Correlated Random Effects for Hurdle Models Applied to Claim Counts

by Jean-Philippe Boucher, Michel Denuit, and Montserrat Guillén

ABSTRACT

New models for panel data that consist of a generalization of the hurdle model are presented and are applied to modeling a panel of claim counts. Correlated random effects are assumed for the two processes involved to allow for dependence among all the contracts held by the same insured. A method to obtain a posteriori distribution of the random effects as well as predictive distributions of the number of claims is presented. A numerical illustration of reported insurance claims shows that if independence between random effects is assumed, then the variance of a priori premiums may be underestimated. If dependence between random effects is considered, then the predicted number of claims given past observations and covariate information and its variance is also larger than the one obtained when independence is assumed.

KEYWORDS

Count data, panel data, random effects, hurdle model, Gaussian copula, posterior distribution

1. Introduction

Many attempts have been made in the actuarial literature to find a model for the distribution of the annual number of claims reported by a given policyholder. Among all possible models, Boucher, Denuit, and Guillén (2007) show that when data exhibit a high number of zero values, hurdle models often provide a good fit for cross-section data. Boucher, Denuit, and Guillén (2008b) report that panel data models inducing serial dependence by means of random effects are well suited to fit observed claim counts. The present paper combines these two ideas and extends the hurdle model to panel count data.

The hurdle model was introduced by Cragg (1971) and reviewed by Mullahy (1986). It is characterized by the processes below and above the hurdle. The most widely used hurdle model sets the hurdle at zero: first, a binary variable allows for the participation to the second process and second, another process specifies the count number if the first process succeeds. See Mullahy (1986), Winkelmann (2003b), Grootendorst (1995), or Gurmu (1998). For a general overview, we refer the reader to Winkelmann (2003a).

The hurdle models have been successfully applied to the modeling of health care demand. It is generally accepted that the demand for certain types of health care services depends on two processes: the need for health care and the intensity of demand. Therefore, subject to certain assumptions, the use of a hurdle model is intuitive and the parameters can have a structural interpretation. See, e.g., Stoddart and Barer (1981), Pohlmeier and Ulrich (1995), Mullahy (1998), or Santos Silva and Windmeijer (2001). Coming back to insurance, a classification of the insured drivers based on two processes seems interesting because the majority of policyholders report less than two claims per year. A dichotomous variable first separates insureds with and without claim. In the former case, another process then generates the number of reported claims. The reluctance of some insured drivers to report their accidents (because they would lose their favorable bonus-malus scheme, i.e., a rebate under an experience rating system) can support the use of a hurdle model for the number of reported claims. The behavior of the insureds is likely to change when they have already reported a claim, confirming the hypothesis concerning the two processes that determine the total number of claims.

We consider our contribution necessary both from a theoretical and a practical point of view. On the theoretical sphere, we believe that hidden factors influencing claiming behavior have not been studied much. Once observable risk factors are taken into account, it is simplistic to assume that claiming is the result of chance. Some risk factors may not be directly observable, but still influence claiming. The nature of unobserved proneness to report claims is difficult to study. Here we address one common assumption about unobserved risk factors that influence a policyholder's claiming behavior. We want to relax the hypothesis that those hidden factors which may induce to report a claim are independent from those that would influence the policyholder to report at least one more claim. In the empirical part, we show that dependence is significant, and even not distinguishable from perfect dependence. This means that random effects influencing the two processes in the hurdle model have much in common. We also see that ignoring random effects dependence implies variance underestimation when calculating a priori premiums and when predicting the number of claims given past observations. An accurate variance estimate is usually needed for calculating premium loadings. Therefore, from the practical point of view, our aim is to show the importance of our methodological approach and to present how it can be implemented effectively.

Boucher, Denuit, and Guillén (2008a) extended hurdle models to panel count data with the help of independent random effects representing unexplained heterogeneity in both components (below and above the hurdle). Here, we allow for correlated random effects in each process. Given that hidden characteristics of the insureds (for instance, swiftness of reflexes, drinking habits, or respect of the highway code) are partly revealed by the number of claims re-

ported by the policyholders, a posteriori distribution of the random effects and predictive distributions of the number of claims are derived. Markov chain Monte Carlo simulations are needed to compute posterior distributions of the random effects. A numerical illustration of reported insurance claims supports the discussion, demonstrating that the dependence between random effects should be considered when computing predictive distribution.

This paper is organized as follows. Section 2 describes the model proposed in this paper and gives a numerical illustration based on a Spanish motor insurance portfolio observed during seven years. Section 3 investigates predictive distributions. Markov chain Monte Carlo simulations are used to generate samples from the predictive distribution for the Spanish data set. Section 4 states the conclusion.

2. Hurdle model for panel count data

2.1. Description of the model

Let us represent the number of claims N reported by a policyholder to the company as the product of an indicator variable J (equal to 1 if the policyholder reported at least 1 claim) and a counting variable $K \geq 1$ (giving the number of claims reported to the company when at least one claim has been filed). Furthermore, J and K are assumed to be independent. Hence,

$$\Pr[N=n]=\Pr[JK=n] = \begin{cases} \Pr[J=0] & \text{for } n=0 \\ \Pr[J=1]\Pr[K=n] & \text{for } n=1,2,\ldots \end{cases}. \quad (2.1)$$

The representation $N = JK$ is similar to the decomposition of the total claim amount in the individual model of risk theory.

Let $N_{i,1}, N_{i,2}, \ldots, N_{i,T}$ be the number of claims reported by policyholder i over period 1 to T. Each $N_{i,t}$ is decomposed into the product $J_{i,t}K_{i,t}$. To allow for serial dependence, it is common since Hausman, Hall, and Griliches (1984) to include random effect modeling unknown individual characteristics. Precisely, the joint distribution of $N_{i,1}, \ldots, N_{i,T}$ is given by

$$\begin{aligned} &\Pr[N_{i,1}=n_{i,1},\ldots,N_{i,T}=n_{i,T}] \\ &= \iint \Pr[N_{i,1}=n_{i,1},\ldots,N_{i,T}=n_{i,T}\,|\,\Theta_{i1}=\theta_{i1},\Theta_{i2}=\theta_{i2}] \\ &\quad g(\theta_{i,1},\theta_{i,2})d\theta_{i,1}d\theta_{i,2} \\ &= \iint \prod_{t=1}^{T}\Big((\Pr[J_{it}=0\,|\,\Theta_{i1}=\theta_{i1}])^{I_{(n_{i,t}=0)}} \\ &\quad (\Pr[J_{it}=1\,|\,\Theta_{i1}=\theta_{i1}]\Pr[K_{it}=n_{it}\,|\,\Theta_{i2}=\theta_{i2}])^{I_{(n_{i,t}>0)}}\Big) \\ &\quad g(\theta_{i,1},\theta_{i,2})d\theta_{i,1}d\theta_{i,2}, \end{aligned}$$

where g is the joint probability density function of $(\Theta_{i1}, \Theta_{i2})$.

Furthermore, we assume that J_{it} is Bernoulli distributed with mean $\Theta_{i,1}$ and that the success probability $\Theta_{i,1}$ is Beta(a_i, b) distributed. Covariates enter the model via $a_i = \exp(x_i'\beta)$. We assume that given $\Theta_{i2} = \theta_{i2}$, K_{it} obeys a shifted Poisson distribution with mean $\gamma_i\theta_{i,2}$ where $\gamma_i = \exp(x_i'\delta)$. The random effect $\Theta_{i,2}$ is Gamma distributed with mean 1 and variance α. Consequently, with these conditional distributions, the joint distribution of all contracts of the same insured is expressed as

$$\begin{aligned} &\Pr[N_{i,1}=n_{i,1},\ldots,N_{i,T}=n_{i,T}] \\ &= \iint \prod_{t=1}^{T}\left(\theta_{i,1}^{I_{(n_{i,t}=0)}}(1-\theta_{i,1})^{I_{(n_{i,t}>0)}}\left(e^{-\gamma_i\theta_{i,2}}\frac{(\gamma_i\theta_{i,2})^{n_{i,t}-1}}{(n_{i,t}-1)!}\right)^{I_{(n_{i,t}>0)}}\right) \\ &\quad g(\theta_{i,1},\theta_{i,2})d\theta_{i,1}d\theta_{i,2} \quad (2.2) \end{aligned}$$

Henceforth, all covariates included in the models are assumed to be time independent so that $\gamma_{i,t} = \gamma_{i,1} = \gamma_i$ and $a_{i,t} = a_{i,1} = a_i$, for all $t = 1, \ldots, T$. This assumption is often made for mathematical convenience. See, e.g., Gourieroux (1999). It has been made in Boucher, Denuit, and Guillén (2008a) for the hurdle model with independent random effects. Here, this assumption is needed to obtain simple equation forms. The inclusion of time-varying covariates does not induce any numerical difficulties. Note also that many variables in our context, such as the policyholder's age or vehicle's age, are changing over time but their evolution is deterministic, so they can be transformed to avoid time variation simply

taking the value at given time reference, i.e., at the beginning of the contract.

Random effects capture the effect of hidden individual characteristics. The two random effects (Θ_{i1}, Θ_{i2}) are likely to be correlated because the same omitted characteristics affect each process. Here, we use a Gaussian copula to represent the joint distribution of the random effects.[1] More precisely, we assume that g can be written as

$$g(\theta_{i,1}, \theta_{i,2}) = C^{GA}(G_1(\theta_{i,1}), G_2(\theta_{i,2}))g_1(\theta_{i,1})g_2(\theta_{i,2}), \tag{2.3}$$

with

$$c^{Ga}(G_1(\theta_{i,1}), G_2(\theta_{i,2})) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\left(\frac{\rho^2\Phi^{-1}(G_1(\theta_{i,1}))^2 + \rho^2\Phi^{-1}(G_2(\theta_{i,2}))^2 - 2\rho\Phi^{-1}(G_1(\theta_{i,1}))\Phi^{-1}(G_2(\theta_{i,2}))}{1-\rho^2} \right) \right) \tag{2.4}$$

where Φ is the standard normal distribution function and the marginal density functions g_1 and g_2 are Beta and Gamma, respectively, whereas G_1 and G_2 are the corresponding distribution functions. If the correlation parameter ρ is equal to 1, then Θ_{i1} and Θ_{i2} are perfectly positively dependent. In this case, the Gaussian copula reduces to the Fréchet-Hoeffding upper bound. Conversely, if the correlation parameter ρ is set to 0, then Θ_{i1} and Θ_{i2} are mutually independent. The latter case has been investigated in Boucher, Denuit, and Guillén (2008a).

The first moments of the hurdle for panel data, when conditioning on the random effects $\Theta_i = \{\theta_{1,i}, \theta_{2,i}\}$ can be expressed as

$$\begin{aligned} E[N_{i,t} \,|\Theta_i] &= \theta_{i,1} \sum_{j=0}^{\infty} (1+j) \Pr(K = j \,|\theta_{i,2}) \\ &= \theta_{i,1} + \gamma_i \theta_{i,1} \theta_{i,2} \end{aligned} \tag{2.5}$$

$$\begin{aligned} E[N_{i,t}^2 \,|\Theta_i] &= \theta_{i,1} \sum_{j=0}^{\infty} (1+j)^2 \Pr(K = j \,|\theta_{i,2}) \\ &= \theta_{i,1}(\gamma_i^2 \theta_{i,2}^2 + 3\gamma_i \theta_{i,2} + 1), \end{aligned} \tag{2.6}$$

which gives

$$E[N_{i,t}] = E[\theta_{i,1}] + \gamma_i E[\theta_{i,1}\theta_{i,2}] \tag{2.7}$$

$$\begin{aligned} Var[N_{i,t}] &= \gamma_i^2 E[\theta_{i,1}\theta_{i,2}^2] + E[\theta_{i,1}\theta_{i,2}][3\gamma_i - 2\gamma_i E[\theta_{i,1}]] \\ &\quad + E[\theta_{i,1}] - E[\theta_{i,1}]^2 - \gamma_i^2 E[\theta_{i,1}\theta_{i,2}]^2 \end{aligned} \tag{2.8}$$

$$\begin{aligned} Cov[N_{i,t}, N_{i,t+j}] &= Var[\theta_{i,1}] + 2\gamma_i (E[\theta_{i,1}^2\theta_{i,2}] - E[\theta_{i,1}\theta_{i,2}]E[\theta_{i,1}]) \\ &\quad + \gamma_i^2 (\mathrm{E}[\theta_{i,1}^2\theta_{i,2}^2] - E[\theta_{i,1}\theta_{i,2}]^2). \end{aligned} \tag{2.9}$$

No closed-form expression is available for the likelihood of the model we defined. Numerical integration techniques or Markov chain Monte Carlo methods can be used. Here, we resort to the NLMIXED procedure from the SAS System and we implement the approach developed by Nelson et al. (2006).

2.2. Empirical illustration

We worked with a sample from 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 contains 15,179 policyholders who remained with the company for seven complete periods. Five exogeneous variables are kept in the panel plus the annual number of claims. The exogenous variables are defined in Table 1.

Table 2 contains the observed annual claim frequency for the whole portfolio, together with the maximum number of claims per policyholder. The average claim frequency is 6.8412% over the whole observation period.

Estimated parameters for the hurdle model with independent random effects are shown in Table 3. We also included the estimations obtained with the

[1]Our approach differs from an existing alternative proposed by Gurmu and Elder (2007) for standard bivariate count data. These authors do not use a copula, but instead they directly specify a bivariate density for the random effects.

Table 1. Exogenous variables

Variable	Description
Sex	equals 1 for women and 0 for men
Years with the company (3–5) (> 5)	equals 1 if the client has been with the company between 3 and 5 years equals 1 if the client has been with the company for more than 5 years
Age	equals 1 if the insured is 30 years old or younger
Vehicle capacity	equals 1 if engine capacity is larger than or equal to 5500 cc

Table 2. Frequency of claims

Period	Frequency (%)	Maximum
1	7.5367	3
2	6.9569	3
3	6.4035	3
4	6.1466	3
5	6.4695	4
6	6.8450	4
7	7.5301	3
Total	6.8412	4

Poisson distribution with Gamma random effects (also called multivariate negative binomial distribution, or MVNB) (Boucher, Denuit, and Guillén 2008b). Analysis of estimations leads to the conclusion that men file fewer claims than women, while new insureds in the company seem to have a worse loss experience than older clients. In the presence of other covariates, we also see that young drivers exhibit worse claim experience, but it is not statistically significant. Finally, insureds with powerful vehicles tend to have more insured periods with accidents than other drivers.

The parameters of the positive part indicate which policyholders are most likely to report a high number of claims in a single time period. Quite surprisingly, only insured drivers who stayed with the company for three to five years tend to be better than the other groups.

The estimated parameters of the hurdle model with correlated random effects are shown in Tables 4 and 5. The estimates of the parameters are very close to the ones obtained with the independent random effects model. The intercepts of the second process (i.e., positive part of the hurdle model) for correlated models are smaller than the one estimated by the independent random effects model. This differ-

Table 3. Estimated parameters for the MVNB model and for the hurdle model with independent (ind.) random effects

			Hurdle Parts (Ind.)	
Variable	Parameter	MVNB	Zero	Positive
Intercept	—	–2.6600 (0.0352)	0.3066 (0.0682)	–2.3688 (0.0525)
Sex	Women	0.1087 (0.0409)	0.1298 (0.0401)	. .
	Men	0	0	. .
Years with the company	3–5	–0.1805 (0.0327)	–0.1726 (0.0325)	–0.1951 (0.0933)
	>5	–0.2103 (0.0370)	–0.2124 (0.0370)	. .
	<3	0	0	0
Age	≤30	0.0471 (0.0346)	0.0616 (0.0341)	. .
	>30	0	0	. .
Vehicle capacity	≥5500 cc	0.0990 (0.0316)	0.0975 (0.0315)	. .
	<5500 cc	0	0	. .
Other	α or b	0.8832 (0.0432)	19.9640 (1.2279)	0.8122 (0.2070)
Loglikelihood		–26,702.98	–26,688.70	

Table 4. Estimated parameters for the hurdle model with Gaussian copula (Gauss.) for random effects (standard errors)

		Hurdle parts (Gauss.)	
Variable	Parameter	Zero	Positive
Intercept	—	0.3043 (0.0531)	–2.9100 (0.0882)
Sex	Women	0.1362 (0.0416)	. .
	Men	0	. .
Years with the company	3–5	–0.1703 (0.0323)	–0.2278 (0.0939)
	>5	–0.2075 (0.0368)	0
	<3	0	0
Age	≤30	0.0600 (0.0440)	. .
	>30	0	. .
Vehicle capacity	≥5500 cc	0.0966 (0.0315)	. .
	<5500 cc	0	. .
Other	α or b	19.9568 (1.0844)	0.7533 (0.2038)
	ρ	0.8424 (0.1165)	. .
Loglikelihood		–26,662.47	

Table 5. Estimated parameters for the hurdle model with Fréchet-Hoeffding copula (F.-H.) for random effects (standard errors)

		Hurdle parts (F.-H.)	
Variable	Parameter	Zero	Positive
Intercept	—	0.3104 (0.0429)	–2.9622 (0.0788)
Sex	Women	0.1368 (0.0399)	. .
	Men	0	. .
Years with the company	3–5	–0.1702 (0.0326)	–0.2330 (0.0927)
	>5	–0.2069 (0.0366)	0
	<3	0	0
Age	≤30	0.0597 (0.0346)	. .
	>30	0	. .
Vehicle capacity	≥5500 cc	0.0964 (0.0315)	. .
	<5500 cc	0	. .
Other	α or b	20.0836 (0.5725)	0.8818 (0.2442)
Loglikelihood		–26,663.28	

ence comes from the correlation added to the model. The second process seems to be affected by the correlated random effects because of the composition of the portfolio. Indeed, for all insureds who did not report a claim (approximately 66% of the portfolio), the model does not need to use the second random effects and its associated correlation.

The model shows a significant value of 0.8424 for the parameter ρ associated with the Gaussian copula. It clearly shows that the insured drivers who report claims also experience time periods with a large number of reported claims. Here, Kendall's tau is equal to $2\arcsin(\rho)/\pi = 0.6377$.

Considering the strong positive dependence existing between the random effects, we also fit the model assuming perfect positive dependence between Θ_{i1} and Θ_{i2}, i.e., we replace the Gaussian copula with the Fréchet-Hoeffding one corresponding to $\rho = 1$. The

results are displayed in Table 5 and are very similar to those of Table 4. This is not unsurprising, as the Gaussian copula model exhibits a strong dependence between its random effects.

Differences between models can be analyzed through the mean and the variance of some insured profiles. Several profiles have been selected and are described in Table 6. The first profile is classified as a good driver, while the last one usually exhibits bad loss experience. The other profile corresponds to medium risk. Table 7 shows that the expected values of all profiles are similar for the four models studied. However, the independent random effects model underestimates the low risk profile and overestimates the riskier one, compared with the correlated hurdle models. The greatest differences between the models lie in the variance estimates because the MVNB model exhibits lower variances than the other models. As for the expected values for the lower risk profiles, the independent random effects model shows variance values that are less than the ones obtained for the correlated random effects models. This does not hold for the bad profile.

Let us now compare the different models. The correlated hurdle model with the Fréchet-Hoeffding copula as well as the independent hurdle model are nested to the correlated hurdle model with the Gaussian copula. For $\rho = 0$ we get the independent copula, while a value $\rho = 1$ results in the Fréchet-Hoeffding copula. This can be tested using versions of the Wald or loglikelihood ratio tests allowing for a null hypothesis on the boundary of the parameter space. Data indicate that the independent copula assumption is rejected against the Gaussian copula, while no statistical difference is shown between the Fréchet-Hoeffding copula and the Gaussian copula.

The MVNB and the hurdle models cannot be compared directly because they are non-nested models. A standard method of comparing non-nested models is through the information criteria, such as the Akaike Information Criteria $(AIC) = -2\log(L) + 2k$ where k is the number of parameters in the model. According to Burnham and Anderson (2002), a difference greater than 10 indicates a significant difference between models. The MVNB model gives an AIC value of 53,419.96 and the hurdle model (with F-H copula) has a value of 53,346.56, which indicates that the latter should be preferred.

3. Predictive distribution

3.1. Updating the random effects distribution

As time passes, more observations become available and the distribution of the random effects $\Theta_{i,1}$ and $\Theta_{i,2}$ can be updated from past experience. This allows the actuary to derive the distribution of the

Table 6. Profiles analyzed

Profile Number	Kind of Profile	Sex	Years 3–5	Years >5	Age	Vehicle
1	Good	0	0	1	0	0
2	Medium	1	1	0	0	0
4	Bad	1	0	0	1	1

Table 7. Expectations and variances of the annual number of claims for the different profiles considered

	Good Profile		Medium Profile		Bad Profile	
Models	Mean	Variance	Mean	Variance	Mean	Variance
MVNB	0.0567	0.0595	0.0651	0.0688	0.0902	0.0974
Hurdle Ind.	0.0570	0.0644	0.0659	0.0717	0.0911	0.0997
Hurdle Gaus.	0.0575	0.0654	0.0663	0.0720	0.0909	0.0985
Hurdle F.-H.	0.0577	0.0655	0.0663	0.0718	0.0909	0.0983

future number of claims $N_{i,T+1}$ from past observations $N_{i,1},\ldots, N_{i,T}$. Formally, the predictive distribution is obtained from

$$\begin{aligned}
&\Pr[N_{i,T+1}=n_{i,T+1}\,|\,N_{i,1}=n_{i,1},\ldots,N_{i,T}=n_{i,T}]\\
&=\frac{\Pr[N_{i,1}=n_{i,1},\ldots,N_{i,T+1}=n_{i,T+1}]}{\Pr[N_{i,1}=n_{i,1},\ldots,N_{i,T}=n_{i,T}]}\\
&=\iint \Pr[N_{i,T+1}=n_{i,T+1}\,|\,\theta_{i,1},\theta_{i,2}]\left(\frac{\Pr[N_{i,1}=n_{i,1},\ldots,N_{i,T}=n_{i,T}\,|\,\theta_{i,1},\theta_{i,2}]g(\theta_{i,1},\theta_{i,2})}{\int\int \Pr[N_{i,1}=n_{i,1},\ldots,N_{i,T}=n_{i,T}\,|\,\theta_{i,1},\theta_{i,2}]g(\theta_{i,1},\theta_{i,2})d\theta_{i,1},\theta_{i,2}}\right)d\theta_{i,1},\theta_{i,2}\\
&=\iint \Pr[N_{i,T+1}=n_{i,T+1}\,|\,\theta_{i,1},\theta_{i,2}]\left(\frac{(\prod_t \Pr[N_{i,t}=n_{i,t}\,|\,\theta_{i,1},\theta_{i,2}])g(\theta_{i,1},\theta_{i,2})}{\int\int \prod_t \Pr[N_{i,t}=n_{i,t}\,|\,\theta_{i,1},\theta_{i,2}])g(\theta_{i,1},\theta_{i,2})d\theta_{i,1},\theta_{i,2}}\right)d\theta_{i,1},\theta_{i,2}\\
&=\iint \Pr[N_{i,T+1}=n_{i,T+1}\,|\,\theta_{i,1},\theta_{i,2}]g(\theta_{i,1},\theta_{i,2}\,|\,n_{i,1},\ldots,n_{i,T})d\,\theta_{i,1},\theta_{i,2},
\end{aligned}\tag{3.1}$$

where $g(\theta_{i,1}, \theta_{i,2}|n_{i,1},\ldots, n_{i,T})$ is the joint posterior distribution of the random effects $\Theta_{i,1}$, $\Theta_{i,2}$), reflecting the past experience of policyholder i.

Exact predictive and posterior distributions for the random effects can only be expressed in closed form for some distributions, such as the hurdle distribution with independent random effects investigated by Boucher, Denuit, and Guillén (2008a). For other models, such as the correlated random effects hurdle models studied here, these distributions cannot be evaluated analytically. Instead, we use Markov chain Monte Carlo (MCMC) simulations to compute posterior and predictive distributions.

3.2. Markov chain Monte Carlo approach

MCMC simulations reproduce realizations from a Markov chain that converges to the joint distribution of the random effects. The resulting random draws are no longer independent, but under mild regularity conditions (as described in the appendix of Smith and Roberts (1993), for example), the value of the draw tends in distribution to that of a random draw from the joint distribution as the number of draws becomes moderately large. See Scollnik (2001) for an introduction to MCMC simulations in actuarial sciences.

The posterior distribution of the random effects can be expressed as

$$\begin{aligned}
&g(\theta_{i,1},\theta_{i,2}\,|\,n_{i,1},\ldots,n_{i,T})\\
&\propto g(n_{i,1},\ldots,n_{i,T}\,|\,\theta_{i,1},\theta_{i,2})g(\theta_{i,1},\theta_{i,2})\\
&\propto c^{Ga}(F_1(\theta_{i,1}),F_2(\theta_{i,2}))\times\theta_{i,1}^{A-1}(1-\theta_{i,1})^{B-1}\\
&\quad\theta_{i,2}^{C-1}\exp(-\theta_{i,2}D)
\end{aligned}\tag{3.2}$$

where

$$A=\sum_t^T I_{(n_{i,t}=0)}+a_i$$

$$B=T-\sum_t^T I_{(n_{i,t}=0)}+b$$

$$C=\sum_t^T (n_{i,t}-I_{(n_{i,t}=0)})+1/\alpha$$

$$D=\lambda_i\sum_t^T I_{(n_{i,t}=0)}+1/\alpha$$

Note that $\sum_t^T I_{(n_{i,t}=0)}$ and $\sum_t^T n_{i,t}$ are respectively the number of insured periods without claims and the total number of reported claims of a specific insured.

Our strategy is to simulate values of $\Theta_{i,1}$ and $\Theta_{i,2}$ using the Metropolis-Hasting algorithm. The most obvious choice for simulating realizations of these two random variables is to simulate independent Beta and Gamma distributed random variables with the parameters A, B and C, D, respectively.

The $(j+1)$th iteration of the Metropolis-Hasting algorithm can be described as follows:

1. Given $\theta_{i,1}^{(j)}, \theta_{i,2}^{(j)}$, simulate $\hat{\theta}_{i,1}, \hat{\theta}_{i,2} \sim g(x,y|\theta_{i,1}^{(j)}, \theta_{i,2}^{(j)})$;
2. Specifying the distribution of interest by

$$\pi(x,y) = c^{Ga}(F_1(x), F_2(y))x^{A-1} (1-x)^{B-1} y^{C-1} \exp(-yD), \quad (3.3)$$

compute the ratio

$$P = \frac{\pi(\hat{\theta}_{i,1}, \hat{\theta}_{i,2})}{\pi(\theta_{i,1}^{(j)}, \theta_{i,2}^{(j)})} \frac{g(\theta_{i,1}^{(j)}, \theta_{i,2}^{(j)} | \hat{\theta}_{i,1}, \hat{\theta}_{i,2})}{g(\hat{\theta}_{i,1}, \hat{\theta}_{i,2} | \theta_{i,1}^{(j)}, \theta_{i,2}^{(j)})} \quad (3.4)$$

3. Simulate $U \sim U(0,1)$
4. Take:

$$\theta_{i,1}^{(j+1)}, \theta_{i,2}^{(j+1)} = \begin{cases} \hat{\theta}_{i,1}, \hat{\theta}_{i,2} & \text{if } U \le P \\ \theta_{i,1}^{(j)}, \theta_{i,2}^{(j)} & \text{otherwise} \end{cases} \quad (3.5)$$

3.3. Empirical illustration

Two jump functions $g(x,y|\theta_{i,1}^{(j)}, \theta_{i,2}^{(j)})$ have been tested for our model:

1. An independent Metropolis-Hastings jump function, where $\hat{\theta}_{i,1}$ and $\hat{\theta}_{i,2}$ are distributed as the product of a Beta(A,B) and a Gamma(C,D). In this situation, the values generated by the proposal distribution (jump function) do not depend on the past realizations. Using this jump function, the acceptance probability can be simplified as

$$P = \frac{c^{Ga}(F_1(\hat{\theta}_{i,1}), F_2(\hat{\theta}_{i,2}))}{c^{Ga}(F_1(\theta_{i,1}^{(j)}), F_2(\theta_{i,2}^{(j)}))} \quad (3.6)$$

where F_1 is the distribution function of the Beta distribution with parameters a and b, while F_2 is the Gamma($1/\alpha$,$1/\alpha$) distribution function.

2. A second jump function has been used where $\hat{\theta}_{i,1}$ and $\hat{\theta}_{i,2}$ are distributed as the product of a Beta and a Gamma distributions, having means equal to $\theta_{i,1}^{(j)}$ and $\theta_{i,2}^{(j)}$, respectively. The second terms of the distributions, needed to compute the second moment, have been set to $T - \sum_t^T k_{i,t} + b$ for the Beta distribution and $1/\alpha$ for the Gamma random effects.

The same process has been used for the hurdle model with a Fréchet-Hoeffding copula.

Table 8 shows the mean of the predictive distribution for a medium risk profile. This mean depends on the sum of reported claims and on the number of insured periods with at least one reported claim $(T - T_0)$. To illustrate, we selected a loss experience of 10 years. The first jump process has been used to compute posterior distributions that are not close to the prior distribution, while the more extreme situations (10 reported claims, 10 insured periods with at least one claim, etc.) required the use of the second jump process.

We simulated five chains of 500,000 runs each, selected a burn-in value of 10,000 draws for our MCMC simulations and a lag of 10 to eliminate the possible correlation between successive draws. Table 8 describes the results of these MCMC simulations, as well as results drawn from analytic computations for the MVNB and the independent random effects models.

Interesting conclusions can be drawn from Table 8. For the independent random effects, the number of insured periods with at least one claim has a greater impact on the predictive mean than the total number of reported claims. For insureds who reported one or fewer claims, the hurdle model shows a decrease in the predictive mean that is less than with the MVNB model. For higher claims reporters, the hurdle model exhibits a wide range of predictive mean values that go from 0.25 to 1.4 times the MVNB ones.

Further, the dependence between the two random effects has a major impact on the predictive mean. The hurdle model's correlated random effects exhibits predictive mean values that are closer to the MVNB ones than with the independent random effects model. Moreover, the impact on the predictive mean of the number of past reported claims compared with the number of past insured periods with claims is different from the hurdle with independent random effects. In fact, it is exactly the opposite: the penalties generated by the model are higher for insured having more reported claims in the same time period.

Table 8. Mean of the predictive distribution

			Sum of claims					
Models	$T - T_0$	A priori	0	1	2	3	4	10
MVNB	.	0.0651	0.0413	0.0778	0.1143	0.1509	0.1874	0.4064
Hurdle Ind.	0	0.0659	0.0448	.	.	.	.	.
	1	0.0659		0.0833	0.0876	0.0920	0.0963	0.1223
	2	0.0659	.	.	0.1246	0.1304	0.1363	0.1715
	3	0.0659	.	.	.	0.1683	0.1755	0.2190
	4	0.0659	.	.	.	.	0.2140	0.2649
	10	0.0659	.	.	.	.	.	0.5177
Hurdle Gaus.	0	0.0663	0.0441	.	.	.	.	.
	1	0.0663	.	0.0776	0.1161	0.1510	0.1848	0.3902
	2	0.0663	.	.	0.1108	0.1495	0.1855	0.3953
	3	0.0663	.	.	.	0.1437	0.1825	0.3965
	4	0.0663	.	.	.	.	0.1761	0.3951
	10	0.0663	.	.	.	.	.	0.3631
Hurdle F.-H.	0	0.0663	0.0441	.	.	.	.	.
	1	0.0663	.	0.0774	0.1225	0.1655	0.2077	0.4640
	2	0.0663	.	.	0.1083	0.1539	0.1965	0.4514
	3	0.0663	.	.	.	0.1427	0.1855	0.4386
	4	0.0663	.	.	.	.	0.1748	0.4256
	10	0.0663	.	.	.	.	.	0.3562

Table 9 shows the predictive variance. Big differences can also be seen between the independent random effects model and the correlated one. In many situations, the predictive variance is smaller for the independent random effects model compared to the other hurdle models. Unlike the expected values, the correlated hurdle models do not exhibit close similarities to the MVNB model for variance values. The difference is greatest for insureds who report often.

4. Conclusion

Even if the a priori and the predictive means of the correlated hurdle model are quite close to the MVNB ones, the corresponding variances greatly differ. This result is not surprising. It seems quite intuitive that the dependence assumption only affects the covariance structure of the model, but not the first-order moments, which should be consistently estimated even under misspecification of the higher-order moments, provided the mean is correctly specified.

This conclusion is similar to what happens in much simpler situations. For instance, when the basic Poisson model for cross-sectional count data is used and conditional over- or under-dispersion is ignored, parameter estimates are consistently estimated if the mean is correctly specified. So, when estimating the negative binomial model, parameters estimates do not differ much from the ones obtained in the Poisson model, while standard errors may be quite different and second-order moment estimates, too.

The method presented in the previous sections shows that it is possible to account for the correlation between the two processes of hurdle count distribution and that estimation is feasible. Moreover, dependence should not be ignored if one is interested

Table 9. Variance of the predictive distribution

			Sum of claims					
Models	$T - T_0$	A priori	0	1	2	3	4	10
MVNB	.	0.0688	0.0428	0.0807	0.1185	0.1564	0.1942	0.4212
Hurdle Ind.	0	0.0717	0.0497	.	.	.	.	.
	1	0.0717	.	0.0841	0.0975	0.1114	0.1258	0.2220
	2	0.0717	.	.	0.1155	0.1332	0.1515	0.2735
	3	0.0717	.	.	.	0.1439	0.1652	0.3066
	4	0.0717	.	.	.	.	0.1697	0.3256
	10	0.0717	.	.	.	.	.	0.2710
Hurdle Gaus.	0	0.0719	0.0466	.	.	.	.	.
	1	0.0720	.	0.0826	0.1289	0.1747	0.2217	0.5486
	2	0.0720	.	.	0.1186	0.1666	0.2148	0.5393
	3	0.0720	.	.	.	0.1542	0.2036	0.5247
	4	0.0720	.	.	.	.	0.1891	0.5066
	10	0.0720	.	.	.	.	.	0.3801
Hurdle F.-H.	0	0.0718	0.0463	.	.	.	.	.
	1	0.0718	.	0.0826	0.1334	0.1840	0.2351	0.5661
	2	0.0718	.	.	0.1176	0.1703	0.2214	0.5492
	3	0.0718	.	.	.	0.1570	0.2080	0.5319
	4	0.0718	.	.	.	.	0.1952	0.5146
	10	0.0718	.	.	.	.	.	0.4228

in studying the variance and not only the mean of the number of claims, given the exogenous characteristics. This is especially relevant in insurance. Ignoring dependence would lead to underestimation of the variance and, therefore, to inefficient pricing. Premium calculation starts from the expected number of claims to obtain a pure premium and it is then usually loaded by a factor which may depend on the estimated variance.

An interesting empirical conclusion results from our analysis too. Our illustration shows empirical evidence that the correlation between the two processes of the hurdle model is significant and that the dependence is quite strong. Therefore, it means that the unobserved characteristics that influence the policyholder's decision to report a claim do not differ much whether or not he has already reported one during that given time-period. So, we can conclude that there is no reason to believe that certain types of policyholders within one risk group, i.e., with some particular unobserved characteristics, would base their claiming behavior on a bonus hunger more intensely than others. By "bonus hunger" we mean that they would refrain from claiming if no claim has already been filed during the year to guarantee that a rebate will be granted in the following year's premium payment, but instead file the claim if the rebate is not granted.

It has sometimes been claimed that experience rating schemes induce in the policyholder a desire not to claim, but if a claim has already taken place, then policyholder would not underreport. In fact, we show that once some observed risk characteristics have been accounted for, there no such switching regime in the model, so that unobserved proneness to claim would still remain once a claim has been filed.

Acknowledgments

Jean-Philippe Boucher would like to thank the financial support from the Natural Sciences and Engineering Research Council of Canada. 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 Guillén would like to thank the Spanish Ministry of Science / ECO2010-21787-C01.

References

Boucher, J.-P., M. Denuit, and M. Guillén, "Risk Classification for Claim Counts: A Comparative Analysis of Various Zero-Inflated Mixed Poisson and Hurdle Models," *North American Actuarial Journal* 11:4, 2007, pp. 110–131.

Boucher, J.-P., M. Denuit, and M. Guillén, "Modeling of Insurance Claim Count with Hurdle Distribution for Panel Data," *Advances in Mathematical and Statistical Modeling*, Boston: Birkhäuser, 2008a.

Boucher, J.-P., M. Denuit, and M. Guillén, "Models of Insurance Claim Counts with Time Dependence Based on Generalization of Poisson and Negative Binomial Distributions," *Variance* 2, 2008b, pp. 135–162.

Burnham, K., and D. Anderson, *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*, New York: Springer, 2002.

Cragg, J., "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods," *Econometrica* 39, 1971, pp. 829–844.

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

Grootendorst, P., "A Comparison of Alternative Models of Prescription Drug Utilization," *Health Economics* 4, 1995, pp. 183–198.

Gurmu, S., "Generalized Hurdle Count Data Regression Models," *Economics Letters* 58, 1998, pp. 263–268.

Gurmu, S. and J. Elder, "A Simple Bivariate Count Data Regression Model," *Economics Bulletin* 3, 2007, pp. 1–10.

Hausman, J., B. Hall, and Z. Griliches, "Econometric Models for Count Data with Application to the Patents–R&D Relationship," *Econometrica* 52, 1984, pp. 909–938.

Mullahy, J., "Specification and Testing in Some Modified Count Data Models," *Journal of Econometrics* 33, 1986, pp. 341–365.

Mullahy, J., "Much Ado about Two: Reconsidering Retransformation and the Two-Part Model in Health Econometrics," *Journal of Health Economics* 17, 1998, pp. 247–281.

Nelson, K., S. Lipsitz, G. Fitzmaurice, J. Ibrahim, M. Parzen, and R. Strawderman, "Use of the Probability Integral Transformation to Fit Nonlinear Mixed-Effects Models with Nonnormal Random Effects," *Journal of Computational and Graphical Statistics* 15, 2006, pp. 39–57.

Pohlmeier, W., and V. Ulrich, "An Econometric Model of the Two-Part Decision Making Process in the Demand for Health Care," *Journal of Human Resources* 30, 1995, pp. 339–361.

Santos Silva, J., and F. Windmeijer, "Two-Part Multiple Spell Models for Health Care Demand," *Journal of Econometrics* 104, 2001, pp. 67–89.

Scollnik, D., "Actuarial Modeling with MCMC and BUGS," *North American Actuarial Journal* 5, 2001, pp. 96–125.

Smith, A., and G. Roberts, "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods (with discussion)," *Journal of the Royal Statistical Society*, Series B 55, 1993, pp. 3–23.

Stoddart, G., and M. Barer, "Analyses of the Demand and Utilization through Episodes of Medical Care" in *Health, Economics and Health Economics: Proceedings of the World Congress on Health Economics*, eds. J. van der Gaag and M. Perlman, Leiden, The Netherlands: Elsevier, 1981.

Winkelmann, R., *Econometric Analysis of Count Data* (4th ed.), New York: Springer, 2003a.

Winkelmann, R., "Health Care Reform and the Number of Doctor Visits: An Econometric Analysis," *Journal of Applied Econometrics* 19, 2003b, pp. 455–472.