−0.1703 (0.0351)	−0.1698 (0.0345)	−0.1419 (0.0337)
β_3	−0.2210 (0.0398)	.	.
β_4	0.0599 (0.0370)	.	.
β_5	0.0980 (0.0339)	.	.
α	0.9407 (0.0508)	0.8304 (0.0526)	0.7678 (0.0693)
Loglike.	-22,619.58	-22,589.59	-22,622.11

Table 6.2: Estimated Parameters and Standard Errors

6.3 A Priori Analysis

6.3.1 Estimated Parameters

Models have been applied to our insurance data. The estimated parameters of these distributions are shown in Table 6.2, where the β_j and γ_j parameters refer to the impact of covariate j on the estimation of the λ and ϕ parameters respectively.

Using the precautions noted in the specification tests in Section 4.1.1, the data indicate that the MZIP-Beta is rejected against the MZIP-BetaGamma model (*p-value* ≤ 0.001), while no statistical difference is shown between the MZIP-Beta and the MZIP-BetaGamma (*p-value* of 0.4988). Consequently, at this point, we kept the MZIP-Gamma model, but did not consider the MZIP-Beta and the MZIP-BetaGamma models. Since the ZI-MVNB is nested to the MVNB model for $\phi \to 0$, we also checked differences between the ZI-MVNB and the MVNB model. No significant difference was shown between them (*p-value* greater than 0.95). This is not altogether surprising given that the same conclusion was obtained in Boucher et al. (2007), where the authors found that the ZI-Poisson model cannot be fitted with any heterogeneity distributions. Nevertheless, we kept this model in our numerical example because we think it could be interesting for smaller panel data sample, since the zero-inflated term needs to be larger for small T.

The analysis of the estimated parameters is particularly interesting. We can see that some covariates are included in the zero-inflated process while others are used only in the count distribution. Because this is a major variable in actuarial classification, we included the sex of the driver in the estimated parameters even if it was not statistically significant for some models. The values of all γ parameters are shown to be very different (MZIP-Gamma vs ZI-MVNB). This is caused by the presence (or absence) of the random effects of the count distribution that directly affects all the probabilities. Yet even if we see that the presence of random effects has an impact on all covariates, it can also be observed that the sign of all estimated parameters remains the same, meaning that the impact of the covariates is fairly similar depending on the model.

Profile Number	Kind of Profile	v1	v2	v3	v4	v5
1	Good	0	0	1	0	0
2	Average	1	1	0	0	0
4	Bad	1	0	0	1	1

Table 6.3: Profiles analyzed

	Good Profile		Average Profile		Bad Profile	
Models	Mean	Variance	Mean	Variance	Mean	Variance
MVNB	0.0549	0.0577	0.0648	0.0687	0.0899	0.0975
MZIP-Gamma	0.0510	0.0577	0.0670	0.0729	0.0882	0.0965
ZI-MVNB	0.0519	0.0540	0.0659	0.0692	0.0776	0.0822

Table 6.4: A priori Premiums

6.3.2 Premiums

Differences between models can be analyzed through the mean and the variance of selected insured profiles. The mean can be seen as the frequency component of the insurance premium, while the other part involves an analysis of the amount of claims reported. Several profiles were selected and are described in Table 6.3. The first profile is classified as a good driver, while the last one usually exhibits bad loss experience. The other profile is medium risk. The results are given in Table 6.4. This table shows that the expected values exhibit small differences for the five models studied. For example, we can see that the ZI-MVNB model offers a smaller insurance premium for the higher risk profile. The same trend can be seen in the analysis of the variance estimates.

6.4 Predictive Distributions

Aside from the fitting of past observations, the comparison of the predictive distributions can also be interesting. Table 6.5 shows the predictive premiums for a medium risk-profile. The values depend on the total number of reported claims and on the number of insured periods with at least one reported claim ($T - T_0$). To illustrate this, we selected a loss experience of 10 years, although other situations can easily be illustrated because closed-form formulas have been found to compute the predictive premiums of each model.

			Sum of claims					
Models	$T - T_0$	A priori	0	1	2	3	4	10
MVNB	1	0.0648	0.0402	0.0781	0.1160	0.1538	0.1917	0.4188
MZIP-Gamma	0	0.0670	0.0434	.	.	.	.	.
	1	.	.	0.0789	0.1151	0.1515	0.1882	0.4150
	2	.	.	.	0.1138	0.1498	0.1860	0.4088
	3	.	.	.	.	0.1482	0.1839	0.4029
	4	.	.	.	.	.	0.1818	0.3972
	10	.	.	.	.	.	.	0.3672
ZI-MVNB	0	0.0659	0.0426	0.0787	0.1129	0.1471	0.1813	0.3864

Table 6.5: Predictive Premiums

Even if the random effects were only introduced into the count distribution, the results of the MZIP-Gamma distribution indicate that T_0 also has an impact on the future premiums. Depending on the pattern of claim reporting, the future premiums show significant differences.

The ZI-MVNB model could be an interesting alternative to the MVNB model because it lessens the impact of the random effects to some extent. Indeed, because another part of the model adjusts the zeros, it reduces the variance of the random effects distribution, leading to predictive premiums that will be lower than those obtained with the MVNB model. This conclusion can be seen in the numerical results because the values of the experimental model are less than those of the MVNB model in both tails of the distribution.

A major difference that can be seen between the models lies in the variance values of the predictive distribution. These results are shown in the Table 6.6.

Models	$T - T_0$	A priori	Sum of claims 0	1	2	3	4	10
MVNB	.	0.0687	0.0418	0.0811	0.1204	0.1596	0.1989	0.4347
MZIP-Gamma	0	0.0729	0.0458	.	.	.	.	.
	1	.	.	0.0840	0.1237	0.1643	0.2059	0.4792
	2	.	.	.	0.1223	0.1623	0.2033	0.4709
	3	.	.	.	.	0.1604	0.2008	0.4631
	4	.	.	.	.	.	0.1983	0.4556
	10	.	.	.	.	.	.	0.4166
ZI-MVNB	.	0.0692	0.0453	0.0799	0.1147	0.1494	0.1841	0.3924

Table 6.6: Predictive Variance

Kind of Insureds	Number of Accidents (Estimated) MZIP-Gamma	ZI-MVNB	Number of Claims (Observed)
$v_2 = 0$	6,623	5,184	4,710
$v_2 = 1$	3,345	2,692	2,559
Total	9,968	7,874	7,269

Table 6.7: Number of Predicted Accidents and Number of Observed Claims

The conclusions that can be drawn about the variance are essentially the same as those made about the predictive premiums. The values of the variance obtained with the models are similar to those obtained with the MVNB model. With regard to the expected predictive value, it should be noted that the number of insured periods with a claim has a greater impact on subsequent premiums than the total number of reported claims.

6.4.1 Distribution of the Number of Accidents

Table 6.2 demonstrated that almost all covariates explained the reporting process, while the only covariate kept in the accident distribution is v2, which corresponds to the time the policyholder has been with the company. Then, for the analysis of the number of accidents, we can separate the portfolio in two parts: those who stayed between 3 and 5 years ($v_2 = 1$, that corresponds to 37.44% of the portfolio) and the remainders ($v_2 = 0$).

Depending on which model is selected, we can also see that the total number of predicted accidents per year can be very different, as shown by Table 6.7. To investigate the idea of finding what we cannot see (number of accidents) from what we see (number of claims) in greater depth, one can use other transformed models of the Poisson distribution and uncensor them to approximate the true frequency. Comparisons between the different obtained results, and analysis of the censorship process may yield interesting conclusions for actuaries about minor accidents, may provide new methods methods of computing the financial impact on deductibles and limits and can be used to perform risk analysis for some classes of risk.

6.5 Models Comparison

As seen in the previous sections, the differences between models produce, for example, different *a priori* and predictive premiums. Since remaining models (MZIP-Gamma and MVNB distributions) are non-nested, the models could not be compared directly and we cannot use the specification tests shown in Section 4.1.1. Consequently, to distinguish between these models, this section will present intuitive comparisons as well as more formal testing procedures which determine the model that is thought to be the best fit for our data.

6.5.1 AIC and Intuitive Comparison

A standard method of comparing non-nested models (and also nested models) is to use the information criteria, such as the Akaike Information Criteria $(AIC) = -2\log(L) + 2k$ or the Bayesian Information Criteria $(BIC) = -2\log(L) + 2\log(n)k$, where k represents the number of parameters of the model and n the total number of observations.

Before performing this AIC/BIC comparison between models, we can perform an intuitive analysis of the log-likelihood by models. The analysis presented here can be seen as a generalization of the comparison of the predictive values. Using the estimated parameters for the first six years, we computed the log-likelihood for each model for all seven years. We also estimated a standard Poisson distribution with the first six years of data. The Poisson distribution does not imply dependence between contracts of the same insured but the distribution can be useful in obtaining comparable results between models because it will smooth the random variations between years by comparing only the difference of fit. The graph in Figure 6.8 shows the values of AIC differences by year (versus the Poisson distribution) for each of the models presented in this paper.

As the number of past insured years grows, we can see that the fits of our models are clearly better than the Poisson distribution. Indeed, as the experience increases, this model shows its limited predictive capacity. For the last year that was not used for the parameters estimation, we can see that the MZIP-Gamma distribution shows the better fit, while the ZI-MVNB model does not show significant differences from the standard MVNB model.

Using only the remaining models (those not rejected by the specification tests), we can compare the MVNB and the MZIP-Gamma models with the information criteria (using the first six years):

Model	Log-likelihood	# Param.	AIC	BIC
MVNB	-22,619.58	7	45,253.16	45,308.59
MZIP-Gamma	-22,589.59	8	45,195.19	45,258.54

In both situations, the MZIP-Gamma shows better information criteria than the MVNB model.

6.5.2 Vuong Test

To see if the differences in the log-likelihood and the information criteria between the models are statistically significant, a test based on the difference in the log-likelihood can be performed. For independent observations, a log-likelihood ratio test for non-nested models, developed by Vuong (1989) and generalized by Rivers and Vuong (2002), can be used to see whether the MZIP-Gamma model is statistically better than the MVNB model. This test cannot be applied directly to our models since some observations - all contracts of the same insured - are not independent. However,

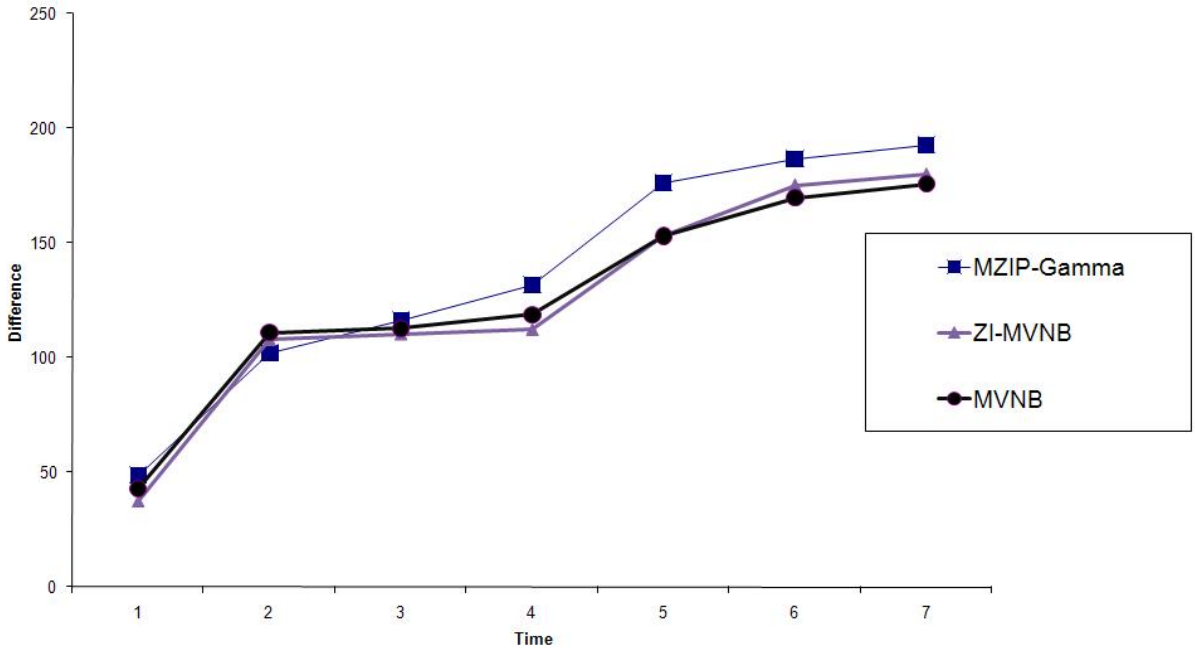

Table 6.8: Log-likelihood Differences (AIC) by year versus Poisson Distribution

as proposed by Golden (2003) and applied for instance in Boucher et al. (2008), this non-nested model test can be generalized fairly directly to correlated observations, as shown in Appendix A.

For the MZIP-Gamma against the MVNB models, the Vuong test, adapted for correlated observations, shows that the MZIP-Gamma model is statistically different from the MVNB model. Indeed, the resulting test involves *p-values* of less than 0.01% for the adapted tests (AIC and BIC).

7 Conclusion

We present new models for panel count data that can be used to model the data exhibiting a large number of zeros. Their joint distribution and their predictive distributions can be expressed analytically, which can be a significant advantage. Specification tests have been derived to distinguish between classic distributions and zero-inflated ones. The new models offer an interesting alternative to the standard multivariate negative binomial. Comparison between models shows that our insurance data were better fitted by the multivariate zero-inflated Poisson gamma distribution.

When applied to insurance data, these models can be used to distinguish claim and accident distributions. Compared with Lemaire's model, the zero-inflated distributions do not put too much weight on heterogeneity and do not heavily rationalize the behavior of insureds when they decide whether or not to report an accident. Analysis of the censorship parameters may help actuaries understand the claim process in greater depth. This new area of research seems very promising for future investigations, such as risk, fraud or marketing analysis.

Appendix A

The Vuong test statistic can be expressed as:

$$T_{LR,NN} = \frac{\left(\sum_i \sum_t^T \ell(f_{i,t}, \hat{\beta}_1) - \ell(g_{i,t}, \hat{\beta}_2)\right)}{\sqrt{nT}\ \sigma_{T_{LR,NN}}}, \tag{7.1}$$

where the log-likelihood function for the distribution f is defined as:

$$\ell(f_{i,t}, \hat{\beta}_1) = -\log\left(Pr_f[N_{i,t} = n_{i,t} | n_{i,t-1}, ..., n_{i,1}]\right). \tag{7.2}$$

Then, for an observation at time t, the conditional log-likelihood must be expressed based on experience $1, 2, ..., t-1$, using predictive distributions. The parameter $\sigma_{T_{LR,NN}}$ is the square root of the *discrepancy variance* and is expressed as:

$$\sigma^2_{T_{LR,NN}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{T}\left(\ell(f_{i,t}, \hat{\beta}_2) - \ell(g_{i,t}, \hat{\beta}_2)\right)\left(\ell(f_{i,j}, \hat{\beta}_2) - \ell(f_{i,j}, \hat{\beta}_2)\right). \tag{7.3}$$

The variance expression is closely related to the expression obtained using the standard Vuong test, except that the covariances between correlated observations are considered. Golden (2003) showed that intermediary tests must be performed to ensure that the test is valid. First, a necessary but not sufficient condition for establishing the asymptotic properties of this test involves the calculation of the following matrix:

$$B = \begin{bmatrix} B_{f,f} & B_{f,g} \\ B_{g,f} & B_{g,g} \end{bmatrix}, \tag{7.4}$$

where

$$B_{f,g} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{T}\nabla\ell(f_{i,t}, \hat{\beta}_1)\nabla\ell(g_{i,j}, \hat{\beta}_2). \tag{7.5}$$

The matrix B must be positive definite or, at least, must converge to it (see Golden (2003) for details). Because the variance of the difference in log-likelihood includes covariance terms, the second intermediary step is performed to avoid a situation in which $\sigma_{T_{LR,NN}}$ could be equal to zero. This involves the calculation of the *estimated discrepancy autocorrelation coefficient* and is defined as:

$$\hat{r} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1,j\neq t}^{T}\left(\ell(f_{i,t}, \hat{\beta}_1) - \ell(g_{i,t}, \hat{\beta}_2)\right)\left(\ell(f_{i,j}, \hat{\beta}_1) - \ell(g_{i,j}, \hat{\beta}_2)\right)}{2T\sum_{i=1}^{n}\sum_{t=1}^{T}(\ell(f_{i,t}, \hat{\beta}_1) - \ell(g_{i,j}, \hat{\beta}_2))^2}. \tag{7.6}$$

Given that Golden's adapted test (Golden (2003)) assumes that $\hat{r} \neq -1/(2T)$, this assumption must be verified to ensure that the test is correctly specified. The last intermediary test is to ascertain that the two models are not equal. Since we work with non-nested models and we already ensure that models cannot be equal by introducing simultaneous parameter restrictions, this step is not needed in our case.

To apply the adapted Vuong test for correlated observation, neither of the two models has to be true. The null hypothesis of the test is that the two models are equivalent, expressed as $H_0 : E[\ell(f, \hat{\beta}_1) - \ell(g, \hat{\beta}_2)] = 0$. Under the null hypothesis, the test converges to a standard normal

distribution. Rejection of the test in favor of the distribution f occurs when $T_{LR,NN} > c$, or in favour of g if $T_{LR,NN} < c$, where c represents the critical value for some significance level. Modification of this test is needed is cases where the compared models do not have the same number of parameters. As proposed by Vuong (1989), we may consider the following adjusted statistic:

$$\hat{C}(\theta) = C(\theta) + K(f, g),$$

where $K(f, g)$ is a correction factor, such as the AIC. Intermediary tests show that the matrix B is positive definite, while the value of the *estimated discrepancy autocorrelation coefficients* is far from $-1/12$, at approximately -0.003.

References

Artis, M., Ayuso, M., and Guillén, M. (1999). Modelling different types of automobile insurance fraud behaviour in the Spanish market. *Insurance Mathematics and Economics*, 24(1-2):67–81.

Boucher, J.-P. and Denuit, M. (2006). Fixed versus Random Effects in Poisson Regression Models for Claim Counts: Case Study with Motor Insurance. *ASTIN Bulletin*, 36:285–301.

Boucher, J.-P. and Denuit, M. (2008). Credibility Premiums for the Zero-Inflated Model and New Hunger for Bonus Interpretation. *Insurance: Mathematics and Economics*, 42:727–735.

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. *North American Actuarial Journal*, 11-4:110–131.

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.

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

Chernoff, H. (1954). On the Distribution of the Log-likelihood Ratio. *Ann. Math. Statist.*, 25:573–578.

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.

Dionne, G. and Vanasse, C. (1989). A Generalization of Automobile Insurance Rating Models: The Negative Binomial Distribution with Regression Component. *ASTIN Bulletin*, 19:199–212.

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

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

Gourieroux, C. (1999). The Econometrics of Risk Classification in Insurance. *The Geneva Papers on Risk and Insurance Theory*, 24:119–137.

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.

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*.

Holla, M. (1966). On a Poisson-Inverse Gaussian Distribution. *Metrika*, 11:115–121.

Hsiao, C. (2003). *Analysis of Panel Data*. Cambridge: Cambridge University Press, 2nd ed.

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

Lemaire, J. (1977). La Soif du Bonus. *ASTIN Bulletin*, 9:181–190.

Lemaire, J. (1995). *Bonus-Malus Systems in Automobile Insurance*. Boston: Kluwer Academic Publishers.

Lemaire, J. and Zi, H. (1994). A Comparative Analysis of 30 Bonus-Malus systems. *ASTIN Bulletin*, 24:287–309.

Moran, P. (1971). Maximum Likelihood Estimation in Non-Standard Conditions. *Proceeding of the Cambridge Philosophical Society*, 70:441–450.

Mullahy, J. (1986). Specification and Testing in some Modified Count Data Models. *Journal of Econometrics*, 33:341–365.

Mullahy, J. (1997). Heterogeneity, Excess Zeros, and the Structure of Count Data Models. *Journal of Applied Econometrics*, 12(3):337–350.

Philipson, C. (1960). The Swedish System of Bonus. *ASTIN Bulletin*, 1:134–141.

Rivers, D. and Vuong, Q. (2002). Model Selection Tests for Nonlinear Dynamic Models. *Econometrics Journal*, 5:1–39.

Shoukri, M., Asyali, M., VanDorp, R., and Kelton, D. (2004). The Poisson Inverse Gaussian Regression Model in the Analysis of Clustered Counts Data. *Journal of Data Science*, 2:17–32.

Song, J. (2005). Zero-Inflated Poisson Regression to Analyze Lenghts of Hospital Stays Adjusting for Intra-Center Correlation. *Communications in Statistics-Simulation and Computation*, 11:59–67.

Vuong, Q. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57:307–333.

Walhin, J.-F. and Paris, J. (2000). The True Claim Amount and Frequency Distributions Within a Bonus-Malus System. *ASTIN Bulletin*, 30:391–403.

Young, V. and De Vylder, F. (2000). Credibility in Favor of Unlucky Insureds. *North American Actuarial Journal*, 4:107–113.